Frontiers in Psychology

TYPE Original Research
PUBLISHED 05 April 2024
DOI 10.3389/fpsyg.2024.1254564

OPEN ACCESS

EDITED BY
Pietro Cipresso,
University of Turin, Italy

REVIEWED BY
Amanda Claire Marshall,
Ludwig Maximilian University of Munich,
Germany
Francesca Borghesi,
University of Turin, Italy

*CORRESPONDENCE
Katja Brand
✉ brand@biomed.ee.ethz.ch

RECEIVED 24 July 2023
ACCEPTED 18 March 2024
PUBLISHED 05 April 2024

CITATION
Brand K, Wise T, Hess AJ, Russell BR, Stephan KE and Harrison OK (2024) Incorporating uncertainty within dynamic interoceptive learning. *Front. Psychol.* 15:1254564.
doi: 10.3389/fpsyg.2024.1254564

# Incorporating uncertainty within dynamic interoceptive learning

Katja Brand[1,2]*, Toby Wise[3], Alexander J. Hess[1], Bruce R. Russell[4], Klaas E. Stephan[1,5] and Olivia K. Harrison[1,2,6]

[1]Translational Neuromodeling Unit, Institute for Biomedical Engineering, University of Zurich and ETH Zurich, Zurich, Switzerland, [2]Department of Psychology, University of Otago, Dunedin, New Zealand, [3]King's College London, Institute of Psychiatry, Psychology and Neuroscience, London, United Kingdom, [4]School of Pharmacy, University of Otago, Dunedin, New Zealand, [5]Max Planck Institute for Metabolism Research, Cologne, Germany, [6]Nuffield Department of Clinical Neurosciences, University of Oxford, Oxford, United Kingdom

**Introduction:** Interoception, the perception of the internal state of the body, has been shown to be closely linked to emotions and mental health. Of particular interest are interoceptive learning processes that capture associations between environmental cues and body signals as a basis for making homeostatically relevant predictions about the future. One method of measuring respiratory interoceptive learning that has shown promising results is the Breathing Learning Task (BLT). While the original BLT required binary predictions regarding the presence or absence of an upcoming inspiratory resistance, here we extended this paradigm to capture continuous measures of prediction (un)certainty.

**Methods:** Sixteen healthy participants completed the continuous version of the BLT, where they were asked to predict the likelihood of breathing resistances on a continuous scale from 0.0 to 10.0. In order to explain participants' responses, a Rescorla-Wagner model of associative learning was combined with suitable observation models for continuous or binary predictions, respectively. For validation, we compared both models against corresponding null models and examined the correlation between observed and modeled predictions. The model was additionally extended to test whether learning rates differed according to stimuli valence. Finally, summary measures of prediction certainty as well as model estimates for learning rates were considered against interoceptive and mental health questionnaire measures.

**Results:** Our results demonstrated that the continuous model fits closely captured participant behavior using empirical data, and the binarised predictions showed excellent replicability compared to previously collected data. However, the model extension indicated that there were no significant differences between learning rates for negative (i.e. breathing resistance) and positive (i.e. no breathing resistance) stimuli. Finally, significant correlations were found between fatigue severity and both prediction certainty and learning rate, as well as between anxiety sensitivity and prediction certainty.

**Discussion:** These results demonstrate the utility of gathering enriched continuous prediction data in interoceptive learning tasks, and suggest that the updated BLT is a promising paradigm for future investigations into interoceptive learning and potential links to mental health.

KEYWORDS

**interoception, learning, breathing, inspiratory resistance, mental health**

# 1 Introduction

Perception goes beyond the path of registration of sensations but involves the active interpretation of sensory inputs. This interpretation is shaped by numerous cognitive factors, such as prior knowledge or expectations, as well as attention. While exteroception refers to the perception of the external environment through the traditional senses of touch, sight, hearing, taste and smell, interoception refers to the perception of the internal state of the body (Khalsa et al., 2018). Interoception includes both conscious and subconscious processes; investigating these processes is critical for understanding brain-body interactions. While one of the main roles of interoception is to drive actions to maintain homeostasis (Pezzulo et al., 2015; Stephan et al., 2016; Petzschner et al., 2017; Quadt et al., 2018), interoception is also thought to play an important role in emotional regulation (Barrett et al., 2004; Critchley et al., 2004; Füstös et al., 2013). Importantly, interoceptive dysfunction has been implicated in a range of psychological disorders such as depression, anxiety and eating disorders (Paulus and Stein, 2010; Khalsa et al., 2018), and is a rapidly expanding field of research (Brewer et al., 2021).

Computational theories of perception suggest that the brain acts as an "inference machine" that uses its probabilistic representations (beliefs) about the state of the world to make predictions about future incoming sensory stimuli (Rao and Ballard, 1999; Friston, 2005). The discrepancy between the actual and predicted stimuli (the prediction error) is then used to continuously update these beliefs according to Bayesian principles (Friston, 2005). Numerous experimental studies have provided empirical evidence that exteroceptive processes such as sight and hearing operate in this manner (Chennu et al., 2013; Lieder et al., 2013; Kok and de Lange, 2014; Stefanics et al., 2018), and these theories are now being extended to interoception in order to explain how the brain creates a predictive model of the internal state (Seth et al., 2012; Gu et al., 2013; Barrett and Simmons, 2015; Pezzulo et al., 2015; Critchley and Garfinkel, 2017). Interoceptive learning thus refers to the updating of beliefs about the internal (bodily) state based on predictions made (and errors received) regarding interoceptive stimuli.

Altered interoceptive learning has been proposed to underpin aspects of psychopathology. Paulus et al. (2019) hypothesized that depression and anxiety are potentially linked to two main dysfunctions in the interoceptive processing pathway: overly strong expectations, which shape the processing of interoceptive stimuli; and difficulty in updating predictions to reflect changes in the external or internal state, which may involve faulty prediction error signaling. There is already evidence that this is the case for exteroceptive processing, with a previous study finding that individuals with anxiety have altered neural activity (as measured by electroencephalography/EEG) when processing predictions and prediction errors in a visual reward-learning task (Hein and Ruiz, 2022). Regarding interoceptive learning, previous work by Harrison et al. (2021) developed an interoceptive learning task that was performed during functional magnetic resonance imaging (fMRI), using inspiratory resistances (which make it harder to breathe in) as an interoceptive stimulus (Rieger et al., 2020; Frässle et al., 2021). The results indicated a link between heightened anxiety and alterations in the processing of interoceptive breathing predictions, finding that more anxious individuals showed altered activity in the anterior insula related to prediction certainty compared to less anxious individuals (Harrison et al., 2021). In this study, binary measures of interoceptive predictions were analyzed by an established associative learning model (a Rescorla-Wagner model; Rescorla et al., 1972), to determine a learning rate for each participant as well as the corresponding trajectories for predictions and prediction errors. While the breathing learning task (BLT) used in this previous study shows great potential for investigating changes in interoceptive learning in different mental health conditions, one limitation is its reliance on model estimates to quantify prediction certainty from binary responses, rather than including a direct measure thereof.

The current study builds on this interoceptive breathing learning task (BLT) (Harrison et al., 2021) by incorporating direct measures of prediction certainty in place of binary predictions. Past research has suggested that mental health disorders such as anxiety may be associated with an altered response to uncertainty (Grupe and Nitschke, 2013). Therefore, to more accurately capture measures of (un)certainty, the BLT from Harrison et al. (2021) was modified to elicit continuous rather than binary prediction data, thus incorporating a direct measure of certainty surrounding predictions. Additionally, there is evidence that individuals learn more quickly (i.e. adapt their predictions more rapidly) in response to negative outcomes (Khdour et al., 2016; Aylward et al., 2019). Therefore, a model extension that incorporated stimuli valence (i.e. the presence of a breathing resistance [negative stimulus] or the absence of a breathing resistance [positive stimulus]) was also employed to examine whether a single or two separate learning rates would better explain the behavioral data. The modified BLT and learning model were then tested on a sample of 16 healthy participants, and the results were compared to the data collected by Harrison et al. (2021). Additionally, the estimated interoceptive learning parameters were compared to measures of anxiety, depression, affect and subjective interoception in an exploratory analysis. Through the incorporation of a direct measure of prediction certainty in the task and model design, the current study provides richer information regarding interoceptive learning and its relationship to mental health.

# 2 Materials and methods

## 2.1 BLT equipment setup

In order to deliver the breathing resistances to participants during the BLT, an inspiratory resistance circuit was utilized (Rieger et al., 2020). Participants were fitted with a silicone facemask that was adjusted to make a tight seal around mouth and nose. This was then connected to a single-use bacterial and viral filter within the circuit, and inspiratory resistances were induced via an automatically controlled solenoid valve. This valve allowed air to be drawn into the circuit either directly from the environment (with no resistance) or through a PowerBreathe device, which was set to deliver the predetermined amount of resistance (30% of a participant's maximal inspiratory pressure). The circuit also included a pressure sampling line and spirometer, allowing for continuous measurements of inspiratory pressure and flow to be

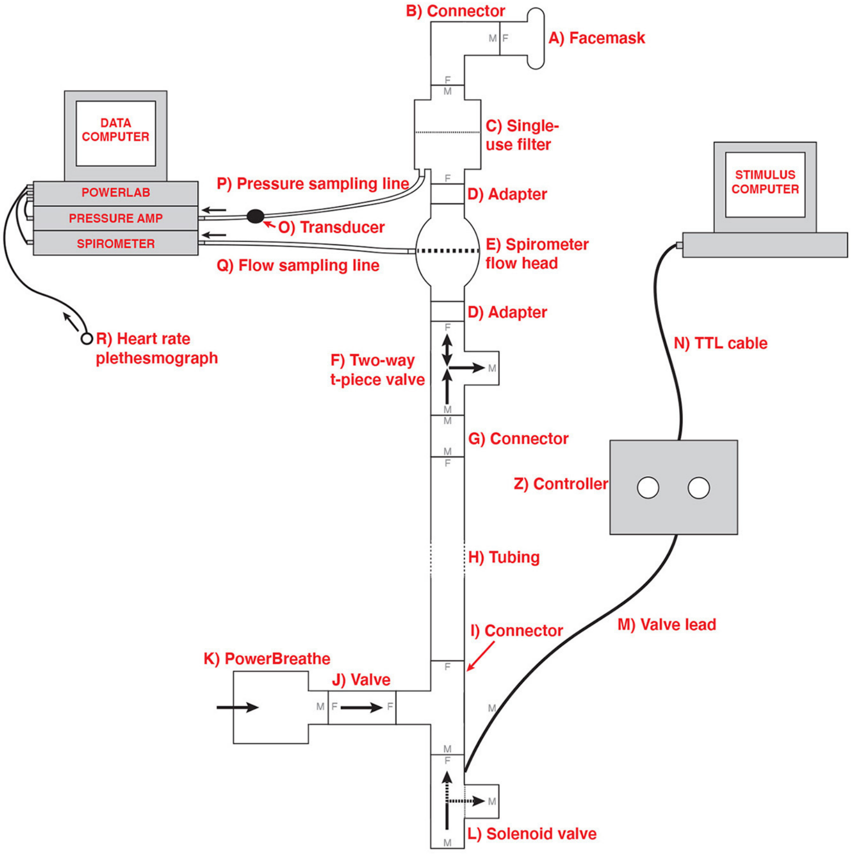


FIGURE 1
Schematic of the inspiratory resistance circuit used. Participants were connected to the circuit via a facemask covering their mouth and nose which was attached to a bacterial and viral filter (labeled C). A sampling line to a pressure transducer (labeled O) and amplifier was used to record inspiratory pressure, and a spirometer (labeled E) was used to measure inspiratory flow. Inspiratory resistances were induced automatically via the stimulus computer, controller box and solenoid valve (labeled L) to redirect unobstructed air from the environment to the PowerBreathe device (labeled K), which provided a set magnitude of resistance. Figure adapted from Rieger et al. (2020) under a CC-BY license.

taken throughout the task. For a diagram of the full circuit used see Figure 1.

## 2.2 Participants and recruitment

In order to test the continuous response version of the BLT, data was gathered from 16 healthy volunteers. Participants were aged 19–42 years (mean age: 23y; 4M, 12F), and were pre-screened according to the following criteria:

- Aged 18–45
- Regularly exercising no more than once per week
- Non-smoker or light smoker (smoking or vaping once per week or less)
- Not on any regular medication at time of study (except the oral contraceptive pill)
- Full color-vision
- Not suffering from any chronic medical conditions, including current or past history of brain injury or breathing disorder
- No past or current diagnoses of schizophrenia, bipolar disorder, drug addiction, or psychosis
- Not pregnant or breastfeeding

Participants were recruited from the community using study advertisements. All participants signed a written, informed consent, and the study was approved by the New Zealand Health and Disability Ethics Committee (HDEC) (Ethics approval 20/CEN/168).

Data from a separate group of eight participants (4M, 4F) were used to determine model priors. These participants had completed an earlier version of the BLT using binarised responses and their data had initially been used by Harrison et al. (2021). All participants signed a written, informed consent, and the study was approved by the Cantonal Ethics Committee Zurich (Ethics approval BASEC-No. 2017-02330).

## 2.3 Procedure

Participants who were selected for the study following online pre-screening were asked to complete a series of questionnaires (details in Section 2.3.1) followed by BLT (see Section 2.3.2). Both tasks required 30–45 min to complete.

### 2.3.1 Questionnaires

Following online pre-screening and informed consent, participants were firstly asked to fill in a number of questionnaires on the lab computer. The questionnaires presented to the participants were designed to capture subjective affective measures as well as general and breathing-specific interoceptive beliefs.

Affective qualities that were measured included state anxiety (measured by the Spielberger Trait Anxiety Inventory; STAI-S; Spielberger et al., 1970), symptoms of generalized anxiety disorder (Generalized Anxiety Disorder Questionnaire; GAD-7; Spitzer et al., 2006), anxiety sensitivity (Anxiety Sensitivity Index; ASI-3; Taylor et al., 2007), symptoms of depression (Center for Epidemiologic Studies Depression Scale; CES-D; Radloff, 1977), as well as general positive and negative affect (Positive Affect Negative Affect Schedule; PANAS; Watson et al., 1988).

Self-reported interoceptive awareness was measured by the Multidimensional Assessment of Interoceptive Awareness Questionnaire (MAIA; Mehling et al., 2012). Two further questionnaires measured the tendency to catastrophise in response to breathlessness (Pain Catastrophising Scale, adapted to replace pain with breathlessness; PCS-B; Sullivan et al., 1995), as well as awareness and vigilance surrounding breathlessness (Pain Vigilance and Awareness Questionnaire, again replacing pain with breathlessness; PVAQ-B; McCracken, 1997), in line with previous research (Herigstad et al., 2017; Harrison et al., 2021).

Additional facets related to mental health were measured by the following questionnaires: General Self Efficacy scale (GSE; Schwarzer et al., 1997) which measured self-efficacy, Connor Davidson Resilience Scale (CD-RISC; Connor and Davidson, 2003) for resilience, and the Fatigue Severity Scale (FSS; Krupp et al., 1989) for fatigue. Finally, trait anxiety was measured by the Spielberger Trait Anxiety Inventory (STAI-T; Spielberger et al., 1970), which participants filled out during the online pre-screening process.

### 2.3.2 BLT procedure

After filling out the questionnaires, participants completed the BLT. In order to set an appropriate level of breathing resistance for this task, the maximum inspiratory pressure (MIP) was first measured and recorded for each participant using a PowerBreathe device (PowerBreathe International Ltd, Warwickshire, UK). The resistance magnitude for the BLT was then set to 30% of the participant's MIP. Once the breathing resistance had been calibrated to the participant, they were fitted with a breathing mask (Hans Rudolph, Kansas City, MO, United States) which was connected to the breathing circuit (see Section 2.1 for details), and tested to ensure that they could feel the resistances. In two cases, the participants were unable to perceive the resistances when operating at their normal tidal volume, and here the resistance was increased to 50% of their MIP to accommodate for this. Written instructions for the BLT were given to the participant to read (see Supplementary Figure S1) and these were repeated verbally, as well as allowing the participant to ask questions to ensure they understood the task. Participants were also given a practice version of the task with six trials before beginning the actual task.

Participants were informed that the goal of the BLT was to measure how participants would learn to predict upcoming breathing resistances, based on their previous associations with visual cues. During the task, participants were asked how certain they were that there would be an inspiratory resistance, given the presentation of one of two visual cues (see Figure 2). The visual cue was presented together with the prompt to make a prediction about the upcoming stimulus phase. Participants were told beforehand that one of the cues had an 80% chance of being followed by a resistance, while the other cue had a 20% chance of being followed by a resistance. They were also told that the pairings of the images with the probabilities could swap during the task, so that the cue previously associated with an 80% chance of resistance now predicted a 20% chance of resistance and vice versa. Participants were not informed of which cue started with which probability, or when the switches would occur.

For each trial, in the cue/predict phase, one of the visual cues was displayed on a computer screen for five seconds, along with a prompt asking the participant to predict the likelihood of a resistance occurring in the upcoming stimulus phase. Participants entered their prediction (along with their certainty in the prediction) by using arrow keys on a keyboard to move a slider on a scale from “definitely yes” to “definitely no”. Immediately following this, a circle was shown on the screen for five seconds, during which time the breathing resistance occurred on resistance trials, and no resistance (i.e. normal breathing) occurred on all other trials (stimulus phase). Participants were able to easily assess whether their prediction was correct due to the unambiguous nature of the resistance stimulus. Following this period, participants were asked whether or not a resistance had occurred (report phase). This report served only to validate that participants were able to identify when breathing resistance occurred. Participants were then given a rest period of between seven and nine seconds before the next trial began. The structure of this task was developed specifically for feasibility when using breathing stimuli, and was tested with both synthetic and pilot data (Harrison et al., 2021). The task protocol used in the current study was adapted from that used by Harrison et al. (2021) to collect continuous rather than binary prediction data. The main modification that was made to the BLT for the current study was the change from a binary prediction (yes/no options) to a continuous prediction (slider from definitely yes to definitely no). Additional modifications included the changing of the report phase from a slider rating of difficulty to a binary yes/no question regarding the presence of a resistance, to simplify this question for participants. Timings of the cue/predict and report phases were modified to accommodate these changes. Finally, the pause before stimulus presentation was removed, as this served to meet requirements of fMRI data analysis, which did not apply in the present study. Figure 2 provides an overview of the structure of each trial and the alterations that were made to gather continuous response data. The code for the updated BLT which includes an option for collecting continuous response data is available in the open-source TAPAS collection (Frässle et al., 2021) (https://www.translationalneuromodeling.org/tapas/).

In total, the task consisted of 80 trials, which took approximately 30 minutes to complete. During the task, physiological recordings of inspiratory pressure, breathing rate,

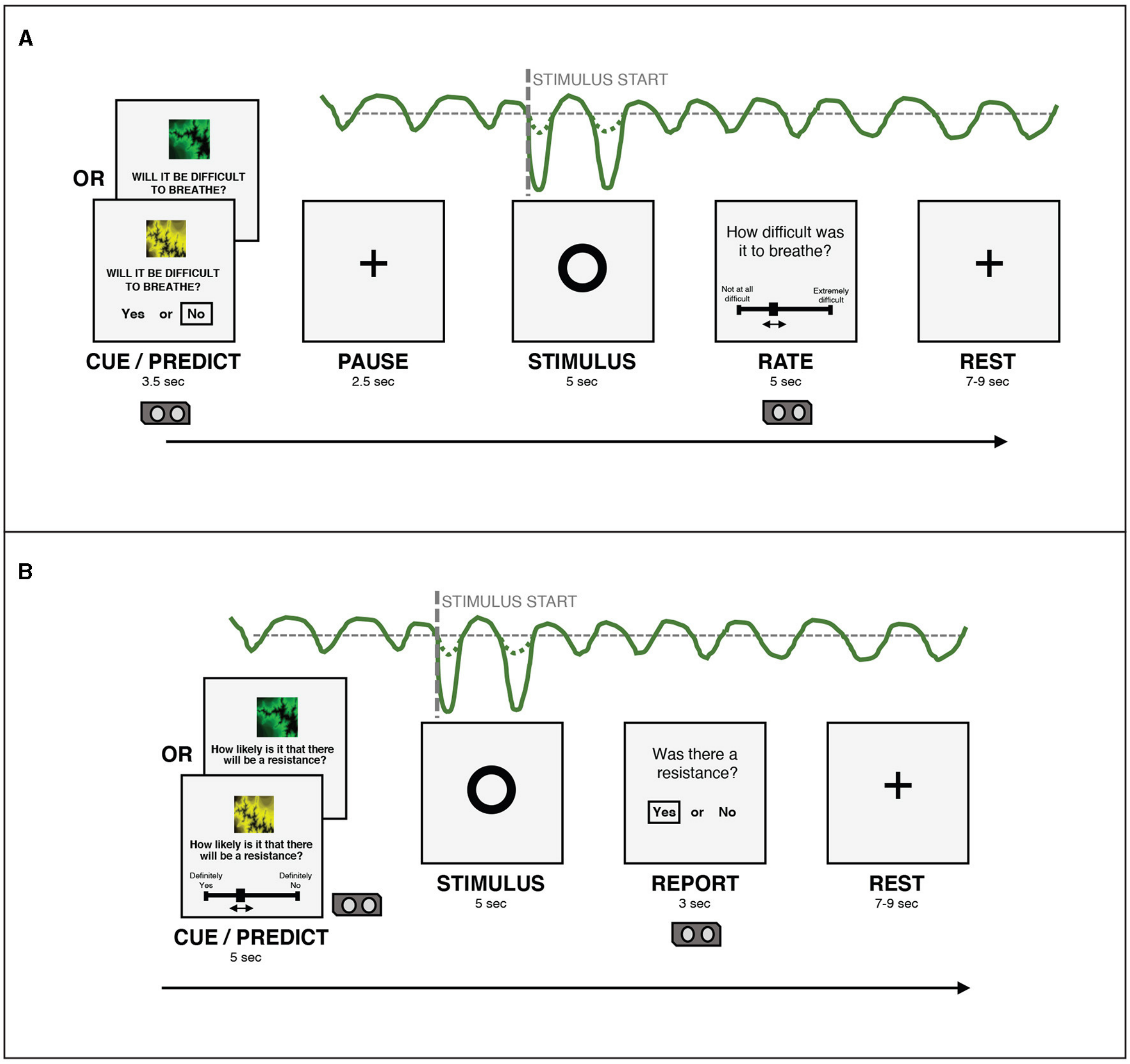


FIGURE 2
**(A)** An overview of the structure of a single trial in the version of the BLT used in Harrison et al. (2021). **(B)** An overview of the structure of a single trial in the version of the BLT used for the current study. The main changes are that the cue/prediction phase was changed to allow continuous predictions via a slider (with a longer timeframe for this phase to compensate), while the report phase was changed to a binary yes/no response to simplify this decision for participants. Additionally, the pause before the stimulus was removed after lengthening the cue/prediction phase. Figure adapted from Harrison et al. (2021) under a CC-BY license.

breathing volume and heart rate were taken using a spirometer and pulse monitor, connected to a PowerLab and recorded using LabChart 8 software (ADInstruments, Dunedin, New Zealand). Participants also wore headphones playing pink noise throughout the task.

The initial pairing of cue-to-resistance was counter-balanced across participants (such that each cue was first paired with an 80% chance of resistance for half of the participants), as well as the position of the "definitely yes" and "definitely no" anchors on the left or right of the screen. The initial pairing was always held constant for the first 30 trials before the pairings were switched (i.e. the cue initially paired with 20% chance of resistance was now paired with 80% chance of resistance and vice versa). The pairings were then switched three more times during the remaining 50 trials at shorter intervals (12-13 trials), for a total of four switches of cue-resistance pairings throughout the task. The number of trials between each switch was held constant for all participants. This is the same protocol as was used and validated in the study by Harrison et al. (2021).

## 2.4 Data processing

Questionnaires were scored according to their respective manuals, with a summary of the relevant scores presented in Section 3.3. For the BLT, data was first checked for missed trials. Data from one participant was excluded from further analysis due to missing more than 10 trials (as predetermined in the analysis plan which can be viewed at https://github.com/IMAGEotago/Katja-BLT-analysisPlan). Next, each participant's average certainty was determined by taking the absolute value of the difference between their response and 0.5 [with responses being values between 0.0 (definitely no resistance) and 1.0 (definitely resistance) and 0.5 thus representing complete indecision] for each trial and averaging across all trials. For the binary model, predictions were then binarised, with each value above 0.5 becoming 1.0, and each value below 0.5 becoming 0.0 (values at exactly 0.5 were treated as missed trials for the binary model). Finally, the outcomes of each trial were adjusted to be in "contingency space": i.e., any trial where cue 1 was paired with a resistance and cue 2 with no resistance was coded as 1, and any trial where the cue-outcome pairing was reversed was coded as 0, as previously described (Iglesias et al., 2013; Harrison et al., 2021). Notably, contingency space coding does not depend on which binary value is assigned to coupled cue-outcome pairs and is equivalent to the case of running two models in parallel, one for each outcome [for details, see Supplementary material to Iglesias et al. (2013)]. This type of coding was possible because of the fixed coupling of contingencies in our task (see Section 2.3.2 for more detail), which the participants were made explicitly aware of. However, this method does assume that learning is the same for resistance and no resistance trials. To address this assumption, we extended our model to separate the learning parameters for resistance and no resistance trials (see Section 2.6.2).

## 2.5 Task validation

In order to validate that participants were completing the task as expected, the proportion of correct (binarised) responses on each trial across all participants was compared with data from Harrison et al. (2021), which investigated a larger cohort of participants using the binary prediction version of the BLT. This allowed us to verify that overall, participants understood the task and behaved in line with previous findings from a larger cohort.

## 2.6 Associative learning model

The previous study by Harrison et al. (2021) utilized a classical associative learning Rescorla-Wagner model (Rescorla et al., 1972) to analyse responses from the BLT. Model comparison was initially performed with two alternative hierarchical Gaussian filter (HGF; Mathys et al., 2011) models that included both a dynamic learning rate (i.e., a two-level HGF), and a measure of volatility (i.e. a three-level HGF). However, as there was no clear winning model when fit to the data, the simplest model was chosen as per the pre-specified analysis plan of Harrison et al. (2021). Therefore, here we chose the same Rescorla-Wagner model for comparable results. In this model, the predicted outcome for a given trial $v_{t+1}$ is based on the predicted outcome for the previous trial $v_t$ as well as the prediction error for the previous trial $\delta_t$ scaled by the learning rate of the participant $\alpha$:

$$v_{t+1} = v_t + \alpha\delta_t \tag{1}$$

where the prediction error is the difference between the actual outcome $o_t$ and the predicted outcome $v_t$ for the previous trial:

$$\delta_t = o_t - v_t \tag{2}$$

Previous research has employed similar models on binary BLT data (see Harrison et al., 2021), where participants had two choices (Resistance/No Resistance) when asked to predict the outcome of each trial. However, in the current version of the BLT, participants were asked to predict the outcome of each trial on a sliding scale with values from 0.0 (Definitely No Resistance) to 1.0 (Definitely Resistance). These continuous data therefore include a direct measure of the participant's certainty in their predictions, rather than requiring this to be inferred from fitted model trajectories. In order to apply the model to the participants' observed responses, the Rescorla-Wagner learning model (Eq. 1) was paired with two different observation models - one using binarised responses (for comparisons with previous versions of the task) and the other continuous responses (henceforth referred to as the binary model and the continuous model respectively). In either case, the participant's responses (i.e., binary or continuous predictions about trial outcome) and outcomes (binary: absence or presence of resistance) from each trial were modeled by estimating a subject-specific learning rate ($\alpha$) for the participant (see Eq. 1). The specifics of each observation model are discussed in Section 2.6.1.

### 2.6.1 Observation models

The observation model translates the predicted outcome, as provided by the learning model, into predicted behavior (i.e., moving the slider to a certain position for the next trial in the case of continuous data, or a binary decision for the binarised data).

#### 2.6.1.1 Binary model

The binary model uses a softmax function as the observation model, which translates the estimates obtained by the learning model into the probability of choosing a given action - in this case, deciding whether a given stimulus predicts resistance or no resistance. The softmax function used by the binary model can be represented as follows (Equation 3):

$$p(y_t|v_t, \beta) = \frac{e^{\beta v_t}}{e^{\beta v_t} + e^{\beta(1-v_t)}} \tag{3}$$

Where $p(y_t|v_t, \beta)$ represents the estimated probability of choosing a given binary prediction response $y_t$ given the predicted outcome $v_t$, $\beta$ is a static parameter that determines the steepness of the gradient of the softmax function, and $v_t$ is the value calculated by the learning model. The parameter $\beta$ can be altered to represent more or less noise in the decision-making process, with a higher $\beta$ resulting in a steeper softmax gradient, and thus more deterministic behavior.

#### 2.6.1.2 Continuous model

As the continuous model allows for predictions to occur on a continuous scale rather than having to make a binary decision, the value obtained from the learning model for a given trial $v_t$ is represented by the observed continuous prediction response $y_t$. The likelihood of the data is derived using a beta distribution, given that the responses lie in the range [0,1], and that the beta distribution is an adequate model for continuous bounded data, such as proportions or probabilities. Here, we use a formulation which reparameterises the usual shape parameters of the beta distribution in terms of parameters for mean and dispersion (Paolino, 2001). Moreover, we make dispersion a group parameter, $\phi$, where higher $\phi$ represents less noise.

Specifically, this is implemented by re-parameterising the shape parameters of the standard beta distribution (specified by $a$ and $b$, both vectors of length $n_{trials} \cdot n_{subjects}$) in terms of mean $\mu$ (a vector of length $n_{trials} \cdot n_{subjects}$, which contains the values obtained from the learning model (see Equation 1) for each subject and trial) and a scalar dispersion parameter $\phi$ (which is constant across subjects and trials), as follows (Equations 4, 5) (Paolino, 2001):

$$a = \mu \cdot \phi \tag{4}$$

$$b = (1 - \mu) \cdot \phi \tag{5}$$

As a result, large values of $\phi$ result in a tighter distribution while smaller values result in a wider distribution. In other words, less consistent (i.e., noisier) responses across subjects are reflected by a smaller value of $\phi$. Here we estimate a single value of $\phi$ across all subjects, though we note that subject-specific values could in principle be estimated to allow for differences in response consistency across subjects.

### 2.6.2 Dual learning rate model

To investigate whether there were learning differences between positive and negative valence stimuli, a variation of the model was created that used an altered version of the Rescorla-Wagner algorithm. This version is equivalent to Equations 1, 2, except that it contains two learning rates: $\alpha_p$ and $\alpha_n$. Which learning rate is used for the update on a given trial depends on the stimulus type $s$ that occurred during that trial - no breathing resistance is considered a positive valence stimulus (represented as $s_t = 0$) while breathing resistance is considered a negative valence stimulus (represented as $s_t = 1$). The state equation for the dual learning rate model is thus as follows (Equation 6):

$$v_{t+1} = \begin{cases} v_t + \alpha_p \delta_t & \text{if } s_t = 0 \\ v_t + \alpha_n \delta_t & \text{if } s_t = 1 \end{cases} \tag{6}$$

where the predicted error $\delta_t$ is calculated as in Equation 2.

## 2.7 Model testing

### 2.7.1 Parameter estimation and prior selection

Both models used maximum a posteriori (MAP) methods to obtain parameter estimates (using single-start optimisation and the L-BFGS-B algorithm; Byrd et al., 1995; Zhu et al., 1997), with the following priors: for $\alpha$ (as well as for $\alpha_p$ and $\alpha_n$ for the dual learning rate model) the prior mean was 0.34 and variance was 0.88 and for $\beta$ the prior mean was 4.21 and variance was 1.75 (all specified in native space using a Gaussian distribution, with $\alpha$ values bounded between 0.0 and 1.0). These prior distributions were calculated using the binary model to obtain Maximum Likelihood Estimates (MLE) from data of a separate group of eight pilot participants, originally used for the Harrison et al. (2021) study. The initial value of $v$ (the predicted outcome) was fixed at 0.5 (representing complete uncertainty of the outcome) for both models. The code used for model inversions can be found at https://github.com/IMAGEotago/Katja-BLT-code.

### 2.7.2 Simulation and parameter recovery

To test and validate the models, we first simulated responses to the BLT for hypothetical participants with a range of different learning rates. This was performed 500 times for each version of the model (as pre-specified in the analysis plan), using a randomly generated learning rate each time—drawn from a truncated normal distribution with a mean of 0.34, a variance of 0.88, with lower and upper bounds of 0.0 and 1.0 respectively. Each simulation was performed using the same sequence of outcomes for each trial as in the BLT. The simulations were repeated for four different $\beta$ values (of the softmax response function) for the binary model. For the continuous model, noise drawn from a Gaussian distribution with a mean of 0.0 and a standard deviation $\sigma$ was added to the simulated prediction responses in order to reflect the noise inherent in real-world behavioral data. Once noise was added, the observed values were constrained such that the simulated prediction responses $y_t$ remain within 0.0 and 1.0 by bounding upper and lower values. Similarly to the binary model, simulations for the continuous model were repeated for four different $\sigma$ values (representing different levels of noise). An example of the simulated trajectories can be seen in Figure 3.

The data generated from each simulation was then fitted to assess how accurately the model parameters could be recovered (Wilson and Collins, 2019). 10 runs of simulation and parameter recovery were completed at each noise level (simulating 500 subjects each time), to ensure consistency was high across simulation runs. The results for parameter recovery are presented in Section 3.2.1.

### 2.7.3 Model inversions on empirical data

Following successful parameter recovery using synthetic data for both models, the empirical data gathered from the BLT for each participant were used to invert each model. The model input consisted of the predictions made by participants on each trial along with the outcomes of each trial (resistance or no resistance, with both predictions and outcomes in contingency space; see Section 2.4). From this input, the models estimate a learning rate ($\alpha$) for each participant, as well as either a $\beta$ value (of the softmax response function) when using the binary model or a group-level dispersion parameter $\phi$ when using the continuous model. The same process was used to obtain estimates for $\alpha_p$ and $\alpha_n$ values (and $\beta$ for the binary version) for the dual learning rate model. The results for each model for all participants are presented in Section 3.2.2.

### 2.7.4 Model validation

The next step for validating the models was to verify that they provided useful information from fitting the participants'

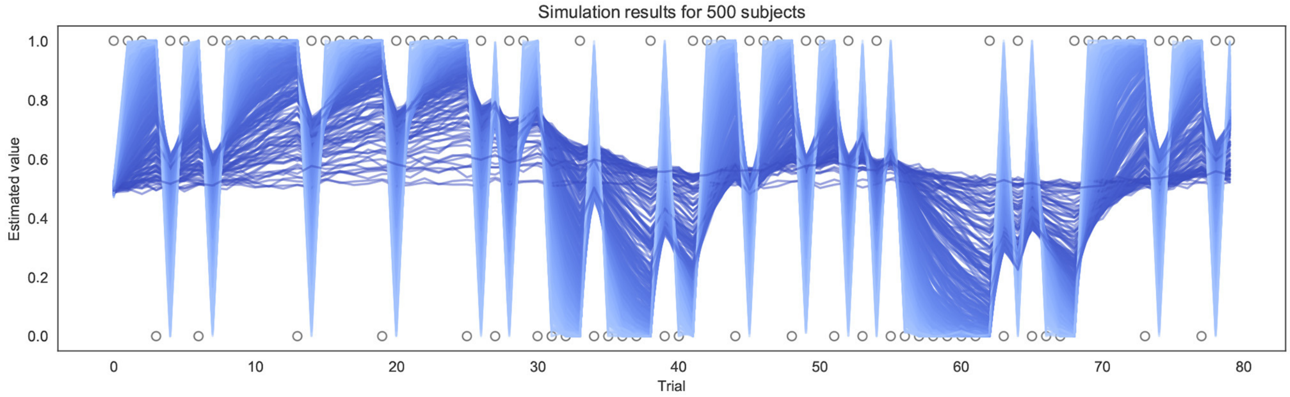


FIGURE 3
Example of simulation results using the continuous model for 500 subjects ($\sigma = 0.01$). Each line represents a different simulated subject, with a lighter blue line indicating a high $\alpha$ and a darker blue line indicating a lower $\alpha$.

data. In order to do this, both the binary and continuous model fits to participant data were compared to a null version of the respective model. This null model works the same as the binary or continuous model except that the learning rate $\alpha$ was fixed at 0.0. It therefore represents a condition where no learning occurs (i.e. $v_t$ is clamped to 0.5) and all input is attributed to noise. The binary and continuous model were validated by comparing the model fits to the respective null model fits using the Bayesian Information Criterion (BIC; Schwarz, 1978) and the Akaike Information Criterion (AIC; Akaike, 1973). Both the BIC and the AIC are ways of determining which model is "best" in terms of the trade-off between goodness of fit and model simplicity (Kuha, 2004). Although this inverts the original definition of BIC (Schwarz, 1978), one way to include the definitions of both BIC and AIC in one expression is:

$$\text{AIC or BIC} = -2ln(\hat{L}) + c\kappa \tag{7}$$

Here $\hat{L}$ is the log likelihood of the data given the parameter estimates, $\kappa$ is the total number of parameters in the model (as a proxy of model complexity), and $c$ is a penalty coefficient. For the BIC, $c = ln(n)$ where $n$ is the number of observations, while for the AIC, $c = 2$ (Vrieze, 2012). Given the above formulation (Equation 7), the model with the smaller BIC or AIC value is considered to be the better model in terms of how well it minimizes information loss. A conversion into Bayes factors can be used to quantify the degree to which one model is preferred (Penny et al., 2004; Wagenmakers, 2007). Since BIC and AIC have different notions of model complexity, we used both of them to compare the binary and continuous models against the respective null models. It should be noted that model comparison by BIC and AIC is only valid when comparing models fit to the same data, therefore they cannot be used to compare directly between the binary and continuous models.

In addition to this, a second validation was performed by analysing how well model fits correlated to actual participant behavior at the group level. To do this, the average prediction trajectories fitted by the model were compared to the actual predictions of the participants (averaged across participants for each trial) using a Pearson correlation to determine how well they were aligned. Results from both of these tests are presented in Section 3.2.3.

### 2.7.5 Comparisons between single and dual learning rate models

To determine whether the dual learning rate models were capturing useful additional information to the single learning rate models, the BIC and AIC were used to compare the binary and continuous versions of the two models (using the same approach as described in Section 2.7.4). Results are presented in Section 3.2.4.

### 2.7.6 Exploratory correlations with questionnaires

Finally, we investigated how the information provided by these models could be used to explore how interoceptive learning may relate to both measures of mental health and subjective interoception. To do this, exploratory non-parametric Spearman rank correlations were performed between the questionnaire scores (STAI-S, STAI-T, GAD-7, ASI-3, CESD, PANAS-P/N, FSS, CD_RISC-25, GSE, MAIA, PCS-B, PVAQ-B), the learning rates estimated by the different models, and the average prediction certainty of each participant using their continuous prediction data. Due to the exploratory nature of these correlations, findings are uncorrected for multiple comparisons. The findings are reported in Section 3.3 below.

# 3 Results

## 3.1 Task validation

To investigate the consistency in task performance in the current study with the previous binary version of the BLT, the proportion of correct responses on each trial across all participants for the current study was firstly compared to the larger cohort of previous data (Harrison et al., 2021) (see Figure 4). A strong correlation was found between the proportion of correct responses at each trial across the two cohorts, with a Pearson's correlation coefficient of $r = 0.85$ ($p = 5.18e^{-}23$). This indicates that overall, participants responded to the new version of the task in a similar manner to the larger cohort of participants used in the previous study.

## 3.2 Model results

### 3.2.1 Simulation and parameter recovery

#### 3.2.1.1 Binary model

For the binary model, simulation and parameter recovery was performed at four different values of $\beta$ (steepness of the gradient of the softmax function, representing decision noise): 1, 2, 4, and 8, with 10 simulation runs for each value of $\beta$ and 500 simulation subjects in each run. The Pearson correlation between the simulated values of $\alpha$ and the recovered values of $\alpha$ was averaged across the 10 runs for each different value of $\beta$ by converting the r-values to z-values, taking the mean, and then converting back to r-values. Figure 5 shows the results for a representative run for each different value of $\beta$.

As shown in Figure 5, parameter recovery was very successful for higher values of $\beta$, with an average correlation between estimated and simulated $\alpha$ of $r = 0.97$ for $\beta = 8$, $r = 0.95$ for $\beta = 4$ and $r = 0.88$ for $\beta = 2$. There was still a moderate correlation between estimated and simulated $\alpha$ of $r = 0.68$ for $\beta = 1$. As can be seen in Figure 5, there appears to be a ceiling effect for recovered $\alpha$ values above $\alpha \approx 0.6$ when there is less noise present. This finding has been reported previously with binary data (Harrison et al., 2021). The simulated $\beta$ values were also successfully recovered, with the exception of $\beta = 8$ where the recovered $\beta$ value was lower than the simulated $\beta$ value. A visualization of this recovery can be found in Section 2 of the Supplementary material.

#### 3.2.1.2 Continuous model

Similarly to the binary model, simulation and parameter recovery was performed at four different values of $\sigma$ (the standard deviation of the added noise) - 0.05, 0.1, 0.2, and 0.4, with 10 simulation runs and 500 simulated subjects for each value of $\sigma$. Mean Pearson's r-scores for each level of noise were calculated as for the binary model above. Examples of results from a representative run for each level of noise are shown in Figure 6.

As shown in Figure 6, parameter recovery showed a significant correlation between simulated and recovered $\alpha$ values at all levels of noise. The average correlation for each noise level was $r = 1.00$ at $\sigma = 0.05$, $r = 0.99$ at $\sigma = 0.1$, $r = 0.97$ at $\sigma = 0.2$, and $r = 0.90$ at $\sigma = 0.4$. At higher noise values, an under-estimating bias became apparent, with recovered values being estimated in a lower range than simulated values. Increasing the Gaussian noise (via increased $\sigma$) in simulations produced the expected reductions in the recovered group-level dispersion parameter $\phi$, and the values used for simulation can be found in Section 2 of the Supplementary material.

#### 3.2.1.3 Dual learning rate model

Simulation and parameter recovery was also carried out for both the binary and continuous versions of the dual learning rate model, using the same method as explained above. Parameter recovery showed similar results as the single learning rate model for the continuous version, with strong correlations between simulated and recovered $\alpha$ values at all levels of noise: at $\sigma = 0.4$ ($r(\alpha_p) = 0.84$, $r(\alpha_n) = 0.88$), $\sigma = 0.2$ ($r(\alpha_p) = 0.94$, $r(\alpha_n) = 0.96$), $\sigma = 0.1$ ($r(\alpha_p) = 0.98$, $r(\alpha_n) = 0.99$), and $\sigma = 0.05$ ($r(\alpha_p) = 0.99$, $r(\alpha_n) = 1.00$). Similar results were obtained with the binary version—see Supplementary material Section 4.1 for further detail and a graphical representation of the parameter recovery results for the dual learning rate model.

### 3.2.2 Model inversions on empirical data

#### 3.2.2.1 Binary and continuous models

Both the binary and continuous models were used to estimate a learning rate ($\alpha$) for each of the participants (as well as a $\beta$ value for the binary model or a $\sigma$ value for the continuous model) and produce a corresponding prediction trajectory. The prediction trajectories for each of the fitted learning rates from both models are displayed in Figure 7. Supplementary Figure S4 shows a comparison of individual response data to model-fitted trajectories for a range of learning rates.

As can be seen in Figure 7, the $\alpha$ values fitted by each of the models varied across participants. For the continuous model, the mean $\alpha$ value was 0.08 with a standard deviation of 0.10 and the value of the group parameter $\phi$ was 0.04. For the binary model, the mean $\alpha$ value was 0.18 with a standard deviation of 0.22, and the mean $\beta$ value was 4.68 with a standard deviation of 2.00. Learning rates fitted by the binary and continuous model were compared using a t-test, which indicated that the learning rates fitted by the continuous model were significantly lower than those fitted by the binary model ($p = 0.03$).

#### 3.2.2.2 Dual learning rate model

For the dual learning rate model, both the binary and continuous versions were used to estimate two learning rates ($\alpha_p$ and $\alpha_n$) for each of the participants. The corresponding prediction trajectories generated for each participant are displayed in Supplementary Figure S7. For the continuous version, the mean value of $\alpha_p$ was 0.09 (standard deviation of 0.11), the mean value of $\alpha_n$ was 0.07 (standard deviation of 0.11), and the group parameter $\phi$ was 0.05. For the binary version, the mean value of $\alpha_p$ was 0.25 (standard deviation of 0.26), the mean value of $\alpha_n$ was 0.19 (standard deviation of 0.26), and the mean value of $\beta$ was 4.31 (standard deviation of 2.49).

### 3.2.3 Comparison against null models and validation

#### 3.2.3.1 Binary and continuous models

To test whether the binary and continuous models were providing useful information when fitting participants' data, they were each compared to their respective null models (see Section 2.7.4 for a full explanation). Both models achieved a smaller BIC and AIC score (indicating a better fit) than their null model

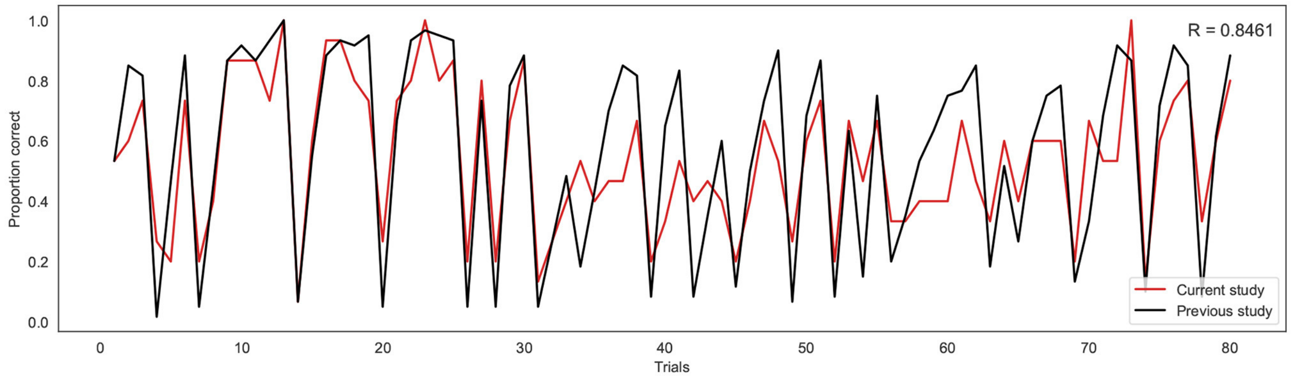


FIGURE 4
Proportion of correct responses for each trial across the current study cohort (red line) and the previous study cohort (black line). R indicates the Pearson correlation coefficient.

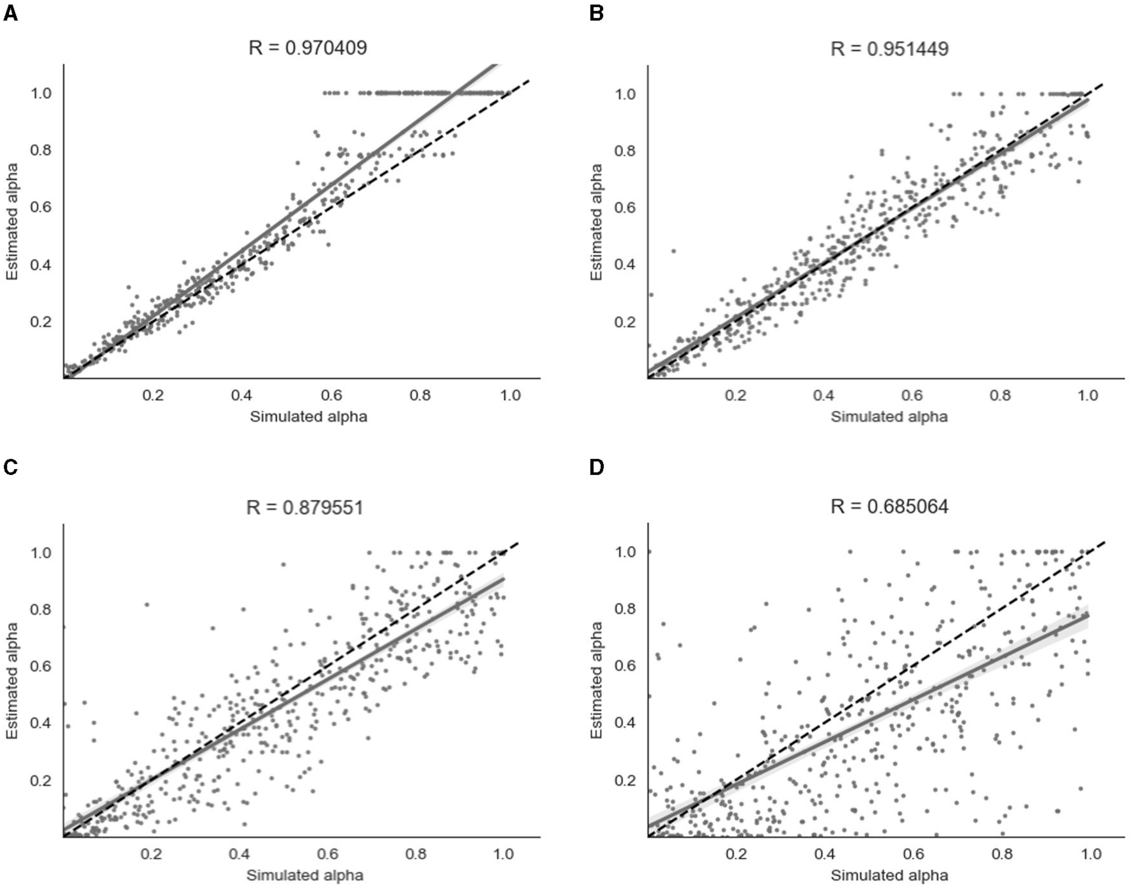


FIGURE 5
Parameter recovery of $\alpha$ for the binary model at four different values of $\beta$. Data points shown were taken from one representative simulation run. R indicates the Pearson correlation coefficient of the simulated values. **(A)** β = 8. **(B)** β = 4. **(C)** β = 2. **(D)** β = 1.

counterparts. The binary model produced a BIC of 1,253 and an AIC of 1,252 compared to the binary null model which produced a BIC and AIC of 1,691. Calculations of the Bayes factor indicate the binary model is preferred over the null model with a Bayes factor of $1.36 \times 10^{95}$. The continuous model produced a BIC of -1,315 and an AIC of -1,316 compared to the continuous null model which produced a BIC and AIC of -1,070, leading to the continuous model being preferred with a Bayes factor of $1.57 \times 10^{53}$. Both of these Bayes factors indicate very strong evidence favoring the binary and continuous models (Raftery, 1995).

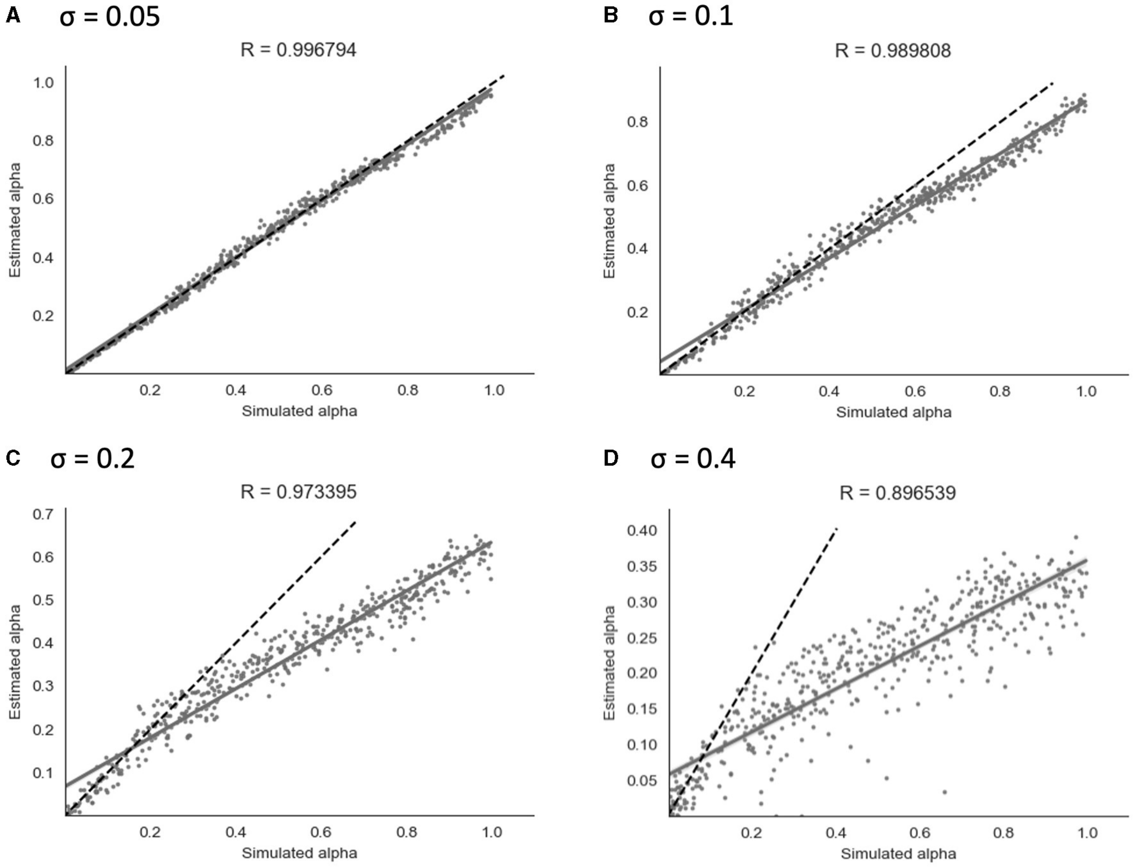


FIGURE 6
Parameter recovery of $\alpha$ for the continuous model at four different values of $\sigma$. Data points shown were taken from one representative simulation run. R indicates the Pearson correlation coefficient of the simulated values. **(A)** $\sigma = 0.05$. **(B)** $\sigma = 0.1$. **(C)** $\sigma = 0.2$. **(D)** $\sigma = 0.4$.

For the group level validation, an average predicted trajectory was created by calculating the mean predictions generated by the model across all participants for each trial. This was then compared to the mean of the observed prediction values entered by each participant on each trial, using a Pearson correlation. This procedure was performed for both the binary and continuous model, with the binary model being compared to the mean of the binarised observed prediction values that were used as the input for that model. The results are presented in Figure 8. The mean observed predictions and mean modeled predictions showed a strong correlation for both the binary ($r = 0.80, p = 8.18e^{-19}$) and continuous ($r = 0.79, p = 1.41e^{-18}$) models.

### 3.2.4 Comparisons between single and dual learning rate models

To determine whether the dual learning rate binary and continuous models were explaining the data better than their single learning rate counterparts, the models were compared using the BIC and AIC metrics. For the continuous models, the dual learning rate version produced a BIC of -1,298 and an AIC of -1,299 when run on the participants' data compared to a BIC of -1,315 and an AIC of -1,316 for the single learning rate version, resulting in a Bayes factor of 6311 in favor of the single learning rate version. Results were similar for the binary models, with the dual learning rate version producing a BIC of 1,262 and an AIC of 1,259, while the single learning rate version had a BIC of 1253 and an AIC of 1,252, resulting in a Bayes factor of 70.1 in favor of the single learning rate version. These comparisons provide strong to very strong evidence that the both the binary and continuous single learning rate models represent a better trade-off between accuracy and complexity than the corresponding dual learning rate models.

## 3.3 Questionnaire results

Participants completed a number of questionnaires measuring affective and interoceptive qualities (see Section 2.3.1 for full list). Median scores and interquartile ranges (IQR) across participants were calculated for each questionnaire and are presented in Table 1.

Exploratory correlations were then performed between the questionnaire scores and the estimated learning rates obtained from the continuous [$\alpha$(c)] and binary [$\alpha$(b)] models, as well as the average certainty across trials for each participant from the continuous prediction data. The results are presented in a correlation matrix in Figure 9. Results are uncorrected as the correlations are exploratory in nature.

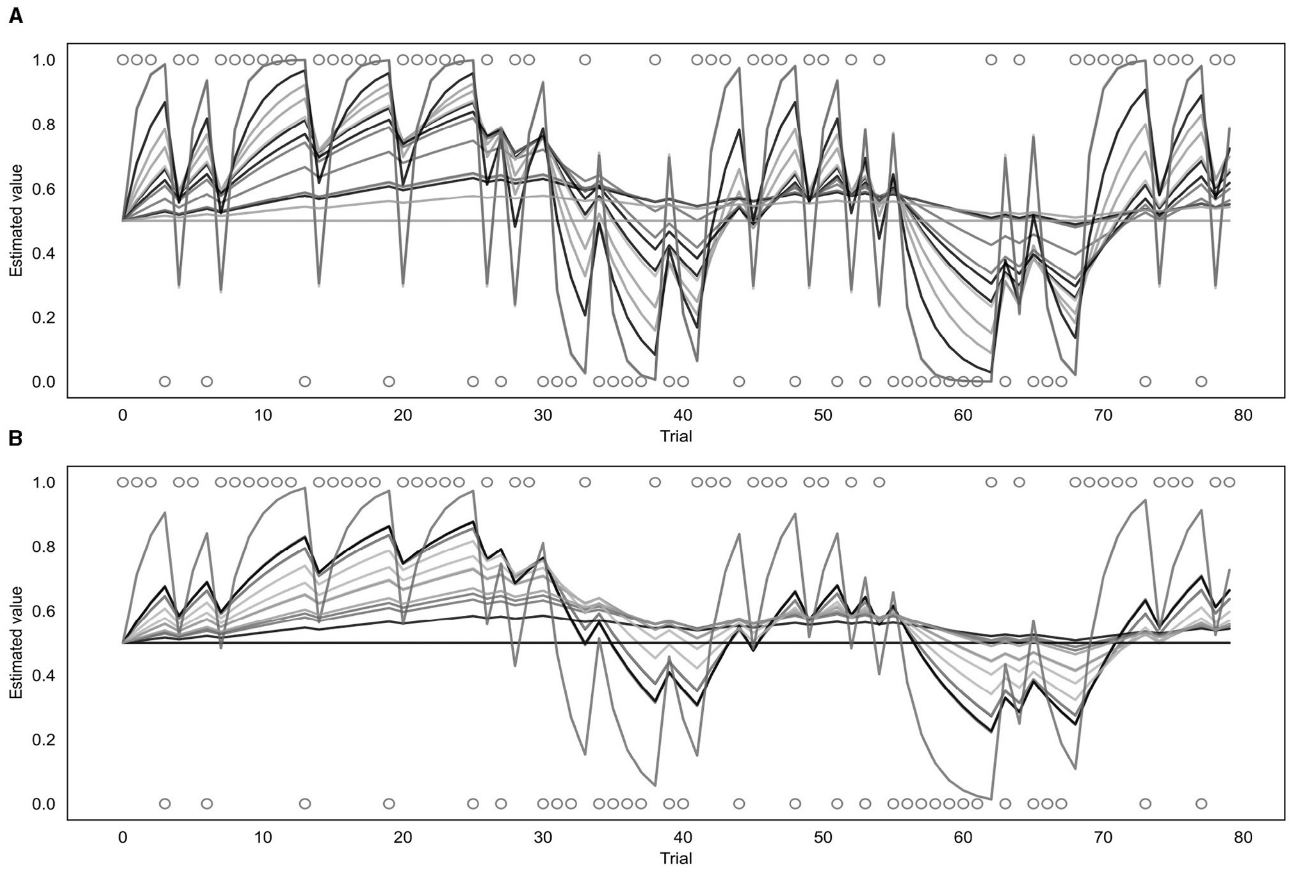


FIGURE 7
Fitted prediction trajectories for each participant. **(A)** Represents results from the binary model, **(B)** represents results from the continuous model. Circles represent trial outcomes.

As can be seen in Figure 9, significant correlations were observed between the questionnaire measures of anxiety, depression and affect (STAI-T, STAI-S, GAD-7, ASI-3, CESD, PANAS-P and PANAS-N), and some of these were additionally related to measures of interoceptive qualities (MAIA, PCS-B, and PVAQ-B) as expected. Fatigue severity (FSS) was positively correlated with anxiety sensitivity (ASI-3, $r_s = 0.67, p = 0.005$), depression symptoms (CESD, $r_s = 0.50, p = 0.05$), and increased vigilance around breathlessness (PVAQ-B, $r_s = 0.63, p = 0.01$). Between the BLT data and questionnaire scores, there was a moderate correlation between FSS scores and both average certainty values ($r_s = 0.66, p = 0.01$) as well as continuous model learning rates ($r_s = 0.60, p = 0.01$). There was also a significant correlation between average certainty values and ASI-3 scores ($r_s = 0.51, p = 0.04$). Binary model learning rates were negatively correlated with STAI-T ($r_s = -0.52, p = 0.04$) and positively correlated with GSE scores ($r_s = 0.52, p = 0.04$).

## 4 Discussion

This work has provided an advancement on the interoceptive learning paradigm (BLT) used in Harrison et al. (2021) by incorporating continuous response data, thus providing a more direct measure of prediction certainty. Additionally, an extension of the computational model tested whether stimuli valence had an effect on learning rate, investigating whether separate learning rates for positive and negative stimuli should be considered when fitting behavioral data. Data from a novel continuous version of the BLT was gathered from a cohort of 16 healthy participants, and behavioral data from this cohort were compared to that gathered by Harrison et al. (2021). Importantly, a close correlation was observed in participant performance (i.e. percentage of correct predictions) between the two cohorts, indicating that participants were similar in their overall task performance when moving from binary to continuous predictions.

By modifying the task to record predictions with a continuous rather than a binary measure, we can collect additional information about the interoceptive learning process. While binary response data reflect only the direction of the prediction (i.e., predicting a resistance or no resistance) and therefore provide one bit of information per trial, requiring participants to indicate the degree of certainty in the prediction provides a continuous readout with more information. Furthermore, in the classical (binary) version of the BLT, information about certainty could only be inferred from the binary data through an appropriate learning model, while the continuous response data from our modified BLT provide a more direct readout from participants, allowing prediction certainty

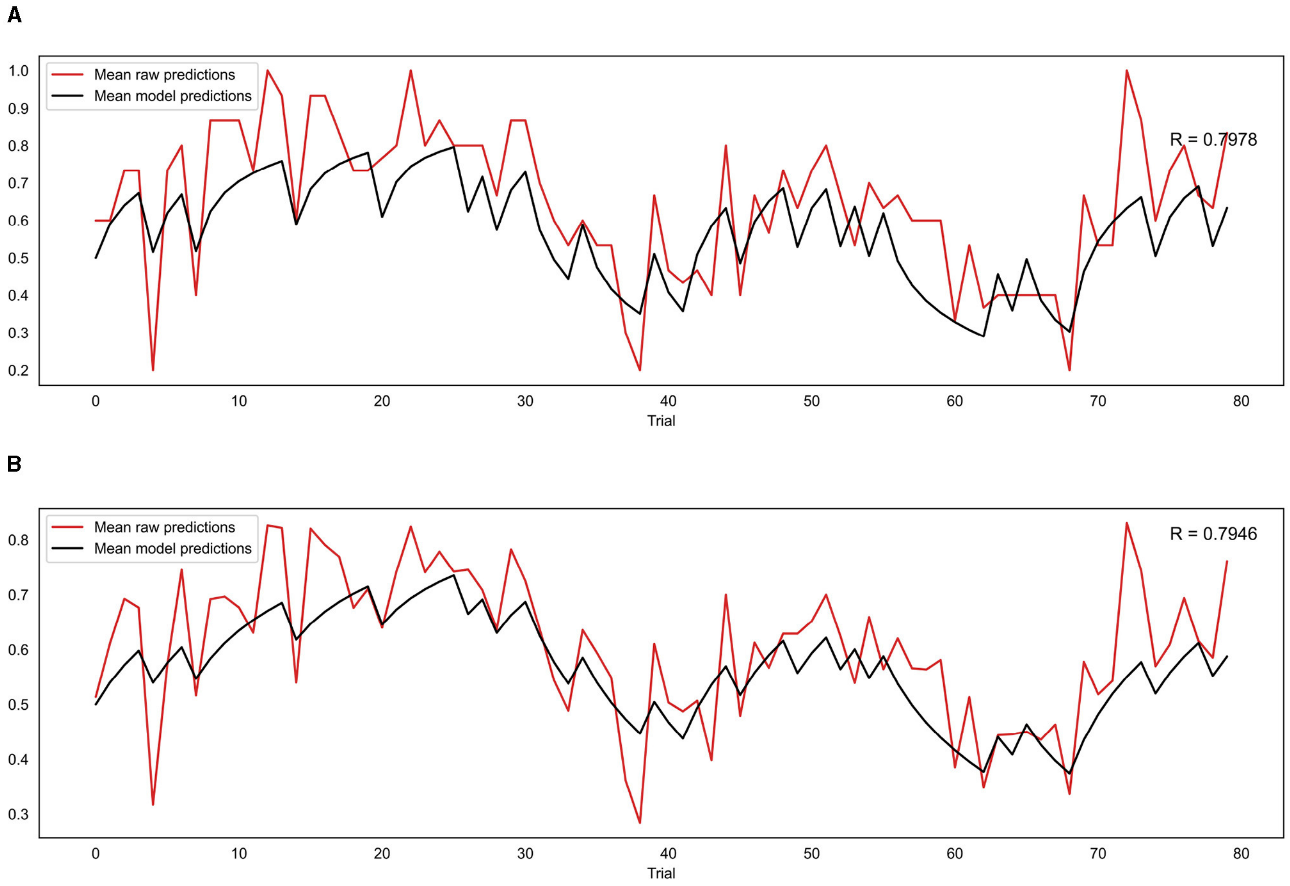


FIGURE 8
**(A)** Correlation between mean model predictions across all participants and mean observed prediction values for the binary model. **(B)** Correlation between mean model predictions across all participants and mean observed prediction values for the continuous model. R indicates the Pearson correlation coefficient.

to be measured independently of any model assumptions. This allows us to directly investigate the role that (un)certainty plays in interoceptive learning, and how this may be altered in mental health disorders such as anxiety. Therefore, to validate our model alterations for continuous data, we compared the continuous model to the previous binary model using several validation techniques.

When considering model validity using simulated data, both the binary and continuous models provided significant parameter recovery across varying levels of noise. It should be noted that for the continuous model, constrained Gaussian noise was added to the simulated trajectories. This noise added variability to simulated responses, but was constrained such that prediction values remained between 0 and 1. However, these Gaussian noise assumptions were not implemented during model inversion (see Section 2.6.1), with variability instead captured by a group-level dispersion parameter, $\phi$.

For both binary and continuous data, comparisons to the respective null models using the BIC and AIC indicated that our trial-wise learning models were greatly superior, indicating that they were capturing useful information from participants' responses. Furthermore, the average modeled prediction trajectories from both models were highly correlated with the average observed predictions, suggesting that both models did well at representing the behavior of participants at a group level.

No significant correlation was observed when comparing the estimated learning rate parameters fit by each of the models to the empirical data. Additionally, the continuous model generally produced smaller $\alpha$ values in comparison to the binary model. This may be influenced by both the under-estimation bias and the reduced variability in the estimated learning rates observed in the continuous vs. binary model simulation results. In addition, two of the participants had estimated learning rates below 0.0001 by one or both of the models, which would indicate that little to no learning occurred during the task. It is therefore possible that these participants did not properly understand the task or failed to follow instructions, which future studies using this task may need to further accommodate for.

In addition to incorporating continuous predictions, we also explicitly tested the model assumption that a single learning rate was able to adequately capture participant behavior across the resistance and no-resistance stimuli. One potential issue with a single learning rate model is that it assumes that participants learn at the same rate from negative (i.e., resistance) as for positive (i.e. no resistance) outcomes, as participants were explicitly told that cues act as a pair. However, previous research has indicated that people may learn differently from negative compared to positive stimuli, and this may be influenced by factors such as anxiety (Khdour et al., 2016; Aylward et al., 2019). To test whether learning differed according to stimuli valence in the BLT, the model was extended to introduce a dual learning rate algorithm that estimated separate learning rates following resistance and no-resistance trials. Overall, both binary and continuous models that incorporated a dual learning rate produced results consistent with their single learning rate counterparts. However, the results from comparing the BIC and AIC for each of the models suggest that the single learning rate models are preferred over the dual learning rate models in terms of accuracy-complexity trade-off. Therefore, as the dual learning rate model did not convey an advantage to explaining participants' behavior, a meaningful difference in learning between positive and negative stimuli for this task is unlikely. Thus, the original assumptions made for the single learning rate model are likely adequate, although a larger sample size would allow for further validation of this result.

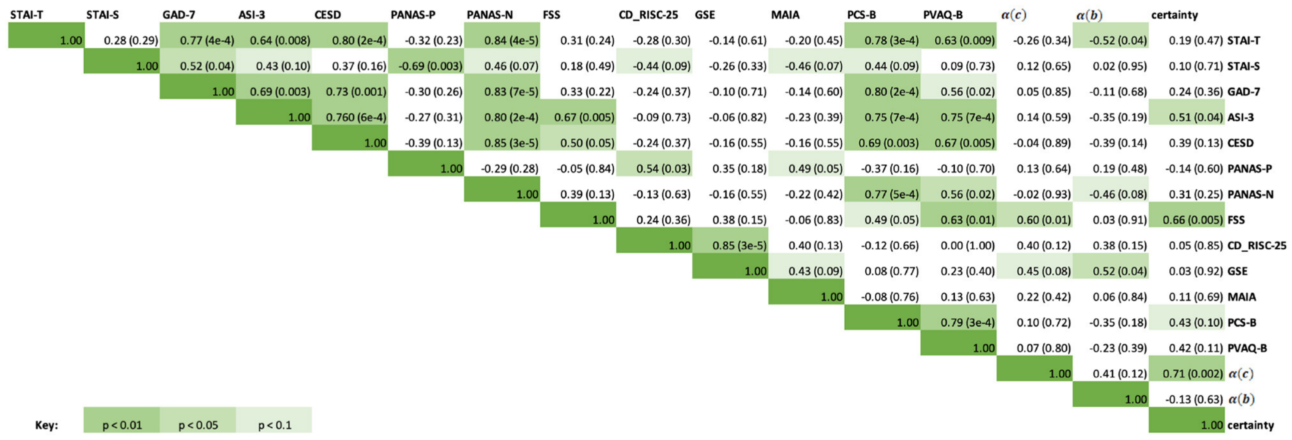
TABLE 1 Questionnaire scores.

| | Median (IQR) |
|---|---|
| STAI-T | 36.5 (18.75) |
| STAI-S | 33 (18.75) |
| GAD-7 | 2.5 (4.75) |
| ASI-3 | 17.5 (24.75) |
| CESD | 10 (7.75) |
| PANAS-P | 33.5 (8.5) |
| PANAS-N | 18 (10.25) |
| FSS | 35 (17.75) |
| CD-RISC 25 | 71 (18.25) |
| GSE | 31 (19) |
| MAIA | 21.29 (6.37) |
| PCS-B | 11.5 (29) |
| PVAQ-B | 32.5 (8.5) |

Finally, we investigated whether learning rate parameters fit by either model and/or the observed prediction certainty from the continuous task were related to questionnaire measures of both affective and interoceptive qualities. Fatigue severity (as measured by the FSS) was found to be positively correlated with both the learning rate fit by the continuous model and with average response certainty, but not the learning rate from the binary model. However, learning rates from the binary model were correlated with questionnaire measures of trait anxiety and self-efficacy. Additionally, there was a significant correlation between average certainty and anxiety sensitivity (as measured by the ASI-3). These results indicate that the continuous data provide us with different information compared to when we only consider binary decisions. However, the binary information is not lost in this version of the task, allowing the data to be analyzed in multiple ways. Overall, these preliminary findings demonstrate that valuable information can be gained from including a direct measure of response certainty within interoceptive learning tasks when investigating the relationship between interoception and mental health.

## 5 Conclusion

This report has presented a method of incorporating continuous response data into the interoceptive learning paradigm (BLT) introduced by Harrison et al. (2021), and presented a suitably extended response model for a Rescorla-Wagner learning model of the measured data. Furthermore, it tested whether assuming a single learning rate, regardless of stimulus valence, was adequate for the BLT, or whether an extended model with two separate learning rates would be advantageous. Both binarised and continuous data from a pilot cohort who completed the modified BLT were fit using a Rescorla-Wagner learning model. Both models performed



| STAI-T | STAI-S | GAD-7 | ASI-3 | CESD | PANAS-P | PANAS-N | FSS | CD_RISC-25 | GSE | MAIA | PCS-B | PVAQ-B | $\alpha(c)$ | $\alpha(b)$ | certainty | |
|---|---|---|---|---|---|---|---|---|---|---|---|---|---|---|---|---|
| 1.00 | 0.28 (0.29) | 0.77 (4e-4) | 0.64 (0.008) | 0.80 (2e-4) | -0.32 (0.23) | 0.84 (4e-5) | 0.31 (0.24) | -0.28 (0.30) | -0.14 (0.61) | -0.20 (0.45) | 0.78 (3e-4) | 0.63 (0.009) | -0.26 (0.34) | -0.52 (0.04) | 0.19 (0.47) | STAI-T |
| | 1.00 | 0.52 (0.04) | 0.43 (0.10) | 0.37 (0.16) | -0.69 (0.003) | 0.46 (0.07) | 0.18 (0.49) | -0.44 (0.09) | -0.26 (0.33) | -0.46 (0.07) | 0.44 (0.09) | 0.09 (0.73) | 0.12 (0.65) | 0.02 (0.95) | 0.10 (0.71) | STAI-S |
| | | 1.00 | 0.69 (0.003) | 0.73 (0.001) | -0.30 (0.26) | 0.83 (7e-5) | 0.33 (0.22) | -0.24 (0.37) | -0.10 (0.71) | -0.14 (0.60) | 0.80 (2e-4) | 0.56 (0.02) | 0.05 (0.85) | -0.11 (0.68) | 0.24 (0.36) | GAD-7 |
| | | | 1.00 | 0.760 (6e-4) | -0.27 (0.31) | 0.80 (2e-4) | 0.67 (0.005) | -0.09 (0.73) | -0.06 (0.82) | -0.23 (0.39) | 0.75 (7e-4) | 0.75 (7e-4) | 0.14 (0.59) | -0.35 (0.19) | 0.51 (0.04) | ASI-3 |
| | | | | 1.00 | -0.39 (0.13) | 0.85 (3e-5) | 0.50 (0.05) | -0.24 (0.37) | -0.16 (0.55) | -0.16 (0.55) | 0.69 (0.003) | 0.67 (0.005) | -0.04 (0.89) | -0.39 (0.14) | 0.39 (0.13) | CESD |
| | | | | | 1.00 | -0.29 (0.28) | -0.05 (0.84) | 0.54 (0.03) | 0.35 (0.18) | 0.49 (0.05) | -0.37 (0.16) | -0.10 (0.70) | 0.13 (0.64) | 0.19 (0.48) | -0.14 (0.60) | PANAS-P |
| | | | | | | 1.00 | 0.39 (0.13) | -0.13 (0.63) | -0.16 (0.55) | -0.22 (0.42) | 0.77 (5e-4) | 0.56 (0.02) | -0.02 (0.93) | -0.46 (0.08) | 0.31 (0.25) | PANAS-N |
| | | | | | | | 1.00 | 0.24 (0.36) | 0.38 (0.15) | -0.06 (0.83) | 0.49 (0.05) | 0.63 (0.01) | 0.60 (0.01) | 0.03 (0.91) | 0.66 (0.005) | FSS |
| | | | | | | | | 1.00 | 0.85 (3e-5) | 0.40 (0.13) | -0.12 (0.66) | 0.00 (1.00) | 0.40 (0.12) | 0.38 (0.15) | 0.05 (0.85) | CD_RISC-25 |
| | | | | | | | | | 1.00 | 0.43 (0.09) | 0.08 (0.77) | 0.23 (0.40) | 0.45 (0.08) | 0.52 (0.04) | 0.03 (0.92) | GSE |
| | | | | | | | | | | 1.00 | -0.08 (0.76) | 0.13 (0.63) | 0.22 (0.42) | 0.06 (0.84) | 0.11 (0.69) | MAIA |
| | | | | | | | | | | | 1.00 | 0.79 (3e-4) | 0.10 (0.72) | -0.35 (0.18) | 0.43 (0.10) | PCS-B |
| | | | | | | | | | | | | 1.00 | 0.07 (0.80) | -0.23 (0.39) | 0.42 (0.11) | PVAQ-B |
| | | | | | | | | | | | | | 1.00 | 0.41 (0.12) | 0.71 (0.002) | $\alpha(c)$ |
| | | | | | | | | | | | | | | 1.00 | -0.13 (0.63) | $\alpha(b)$ |
| | | | | | | | | | | | | | | | 1.00 | certainty |

Key: p < 0.01 p < 0.05 p < 0.1

FIGURE 9
Correlation matrix containing the Spearman correlation coefficients for questionnaire scores, learning rates from the continuous and binary models, and average certainty. The upper right half of the matrix contains the Spearman correlation coefficients, while the italicized fields represent the corresponding $p$-values for each correlation score. Green highlighted fields indicate significant results at $p < 0.1$, $p < 0.05$, $p < 0.01$, with a darker green indicating a more significant result. Specific p-values are included in brackets.

well on simulations (recovery analyses) and fitting the empirical data. While there was some variability of individual model fits, the continuous model accurately captured behavior at the group level. Our preliminary analyses indicate that collecting and analysing continuous prediction data from the BLT is a valid extension from previous binary predictions, and may be helpful for investigating the relationship between mental health and interoceptive learning. Specifically, the additional information to quantify prediction certainty demonstrated significant relationships with both fatigue and anxiety sensitivity scores. Therefore, our extensions provide an important development for understanding interoceptive learning, and how this may be altered with mental health conditions.

## Data availability statement

The raw data supporting the conclusions of this article will be made available by the authors, without undue reservation.

## Ethics statement

The studies involving humans were approved by the New Zealand Health and Disability Ethics Committee and the Cantonal Ethics Committee Zurich. The studies were conducted in accordance with the local legislation and institutional requirements. The participants provided their written informed consent to participate in this study.

## Author contributions

KB: Formal analysis, Investigation, Software, Visualization, Writing – original draft, Writing – review & editing, Conceptualization, Data curation, Methodology, Project administration. TW: Software, Writing – review & editing, Resources. AH: Methodology, Software, Validation, Writing – review & editing. BR: Conceptualization, Supervision, Writing – review & editing. KS: Methodology, Writing – review & editing. OH: Conceptualization, Formal analysis, Funding acquisition, Investigation, Methodology, Project administration, Resources, Software, Supervision, Visualization, Writing – original draft, Writing – review & editing, Data curation.

## Funding

The author(s) declare financial support was received for the research, authorship, and/or publication of this article. This study was supported by a Rutherford Discovery Fellowship awarded by the Royal Society of New Zealand to OH (grant number RDF-UOO1902), with additional support provided by the School of Pharmacy and Department of Psychology at the University of Otago. KB was supported by a University of Otago Master's scholarship and the writing of the manuscript was supported by the University of Otago Postgraduate Publishing Bursary (Master's). KS acknowledges support from the René and Susanne Braginsky Foundation and the ETH Foundation. Open access funding by ETH Zurich.

## Acknowledgments

The authors would like to thank Sophie Toohey for her assistance with participant recruitment and data collection as part of a wider study. Thanks also to our clinical Pharmacists Emma Smith, Lauren Smith, and Tara Wheeler for their support. This manuscript includes material that first appeared in a Master's thesis (Brand, 2022). This manuscript has been released as a pre-print at BioRxiv.org (https://www.biorxiv.org/content/10.1101/2023.05.19.538717v1) (Brand et al., 2023).

## Conflict of interest

The authors declare that the research was conducted in the absence of any commercial or financial relationships that could be construed as a potential conflict of interest.

## Publisher's note

All claims expressed in this article are solely those of the authors and do not necessarily represent those of their affiliated organizations, or those of the publisher, the editors and the reviewers. Any product that may be evaluated in this article, or claim that may be made by its manufacturer, is not guaranteed or endorsed by the publisher.

## Supplementary material

The Supplementary Material for this article can be found online at: https://www.frontiersin.org/articles/10.3389/fpsyg.2024.1254564/full#supplementary-material

## References

Akaike, H. (1973). "Information theory and an extension of the maximum likelihood principle," in *Second International Symposium on Information Theory*, eds. B. N. Petrov, and F. Csaki. Budapest: Akademiai Kaido.

Aylward, J., Valton, V., Ahn, W.-Y., Bond, R. L., Dayan, P., Roiser, J. P., et al. (2019). Altered learning under uncertainty in unmedicated mood and anxiety disorders. *Nat. Human Behav.* 3, 1116–1123. doi: 10.1038/s41562-019-0628-0

Barrett, L. F., Quigley, K. S., Bliss-Moreau, E., and Aronson, K. R. (2004). Interoceptive sensitivity and self-reports of emotional experience. *J. Pers. Soc. Psychol.* 87:684. doi: 10.1037/0022-3514.87.5.684

Barrett, L. F., and Simmons, W. K. (2015). Interoceptive predictions in the brain. *Nature Revi. Neurosci.* 16, 419–429. doi: 10.1038/nrn 3950

Brand, K. (2022). *Incorporating Uncertainty and Valence Within Dynamic interoceptive Learning Models to Better Understand Anxiety* (master's thesis). Dunedin: University of Otago. Available online at: http://hdl.handle.net/10523/12945 (accessed June 27, 2023).

Brand, K., Wise, T., Hess, A. J., Russell, B. R., Stephan, K. E., and Harrison, O. K. (2023). Incorporating uncertainty within dynamic interoceptive learning. *bioRxiv*. doi: 10.1101/2023.05.19.538717

Brewer, R., Murphy, J., and Bird, G. (2021). Atypical interoception as a common risk factor for psychopathology: a review. *Neurosci. Biobehav. Rev*. 130, 470–508. doi: 10.1016/j.neubiorev.2021.07.036

Byrd, R. H., Lu, P., Nocedal, J., and Zhu, C. (1995). A limited memory algorithm for bound constrained optimization. *SIAM J Scient. Comput*. 16, 1190–1208. doi: 10.1137/0916069

Chennu, S., Noreika, V., Gueorguiev, D., Blenkmann, A., Kochen, S., Ibánez, A., et al. (2013). Expectation and attention in hierarchical auditory prediction. *J. Neurosci*. 33, 11194–11205. doi: 10.1523/JNEUROSCI.0114-13.2013

Connor, K. M., and Davidson, J. R. (2003). Development of a new resilience scale: the Connor-Davidson resilience scale (CD-RISC). *Depress. Anxiety* 18, 76–82. doi: 10.1002/da.10113

Critchley, H. D., and Garfinkel, S. N. (2017). Interoception and emotion. *Curr. Opini. Psychol*. 17, 7–14. doi: 10.1016/j.copsyc.2017.04.020

Critchley, H. D., Wiens, S., Rotshtein, P., Öhman, A., and Dolan, R. J. (2004). Neural systems supporting interoceptive awareness. *Nat. Neurosci*. 7, 189–195. doi: 10.1038/nn1176

Frässle, S., Aponte, E. A., Bollmann, S., Brodersen, K. H., Do, C. T., Harrison, O. K., et al. (2021). Tapas: an open-source software package for translational neuromodeling and computational psychiatry. *Front. Psychiatry* 12, 680811. doi: 10.3389/fpsyt.2021.680811

Friston, K. (2005). A theory of cortical responses. *Philos. Trans. R. Soc. Lond. B Biol. Sci*. 360, 815–836. doi: 10.1098/rstb.2005.1622

Füstös, J., Gramann, K., Herbert, B. M., and Pollatos, O. (2013). On the embodiment of emotion regulation: interoceptive awareness facilitates reappraisal. *Soc. Cogn. Affect. Neurosci*. 8, 911–917. doi: 10.1093/scan/nss089

Grupe, D. W., and Nitschke, J. B. (2013). Uncertainty and anticipation in anxiety: an integrated neurobiological and psychological perspective. *Nat. Rev. Neurosci*. 14, 488–501. doi: 10.1038/nrn3524

Gu, X., Hof, P. R., Friston, K. J., and Fan, J. (2013). Anterior insular cortex and emotional awareness. *J. Comparat. Neurol*. 521, 3371–3388. doi: 10.1002/cne.23368

Harrison, O. K., Köchli, L., Marino, S., Luechinger, R., Hennel, F., Brand, K., et al. (2021). Interoception of breathing and its relationship with anxiety. *Neuron* 109, 4080–4093. doi: 10.1016/j.neuron.2021.09.045

Hein, T. P., and Ruiz, M. H. (2022). State anxiety alters the neural oscillatory correlates of predictions and prediction errors during reward-based learning. *NeuroImage*. 11, 8895. doi: 10.1016/j.neuroimage.2022.118895

Herigstad, M., Faull, O. K., Hayen, A., Evans, E., Hardinge, F. M., Wiech, K., et al. (2017). Treating breathlessness via the brain: changes in brain activity over a course of pulmonary rehabilitation. *Eur. Respirat. J*. 50, 2017. doi: 10.1183/13993003.01029-2017

Iglesias, S., Mathys, C., Brodersen, K. H., Kasper, L., Piccirelli, M., den Ouden, H. E., et al. (2013). Hierarchical prediction errors in midbrain and basal forebrain during sensory learning. *Neuron* 80, 519–530. doi: 10.1016/j.neuron.2013.09.009

Khalsa, S. S., Adolphs, R., Cameron, O. G., Critchley, H. D., Davenport, P. W., Feinstein, J. S., et al. (2018). Interoception and mental health: a roadmap. *Biol. Psychiat*. 3, 501–513. doi: 10.1016/j.bpsc.2018.04.007

Khdour, H. Y., Abushalbaq, O. M., Mughrabi, I. T., Imam, A. F., Gluck, M. A., Herzallah, M. M., et al. (2016). Generalized anxiety disorder and social anxiety disorder, but not panic anxiety disorder, are associated with higher sensitivity to learning from negative feedback: behavioral and computational investigation. *Front. Integr. Neurosci*. 10, 20. doi: 10.3389/fnint.2016.00020

Kok, P., and de Lange, F. P. (2014). Shape perception simultaneously up-and downregulates neural activity in the primary visual cortex. *Curr. Biol*. 24, 1531–1535. doi: 10.1016/j.cub.2014.05.042

Krupp, L. B., LaRocca, N. G., Muir-Nash, J., and Steinberg, A. D. (1989). The fatigue severity scale: application to patients with multiple sclerosis and systemic lupus erythematosus. *Arch. Neurol*. 46, 1121–1123. doi: 10.1001/archneur.1989.00520460115022

Kuha, J. (2004). AIC and BIC: Comparisons of assumptions and performance. *Sociol. Methods* & *Res*. 33, 188–229. doi: 10.1177/0049124103262065

Lieder, F., Daunizeau, J., Garrido, M. I., Friston, K. J., and Stephan, K. E. (2013). Modelling trial-by-trial changes in the mismatch negativity. *PLoS Comput. Biol*. 9, e1002911. doi: 10.1371/journal.pcbi.1002911

Mathys, C., Daunizeau, J., Friston, K. J., and Stephan, K. E. (2011). A bayesian foundation for individual learning under uncertainty. *Front. Hum. Neurosci*. 5, 39. doi: 10.3389/fnhum.2011.00039

McCracken, L. M. (1997). "Attention" to pain in persons with chronic pain: a behavioral approach. *Behav. Ther*. 28, 271–284. doi: 10.1016/S0005-7894(97)80047-0

Mehling, W. E., Price, C., Daubenmier, J. J., Acree, M., Bartmess, E., and Stewart, A. (2012). The multidimensional assessment of interoceptive awareness (MAIA). *PLoS ONE* 7, e48230. doi: 10.1371/journal.pone.0048230

Paolino, P. (2001). Maximum likelihood estimation of models with beta-distributed dependent variables. *Polit. Anal*. 9, 325–346. doi: 10.1093/oxfordjournals.pan.a004873

Paulus, M. P., Feinstein, J. S., and Khalsa, S. S. (2019). An active inference approach to interoceptive psychopathology. *Annu. Rev. Clin. Psychol*. 15, 97–122. doi: 10.1146/annurev-clinpsy-050718-095617

Paulus, M. P., and Stein, M. B. (2010). Interoception in anxiety and depression. *Brain Struct. Funct*. 214, 451–463. doi: 10.1007/s00429-010-0258-9

Penny, W. D., Stephan, K. E., Mechelli, A., and Friston, K. J. (2004). Comparing dynamic causal models. *Neuroimage* 22, 1157–1172. doi: 10.1016/j.neuroimage.2004.03.026

Petzschner, F. H., Weber, L. A., Gard, T., and Stephan, K. E. (2017). Computational psychosomatics and computational psychiatry: toward a joint framework for differential diagnosis. *Biol. Psychiatry* 82, 421–430. doi: 10.1016/j.biopsych.2017.05.012

Pezzulo, G., Rigoli, F., and Friston, K. (2015). Active inference, homeostatic regulation and adaptive behavioural control. *Prog. Neurobiol*. 134, 17–35. doi: 10.1016/j.pneurobio.2015.09.001

Quadt, L., Critchley, H. D., and Garfinkel, S. N. (2018). The neurobiology of interoception in health and disease. *Ann. N. Y. Acad. Sci*. 1428, 112–128. doi: 10.1111/nyas.13915

Radloff, L. S. (1977). The CES-D scale: a self-report depression scale for research in the general population. *Appl. Psychol. Meas*. 1, 385–401. doi: 10.1177/014662167700100306

Raftery, A. E. (1995). Bayesian model selection in social research. *Sociological methodology, pages* 111–163. doi: 10.2307/271063

Rao, R. P., and Ballard, D. H. (1999). Predictive coding in the visual cortex: a functional interpretation of some extra-classical receptive-field effects. *Nat. Neurosci*. 2, 79–87. doi: 10.1038/4580

Rescorla, R. A., Wagner, A., Black, A., and Prokasy, W. (1972). *Classical Conditioning II: Current Research and Theory*. New York: Appleton-Century-Crofts.

Rieger, S. W., Stephan, K. E., and Harrison, O. K. (2020). Remote, automated, and MRI-compatible administration of interoceptive inspiratory resistive loading. *Front. Hum. Neurosci*. 14:161. doi: 10.3389/fnhum.2020.00161

Schwarz, G. (1978). Estimating the dimension of a model. *Ann. Statist*. 11, 461–464. doi: 10.1214/aos/1176344136

Schwarzer, R., Bässler, J., Kwiatek, P., Schröder, K., and Zhang, J. X. (1997). The assessment of optimistic self-beliefs: comparison of the German, Spanish, and Chinese versions of the general self-efficacy scale. *Appl. Psychol*. 46, 69–88. doi: 10.1111/j.1464-0597.1997.tb01096.x

Seth, A. K., Suzuki, K., and Critchley, H. D. (2012). An interoceptive predictive coding model of conscious presence. *Front. Psychol*. 2, 395. doi: 10.3389/fpsyg.2011.00395

Spielberger, C. D., Gorsuch, R. L., and Lushene, R. E. (1970). *State-Trait anxiety (STAI) Manual*. Palo Alto: Consulting Psychologists Press.

Spitzer, R. L., Kroenke, K., Williams, J. B., and Löwe, B. (2006). A brief measure for assessing generalized anxiety disorder: the GAD-7. *Arch. Intern. Med*. 166, 1092–1097. doi: 10.1001/archinte.166.10.1092

Stefanics, G., Heinzle, J., Horváth, A. A., and Stephan, K. E. (2018). Visual mismatch and predictive coding: a computational single-trial erp study. *J. Neurosci*. 38, 4020–4030. doi: 10.1523/JNEUROSCI.3365-17.2018

Stephan, K. E., Manjaly, Z. M., Mathys, C. D., Weber, L. A., Paliwal, S., Gard, T., et al. (2016). Allostatic self-efficacy: a metacognitive theory of dyshomeostasis-induced fatigue and depression. *Front. Hum. Neurosci*. 10:550. doi: 10.3389/fnhum.2016.00550

Sullivan, M. J., Bishop, S. R., and Pivik, J. (1995). The pain catastrophizing scale: development and validation. *Psychol. Assess*. 7, 524. doi: 10.1037//1040-3590.7.4.524

Taylor, S., Zvolensky, M. J., Cox, B. J., Deacon, B., Heimberg, R. G., Ledley, D. R., et al. (2007). Robust dimensions of anxiety sensitivity: development and initial validation of the Anxiety Sensitivity Index-3. *Psychol. Assess*. 19, 176. doi: 10.1037/1040-3590.19.2.176

Vrieze, S. I. (2012). Model selection and psychological theory: a discussion of the differences between the Akaike information criterion (AIC) and the Bayesian information criterion (BIC). *Psychol. Methods* 17, 228. doi: 10.1037/a0027127

Wagenmakers, E.-J. (2007). A practical solution to the pervasive problems of p values. *Psychonomic bullet. Rev*. 14, 779–804. doi: 10.3758/BF03194105

Watson, D., Clark, L. A., and Tellegen, A. (1988). Development and validation of brief measures of positive and negative affect: the PANAS scales. *J. Pers. Soc. Psychol*. 54, 1063. doi: 10.1037//0022-3514.54.6.1063

Wilson, R. C., and Collins, A. G. (2019). Ten simple rules for the computational modeling of behavioral data. *Elife* 8, e49547. doi: 10.7554/eLife.49547

Zhu, C., Byrd, R. H., Lu, P., and Nocedal, J. (1997). Algorithm 778: L-bfgs-b: Fortran subroutines for large-scale bound-constrained optimization. *ACM Trans. Mathemat. Softw*. (*TOMS*) 23, 550–560. doi: 10.1145/279232.279236