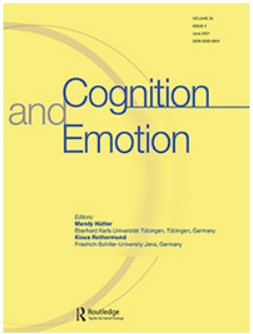

Cognition and Emotion

ISSN: (Print) (Online) Journal homepage: https://www.tandfonline.com/loi/pcem20

Testing a computational model of subjective well-being: a preregistered replication of Rutledge et al. (2014)

Niels Vanhasbroeck, Levi Devos, Sebastiaan Pessers, Peter Kuppens, Wolf Vanpaemel, Agnes Moors & Francis Tuerlinckx

To cite this article: Niels Vanhasbroeck, Levi Devos, Sebastiaan Pessers, Peter Kuppens, Wolf Vanpaemel, Agnes Moors & Francis Tuerlinckx (2021) Testing a computational model of subjective well-being: a preregistered replication of Rutledge et al. (2014), Cognition and Emotion, 35:4, 822-835, DOI: 10.1080/02699931.2021.1891863

To link to this article: https://doi.org/10.1080/02699931.2021.1891863

COGNITION AND EMOTION
2021, VOL. 35, NO. 4, 822–835
https://doi.org/10.1080/02699931.2021.1891863

Routledge
Taylor & Francis Group

Testing a computational model of subjective well-being: a preregistered replication of Rutledge et al. (2014)

Niels Vanhasbroeck, Levi Devos, Sebastiaan Pessers, Peter Kuppens, Wolf Vanpaemel, Agnes Moors [b] and Francis Tuerlinckx

[a]Research Group of Quantitative Psychology and Individual Differences, KU Leuven, Leuven, Belgium; [b]Center for Social and Cultural Psychology, KU Leuven, Leuven, Belgium

ABSTRACT

Subjective well-being changes over time. While the causes of these changes have been investigated extensively, few attempts have been made to capture these changes through computational modelling. One notable exception is the study by Rutledge et al. [Rutledge, R. B., Skandali, N., Dayan, P., & Dolan, R. J. (2014). A computational and neural model of momentary subjective well-being. *Proceedings of the National Academy of Sciences*, 111(33), 12252–12257. https://doi.org/10.1073/pnas.1407535111], in which a model that captures momentary changes in subjective well-being was proposed. The model incorporates how an individual processes rewards and punishments in a decision context. Using this model, the authors were able to successfully explain fluctuations in subjective well-being observed in a gambling paradigm. Although Rutledge et al. reported an in-paper replication, a successful independent replication would further increase the credibility of their results. In this paper, we report a preregistered close replication of the behavioural experiment and analyses by Rutledge et al. The results of Rutledge et al. were mostly confirmed, providing further evidence for the role of rewards and punishments in subjective well-being fluctuations. Additionally, the association between personality traits and the way people process rewards and punishments was examined. No evidence for such associations was found, leaving this an open question for future research.

ARTICLE HISTORY
Received 17 July 2020
Revised 4 February 2021
Accepted 15 February 2021

KEYWORDS
Momentary subjective well-being; computational model; happiness; prediction error

Happiness concerns us all: everyone likes to be happy, even if one does not actively pursue it (Diener et al., 2009). And we have good reason to. Happiness or subjective well-being (SWB) has been associated with greater professional and academic success (not only as a consequence, but also as a cause; Lyobomirsky & Layous, 2013; Lyobomirsky et al., 2005), with greater resilience to and recovery from negative events (Cohn et al., 2009; Papousek et al., 2010), and with greater physical health outcomes (Lawrence et al., 2015; Veenhoven, 2008).

However, happiness is not stable, but fluctuates over time (Brickman & Campbell, 1971; Diener et al., 2009; Diener et al., 2006). Despite the recognition and acceptance of these momentary changes, and despite the interest in factors eliciting them (e.g. daily need satisfaction; Oishi et al., 2007), there has been little effort to capture them using a computational model. An important exception is provided by Rutledge et al. (2014). They proposed and tested a model that aims to capture momentary changes in SWB using information about the way people process rewards and punishments in a decision context. It is thus assumed that fluctuations in SWB are a function of how people perceive rewards and punishments that result from their decisions. To test their model, Rutledge et al. (2014) presented individuals with a probabilistic gambling task. Each trial, participants were asked to choose between receiving a certain outcome or an uncertain outcome. When

CONTACT Niels Vanhasbroeck niels.vanhasbroeck@kuleuven.be Research Group of Quantitative Psychology and Individual Differences, KU Leuven, Tiensestraat 102, Leuven 3000, Belgium

Supplemental data for this article can be accessed at https://doi.org/10.1080/02699931.2021.1891863

they chose the latter option, they had a 50/50 probability of receiving either a better or a worse outcome than the certain outcome. Throughout the task, participants could not only win, but also lose money, so that participants experienced both positive and negative changes in SWB. In this paper, our primary aim was to replicate two main findings of Rutledge et al. (2014) related to their self-report measure of SWB.[1] First, they found that their model was able to capture momentary SWB well, which is consistent with the idea that fluctuations in SWB are, at least in part, a function of cumulative rewards and punishments. Second, when the outcome was uncertain, these fluctuations were more strongly associated with the discrepancy between the actual outcome and the expected value of the possible outcomes than with the expected value itself.[2] This suggests an important role of prediction errors in momentary changes of SWB.

The authors provided an in-paper large-sample replication of their findings using a smartphone app (the Great Brain Experiment, http://www.thegreatbrainexperiment.com/; Rutledge et al., 2014). However, a successful replication by an independent lab would further increase the credibility of their results. Therefore, we aim to replicate the findings reported by Rutledge et al. (2014) using a preregistered close replication of both the experimental procedure and the analyses related to the self-report measure of happiness (LeBel et al., 2018).

The second aim of our study was to explore the correlates of the individual differences in reactivity towards rewards and punishments (as operationalised by the parameters of the model). Rutledge et al. (2014) observed such individual differences, but could not account for them. Based on previous research, we examined neuroticism, extraversion, the behavioural inhibition system (BIS), and the behavioural activation system (BAS) in this study. Neuroticism has previously been linked to negative emotional reactivity (Dauvier et al., 2019; Thake & Zelenski, 2013), whereas extraversion has been linked to increased reactivity towards rewards (Diener et al., 2003). Similarly, individuals scoring high on BIS are expected to be more sensitive to punishment cues, while people scoring high on BAS are expected to react more strongly to reward cues (Carver & White, 1994; Gray, 1972).

The work reported in this paper was partly preregistered. The preregistration was done on the Open Science Framework and the preregistration document can be found through the following link: https://osf.io/krhyz/. We preregistered on the 28th of March 2016 and data collection started on the 30th of March 2016. All deviations from the preregistration are discussed below.

Method

Participants

Participants were recruited through social media channels. The final sample consists of 50 participants (22 female) with a mean age of 23.3 ($SD = 11.4$). The total spent budget was €544.6, which is consistent with our preregistered stopping rule. Due to technical problems occurring in the beginning stages of the experiment, two participants were asked to restart the experiment. One participant suffered from similar technical issues later in the experiment and was not asked to start again. All three participants are included in the final sample and their accompanying data were used in all analyses.

Participants were rewarded with their total earnings. These varied across participants because they depended on the choices made during the experiment. Participants left with an average reward of €10.9 ($SD = €3.3$).

Materials

Two questionnaires were used to assess the personality traits of interest.

NEO Five-Factor Inventory (NEO-FFI; McCrae & Costa, 2004). This questionnaire consists of 60 items, measuring 5 dimensions, namely neuroticism, extraversion, openness, agreeableness, and conscientiousness. Because we were interested in the traits neuroticism and extraversion, only their associated items were used (24 in total), extended with 10 filler items, making a total of 34 questions. The filler items were selected from the remainder of the NEO-FFI. Reliability for the scales under investigation is good, with Cronbach's α of 0.82 for neuroticism and 0.80 for extraversion (McCrae & Costa, 2004).

BIS/BAS scales (Carver & White, 1994). This questionnaire uses 24 items to measure 4 different scales, namely *BIS*, *BAS-Drive*, *BAS-Fun Seeking*, and *BAS-Reward Responsiveness*. In addition to these four scales, we calculated an overall *BAS*-score by averaging over its components. The separate scales have acceptable reliability with Cronbach's α's of 0.74,

0.73, 0.76, and 0.66, respectively (Carver & White, 1994).

Procedure

The code used to run the experimental task is based on the code of the original experiment, provided to us by the original researchers. The original code for the experiment is available via the following the link: https://www.github.com/RutledgeLab/Rutledge-happiness-task.

Participants voluntarily registered for the experiment and were tested at the second author's home or in another quiet location. Upon their arrival, participants were informed about the experiment and completed an informed consent form.

After reading the instructions on the screen, participants received a beginning capital of €5 and started with 20 mandatory practice trials. During the practice trials, participants were aware that their net total would reset to €5 when the experimental trials started, making every monetary change during the practice trials only temporary. After these initial 20 trials, the real gambling task started.

Figure 1 shows a single trial of the gambling task. Participants were asked to choose between a certain outcome (e.g. €0.15) and a gamble between two equally probable outcomes (e.g. between €0 and €0.4). One of the gamble outcomes was always better than the certain outcome; the other was always worse (see example). When participants had made their decision, their choice remained visible on screen for 4s. When the gamble was chosen, the gamble outcome became visible for 1s after these 4s. Three types of trials were used: gain trials (i.e. both the certain outcome and gamble lead to a monetary increase), loss trials (i.e. both the certain outcome and gamble lead to a monetary decrease), and mixed trials (i.e. a certain outcome of €0 and a gamble with possible gain and loss). Each type of trial was presented 50 times, resulting in a total of 150 trials.

Every two to three trials, participants had to answer the question "How happy are you at this moment?" (Dutch: "Hoe gelukkig ben je op dit moment?") on a visual analogue scale ranging from "very unhappy" (0) to "very happy" (1). This always occurred after the trial outcome was shown. The question appeared first without a rating line and after a 5s delay, the rating line became visible. From this moment onward, participants had 4s to answer. This question appeared 60 times throughout the experiment. Throughout the experiment, the total earnings were shown at the bottom of the screen, except when participants were making an SWB judgement. This total was updated when the trial outcomes were known.

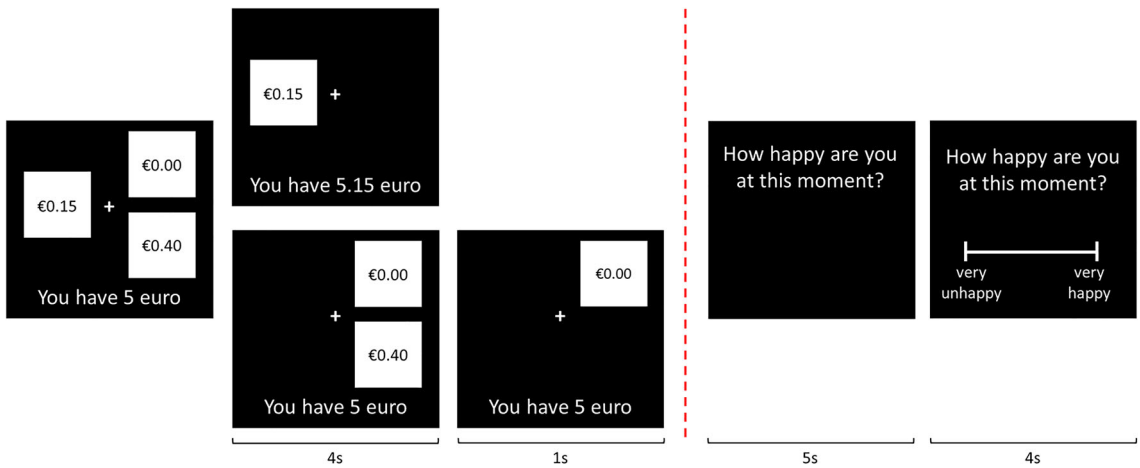

Figure 1. Visualisation of the experimental trials. On the left side of the dotted line, a typical trial is shown. First, participants are shown a choice between a certain outcome and a gamble between two outcomes (respectively on the left and on the right in the first picture). After deciding on an alternative, their choice is shown for 4s. The total amount of money is always shown at the bottom and changes whenever the trial outcome is presented. If the certain outcome is chosen (top), the total is immediately changed. If the gamble is chosen (bottom), the total is only changed after the 4s delay and shown for 1s, together with the gamble outcome. Every two to three trials, participants are asked to give a SWB rating when the outcome is known (right side of the dotted line). First they see the question "How happy are you at this moment?" for 5s, after which a rating line appears. Participants have 4s to answer the question before moving on to the next trial.

After the experiment, participants filled out the two questionnaires, translated to Dutch, on Survey Gizmo (http://www.surveygizmo.com/).

Analysis

Replication of original findings

Concerning the primary aim of our study, we evaluated the success of replication with three different analyses. The first and last, but not the second, analyses were also performed by Rutledge et al. (2014). The first, preregistered, analysis concerns fitting the original model of Rutledge et al. (2014) to individual data using least-squares estimation. The two other, non-preregistered analyses concern a hierarchical extension of the model (estimated within a Bayesian framework) and the other a model comparison procedure.

Individual analysis. Rutledge et al. (2014) formulated the following model to capture fluctuations in SWB for a given individual

$$\widehat{H}_t = w_0 + w_1 \sum_{j=1}^{t} \gamma^{t-j} CR_j + w_2 \sum_{j=1}^{t} \gamma^{t-j} EV_j + w_3 \sum_{j=1}^{t} \gamma^{t-j} RPE_j \quad (1)$$

In this equation, $\widehat{H}_t$ is the predicted momentary SWB score at trial *t*. The variable CR_j represents the value of the certain outcome at trial *j*. The variable EV_j represents the possible gambling outcomes at trial *j* (e.g. for a gamble between €0 and €0.4, the *EV* would be equal to €0.2). The variable RPE_j represents the reward prediction error at trial *j*, defined as the difference between the gamble outcome at trial *j* and the EV_j (e.g. if the outcome of the gamble is €0, the *RPE* would be equal to -€0.2). The three variables CR_j, EV_j, and RPE_j thus represent the outcomes of the choices that participants make throughout the experimental task. When a participants chooses the certain outcome, the value of this outcome will (according to the equation) influence momentary SWB through the variable CR_j. When a participant chooses the gamble, the change in SWB will depend on the expected value of the gamble (i.e. does one expect to win or lose money; EV_j) and the prediction error (i.e. was the outcome better or worse than expected; RPE_j). When a participant chooses the certain outcome at a given trial *j*, it makes no sense to talk about the EV_j and RPE_j, since they only represent the effect of the gamble on SWB. Therefore, these variables are set to 0 when the certain outcome has been chosen at trial *j*. Similarly, CR_j becomes 0 if the gamble has been chosen at trial *j*.

The parameter w_0 is a constant term that reflects the baseline SWB from which an individual fluctuates throughout the experiment. The parameters w_1, w_2, and w_3 are the weights attached to the variables CR_j, EV_j, and RPE_j, respectively. These weights reflect the strength of the influence of these variables on momentary SWB for a given individual. Finally, γ (ranging from 0 to 1) is a forgetting factor which makes the outcomes of recent trials more influential on momentary SWB. For simplicity, the variables CR_j, EV_j, and RPE_j will be represented as *CR*, *EV*, and *RPE* in the remainder of the article. Furthermore, $\widehat{H}_t$ will represent the predicted momentary SWB score at trial *t*, while H_t is reserved for the observed momentary SWB score.

We considered the replication as successful when:

(a) the model explained fluctuations in SWB approximately as well as in the original study (i.e. $R^2 = 0.47 \pm 0.21$;$M \pm SD$);
(b) the estimates for all model weights were significantly greater than 0 (i.e. w_1, w_2, and $w_3 > 0$);
(c) the estimated weight of *RPE* was significantly greater than the estimated weight of *EV* (i.e. $w_3 > w_2$).

To evaluate the first criterion, we compared the fit of the model to our data to the fit reported by the original authors (in terms of mean R^2 and its standard deviation). To evaluate the second criterion, we used frequentist one-sample *t*-tests. Finally, a frequentist paired-sample *t*-test was used to evaluate the last criterion. The latter two analyses were done between-person and correspond to the relevant analyses done by Rutledge et al. (2014).

Following the original analyses performed by Rutledge et al. (2014), we used a least-squares estimation procedure to estimate parameters for each individual separately. The parameters were bound in the same way as the original authors did, such that w_1,w_2,w_3 are constrained by −1 and 1, w_0 by 0.01 and 0.99, and γ by 0 and 1. Furthermore, we followed Rutledge et al. (2014) in using raw SWB ratings for these analyses. In the Supplementary Materials, we also report on the same analyses while making use of standardised SWB scores to check robustness of the results.

Hierarchical analysis. In the original study of Rutledge et al. (2014), all individuals were fitted separately with the model as previously defined. However, hierarchical or multilevel modelling has been advocated as a better alternative when dealing with individual differences. Therefore, in addition to fitting the individual data separately, we applied a hierarchical extension of the model by Rutledge et al. (2014). This analysis was not preregistered.

Hierarchical models offer some important advantages compared to idiosyncratic models when studying individual differences. First, individual parameters obtained through separate analyses may turn out to be extreme compared to estimates from others. In hierarchical analyses, this extremity is corrected by assuming that all parameters are generated from a population distribution, eventually pulling parameter estimates closer to the population mean (a phenomenon called shrinkage; see Gelman & Hill, 2006). Second, instead of ignoring group-level information as in individual analysis, all data are considered and thus taken into account using hierarchical modelling (Gelman, 2006; Gelman & Hill, 2006). This leads to a more principled treatment of uncertainty of the parameter estimates and more robust analysis of population properties. Lastly, hierarchical modelling provides a more parsimonious result than individually estimated models. Taken together, the hierarchical extension of the model was added to ensure that the results, as reported by us, are robust.

The hierarchical model was estimated within a Bayesian framework (Kruschke, 2015). The full model is summarised in Figure 2. At the lowest level, we modelled the observed SWB score H_{it} on trial t for individual i using a t distribution. The t distribution has three specified parameters, being its location (μ_{it}), scale (σ_i), and degrees of freedom (ν). The person-specific location μ_{it} changes with every trial, depending on the contextual input from the gambling paradigm and the person-specific parameters of the model (see Equation 1). The scale parameter σ_i is person-specific and assumed to be drawn from a population-level gamma distribution (with uniform priors on its parameters). The degrees of freedom ν were assumed to be the same for all participants and received an exponential prior. We opted for the t distribution over the traditionally chosen normal distribution because it is a more robust choice, being less sensitive to outliers (Gelman & Hill, 2006).

The person-specific parameters w were assumed to be drawn from different population-level normal distributions (one for each w; the subscript $p \in \{0,1,2,3\}$ is used to differentiate between these different distributions). The means of these population-level distributions μ_p are assigned a normal prior, and the standard deviations σ_p a uniform prior. Lastly, the individual-specific forgetting factor γ_i was assumed to be drawn from a population-level beta distribution (with gamma priors on its parameters). As can be seen in Figure 2, priors were kept relatively vague.

The model was coded in Stan (Carpenter et al., 2017) and called from R using the rstan package (Stan Development Team, 2018). Four chains were sampled, each with a burnin period of 5000 iterations and a sampling period of 20,000 iterations.

Because the hierarchical analysis also aimed at replicating the original findings of Rutledge et al. (2014), we considered the replication successful when:

(a) the model explained fluctuations in SWB approximately as well as in the original study. To make the R^2 of the hierarchical model comparable to the R^2 of the original model, we calculated an R^2 measure based on the individual-specific parameters (as in the original model), which are derived from their posterior means;
(b) the 95% credibility intervals (*CrI*) of all model weights did not contain 0[3];
(c) the 95% *CrI* of $RPE - EV$ did not contain 0.

Model comparison. In their Supporting Information, Rutledge et al. (2014) reported results from a model comparison procedure. They compared nine different computational models with the purpose of evaluating how well the model in Equation 1 describes the data compared to a set of competing models. In a second non-preregistered analysis, we fit all these models to each individual's data separately using the least-squares procedure, as done in the individual analysis.

All nine models are briefly described next. For a more elaborate discussion of these models, we refer the reader to the Supplementary Materials of this paper or the Supporting Information of Rutledge et al. (2014). The models under consideration are:

- M1: The model described by Equation 1;
- M2: A model that only accounts for the trial outcomes, without EV and RPE (i.e. it accounts for certain reward and gamble outcomes);

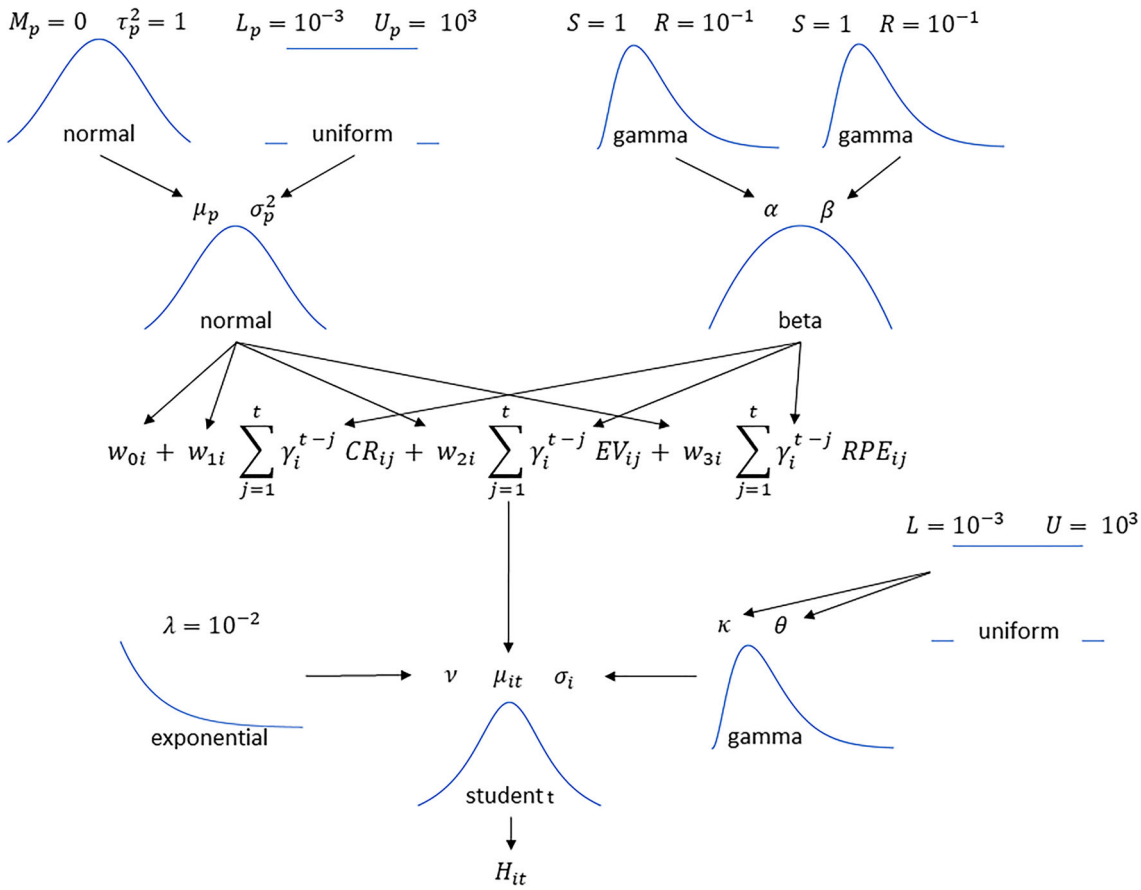

Figure 2. Hierarchical model fitted to the data using Bayesian estimation techniques. As explained in the text, it is assumed that each measurement of SWB is drawn from a Student *t* distribution. The location, but not shape, of this distribution changes with every trial. The model of Rutledge et al. (2014) is assumed to predict this trial-dependent location, based on the outcomes of the experiment and the individual-specific parameters. All parameters *w* are assumed to be drawn from different population-level normal distributions (one for each parameter, hence the subscript *p*). At last, the individual-specific forgetting factors γ_i are assumed to be drawn from a population-level beta distribution. As can be seen, the prior distributions for each of parameters are kept vague.

- M3: Model M1 in which the forgetting factor is the sum of two independent exponential functions (i.e. instead of γ^{t-j}, it uses $({\gamma_1}^{t-j} + {\gamma_2}^{t-j})$);
- M4: Model M1 in which each variable has its own estimated forgetting factor;
- M5, M6, M7: Three linear regressions between happiness and outcomes at a differing number of previous trials (one, two, or seven trials);
- M8: Model M1 with an extra term defining the difference between the attained and unattained outcome, so as to take regret or relief into account (i.e. regret if the attained outcome is lower than the unattained outcome and vice-versa for relief);
- M9: Model M1 with an extra term defining the difference between the attained outcome and the best possible outcome.

Model M2 was introduced by Rutledge et al. (2014) to test whether a model with prediction error would provide a better description of the data than a model that only accounts for trial rewards. Models M3 and M4 were used to test whether one forgetting factor, as introduced in Equation (1), is enough to capture the decaying influence of rewards on SWB. Finally, the aim of models M8 and M9 was to test whether additional variables may also play a role in explaining fluctuations in SWB.

To evaluate model performance, we used the Bayesian Information Criterion (BIC), following Rutledge et al. (2014). The BIC is a frequently used measure of fit which penalises for model complexity, thus avoiding overfit of the model to the data (Schwarz, 1978). Again following Rutledge et al. (2014), we summed

BIC values across participants. Unlike Rutledge et al. (2014), we also included a within-subject analysis where we identify the best-performing model for each individual separately.

Again following Rutledge et al. (2014), we standardised SWB scores before estimation so that participants with a larger variability would not bias the results. This also lead to discarding the intercept of the models. To examine robustness of results, we also performed this analysis while making use of raw SWB scores and when including the intercept for the standardised scores. The results of these analyses are reported in the Supplementary Materials.

Individual differences

The second aim of this study involved evaluating whether an association between personality traits and reactivity to rewards and punishments could be observed. This was evaluated using two analyses, one of which was preregistered.

Both analyses involved computing the correlations between personality trait scores and parameter estimates. Given their hypothesised increased reactivity towards rewards and punishments, we predicted that individuals who score high on neuroticism, extraversion, BIS, or BAS would have higher estimates of w_1, w_2, and w_3, suggesting a greater influence of the variables *CR*, *EV*, and *RPE* on momentary SWB for these individuals. The preregistered analysis made use of the parameter estimates from the individual analysis, whereas the non-preregistered analysis used the estimates from the hierarchical analysis.

Differences with the original study

First, we translated the instructions from English to Dutch and transformed gambling outcomes from pounds to euros. Second, the beginning capital was changed from £20 to €5 and the magnitude of the outcomes was divided by two. Third, some of our participants received their instructions simultaneously, though they were tested in separate experimental rooms. The original researchers always used individual assessments. Fourth, our participants went through fewer practice trials than in the original experiment (20 practice trials instead of 60).

The most important difference, however, lies in the availability of current total earnings. In our experiment, current total earnings were always shown, except when participants were making a SWB judgement. Showing the total earnings makes it possible for participants to evaluate how "well" they are performing, which may lead to increased influence of distal trials to current SWB. Rutledge et al. (2014) reported a similar variant of their experiment, but in their case, total earnings were also shown during SWB judgements. Based on the results of this experiment, we did not expect the results regarding the replication criteria to be different (Rutledge et al., 2014). However, since participants are able to evaluate their overall performance, one possible difference could occur for the forgetting factor γ, which may on average be higher or lower than the average γ reported by the original authors (Rutledge, personal communication, July 1, 2020).

Deviations from the preregistration

When performing our research, we deviated from the preregistered plan.[4] We list all of the deviations below, and if possible, discuss the likely effect on the result. First, in the original preregistration plan, it was indicated that participants would be recruited through the local participant recruitment system and KU Leuven social media groups. However, participants for this study were recruited through social media channels only. Second, while the original plan was to test participants inside computer labs owned by the KU Leuven, participants were tested elsewhere due to recurring technical difficulties. Although we attempted to provide a stable environment for the experiment, it might have been less controlled than in the laboratory. Therefore, data quality may be lower. Third, some non-preregistered analyses were carried out, as indicated above. Finally, the preregistered R code was changed to include the individual fitting procedure, the hierarchical model, and the model comparison procedure. Furthermore, while the original R code contains multiple regressions of the parameters on the personality traits, we decided not to use these analyses, given we already calculated correlations for this purpose.

Results

The raw data and code used for the analyses are available on https://osf.io/9g2zw/.

Descriptive results

SWB

Across individuals, the mean SWB was 0.6, which showed some interindividual differences ($SD = 0.1$). Variability around the individual-specific means also showed differences across individuals (mean $SD = 0.1$; range $SD = [0.0, 0.3]$).

Questionnaires

For the *NEO-FFI*, participants scored an average of 2.9 on neuroticism ($SD = 0.7$) and an average of 3.7 on extraversion ($SD = 0.5$).

For the *BIS/BAS scales*, participants scored an average 2.8 on BIS ($SD = 0.4$) and an average of 3.0 on BAS ($SD = 0.4$), and more specifically 2.7 on *BAS-Drive* ($SD = 0.6$), 3.0 on *BAS-Fun Seeking* ($SD = 0.5$), and 3.3 on *BAS-Reward Responsiveness* ($SD = 0.4$).

Individual parameter estimates

Using the frequentist least-squares estimation procedure, we obtained the following descriptive results (visualised in Figure 3). The mean estimate for w_1 was 0.14 ($SD = 0.14$), for w_2, it was 0.06 ($SD = 0.08$), and for w_3, it was 0.15 ($SD = 0.14$). The mean of the estimates for the forgetting factor γ was 0.81 ($SD = 0.19$). These estimates resemble the ones found by Rutledge et al. (2014), with exception of the mean estimate for γ, which was higher in our study (original estimates: 0.61 ± 0.30; $t(49) = 7.41$, $p < 0.001$, $95\%CI = [0.76,0.87]$, Hedges' $g = 0.85$). This suggests that previous trials had a larger influence on momentary SWB in our sample than in the sample of Rutledge et al. (2014). This could be a result of showing the total earnings, which makes it easier for participants to see the long-term consequences of their decisions.

To increase model clarity, Equation (2) represents the model of Rutledge et al. (2014) with mean estimates inserted:

$$\widehat{H}_t = 0.60 + 0.14 \sum_{j=1}^{t} 0.81^{t-j} CR_j + 0.06 \sum_{j=1}^{t} 0.81^{t-j} EV_j + 0.15 \sum_{j=1}^{t} 0.81^{t-j} RPE_j \quad (2)$$

These estimates may seem small, but it is important to note that they were made for H_t ranging from 0 to 1. This means that, for example, for the *CR* component, each addition of one euro elicited a change in happiness scores of 0.14, or a fluctuation crossing more than a tenth of the SWB-scale.

Hierarchical parameter estimates

Convergence was reached for all parameters in the Bayesian model (all individual $\hat{R} < 1.1$; $\hat{R} = 1.00 \pm 0.02$; $M \pm SD$; Multivariate $\hat{R}. = 1.01$). The posterior population mean estimate[5] of w_1 was 0.08 ($SD = 0.08$), of w_2, it was 0.02 ($SD = 0.03$), and of w_3, it was 0.07 ($SD = 0.05$). The posterior population mean estimate of the forgetting factor γ was 0.85 ($SD = 0.17$). As can be seen, the posterior population means were overall lower than the average of the individual frequentist parameter estimates. A notable exception is the forgetting parameter γ, for which the posterior population mean was greater than the frequentist average.

Another difference is that the posterior standard deviations were generally smaller than the standard deviations of the individual frequentist estimates. This is due to shrinkage: Individual estimates can be more extreme, which leads to a larger *SD*. This is most easily seen for the γ parameter. For the individual estimation procedure, this parameter showed a larger range of individually estimated values (from 0.31 to 1.00) than for the hierarchical estimation procedure (from 0.51 to 1.00).

Replication results

Individual analysis

First, mean fit of the model to the data, as measured by R^2, was equal to 0.54 ($SD = 0.21$), which closely resembles the results reported by Rutledge et al. (2014) ($R^2 = 0.47 \pm 0.21$;$M \pm SD$). Second, all parameter estimates were significantly greater than 0 (w_1: $t(49) = 7.29$, $p < 0.001$, $95\%CI = [0.10,0.18]$; w_2: $t(49) = 4.76$, $p < 0.001$, $95\%CI = [0.03,0.08]$; w_3: $t(49) = 7.63$, $p < 0.001$, $95\%CI = [0.11,0.19]$). Finally, the estimate for the *RPE* component was significantly greater than the estimate for the *EV* component ($t(49) = 4.92$, $p < 0.001$, $95\%CI = [0.06,0.14]$). This analysis thus confirms the findings of Rutledge et al. (2014), suggesting that the model fits well to the data and behaves the same as previously reported. More specifically, these results suggest that when the certain outcome leads to a gain, when the expected value of a gamble is positive, and when the outcome of a gamble is better than expected, SWB increases, and vice-versa for certain losses, negative expected values, and worse than expected outcomes. Furthermore, the prediction error had a greater influence on SWB than the expected value,

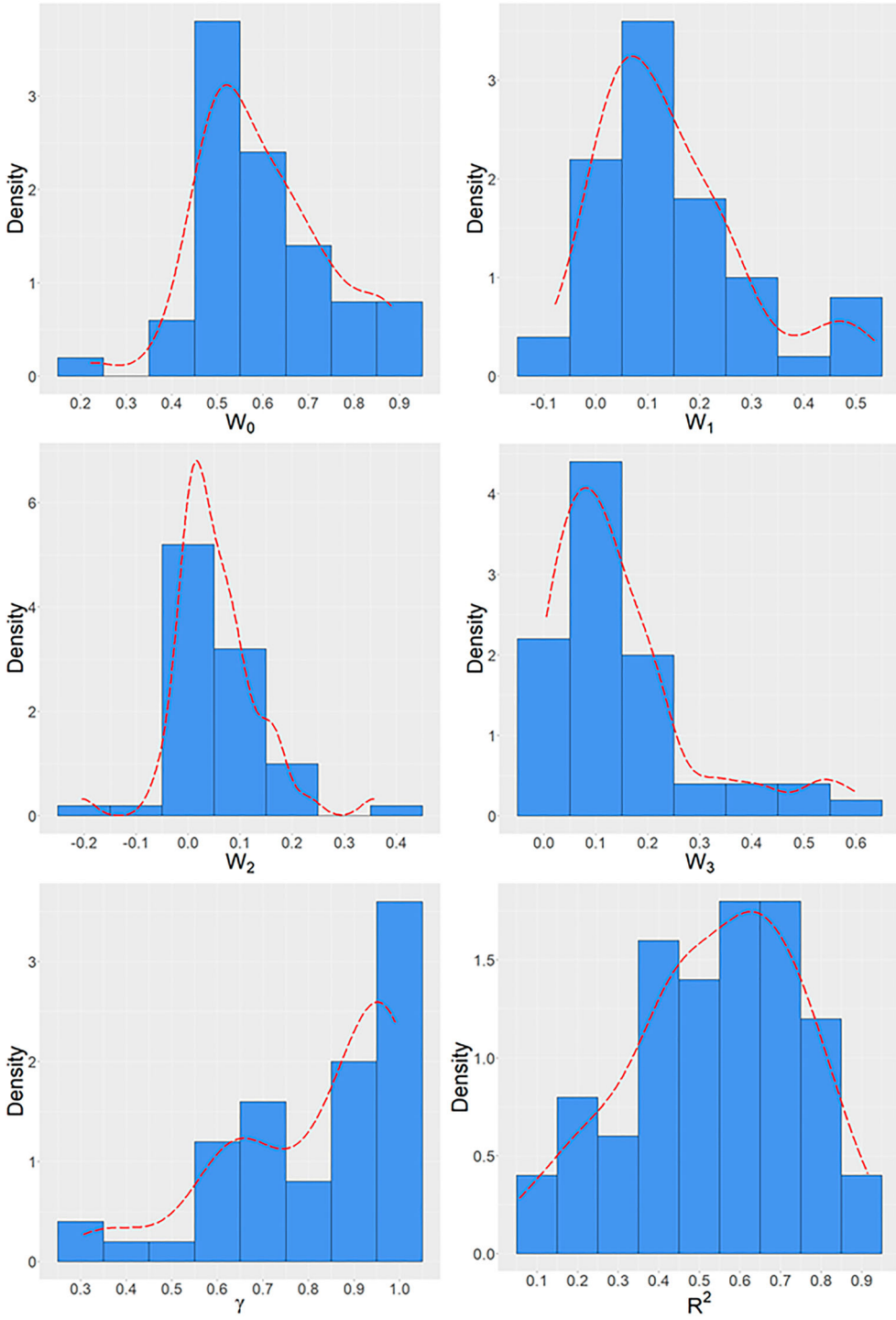

Figure 3. Histograms of the estimated values of the parameters, as well as the R^2 from the frequentist estimation procedure. Their estimated density is shown with the red dotted line. The estimated weights of the model (w_1 for variable *CR*, w_2 for variable *EV*, and w_3 for variable *RPE*) were greater than 0 in most cases, indicating a positive association between a certain reward, expectation of reward, or discrepancy with this expectation and momentary SWB scores.

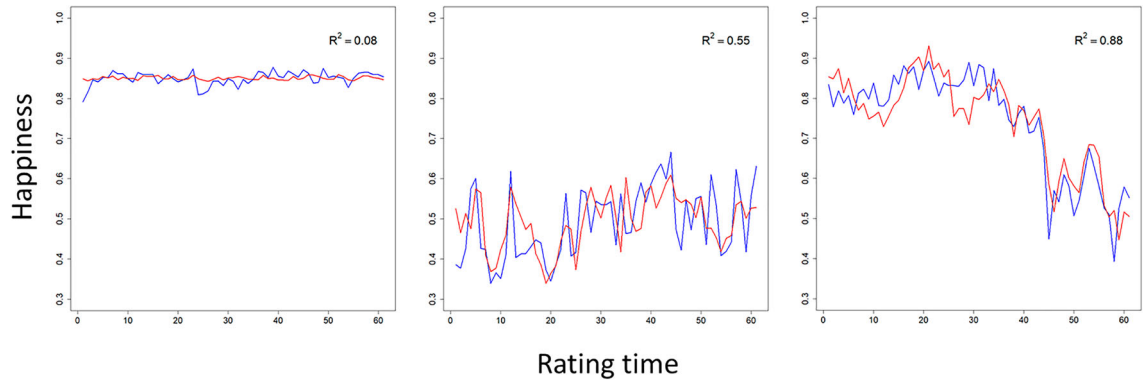

Figure 4. Three plots showing observed SWB scores H_t (in blue) versus model-based SWB scores $H^{\wedge}_t$ (in red). The model fitted some participants better than others, which is probably related to participant's use of the measurement scale. The participant on the left used only little of the scale, resulting in a low variance. This makes it difficult for the model to pick up the little variability there is, resulting in the prediction of an almost straight line. Other participants had a greater variance, resulting in a better fit. The participant on the right shows a decreasing trend in his data, reporting lower SWB in later trials. This trend is picked up by the model. The participant in the middle has the median R^2, providing a good comparison to the other two participants in this figure. Note that the happiness scale on the Y-axis is cropped.

underlining the importance of prediction error in understanding fluctuations in SWB.

Figure 4 shows the real momentary SWB scores for a few participants (blue) as well as calculated momentary SWB scores for these same participants (red) using their individual parameter estimates. Participants were chosen based on their R^2 to give an overview of model fits. As can be seen, the model is not able to fit every participant's data well, which could be related to their use of the measurement scale. For example, the participant on the leftmost panel in Figure 4 had only little variance in SWB, which makes it more difficult for the model to pick up this variation.

Hierarchical analysis

Turning to the replication using the hierarchical model, we observed that the mean R^2 of the model was remarkably lower than the previously reported fit, namely a mean R^2 of 0.32 ($SD = 0.22$). This result can be explained by shrinkage: As individual estimates are shrunken towards the group mean, they may become less ideal at the individual level. Second, 0 was not part of the 95% CrI for any of the parameters ([0.06,0.11] for w_1; [0.01,0.04] for w_2; [0.06,0.09] for w_3), indicating they are reliably greater than 0. Finally, the estimates for the *RPE* component were reliably greater than the estimates for the *EV* component (95% CrI = [0.03,0.07] for $w_3 - w_2$). Thus, despite a lower individual fit to the data, the Bayesian hierarchical analysis also confirms the findings of Rutledge et al. (2014).

Model comparison

Table 1 shows the mean R^2 and summed BIC values found for the nine competing models. There was not really one model that outperformed all others, in contrast to the results reported by Rutledge et al. (2014), where they concluded that the evidence was in favour of model M1. Rather, several models had a similar performance. More specifically, models M1, M3, M4, M8, and M9 showed BIC values that are close to each other, with models M4 and M9 even slightly outperforming M1 on both R^2 and BIC. Based on these results, we cannot conclude that we successfully replicated the model comparison results of Rutledge et al. (2014).

Table 1 also shows the percentage of times each model performed best for an individual is displayed. As can be seen, there is considerable variability of model fit across participants. M1 provides the best fit in only a quarter of the participants, followed by M2 and M4 for about 20% and 18% of the individuals resp. When taking these results together, it seems that the majority of the participants benefited from a model that includes an RPE term, although for 20% of the participants, trial outcomes were the primary drivers of SWB fluctuations, without prediction error playing an important role. In a similar vein, some participants benefited from multiple forgetting factors (about 30%) or from the inclusion of extra parameters (about 22%).

The results are mixed: On the one hand, model M1 certainly performs well compared to the other eight models. However, we cannot conclude that it

Table 1. Results from the model comparison procedure.

Model no.	Number of parameters	Mean R^2	Median R^2	BIC	BIC-BIC$_{M1}$	%best
1	4	50	50	−1530	0	26
2	3	43	43	−1275	255	20
3	5	51	51	−1435	95	12
4	6	56	60	−1539	−9	18
5	3	21	15	−167	1363	2
6	6	31	23	−20	1510	0
7	21	55	55	1693	3223	0
8	5	52	53	−1523	7	8
9	5	53	57	−1573	−43	14

Notes: This table contains the number of parameters for each model, together with two measures of fit, namely the R^2 and the BIC values. The BIC values are summed over all 50 participants. In the second-to-last column, BIC values are related to the first model by subtracting the BIC of the first model with the BIC of the respective model with which it is compared, so that values greater than 0 indicate that M1 outperforms the other model (and vice-versa for values lower than 0). The last column reports the percentage of times each model has been selected as best model, based on a within-person BIC values.

unequivocally outperforms the other models (a conclusion further backed up by the analyses in the Supplementary Materials), as there are compelling individual differences in how momentary SWB can best be described.

Individual differences

Table 2 contains the correlations between the estimated model parameters of the original model and the personality trait scores. None of the correlations reached significance. A same conclusion is reached for the correlations between the individual parameters of the hierarchical model and personality trait scores (see Table 3). These results thus do not warrant conclusions about the relation between the model parameters and personality.

Discussion

The goals of this study were twofold. First, we aimed to replicate the behavioural results reported by Rutledge et al. (2014). In general, this study can be considered a successful replication. We specified several criteria to evaluate the success of replication and almost all criteria were met. Overall, the results suggest that fluctuations of an individual's SWB are (at least partly) driven by rewards and punishments, both when they are certain or uncertain. More specifically, a gain, a positive expected value, and a better than expected outcome reliably increased SWB, whereas a loss, a negative expected value, and a worse than expected scenario reliably decreased SWB. Furthermore, when participants chose a gamble, the deviation with the expected value was found to more strongly predict SWB than the expected value itself, suggesting a larger role for prediction error in the fluctuations of SWB. These results were further corroborated by the application of a hierarchical extension of the original model.

A notable difference between the results of our study and the results reported by Rutledge et al. (2014) resides in the estimated mean of the forgetting factor. More specifically, the forgetting factor was larger in our study, suggesting that for our participants the outcomes of previous trials had a larger influence on current SWB than for participants in the study of Rutledge et al. (2014). The possibility of this result was already suggested by Rutledge (personal communication, July 1, 2020) and can be explained by our choice to keep total earnings visible on screen throughout the experiment. This

Table 2. Correlations between the parameter estimates of the original model and personality trait scores.

	N	E	BIS	BAS	BAS: D	BAS: FS	BAS: RR
w_0	−0.25	0.19	−0.05	0.11	0.10	0.08	0.09
w_1 (CR)	0.17	0.11	0.17	0.14	0.01	0.18	0.16
w_2 (EV)	0.01	−0.02	−0.20	0.12	0.07	0.15	0.04
w_3 (RPE)	0.07	0.10	0.04	0.15	0.04	0.25	0.08
γ	−0.18	0.02	−0.15	−0.08	0.00	−0.09	−0.12
R^2	−0.02	0.22	0.05	0.25	0.26	0.18	0.14

Note: Additional to the parameter estimates, correlations between the fit of the model (R^2) and personality traits are reported. None of the correlations reached significance.

Table 3. Correlations between the parameter estimates of the hierarchical model and personality trait scores.

	N	E	BIS	BAS	BAS: D	BAS: FS	BAS: RR
w_0	−0.13	0.14	0.02	0.12	0.13	0.06	0.09
w_1 (CR)	0.16	0.07	0.15	0.01	−0.09	0.03	0.13
w_2 (EV)	−0.11	0.26	−0.22	0.18	0.06	0.22	0.16
w_3 (RPE)	0.03	0.05	−0.05	0.14	0.12	0.10	0.10
γ	−0.08	−0.05	0.01	−0.06	−0.01	0.04	−0.21
R^2	−0.04	0.07	−0.04	0.05	0.19	−0.05	−0.05

Note: None of the correlations reached significance.

makes it easier for participants to see the long-term consequences of their decisions, making it possible for earlier trials to influence current SWB more strongly than when total earnings are not shown (cf. Rutledge et al., 2014).

While the estimation-based replication criteria were met, we could not fully replicate the results from the model comparison. More specifically, we did not find that the model of Equation (1) described the data best compared to some competing models. This procedure, however, lent support to the notion that prediction error influences SWB to an important extent, as in the majority of cases such models were preferred over a much simpler, reward-based model. Furthermore, we found substantial individual differences in the preferred model, justifying an idiosyncratic approach to modelling fluctuations in SWB. Thus, despite the failure of replication, we still consider the conclusions reached by Rutledge et al. (2014) to be favoured by the evidence, be it with some nuance regarding individual differences in the drivers of SWB fluctuations.

As a second aim to our study, we tried to account for the individual differences in parameter estimates of the model. However, these individual differences could not be meaningfully related to personality traits. One reason for this null result may be inadequate power to reliably detect small correlations, such as the ones found in our study. Another reason may be that the parameter estimates are based on both rewards and punishments, instead of only on rewards or only on punishments. Therefore, it may not be possible to assess reactivity towards a specific type of event with the model, which is a necessary condition to be able to relate the parameter estimates to the studied personality traits. However, the parameters of the model can be interpreted as a measure of reactivity towards a certain versus an uncertain event. This leaves an open avenue for future research to identify other correlates for the parameter estimates of the model.

Despite the (mostly) successful replication, a cautionary note is in place. The model seems to capture momentary SWB well, but it can only do so in the context of the used paradigm. The model parameters are inherently linked to a gambling paradigm in which there is a choice between a certain and an uncertain outcome. This kind of choice may not reflect real-life decisions, where the outcome of each choice may always be uncertain. This limits the generalizability of the model and its usefulness to examine the outcomes of real-world decision processes and how they elicit changes in SWB. Future research should therefore try to identify a more general computational model that captures SWB, which could then prove useful for multiple paradigms (e.g. Bennett et al., 2020, August 3).

Despite these limitations, we believe that the model proposed by Rutledge et al. (2014) is a useful tool to study and understand how rewards and punishments influence momentary SWB in the laboratory. Understanding the nature of the individual differences in processing these rewards and punishments, however, remains a challenge for future research.

Notes

1. Rutledge et al. (2014) also measured task-dependent neural activity in the ventral striatum when participants performed this task. However, this neurological part of the study is not part of our replication study and we will thus leave it undiscussed.
2. In this replication paper, we not only adopt the methods, but also the terminology of the original study.
3. Note that we do not consider this to be a formal analysis procedure. Instead, we evaluate the estimates of the different parameters to reach our conclusions.
4. When comparing the current paper to the preregistration, it may be helpful to know that different parts of the method section from the preregistration were relabelled and reorganised. More specifically, the subsection *Sampling plan* was changed to *Participants* (completed with demographic information), *Individual difference variables* was changed to *Materials*, and *Experimental task* was placed under *Procedure*. Furthermore, the *Analysis* subsection in this paper consists of the subsection

Confirmatory analysis plan of the preregistration, and the parts of the *Introduction* that contained the mathematical information about the computational model and general information about the analysis plan for the individual difference variables.

5. The actual term that should be used here is "the estimated posterior mean of the population mean". We opted for a shorter version to increase readability.

Acknowledgements

We would like to thank Robb Rutledge for sharing the code of his experiment (now available on Github: https://www.github.com/RutledgeLab/Rutledge-happiness-task) and analyses, and for providing useful comments on an earlier draft of this paper.

Disclosure statement

No potential conflict of interest was reported by the author(s).

Funding

This work was supported by the Research Fund of KU Leuven under Grant C14/19/054 and C14/17/047; and Fonds Wetenschappelijk Onderzoek under Grant G073317N and G074219N. The funders had no role in study design, data collection and analysis, decision to publish or preparation of the manuscript.

Ethics

This study was approved by the local ethics committee (Social and Societal Ethics Committee at the KU Leuven). All participants had to complete an informed consent before their participation to the experiment.

Data availability statement

As specified in the article, the preregistration to this article can be found on: https://osf.io/krhyz/. Raw and preprocessed data, and R scripts containing the analyses can be found on OSF using the following link: https://osf.io/9g2zw/.

Authors' contributions

Levi Devos ran this study for his master's thesis, under supervision of Peter Kuppens and Francis Tuerlinckx. Sebastiaan Pessers translated the estimation code from Matlab to R and created estimation code for the hierarchical model. Niels Vanhasbroeck confirmed the analyses, rewrote the analysis code, and wrote the first version of this manuscript. All authors, with exception of Sebastiaan Pessers, contributed to all subsequent versions of the manuscript and approved the final version.

ORCID

Niels Vanhasbroeck http://orcid.org/0000-0002-0056-3183
Peter Kuppens http://orcid.org/0000-0002-2363-2356
Wolf Vanpaemel http://orcid.org/0000-0002-5855-3885
Agnes Moors http://orcid.org/0000-0002-5137-557X
Francis Tuerlinckx http://orcid.org/0000-0002-1775-7654

References

Bennett, D., Davidson, G., & Niv, Y. (2020, August 3). *A model of mood as integrated advantage*. https://10.31234/osf.io/y2vzu.

Brickman, P., & Campbell, D. T. (1971). Hedonic relativism and planning the good society. In M. H. Appley (Ed.), *Adaptation-level theory* (pp. 287–305). Academic Press.

Carpenter, B., Gelman, A., Hoffman, M. D., Lee, D., Goodrich, B., Betancourt, M., Brubaker, M., Guo, J., Li, P., & Riddell, A. (2017). Stan: A probabilistic programming language. *Journal of Statistical Software*, *76*(1), https://doi.org/10.18637/jss.v076.i01

Carver, C. S., & White, T. L. (1994). Behavioral inhibition, behavioral activation, and affective responses to impending reward and punishment: The BIS/BAS scales. *Journal of Personality and Social Psychology*, *67*(2), 319–333. https://doi.org/10.1037/0022-3514.67.2.319

Cohn, M. A., Fredrickson, B. L., Brown, S., Mikels, J. A., & Conway, A. M. (2009). Happiness unpacked: Positive emotions increase life satisfaction by building resilience. *Emotion*, *9*(3), 361–368. https://doi.org/10.1037/a0015952

Dauvier, B., Pavani, J.-B., Le Vigouroux, S., Kop, J.-L., & Congard, A. (2019). The interactive effect of neuroticism and extraversion on the daily variability of affective states. *Journal of Research in Personality*, *78*, 1–15. https://doi.org/10.1016/j.jrp.2018.10.007

Diener, E., Lucas, R. E., & Scollon, C. N. (2006). Beyond the hedonic treadmill: Revisiting the adaptation theory of well-being. *American Psychologist*, *61*(4), 305–314. https://doi.org/10.1037/0003-066X.61.4.305

Diener, E., Oishi, S., & Lucas, R. E. (2003). Personality, culture, and subjective well-being. *Annual Review of Psychology*, *54*(1), 403–425. https://doi.org/10.1146/annurev.psych.54.101601.145056

Diener, E., Scollon, C. N., & Lucas, R. E. (2009). The evolving concept of subjective well-being: The multifaceted nature of happiness. In E. Diener (Ed.), *Assessing well-being: The collected works of Ed diener* (pp. 67–100). Springer.

Gelman, A. (2006a). Multilevel (hierarchical) modeling: What it can and cannot do. *Technometrics*, *48*(3), 432–435. https://doi.org/10.1198/004017005000000661

Gelman, A., & Hill, J. (2006b). *Data analysis using regression and multilevel/hierarchical models*. Cambridge University Press.

Gray, J. A. (1972). The psychophysiological nature of introversion-extraversion: A modification of eysenck's theory. In V. D. Nebylitsyn & J. A. Gray (Eds.), *The biological bases of individual behaviour* (pp. 182–205). Academic Press.

Kruschke, J. K. (2015). *Doing Bayesian data analysis: A tutorial with R, JAGS, and stan*. Academic Press.

Lawrence, E. M., Rogers, R. G., & Wadsworth, T. (2015). Happiness and longevity in the United States. *Social Sciences & Medicine*,

145, 115–119. https://doi.org/10.1016/j.socscimed.2015.09.020

LeBel, E. P., McCarthy, R. J., Earp, B. D., Elson, M., & Vanpaemel, W. (2018). A unified framework to quantify the credibility of scientific findings. *Advances in Methods and Practices in Psychological Science*, *1*(3), 389–402. https://doi.org/10.1177/2515245918787489

Lyobomirsky, S., & Layous, K. (2013). How do simple positive activities increase well-being? *Current Directions in Psychological Science*, *22*(1), 57–62. https://doi.org/10.1177/0963721412469808

Lyobomirsky, S., King, L., & Diener, E. (2005). The benefits of frequent positive affect: Does happiness lead to success? *Psychological Bulletin*, *131*(6), 803–855. https://doi.org/10.1037/0033-2909.131.6.803

McCrae, R. R., & Costa, P. T. (2004). A contemplated revision of the NEO Five-Factor inventory. *Personality and Individual Differences*, *36*(3), 587–596. https://doi.org/10.1016/S0191-8869(03)00118-1

Oishi, S., Diener, E., Choi, D.-W., Kim-Prieto, C., & Choi, I. (2007). The dynamics of daily events and well-being across cultures: When less is more. *Journal of Personality and Social Psychology*, *93*(4), 685–698. https://doi.org/10.1037/0022-3514.93.4.685

Papousek, I., Nauschnegg, K., Paechter, M., Lackner, H. K., Goswami, N., & Schulter, G. (2010). Trait and state positive affect and cardiovascular recovery from experimental academic stress. *Biological Psychology*, *83*(2), 108–115. https://doi.org/10.1016/j_biopsycho.2009.11.008

Rutledge, R. B., Skandali, N., Dayan, P., & Dolan, R. J. (2014). A computational and neural model of momentary subjective well-being. *Proceedings of the National Academy of Sciences*, *111*(33), 12252–12257. https://doi.org/10.1073/pnas.1407535111

Schwarz, G. (1978). Estimating the dimension of a model. *The Annals of Statistics*, *6*(2), 61–464. https://doi.org/10.1214/aos/1176344136

Stan Development Team. (2018). *RStan: The R interface to Stan*. http://mc-stan.org.

Thake, J., & Zelenski, J. M. (2013). Neuroticism, BIS, and reactivity to discrete negative mood inductions. *Personality and Individual Differences*, *54*(2), 208–213. https://doi.org/10.1016/j.paid.2012.08.041

Veenhoven, R. (2008). Healthy happiness: Effects of happiness on physical health and consequences for preventive health care. *Journal of Happiness Studies*, *9*(3), 449–469. https://doi.org/10.1007/s10908-006-9042-1