The Filter Detection Task for measurement of breathing-related interoception and metacognition

Olivia K. Harrison[1], Sarah N. Garfinkel[2], Lucy Marlow[3], Sarah Finnegan[3], Stephanie Marino[1], Laura Nanz[1], Micah Allen[4,5,6], Johanna Finnemann[6], Laura Keur-Huizinga[6], Samuel J. Harrison[1], Klaas Enno Stephan[1,7,8], Kyle Pattinson[3], Stephen M. Fleming[7,9,10]

[1] Translational Neuromodeling Unit, Institute for Biomedical Engineering, University of Zurich and ETH Zurich, Switzerland
[2] Brighton and Sussex Medical School, University of Sussex, United Kingdom
[3] Nuffield Department of Clinical Neurosciences, University of Oxford, United Kingdom
[4] Aarhus Institute of Advanced Studies, Aarhus University, Denmark
[5] Center of Functionally Integrative Neuroscience, Aarhus University Hospital, Denmark
[6] Cambridge Psychiatry, University of Cambridge, United Kingdom
[7] Wellcome Centre for Human Neuroimaging, University College London, United Kingdom
[8] Max Planck Institute for Metabolism Research, Cologne, Germany
[9] Max Planck Centre for Computational Psychiatry and Ageing Research, University College London, United Kingdom
[10] Department of Experimental Psychology, University College London, United Kingdom

Key words: interoception, breathing, inspiratory resistance, metacognition

Corresponding author:
Dr Olivia Harrison (née Faull)
Translational Neuromodeling Unit
Institute for Biomedical Engineering
University of Zurich and ETH Zurich

Abstract

The study of the brain's processing of sensory inputs from within the body ('interoception') has been gaining rapid popularity in neuroscience, where interoceptive disturbances have been postulated to exist across a wide range of chronic physiological and psychological conditions. Here we present a task and analysis procedure to quantify specific dimensions of breathing-related interoception, including interoceptive sensitivity (accuracy), decision bias, metacognitive bias, and metacognitive performance. We describe a task that is tailored to methods for assessing respiratory interoceptive accuracy and metacognition, and pair this with an established hierarchical statistical model of metacognition (HMeta-d) to overcome significant challenges associated with the low trial numbers often present in interoceptive experiments. Two major new developments have been incorporated into this task analysis by pairing: (i) an adaptive algorithm to maintain task performance at 70-75% accuracy, and (ii) an extended metacognitive model developed to hierarchically estimate multiple regression parameters linking metacognitive performance to relevant (e.g. clinical) variables. We demonstrate the utility of both developments, using both simulated and empirical data from three separate studies. This methodology represents an important step towards accurately quantifying interoceptive dimensions from a simple experimental procedure that is compatible with the practical constraints in clinical settings.

Introduction

Understanding how the brain integrates sensory information to guide perception and action is a core component of neuroscientific research. Whilst the mapping of sensory pathways and perceptual phenomena have seen major developments in our understanding of the 'exteroceptive' domain (such as vision, audition, touch etc.), the study of 'interoception' (or the brain's processing of sensory inputs from within the body) has begun receiving attention only relatively recently[1]. While theoretical concepts of the dynamic interplay of brain and body – including interoception, homeostatic and allostatic control[2-5] – exist, empirical investigations have lagged behind. However, empirical studies of interoception have been recently boosted by a surge of interest in multiple neuroscientific fields, given that impairments in interoceptive processing have been proposed to play a role in emotions, decision making, consciousness and mental health[1,6].

Perceptual processing is a complex phenomenon, and one that is highly integrated with other domains of brain function. For example, visual perception can be manipulated via changes in factors such as attention[7], emotional state[8] or expectation[9]. Furthermore, objective performance in perceptual detection tasks (i.e. accuracy and sensitivity towards stimulus detection, as often measured with classic psychophysics experiments[10]) can be differentiated from more 'metacognitive' dimensions, where metacognition refers to the ability to accurately reflect and monitor cognitive or perceptual processes[11-13]. To quantify aspects of metacognition, measures of task performance are often paired with judgements of the confidence assigned to a decision[12,14]. From these metrics, average confidence can be thought of as a 'metacognitive bias', or a tendency towards a certain level of confidence, while 'metacognitive performance' (or 'metacognitive sensitivity') reflects how well confidence measures align with actual task performance[11-13].

These dimensions – task performance, metacognitive bias and metacognitive performance – have been distinguished within an interoceptive model by Garfinkel and colleagues[14]. Here, the authors demonstrated that these domains appear to be both quantifiable and distinct, and potentially related to traits such as anxiety. While their study was focussed on cardiac-related body signals, initial work has hinted at potential cross-talk across different interoceptive 'channels' in the metacognitive domain, where corresponding interoceptive metacognition (but not task performance) was observed across cardiac and respiratory tasks[15]. Interestingly, this study also reported significantly elevated confidence in breathing-related perceptual decisions when compared to judgements of cardiac and tactile performance[15]. Whilst breathing is more consciously accessible for both perception and control than the cardiac domain, this elevated confidence also highlights the importance and relevance of breathing-related symptoms in the maintenance of homeostasis, whereby even a single breath of restricted or occluded breathing can be perceived as extremely unpleasant and frightening[16].

Central for the further development of interoceptive research is the requirement to develop robust methodologies that can quantify interoceptive and metacognitive dimensions. Breathing is often considered to lie at the border of interoception and exteroception, combining cues from sensory avenues such as tactile and skeleto-muscular sensations across the chest wall, muscular effort, blood-gas signals representing bodily respiratory status, and air temperature and humidity, to name a few. However, the accessibility of breathing to voluntary alterations and conscious perceptions lends itself to an array of experimental paradigms, including those that do not require exteroceptive cues. Therefore, in this paper we combine an interoceptive breathing task based on resistive loads[15] with established computational models of metacognition[11-13], utilizing a robust hierarchical statistical methodology to overcome the difficulties associated with low numbers of trials available from interoceptive tasks. Furthermore, we present an extension to a hierarchical Bayesian model of metacognitive efficiency, HMeta-d[13], that allows measures of metacognitive performance to be directly regressed against external variables of interest within the hierarchical model. This is an advantage over standard approaches as it capitalises on the power of hierarchical estimation, especially when trial numbers are low, but avoids the problems

encountered by post-hoc regressions on hierarchical model parameters such as unwanted shrinkage to the group mean. Lastly, we present an adaptive task-performance algorithm that directly targets a perceptual threshold accuracy of ~70%, to aid the collection of a maximal number of usable trials within an individual. The utility of the analysis models and task-performance algorithm are established using both simulations and empirical data.

Methods

Filter detection task overview

To systematically test breathing perception as one form of interoception[1], we have developed a perceptual threshold breathing task (the Filter Detection Task; FDT) based on a previously-reported perceptual breathing task[15]. This task is a perceptual discrimination task, and can either be completed as a 'Yes/No' decision task, or a two-interval forced choice (2IFC) task.

For the 'Yes/No' version of the task, a standard trial structure consists of participants first taking three 'baseline' breaths through a mouthpiece connected to a simple breathing system (outlined in Figure 1). Following the baseline breaths, three breaths take place under 'resistance' or 'sham' conditions: either an inspiratory load is created via the addition of combinations of clinical breathing filters (signal trials), or an empty filter (sham trials) is added to the system. After each trial, participants are asked to decide whether or not a load had been added. Alternatively, for the 2IFC task, the inspiratory load is either added in the first or second set of three breaths, and participants are asked to choose in which of the two intervals the inspiratory load was present. In either task, after each trial participants are asked to report their confidence in this decision on a user-defined scale, e.g. from 1-10 (1 = not at all confident in decision, 10 = extremely confident in decision).

The responses from the FDT can then be used to determine a variety of interoceptive measures (outlined below), including inspiratory load-related perceptual sensitivity, bias in symptom reporting, perceptual confidence and metacognition (the ability to accurately reflect upon cognitive or perceptual processes).

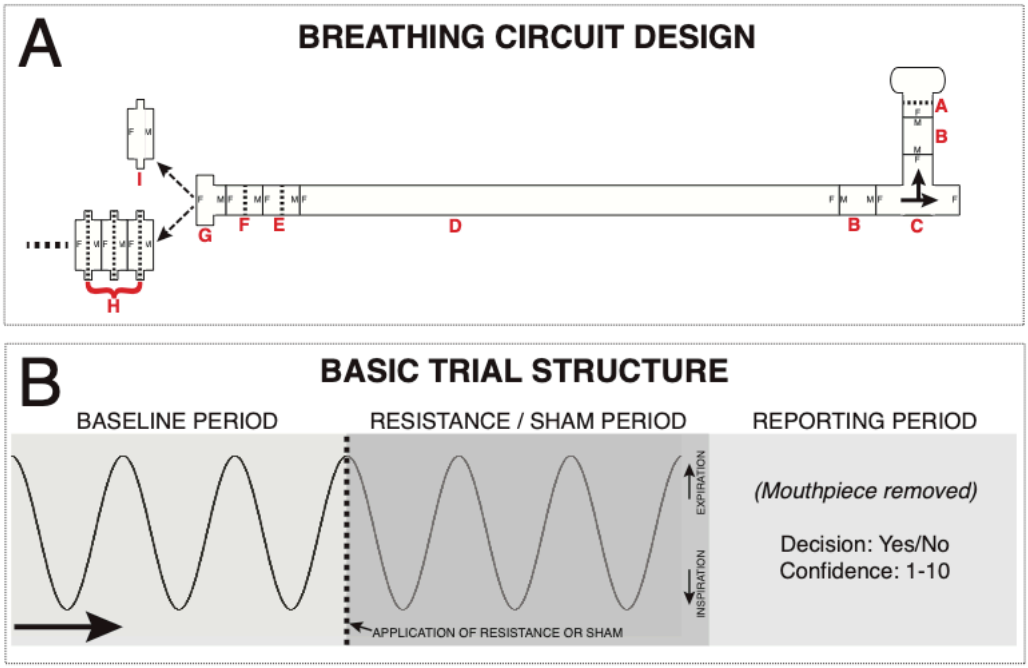


Figure 1. A) Diagram of circuitry for the filter detection task. A single-use, bacterial and viral mouthpiece (A: Powerbreathe International Ltd., Warwickshire, UK - Product SKU PBF03) is attached to a 22 mm diameter connector (B: Intersurgical Ltd., Berkshire, UK - Product 1960000) and a t-shaped inspiratory valve (C: Hans Rudolf, Kansas City, MO, USA - Product 1410/112622), connected to a 2 m length of 22 mm diameter flexible tubing (D: Intersurgical Ltd. - Product 1573000) and two additional baseline filters (E: Intersurgical Ltd. - Product 1541000, and F: GVS, Lancashire, UK - Product 4222/03BAUA). A 22-30 mm (G: Intersugical Ltd. - Product 197100) adapter then allows the attachment of either a series of connected spirometry filters (H: GVS - Product 2800/17BAUF, Resistance at 30 L/min = 0.3 cm H_2O) or a sham 'dummy' filter – a spirometry filter shell with the inner bacterial protection pad removed (I). B) Overview of the basic trial structure for a Yes/No formulation of the task. Participants take three normal size/pace breaths (with the sham filter attached), and during the third exhalation (indicated by the participant raising their hand and the dotted line in panel B) the experimenter either swaps the sham for a number of stacked filters (to provide a very small inspiratory resistance) or removes and replaces the sham filter. Following three more breaths, the participant removes the mouthpiece and reports whether they thought it a resistance was added ('Yes') or not ('No'), and how confident they are in their decision on any scale (here 1-10 used, with 1 = guessing and 10 = maximally confident in their decision). If a two-interval forced choice (2IFC) formulation of the task is used, the filters (resistance) are either placed on the circuit for the first three breaths or the second three breaths according to the FDT algorithm, with the sham filter on the system during the alternate period. The reported decision from the participant is whether they thought the resistance was on in either the first set or the second set of three breaths, and also again the confidence in their decision.

Breathing-related interoceptive measures

In the original version of the filter detection task[15], three domains of breathing-related interoception were quantified: interoceptive sensitivity (number of filters required for a discrimination threshold of approximately 75%), interoceptive sensibility (average confidence over threshold trials) and metacognition (correspondence between accuracy and confidence using a Type 2 receiver operating characteristic curve (ROC curve)). Here, we aimed to extend these measures to incorporate a more thorough and nuanced overview of a range of interoceptive dimensions. These measures include interoceptive sensitivity, decision bias, metacognitive bias, and metacognitive performance.

Interoceptive sensitivity. This measure is analogous to interoceptive accuracy, and in this task aims to quantify the 'perceptual threshold' as a means for determining breathing-related interoceptive sensitivity. In other words, interoceptive sensitivity here is the level (i.e. number of filters) at which a participant is able to consciously detect an inspiratory loading stimulus. However, to allow for the quantification of higher-order interoceptive measures (such as metacognition), the number of filters used must elicit a performance that is both significantly above chance (or guessing) and lower than 100%, so that perceptual errors can be used to quantify the correspondence between accuracy and confidence (as explained below in 'metacognitive performance'). To achieve a task difficulty that elicits this required performance, the original publication of this task[15] utilised a descending accuracy staircase protocol, whereby a large starting filter number (for example 7 filters), and 20 trials were completed at descending filters until a final filter level when performance first dropped below 70%. While many staircase protocol options could be adopted to achieve a desired task difficulty, the time to complete one trial (approximately 30-60 seconds), the natural variability in resting tidal breaths[17], the inherent bias associated with the Yes/No task formulation[18] and the fixed filter intervals render many traditionally-employed staircase protocols (such as the two/three-down-one-up[10]) less straightforward with the current methodology. Therefore, we have developed a custom staircase algorithm (explained in 'Task performance algorithm') that employs probability metrics to objectively assess both task performance and trajectory towards the required perceptual threshold (see 'Task performance algorithm' section for full explanation of the algorithm).

Interoceptive decision bias. If using the FDT as a Yes/No task, a quantifiable measure of behaviour is the 'bias' towards reporting 'Yes' or 'No'. This bias represents the placement of a criterion

value above which the presence of a resistance is reported, reflecting an individual's inherent tendency to report the presence of an inspiratory resistance. Importantly, this bias can be quantified using Signal Detection Theory (SDT[19,20]), and may represent an important cognitive trait regarding the experience of respiratory symptoms. Using SDT, stimulus sensitivity (d') can be separated from bias (or the placement of a criterion, c)[19,20]. Using this theory, we are able to disentangle the components of measures that may be confounded by a mix of task sensitivity and bias, such as performance accuracy. When using the FDT as a 2IFC task, the measured 'bias' will instead be the tendency to report the resistance on the first or second interval. While this may have limited relevance to real world scenarios, quantification of d' using this task design is likely to also allow for a more accurate representation of task sensitivity and a possibly more translatable measure of metacognitive performance[21], which can both be confounded by variations in criterion placement.

Interoceptive metacognitive bias. Average subjective confidence or metacognitive bias in interoceptive decisions has also previously been referred to as interoceptive "sensibility"[14]. In this task, we take an overall average of the confidence scores (measured across the perceptual threshold trials) to represent interoceptive sensibility that directly corresponds to the task at hand. Additional trait-like, global perceptual measures of sensibility could also be gathered by using separate interoceptive questionnaires such as the Porges Body Questionnaire[22]. Interoceptive sensibility is also referred to as "metacognitive bias"[12,13,23], as it represents the tendency to give higher or lower confidence ratings.

Interoceptive metacognitive performance. Breathing-related metacognitive performance (also termed 'interoceptive awareness'[14] and 'interoceptive insight'[1] in the literature) in this instance is considered to be the correspondence between task accuracy and confidence[14], or the ability to recognise successful perceptual processing[12]. Previous reports of interoceptive metacognition have utilised the area under a type 2 ROC curve[14,15,24], resulting in a measure of *absolute* metacognition where the effect of underlying task performance also influences the final score[12]. However, more recent model-based approaches have developed a metric of metacognitive performance that takes into account task performance known as meta-d'[11]. Meta-d' represents "the sensory evidence available for metacognition in signal-to-noise ratio units"[12], which, because it is in the same units as d', can be straightforwardly compared to task performance as a ratio (Mratio: meta-d'/d'; the log of the ratio, logMratio, is also often used to meet Gaussian assumptions: log(meta-d'/d')). This ratio is a *relative* measure of metacognitive performance (often termed 'metacognitive efficiency'). In order to employ this model-based approach for the FDT, a hierarchical formulation of the meta-d' model (HMeta-d) is employed that allows efficient pooling of data from multiple subjects[13], allowing us to estimate model parameters on a relatively small number of threshold trials (>= 40 trials).

Metacognitive model simulations

To demonstrate the feasibility of utilising the meta-d' model for interoceptive tasks with low trial numbers, we first present simulated results to establish the recoverability of group metacognitive performance (Mratio) values using the original maximum likelihood estimation algorithm (MLE[11]), a single-subject Bayesian model and a hierarchical (group) Bayesian model (HMeta-d), which are both provided in the HMeta-d toolbox[13]. Parameter inference in the HMeta-d toolbox rests on a Markov chain Monte Carlo (MCMC) sampling procedure, implemented using the JAGS software package (http://mcmc-jags.sourceforge.net). The extent of the group Mratio recoverability is demonstrated for 20, 40 and 60 trials per subject, with 30 simulated participants. Simulations were generated from a set of values $N(\mu, \sigma)$, which refer to Gaussians parameterised by a mean and standard deviation, using the metad_sim function provided in the HMeta-d toolbox[13]. The meta-d' values for the first set of simulations were generated from seven group Mratio distributions (meta-d' / d') with parameters N([0.25 0.5 0.75 1.0 1.2 1.5 1.75 2], 0.1), where d' ~ N(1, 0.1) and c ~ N(0, 0.1).

Second, we developed and simulated an extension of the HMeta-d model (RHMeta-d), which incorporates a simultaneous hierarchical estimation of a regression parameter (beta) that controls

variation in logMratio values in relation to a subject-level predictor (such as a clinical score). The model was adjusted as follows (for full details of the original model please see the original publication[13]). $N(\mu, \sigma)$ and $HN(\mu, \sigma)$ refer to Gaussians and Half-Gaussians parameterised by a mean and standard deviation, while $T(\mu, \sigma, \nu)$ refers to a T-distribution parameterised by a mean, standard deviation and degrees of freedom:

$$M_0 \sim N(0,1)$$
$$\beta \sim N(0,1)$$
$$\sigma_\delta \sim HN(1)$$
$$\zeta \sim Beta(1,1)$$
$$\delta_s \sim T(0, \sigma_\delta, 5)$$
$$\varepsilon_s = \zeta * \delta_s$$

$$M_s = M_0 + \beta * X_s + \varepsilon_s$$

where M_s refers to the log(meta-d'/d') value for subject *s*, M_0 is the baseline logMratio for the group (i.e. the intercept of the regression), and X_s is a vector of predictor values (e.g. clinical scores) for each subject. This formulation embeds the estimation of psychopathology-cognition relationships into the parameter inference routine, such that the group-level posterior over β reflects the influence of individual differences in *X* on metacognitive performance[25]. To ensure that the regression is robust to outliers, the noise ε_s is drawn from a T-distribution with a standard deviation of $|\zeta|\sigma_\delta$ and 5 degrees of freedom[26]. Consistent with the original HMeta-d model[13], a redundant multiplicative parameter ζ is used to introduce an additional random component in the sampling process to aid the recovery of the posterior on the noise scale.

For these simulations, group baseline logMratio (M_0) values were generated from a distribution ~ *N*(log(0.8), 0.1). Then, to simulate data from each subject, a random value of the covariate X_s was drawn from a standardised distribution *N*(0,1), multiplied by a group regression coefficient that was one of a fixed set of β values (β ~ [-0.5:0.5]) and added to this baseline logMratio together with zero-mean noise sampled from *N*(0, 0.1). Values used for d' and c' were consistent with previous simulations (d' ~ *N*(1, 0.1) and c ~ *N*(0, 0.1)), and data were generated using either 20, 40 or 60 trials per subject and a confidence scale of 10 rating points. From these simulated data, parameter estimates were then obtained either using the original HMeta-d model combined with a post-fit linear regression (i.e. a standard linear regression conducted on the per-subject point estimates obtained from the group-level fit, which we denote HMeta-d+R), as well as the extended hierarchical regression model (RHMeta-d) described above. An average for ten sets of simulations for each group β value was calculated for the final results.

Analysis methods: Example dataset

To demonstrate the utility of the task and analysis methods, we utilise an example dataset that includes a group of individuals who have an elevated frequency of breathing symptoms. 63 individuals with asthma (39 females, mean age (± sd) 43.7 ± 12.2 years, recruited through general practitioner clinics and public advertisements) and 30 healthy controls (19 females, mean age (± sd) 44.2 ± 12.2 years, recruited through public advertisements) took part in a study approved by the Oxfordshire Clinical Research Ethics Committee. Participants underwent the FDT and completed the Dyspnea 12[27] questionnaire as a subjective assessment of their breathlessness severity. Additionally, participants completed a further set of questionnaires and additional physiological and behavioural measures that will be addressed elsewhere. Seven individuals with asthma were excluded from the analysis due to insufficient data (10 trials or less of the FDT, n = 4), or performance of less than 50% correct (n = 3),

as determined in the preregistered analysis plan (https://gitlab.ethz.ch/tnu/analysis-plans/harrisonetal_fdt_methods_2020).

During the FDT in this study, the number of filters required for each participant to induce task performance at perceptual threshold was determined manually, using a step-wise method adjusted from a previous publication of the task[15]. In the present protocol, participants first completed 10 trials at 4 filters. If the task accuracy was below or above 70%, the number of filters was adjusted up or down accordingly. Performance accuracy was assessed again at 20, 30 and 40 trials, with adjustments made if the accuracy moved outside of the 65-75% range (with an acceptable range of 60-80% in later trials). The aim was to complete 40-60 trials, which was limited by time and attention constraints of each participant. A 0-100 confidence rating scale was employed, with 0 = not at all confident, and 100 = maximal confidence. These confidence scores were down-sampled into 10 rating bins prior to analysis with the HMeta-d model.

To demonstrate the type of questions that could be answered using the FDT and metacognitive models, we formulated three directed hypotheses for this dataset, and two exploratory questions (https://gitlab.ethz.ch/tnu/analysis-plans/harrisonetal_fdt_methods_2020). The directed hypotheses were that individuals with asthma would have: 1) greater interoceptive bias (towards symptom over-report); 2) lower metacognitive performance (reduced correspondence between accuracy and confidence); and 3) metacognitive performance will be negatively associated with prevalence of reported breathing-related symptoms. The exploratory questions we proposed were whether there were any differences between asthma and controls in the measures of: 1) interoceptive sensitivity (number of filters that elicited a task performance between ~60-80%); and 2) metacognitive bias (average confidence).

Analysis of hypothesis 1: Decision bias in asthma vs healthy controls. To address the first hypothesis, a group comparison was conducted between decision bias parameters from the participants with asthma and healthy controls. Tests for data normality were first conducted using Anderson-Darling tests, with an alpha value of $p < 0.05$ required to reject the null hypothesis of normally distributed data. A directed significant group difference in decision bias was then tested using a one-tailed, non-parametric Wilcoxon rank-sum test, where it was hypothesised that individuals with asthma would have a greater bias towards reporting the presence of resistance. Significance was denoted by $p < 0.017$ ($p < 0.05$ Bonferroni corrected for the three main hypothesis tests).

Analysis of hypothesis 2: Metacognition in asthma vs healthy controls. A group difference in metacognitive performance (the Mratio parameter from the HMeta-d model output) was assessed by first calculating the distribution of differences in posterior parameter samples from each group (control Mratio samples > asthma Mratio samples), and then determining the highest density interval (HDI) for this distribution. The HDI employed was a one-tailed 95% (Bonferroni corrected to 98.3% for three tests) confidence interval, where a significant difference between groups was denoted if the resulting HDI did not span zero.

Analysis of hypothesis 3: Metacognition in asthma according to symptom severity. To address the final hypothesis (that worsened metacognition will be associated with greater prevalence of reported breathing-related symptoms), individuals within the asthma group were first divided into a high-symptom and low-symptom group via a median-split on the D12 scores. A group difference in metacognitive performance (Mratio) was assessed by first calculating the distribution of differences in posterior parameter samples from each group (low symptom Mratio samples > high symptom Mratio samples), and then determining the highest density interval (HDI) for this distribution. The HDI employed was a one-tailed 95% (Bonferroni corrected to 98.3% for three tests) confidence interval, where a significant difference between groups was denoted when the resulting HDI did not span zero. Secondly, a direct test of the relationship between metacognitive performance and D12 was performed using a linear regression model in the individuals with asthma. To achieve this, the HMeta-d model was extended to include a hierarchical estimation of a linear regression parameter, whereby a group

regression coefficient (beta) was simultaneously fit within the model to determine the relationship between logMratio and D12. The significance of the beta parameter was then also determined using its posterior samples, with a one-tailed 95% HDI (Bonferroni corrected to 98.3%) that did not span zero determining a significant relationship between more severe D12 and worsened metacognitive performance.

Analysis of exploratory tests: Two exploratory tests were also conducted, investigating any potential group difference in interoceptive sensitivity and metacognitive bias between asthma and controls. Tests for data normality were first conducted using Anderson-Darling tests, with an alpha value of $p < 0.05$ required to reject the null hypothesis of normally distributed data. Exploratory two-tailed Wilcoxon rank-sum tests were then used to assess group differences in sensitivity and metacognitive bias, with significance denoted by $p < 0.025$ ($p < 0.05$ Bonferroni corrected for two exploratory tests).

Task performance algorithm: Simulations and empirical data

Lastly, to assist in the collection of a greater number of usable trials for further instances of the FDT, we created a novel staircase protocol (within a MATLAB toolbox package) to aid the selection of the appropriate number of filters for each participant. This staircase begins with two practice trials and a short calibration. In the practice trials, an 'explicit dummy' is first applied, whereby participants are told that it is a dummy. This is followed by a trial of 7 filters where no feedback is given (no feedback is maintained for the rest of the protocol). The practice is immediately followed by the calibration trials, where participants are subjected to an increase of one filter each trial (beginning with a dummy) until they have correctly reported the resistance for two consecutively increasing filter numbers. A final calibration trial is then given, where the number of filters is dropped by one from the last trial. If participants correctly report the final calibration trial, they begin on that number of filters, whereas if they are incorrect they begin with one additional filter. A diagram of the basic trial structure is presented in Figure 1, and the practice, calibration and real trial trajectory is provided in the supplementary material (Supplementary Figure 1).

Once the calibration is complete (or alternatively, a manual starting point can be provided), the main task trials begin. The target number of trials is specified (recommendation of >= 60 trials), and a pseudo-randomised sequence of trials are presented (trials are balanced between present/absent for a Yes/No task or between first and second interval for 2IFC). The target is for participants to be within a 65-80% accuracy band. Given a set of binary trials (and an appropriate prior), we use the fact that the posterior distribution over the underlying accuracy follows a beta distribution. We use a weak prior on the accuracy itself (beta distribution with the parameters $\alpha = 2$ and $\beta = 1$; prior mean = 67% accuracy, interquartile range = 37%). After 5 trials at one filter level (with at least one resistance present for the Yes/No task), the posterior probability that the underlying performance accuracy for the current task difficulty (i.e. the current number of filters) falls between 65-80% is calculated using the difference in beta cumulative distribution functions for 80% and 65%, in a similar vein to the QUEST algorithm[28]. If the probability that the underlying accuracy is between 65-80% falls below a threshold of 20%, an addition or removal of a filter is automatically suggested to decrease or increase task difficulty respectively. If a new filter number is started, the trial count will begin again and 5 trials (with at least one resistance) must be completed before the algorithm will suggest any changes. If the filter change moves the filter number back to that of previous trials, the trial count will pick up again from the last trial at this level. In this instance, 3 trials (with at least one resistance) must be completed before any changes are suggested.

To demonstrate the utility of this task performance algorithm, we firstly present simulation results from a range of possible participant performances, characterised by a distribution of potential psychometric functions. These psychometric functions (n = 350) were constructed from an underlying logistic sigmoid with a lower asymptote at 0.5 (to account for chance answers with the two answer options of 'yes' and 'no'), a slope k = [0.7:1.2], the number of filters at which the 75% threshold is

obtained t = [1:7] and added Gaussian noise $\varepsilon = N(0, [0.05:0.015])$. We then ran each of the sigmoids generated from each of 5 starting points – from two filters below to two filters above the t parameter, totalling 7000 simulations. Second, we provide data metrics (number of trials, performance accuracy, number of filters) for two collected datasets using the Yes/No version of the task. The first of these collected datasets stems from the first 50 participants measured as part of a wider study approved by the Cantonal Ethics Committee Zurich (Ethics approval BASEC-No. 2017-02330). For this study, we employed a 'constant' staircase formulation of the algorithm, with the aim of collecting 60 trials at a single level of filters that elicited a task performance between ~60-85%. The second empirical dataset includes the first 22 participants measured as part of a wider study approved by the Cambridge ethics committee (Ethics approval PRE.2018.092). This study employed a 'roving' staircase with 60 trials total, where the aim was to simply collect 60 trials regardless of the number of filters. In this scenario, the 'threshold' filter becomes the average of the number of filters employed across the task.

Results

Metacognitive model simulations

The simulation results firstly demonstrate that utilising the hierarchical Bayesian HMeta-d model fit allows adequate recovery of group Mratio values (Figure 2). This recovery is possible even using as few as 20 trials per subject, with slightly larger uncertainties (demonstrated by the width of the highest density intervals) than those obtained for 40 and 60 trials. It is instructive to compare this to the alternative estimation methods: while MLE is able to recover an average Mratio value that is indeed representative of the simulated value, the uncertainty around these estimates (demonstrated by the width of the confidence intervals) is large when using even 60 trials per subject. Moreover, the confidence interval around these MLE estimates incorrectly encompasses zero for all group Mratio values below 1. The recovery of the group Mratio using the Bayesian single subject fit also has large uncertainties and shows shrinkage effects towards zero, recovering Mratio values below the simulated values across all trial numbers tested here.

The second set of simulations were designed to probe the recoverability of single-subject Mratio values, for possible use in analyses comparing individual metacognitive performance against an external variable. Using the original HMeta-d model, we demonstrated that a post-hoc regression on the single-subject values was unable to accurately recover a group regression parameter simulated from the range β = [-0.5:0.5] even when using 60 trials, with all confidence intervals on the regression parameters encompassing zero (Figure 3). This is unsurprising given that the hierarchical model naturally shrinks single-subject estimates towards the group mean, losing information about individual differences. In contrast, the R-HMeta-d model was able to significantly recover beta values of $\pm >0.2$ with 60 trials, $\pm >0.25$ with 40 trials, and $\pm >0.3$ with 20 trials (Figure 3). The results for multiple regression models (with up to three covariates) at each of 20, 40 and 60 trials are presented in the supplementary material (Supplementary Figure 2).

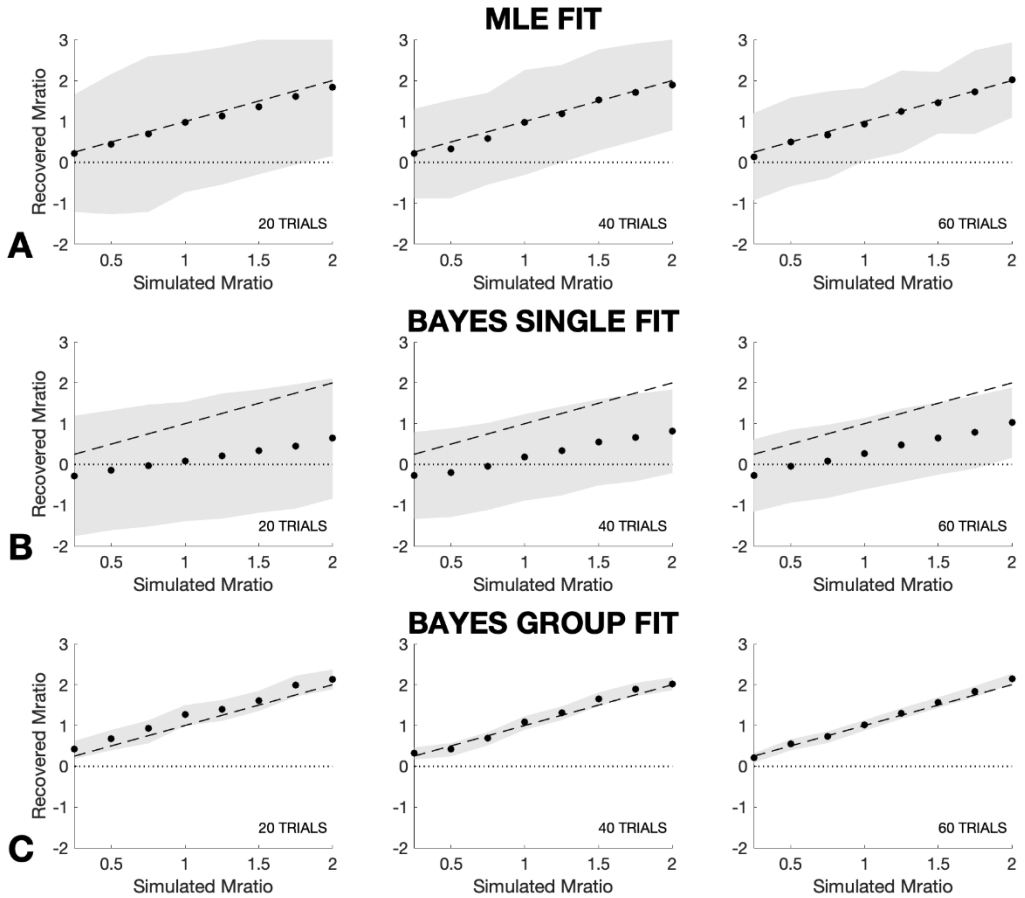


Figure 2. Group Mratio recovery for 20, 40 and 60 trials using three different meta-d' models. Data were simulated from 8 groups of 60 subjects with group mean Mratio (meta-d' / d') values set to [0.25 0.5 0.75 1.0 1.25 1.5 1.75 2] ± 0.1 (sd). All simulated values were generated from data where d' ~ N(1, 0.1) and c ~ N(0, 0.1), and a confidence scale of 10 rating points was used. A) Simulated vs. recovered Mratio values using maximum likelihood estimation[11], where the shaded grey areas denote the 95% confidence interval of the estimate. B) Simulated vs. recovered Mratio values using a Bayesian single-subject fit (provided in the HMeta-d toolbox[13]), where the grey areas denote the 95% highest density interval (equivalent to a 95% credible interval) of the sampled estimate. C) Simulated vs. recovered Mratio values using a hierarchical Bayesian group fit (provided in the HMeta-d toolbox[13]), where the grey areas denote the 95% highest density interval of the sampled estimate. Dashed lines represent ideal recovery, with dotted lines at zero demonstrating the ability of the model fit to significantly recover group estimates (i.e. when confidence or highest density intervals do not include zero).

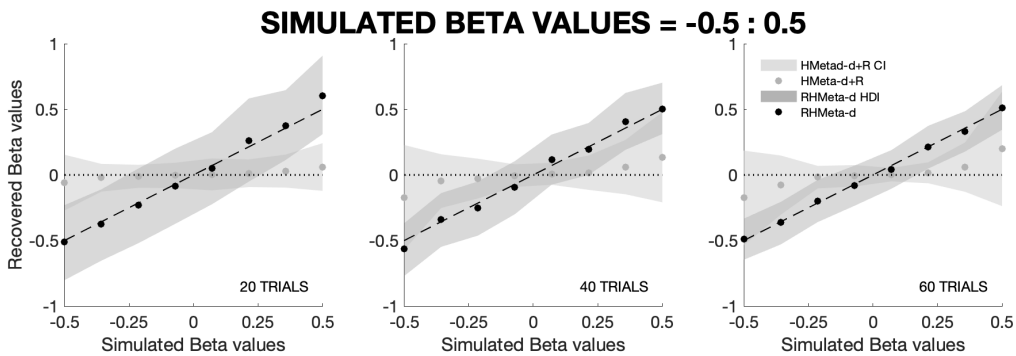


*Figure 3. Demonstration of the recovery of group regression parameters (β ~ [-0.5:0.5]) using either the original HMeta-d model combined with a post-fit linear regression (HMeta-d+R), or the extended regression HMeta-d (RHMeta-d) model. Ten sets of simulations were performed and results were averaged, where each simulation set included 60 simulated 'subjects' where logMratio = $logMratio_{baseline}$ + β*covariate + noise, where $logMratio_{baseline}$ ~ N(log(0.8), 0.1), covariate ~ N(0, 1), β ~ [-0.5:0.5], noise ~ N(0, 0.1), and with d' ~ N(1, 0.1), c ~ N(0, 0.1). Grey areas denote the 95% highest density interval of the sampled estimate. Dashed lines represent ideal recovery of group beta values, and dotted lines at zero demonstrate the ability of the model fit to significantly recover group estimates (i.e. highest density intervals that do not including zero).*

Asthma vs. healthy control comparisons

While the data demonstrated a tendency for the hypothesised differences in decision bias (towards responding 'Yes') (c = mean ± se: controls = 0.01 ± 0.07, asthma = -0.01 ± 0.06) in asthma (Figure 4), no significant difference was found for decision bias nor for metacognitive performance (Mratio = mean ± se: controls = 0.83 ± 0.14, asthma = 0.79 ± 0.12) between groups. For the exploratory tests, no significant difference was found between groups for interoceptive sensitivity (number of filters = mean ± se: controls = 2.87 ± 0.29, asthma = 2.80 ± 0.20) nor metacognitive bias (average confidence % = mean ± se: controls = 66.57 ± 2.99, asthma = 69.74 ± 2.13) (Figure 4).

Relationship between breathing symptoms and metacognition

While the data also demonstrated a tendency for reduced metacognitive performance with higher symptom loads using both a median-split and a hierarchical regression analysis (Figure 5), neither was statistically significant (determined by a HDI that does not encompass zero). The mean difference between the high and low symptom groups was 0.22 ± 0.24 (se), with the HDI on this difference in the range [-1.02, 0.34]. Using a hierarchical regression approach, the beta parameter mean was estimated as -0.19 ± 0.19 (se), with the beta HDI in the range [-1.24, 0.20].

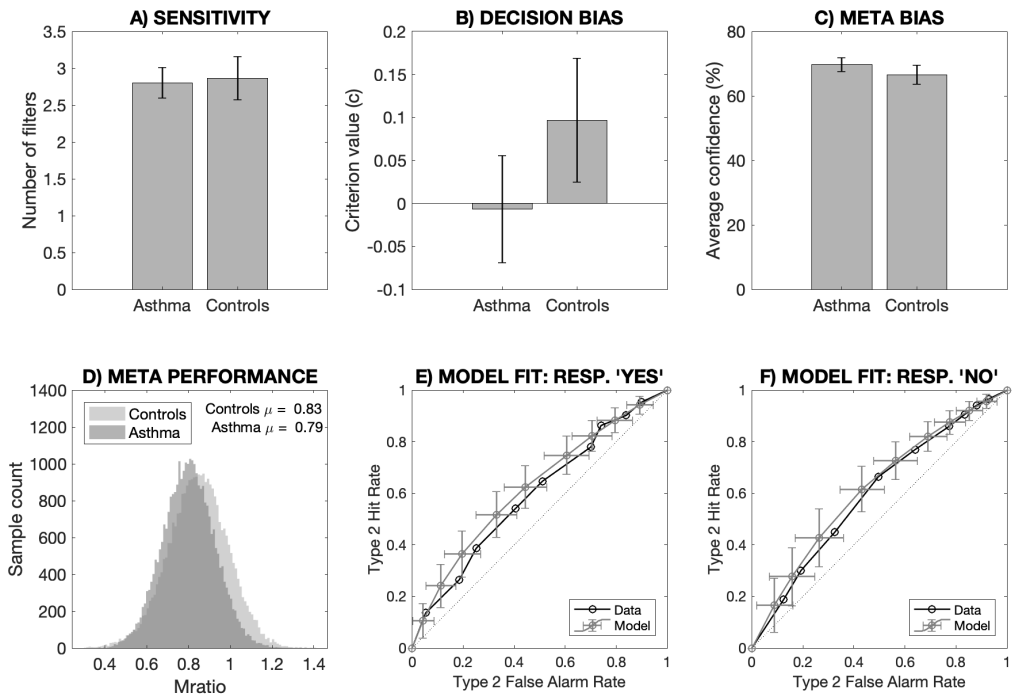


Figure 4. Group difference comparisons between individuals with asthma and healthy controls. Panels A, B and C denote group mean and standard error for each group regarding interoceptive sensitivity, decision bias and metacognitive bias, with no significant differences found between the groups. Panel D demonstrates the sampled posteriors for the group estimates of metacognitive performance (Mratio). Panels E and F demonstrate the model fit by comparing the observed and model estimates of the Type 2 ROC curves for both 'Yes' and 'No' responses (regarding the presence of an added inspiratory resistance). Model fits for all other models are presented in the supplementary material (Supplementary Figure 3).

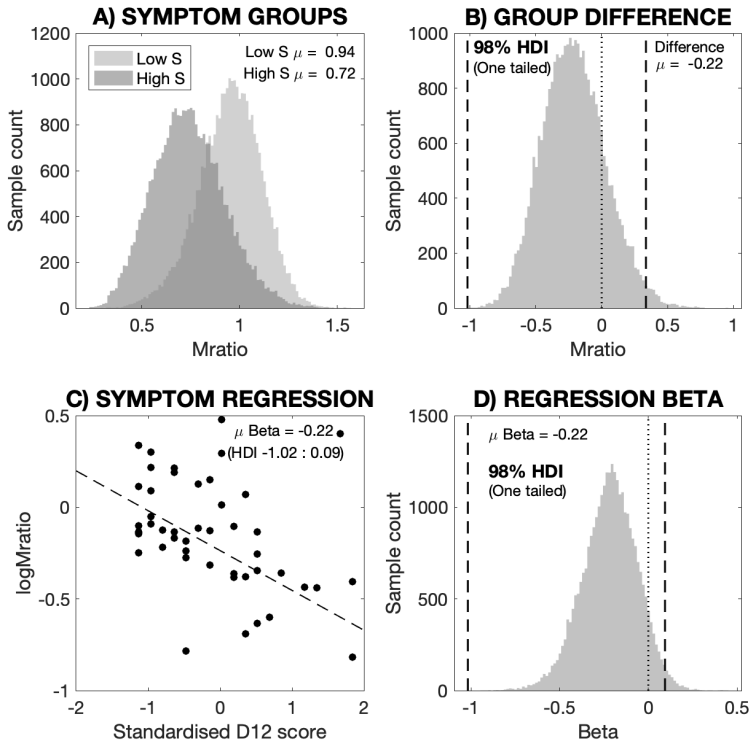


Figure 5. Comparisons between metacognitive performance (logMratio) and symptom load in asthma, using either a median-split analysis for high and low symptoms (panels A and B), or a hierarchical regression analysis (panels C and D). A) The sampled posteriors of the hierarchically-estimated metacognitive performance parameter (Mratio) for individuals with > median D12 scores (high symptom load group) or < median D12 score (low symptom load group). B) The distribution resulting from the difference between the group Mratio distributions, where the dashed lines represent the one-tailed 98% highest density interval (HDI), and the dotted line denotes zero. A HDI encompassing zero indicates a non-significant difference between the groups. C) A hierarchical regression predicting logMratio from the standardised D12 scores within asthma participants. The regression was fit using an extension of the HMeta-d model (R-HMeta-d) in which the beta regression coefficient was fit simultaneously together with the logMratio scores. Dashed line represents the regression line from the model fit. D) The distribution of samples over the regression beta parameter (from panel C) fit using the R-HMeta-d model. Dashed lines represent the one-tailed 98% HDI which encompasses zero (dotted line) consistent with no significant relationship between D12 score and metacognitive efficiency in this dataset.

Task performance algorithm: Simulations and empirical data

The results from both simulated and empirical data demonstrate the ability of the task algorithm to target performance towards a perceptual threshold that lies above chance (50%) and below ceiling (100%) performance (Figure 6). Simulations conducted using the task algorithm (with a 'constant' staircase formulation) produced task accuracy scores with a mean of 74.1 ± 8.7% (sd) and an accurate recovery of the 75% filter number, irrespective of the starting filter value (Figure 6A). Empirical data collected

using the Yes/No formulation of the task (with a constant staircase) produced a task accuracy with a mean of 68.9 ± 7.4% (sd), with the threshold filter number spread between 1 and 8 filters (Figure 6B). Both the simulations and real data demonstrate a feasible number of trials (70.4 ± 10.3 (sd) trials for simulated results with 60 threshold trials, 69.6 ± 10.0 (sd) trials for empirical data) required to complete 60 trials at the threshold filter. In real terms, this indicates that it is possible to reliably measure respiratory interoception and metacognition in approximately one hour or less (assuming approximately 45-60 seconds to complete each trial). These estimates include a constraint whereby the algorithm was additionally programmed to continue until 30 trials are completed at the threshold filter number, followed by the option for manual intervention to instigate filter changes every ten trials if the accuracy moves out of acceptable bounds (experimenter decision required, depending on time taken and participant).

When utilising a 'roving' staircase experimental design, simulations of the task algorithm produced task accuracy scores with a mean of 75.4 ± 8.0% (sd) and an accurate recovery of the 75% filter number, irrespective of the starting filter value (Supplementary Figure 4A). When the first 60 trials from the constant staircase empirical data discussed above were analysed as a roving staircase (i.e. the first 60 trials analysed, regardless of filter intensity), the mean task accuracy was slightly reduced from 68.9% to 67.7%, and the variance of the scores increased from 7.4% to 8.5% (sd) (Supplementary Figure 4B). A final dataset, collected with explicit use of the roving staircase paradigm, demonstrated a mean accuracy of 69.7 ± 11.7% (sd), and threshold filter numbers that ranged from 1-8 filters (Supplementary Figure 4C). No additional trials are required when utilising a roving staircase design, as all trials following the calibration step are included in the analysis. The simulated and empirical results from the calibration algorithm are presented in Supplementary Figure 5.

Direct comparison between the empirical data collected using the three methods (i.e. manual accuracy calculations every 10 trials, the constant staircase design and the roving staircase design) is provided in Figure 7. Data collected using a manual accuracy calibration (asthma and controls) produced task accuracy scores with a mean of 66.4 ± 8.2% (sd), with a large number of additional trials (percentage of the number of threshold trials collected = 71.9 ± 17.7% (sd)). No difference in accuracy was found (Wilcoxon rank-sum tests) between any of the task designs (all $p > 0.05$), however the roving task design produced the largest standard deviation in the task accuracy across the methods. Additionally, while a significant number of additional trials were required when using the manual and constant staircase methods (Wilcoxon signed-rank tests, $p < 0.001$ for both tests against zero), the constant staircase significantly reduced the additional number of trials required from the manual method, both as an absolute number of trials and as a percentage of the number of threshold trials collected (Wilcoxon rank-sum tests, $p < 0.001$ for both tests).

Finally, while using a roving staircase design removes the possibility that any additional (non-analysed) trials will be required, a tighter target accuracy band could be employed to better control the variance in performance accuracy, and the algorithm can be programmed to continue throughout all trials to better control task performance. Therefore, we re-ran the task simulations using a difference in beta cumulative distribution functions between 75-70% (reduced from 80-65%), with a lower bound on the acceptable probabilities increased from 20% to 30%. The results of these simulations compared to the original thresholds can be seen in Supplementary Figure 6, and these changes reduced the simulated standard deviation in task performance accuracy from 8.0 to 6.7.

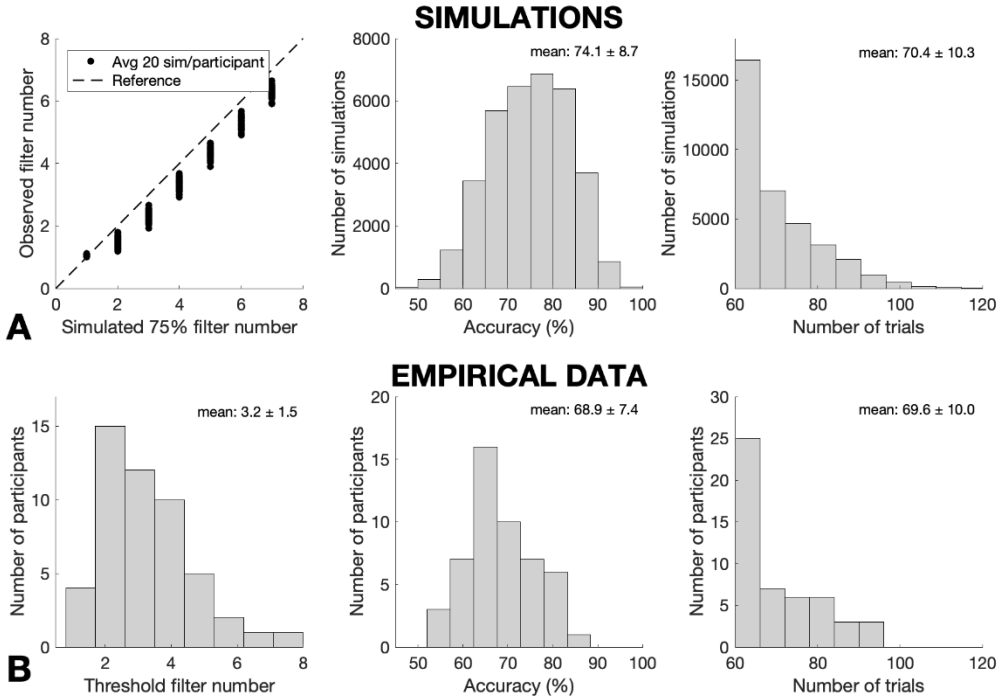


Figure 6. Results demonstrating the use of an adapted staircase algorithm for targeting a specified level of task difficulty over 60 trials. A) Simulation results, where data were generated from a range of logistic sigmoid functions bounded between 0.5 and 1, with 20 simulations for each sigmoid ('participant') from each of five starting points – from two filters below to two filters above the 75% threshold filter. Left: Simulated and recovered 75% filter number for each simulated 'participant'. Middle: Histogram of the task accuracy scores for the 60 threshold trials for all simulations. Right: Histogram of the total number of trials required to complete 60 threshold trials for each simulation. B) Data collected using a Yes/No version of the task (with a constant staircase), where 50 participants each completed 60 threshold trials. Left: Histogram of the measured threshold filter number for each participant. Middle: Histogram of the task accuracy scores for the 60 threshold trials for the 50 measured participants. Right: Histogram of the total number of trials required to complete 60 threshold trials for each participant. All histograms are reported with mean ± standard deviation.

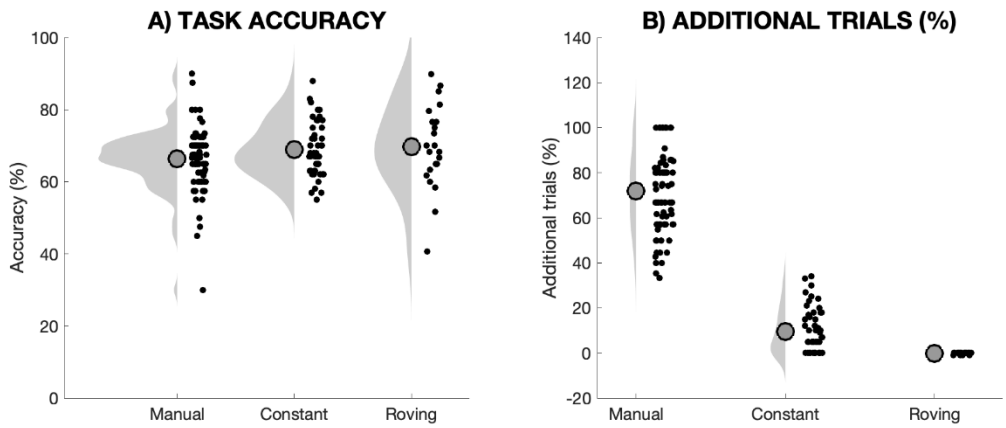


Figure 7. Comparison of the three empirical datasets collected using different methods: 'Manual' = Manual staircase adjustment of filters via accuracy calculations every 10 trials; 'Constant' = Constant formulation of the staircase, where all analysed trials are collected at the same filter number; and 'Roving' = Roving staircase, where all trials (across different filter numbers) are used for data analysis. A) Comparison of task accuracy across the data collection methods, where no difference in accuracy was observed between any of the methods (all $p > 0.05$). B) Comparison of the additional trials required during data collection as percentage of the analysed

threshold trials (N.B., no extra trials are collected for the roving staircase design). In Panel B, both the manual and constant staircase methods produced a significant percentage of additional trials (above zero, both $p < 0.001$), and the constant staircase significantly reduced the number / percentage of these trials compared to the manual adjustment method (both $p < 0.001$). Black dots are individual data points, while grey areas represent the distribution of values. Large circles denote the group mean in each condition.

Discussion

Main findings

In this manuscript we demonstrate the utility of pairing a breathing perception task – the 'Filter detection task' – with the HMeta-d model of metacognition, to quantify the interoceptive domains of sensitivity, decision bias, metacognitive bias and metacognitive performance. Using simulations we have shown how the use of a hierarchical model formulation (HMeta-d[13]) can help overcome the challenge of low trial numbers when calculating metacognitive performance metrics. We have also demonstrated how this hierarchical model can be extended to include a simultaneous hierarchical estimation of regression parameters linking metacognitive performance to individual difference variables. We also demonstrate the use of the model and appropriate statistics to answer research hypotheses in an empirical dataset of healthy controls and individuals with asthma. Lastly, we introduce a task algorithm to help target performance accuracy towards an ideal band of 70-75%, to reduce the number of unused trials performed outside of this accuracy band. We demonstrate the effectiveness of this algorithm using both simulations and via empirical data comparisons when using either manual adjustment strategies or the staircase options provided by the toolbox (constant or roving staircase).

Computational models of breathing-related interoception and metacognition

As interest in interoception-related research grows across neuroscience, psychiatry, physiology and other scientific communities, the importance of developing robust methodologies for quantification of interoceptive dimensions is paramount. While discussions regarding the validity of tasks such as heartbeat counting in the cardiac domain highlight the need for robust measures of interoception[29], the FDT offers one route to overcoming some of these issues within the domain of respiration. Here, we highlight the feasibility of applying signal detection theory-derived computational models of both task and metacognitive performance, first introduced by Maniscalco and Lau[11] (the meta-d' model) and derived from theories of 'Type 2' performance (distinguishing between one's own correct and incorrect decisions[30]). Utilising these signal detection theory models firstly allows us to separate interoceptive sensitivity from decision biases within task (or 'Type 1') performance, both of which may be highly informative in disentangling drivers of altered interoception. For example, while there have been reports of possible blunted sensitivity to inspiratory resistive loads with anxiety disorders[31], there is also an established prevalence of reporting medically unexplained symptoms with anxiety[32], and even early evidence for a potential relationship between symptom over-report and reduced interoceptive accuracy in healthy individuals[33]. Therefore, as a criterion shift may manifest as differences in interoceptive sensitivity, it is imperative to separate these measures both in healthy individuals and within clinical populations.

While perceptual sensitivity and bias metrics can be directly calculated from behavioural data, the estimation of metacognitive parameters such as meta-d' often require optimising a model's predicted responses to match those observed within the data[11]. However, here we demonstrate that the original meta-d' model formulation (using maximum likelihood parameter estimates) is not able to significantly recover group Mratio values below 1 or reliable estimates of individual subject scores when using the low number of trials that are practically feasible within interoceptive experiments (Figure 2). Due to these constraints here we instead explore the utility of hierarchical formulations of the meta-d' model

(HMeta-d) derived by Fleming[13], which can achieve good recovery of metacognitive performance parameters (e.g. Mratio) using as few as 20 trials per subject (Figure 2). Importantly, the meta-d' model allows us to differentiate *relative* metacognitive performance (i.e. metacognitive efficiency controlling for task performance) from *absolute* measures of metacognition, such as that calculated from the area under a type 2 ROC curve[14,15,24]. This is important because it is well-established that absolute measures of metacognition may be biased by differences in underlying task performance between individuals or conditions[11].

Beyond estimating group metrics of metacognition, often it may be desirable to estimate the relationship between individual metacognitive performance and an external measure of interest, for example a clinical score or other behavioural variable. While post-hoc regressions on single-subject parameter estimates are possible, hierarchical models tend to shrink single-subject estimates towards the group mean, thus losing information regarding individual differences and reducing the power of these types of analyses. To this end, we have developed and tested a hierarchical regression model, whereby multiple regression parameters can be simultaneously fit alongside the group logMratio within the HMeta-d model (referred to throughout as RHMeta-d). We find that the sensitivity of the regression model in being able to accurately recover simulated beta coefficients is greatly enhanced when increasing from 20 to 40 and 60 trials, with the width of the posterior (represented by the HDI) notably reducing when trial number is increased (Figure 3). We have also demonstrated the use of this regression approach in an empirical dataset in which interoceptive metacognitive performance was compared against breathlessness symptom reports (measured via the D12 questionnaire) in individuals with asthma (Figure 4)[13].

FDT toolbox

To aid the use of computational models within interoceptive experiments, we have developed a toolbox to run the FDT according to an accuracy-targeted performance algorithm (freely available for download: https://github.com/ofaull/FDT). While practicalities regarding the step-size of each of the inspiratory resistance filters prevents us from utilising established psychophysics staircases, we have instead developed an adapted staircase protocol which prompts adjustment of the filter load once the probability falls too far beyond our desired range of 70-75%. Both simulations and empirical data show that this algorithm produces performances within the desired range required for employing the computational models described above, where participants need to be performing above chance but below 100% accuracy.

The demonstration of the FDT in the current manuscript utilised a Yes/No task formulation, where a participant is required to answer whether or not a resistance was added to the system ('Yes') or stayed the same ('No'). However, the toolbox also provides the option to employ a two-interval forced choice (2IFC) alternative if desired. While criticisms exist of the application of equal-variance signal detection theory metrics in Yes/No tasks (discussed previously[18]), we also see potential practical utility in using these task variants, for example to quantify measures akin to symptom over- or under-report by estimating the criterion parameter. However, if the metrics calculated from the FDT are to be compared with other perceptual tasks that are run as a 2-interval/alternative forced choice, then the 2IFC option may be desirable, allowing for comparable model assumptions across tasks.

Lastly, the toolbox also offers two alternative staircase options to control task performance in either a Yes/No or 2IFC formulation. In the original protocol presented by Garfinkel and colleagues[15], the aim was to collect 20 usable trials at a specific number of filters where performance first fell below 75%, thus corresponding to the participant's perceptual threshold and providing a measure of interoceptive sensitivity. Whilst the number of additional (unused) trials required can be greatly reduced by employing the adapted staircase algorithm in the toolbox (Figure 7), an alternative approach is to employ a 'roving' staircase, whereby all trials are used in the calculation of interoceptive measures, and interoceptive sensitivity is taken as an average of the filter numbers used across trials. As the risk of

needing additional trials is removed, this approach allows experimenters to tighten accuracy thresholds to improve task performance control, as the aim of this staircase is no longer to find a single filter that elicits the desired accuracy. This roving staircase option would likely prove a more viable alternative if using the FDT in a clinical setting, removing the possibility of any additional trials while maintaining adequate representations of interoceptive sensitivity. We note however that roving staircases also have potential downsides in artificially inflating estimates of metacognitive sensitivity when compared to constant-stimulus designs (see Rahnev and Fleming[34] for further discussion of this issue).

Conclusions

Here we present a breathing-related interoceptive application of a computational model designed to tease apart important aspects of perception: sensitivity, decision bias, metacognitive bias and metacognitive performance. Whilst interoceptive experiments often suffer from low trial numbers, by combining a breathing perception task with a hierarchical statistical model we were able to develop a robust algorithm to control task performance while maximising the number of useful trials for analysis.

The FDT toolbox is freely available for download (https://github.com/ofaull/FDT), as are the statistical methods employed (MLE model: http://www.columbia.edu/~bsm2105/type2sdt/; HMeta-d: https://github.com/metacoglab/HMeta-d/).

Acknowledgements

OKH (née Faull) is a Marie Skłodowska-Curie Postdoctoral Fellow that is supported by the European Union's Horizon 2020 research and innovation programme under the Grant Agreement No 793580. MA is supported by a Lundbeckfonden Fellowship (R272-2017-4345), and an AIAS-COFUND II fellowship, which is supported by the Marie Skłodowska-Curie actions under the European Union's Horizon 2020 (Grant agreement no 754513), and the Aarhus University Research Foundation. SJH was supported by the grant #2017-403 of the Strategic Focal Area "Personalized Health and Related Technologies (PHRT)" of the ETH Domain. KES is supported by the René and Susanne Braginsky Foundation and the University of Zurich. KTSP was supported by the JABBS Foundation for this work. SMF is supported by a Sir Henry Dale Fellowship jointly funded by the Wellcome Trust and Royal Society (206648/Z/17/Z).

References

1. Khalsa, S. S. *et al.* Interoception and Mental Health: a Roadmap. *Biological Psychiatry: Cognitive Neuroscience and Neuroimaging* 1–57 (2017). doi:10.1016/j.bpsc.2017.12.004
2. Pezzulo, G., Rigoli, F. & Friston, K. Active Inference, homeostatic regulation and adaptive behavioural control. *Progress in neurobiology* **134,** 17–35 (2015).
3. Stephan, K. E. *et al.* Allostatic Self-efficacy: A Metacognitive Theory of Dyshomeostasis-Induced Fatigue and Depression. *Front Hum Neurosci* **10,** 49–27 (2016).
4. Petzschner, F. H., Weber, L. A. E., Gard, T. & Stephan, K. E. Computational Psychosomatics and Computational Psychiatry: Toward a Joint Framework for Differential Diagnosis. *Biological Psychiatry* 1–10 (2017). doi:10.1016/j.biopsych.2017.05.012
5. Allen, M. Unravelling the Neurobiology of Interoceptive Inference. *Trends in Cognitive Sciences* **24,** 265–266 (2020).
6. Owens, A. P., Allen, M., Ondobaka, S. & Friston, K. J. Interoceptive inference: From computational neuroscience to clinic. *Neuroscience & Biobehavioral Reviews* **90,** 174–183 (2018).
7. Brefczynski, J. A. & DeYoe, E. A. A physiological correlate of the 'spotlight' of visual attention. *Nat Neurosci* **2,** 370–374 (1999).

8. Bocanegra, B. R. & Zeelenberg, R. Emotion improves and impairs early vision. *Psychol Sci* **20,** 707–713 (2009).
9. Summerfield, C. & Egner, T. Expectation (and attention) in visual cognition. *Trends in Cognitive Sciences* **13,** 403–409 (2009).
10. Kingdom, F. & Prins, N. *Psychophysics: a practical introduction*. (Elsevier, 2016).
11. Maniscalco, B. & Lau, H. A signal detection theoretic approach for estimating metacognitive sensitivity from confidence ratings. *Consciousness and cognition* **21,** 422–430 (2012).
12. Fleming, S. M. & Lau, H. C. How to measure metacognition. *Front Hum Neurosci* **8,** 443 (2014).
13. Fleming, S. M. HMeta-d: hierarchical Bayesian estimation of metacognitive efficiency from confidence ratings. *Neuroscience of Consciousness* 377–14 (2017). doi:10.1093/nc/nix007
14. Garfinkel, S. N., Seth, A. K., Barrett, A. B., Suzuki, K. & Critchley, H. D. Knowing your own heart: Distinguishing interoceptive accuracy from interoceptive awareness. *Biological Psychology* **104,** 65–74 (2015).
15. Garfinkel, S. N. *et al.* Interoceptive dimensions across cardiac and respiratory axes. *Philos. Trans. R. Soc. Lond., B, Biol. Sci.* **371,** 20160014–10 (2016).
16. Paulus, M. P. The breathing conundrum - Interoceptive sensitivity and anxiety. *Depress. Anxiety* **30,** 315–320 (2013).
17. Khatib, M. F., Oku, Y. & Bruce, E. N. Contribution of Chemical Feedback Loops to Breath-to-Breath Variability of Tidal Volume. *Respiration Physiology* **83,** 115–127 (1991).
18. Peters, M. A. K., Ro, T. & Lau, H. Who's afraid of response bias? *Neuroscience of Consciousness* **2016,** niw001–8 (2016).
19. Green, D. M. & Swets, J. A. *Signal detection theory and psychophysics*. (1966).
20. Stanislaw, H. & Todorov, N. Calculation of signal detection theory measures. *Behav Res Methods Instrum Comput* **31,** 137–149 (1999).
21. Lee, A. L. F., Ruby, E., Giles, N. & Lau, H. Cross-Domain Association in Metacognitive Efficiency Depends on First-Order Task Types. *Front. Psychol.* **9,** 440–10 (2018).
22. Porges, S. W. Body Perception Questionnaire. 1–3 (1993).
23. Rouault, M., Seow, T., Gillan, C. M. & Fleming, S. M. Psychiatric Symptom Dimensions Are Associated With Dissociable Shifts in Metacognition but Not Task Performance. *Biological Psychiatry* 1–9 (2018). doi:10.1016/j.biopsych.2017.12.017
24. Garfinkel, S. N. *et al.* Discrepancies between dimensions of interoception in autism: Implications for emotion and anxiety. *Biological Psychology* **114,** 117–126 (2016).
25. Moutoussis, M., Hopkins, A. K. & Dolan, R. J. Hypotheses About the Relationship of Cognition With Psychopathology Should be Tested by Embedding Them Into Empirical Priors. *Front. Psychol.* **9,** 1–3 (2018).
26. Gelman, A. *et al.* in *Bayesian Data Analysis* 435–446 (2020).
27. Yorke, J., Moosavi, S. H., Shuldham, C. & Jones, P. W. Quantification of dyspnoea using descriptors: development and initial testing of the Dyspnoea-12. *Thorax* **65,** 21–26 (2010).
28. Watson, A. B. & Pelli, D. G. QUEST: a Bayesian adaptive psychometric method. *Percept Psychophys* **33,** 113–120 (1983).
29. Corneille, O., Desmedt, O., Zamariola, G., Luminet, O. & Maurage, P. A heartfelt response to Zimprich et al. (2020), and Ainley et al. (2020)'s commentaries_ Acknowledging issues with the HCT would benefit interoception research. *Biological Psychology* **152,** 107869 (2020).
30. Galvin, S. J., Podd, J. V., Drga, V. & Whitmore, J. Type 2 tasks in the theory of signal detectability: discrimination between correct and incorrect decisions. *Psychonomic bulletin \& review* **10,** 843–876 (2003).
31. B, J. T. M. B. C., Michael Pain M D, F. C. C. P. & Sc, N. B. B. Anxiety Disorder and Perception of Inspiratory Resistive Loads. *Chest* **91,** 547–551 (1987).
32. Steinbrecher, N., Koerber, S., Frieser, D. & Hiller, W. The Prevalence of Medically Unexplained Symptoms in Primary Care. *PSYM* **52,** 263–271 (2011).
33. Bogaerts, K. *et al.* High symptom reporters are less interoceptively accurate in a symptom-related context. *J Psychosom Res* **65,** 417–424 (2008).
34. Rahnev, D. & Fleming, S. M. How experimental procedures influence estimates of metacognitive ability. *Neuroscience of Consciousness* **2019,** 415–9 (2019).

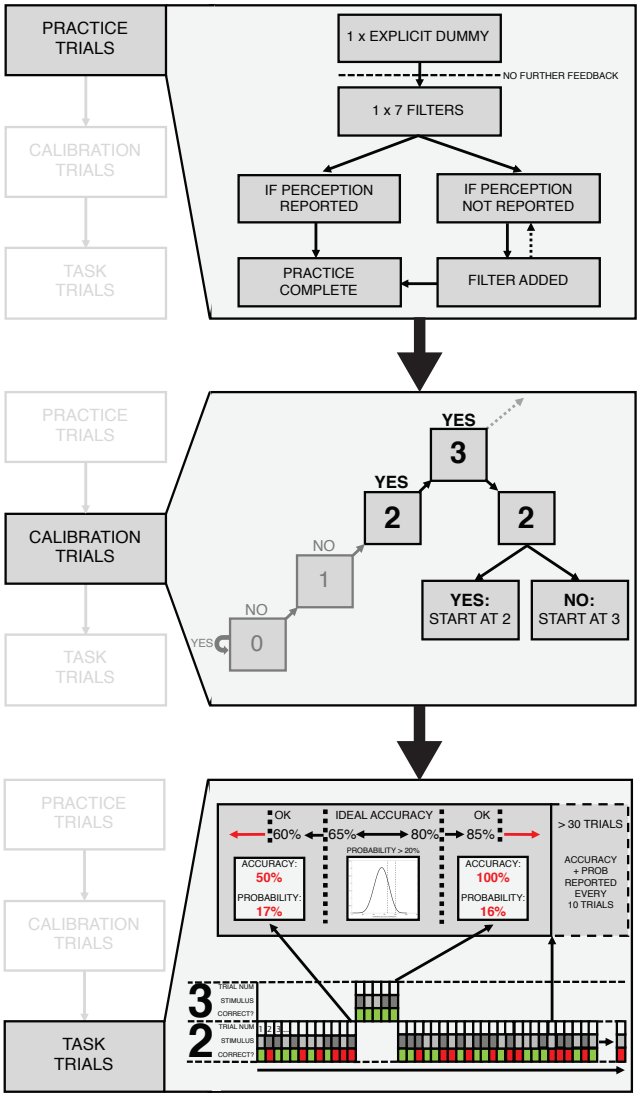
Supplementary Material:
The Filter Detection Task for measurement of breathing-related interoception and metacognition

Olivia K. Harrison[1], Sarah N. Garfinkel[2], Lucy Marlow[3], Sarah Finnegan[3], Stephanie Marino[1], Laura Nanz[1], Micah Allen[4,5,6], Johanna Finnemann[6], Laura Keur-Huizinga[6], Samuel J. Harrison[1], Klaas Enno Stephan[1,7,8], Kyle Pattinson[3], Stephen M. Fleming[7,9,10]

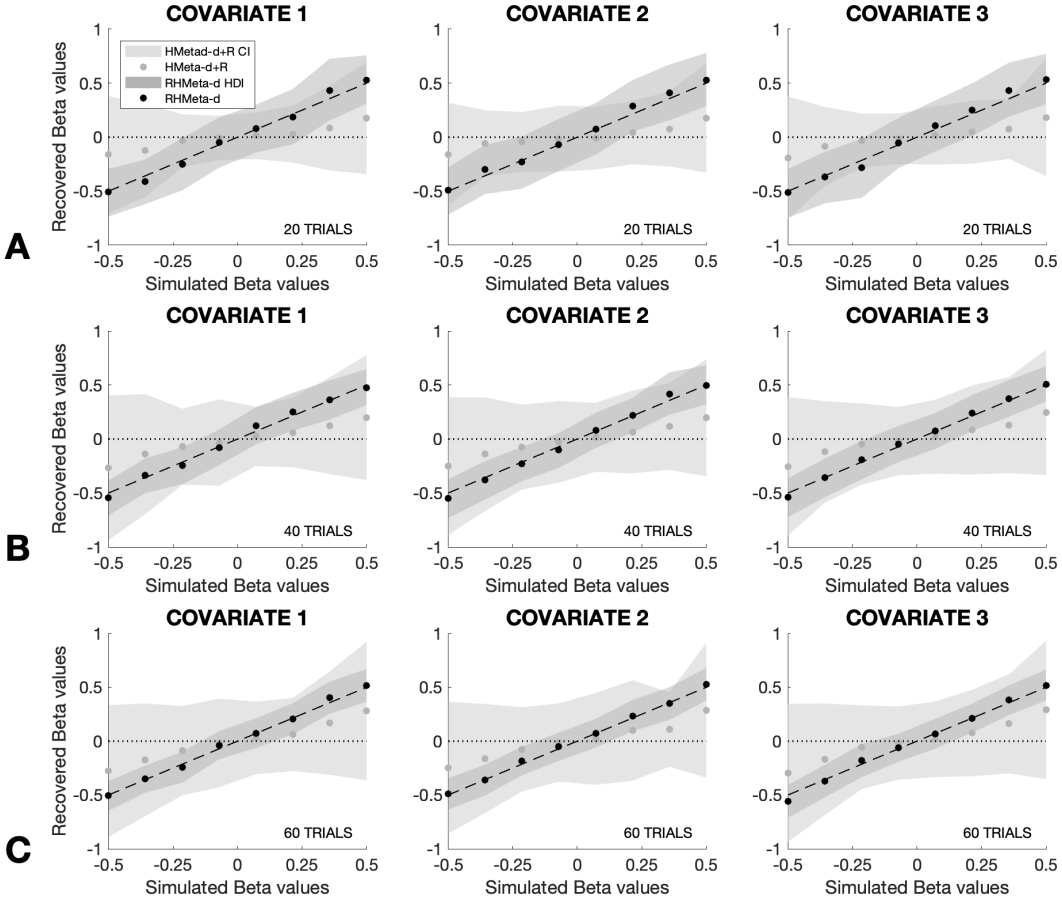
[1] Translational Neuromodeling Unit, Institute for Biomedical Engineering, University of Zurich and ETH Zurich, Switzerland
[2] Brighton and Sussex Medical School, University of Sussex, United Kingdom
[3] Nuffield Department of Clinical Neurosciences, University of Oxford, United Kingdom
[4] Aarhus Institute of Advanced Studies, Aarhus University, Denmark
[5] Center of Functionally Integrative Neuroscience, Aarhus University Hospital, Denmark
[6] Cambridge Psychiatry, University of Cambridge, United Kingdom
[7] Wellcome Centre for Human Neuroimaging, University College London, United Kingdom
[8] Max Planck Institute for Metabolism Research, Cologne, Germany
[9] Max Planck Centre for Computational Psychiatry and Ageing Research, University College London, United Kingdom
[10] Department of Experimental Psychology, University College London, United Kingdom

Key words: interoception, breathing, inspiratory resistance, metacognition

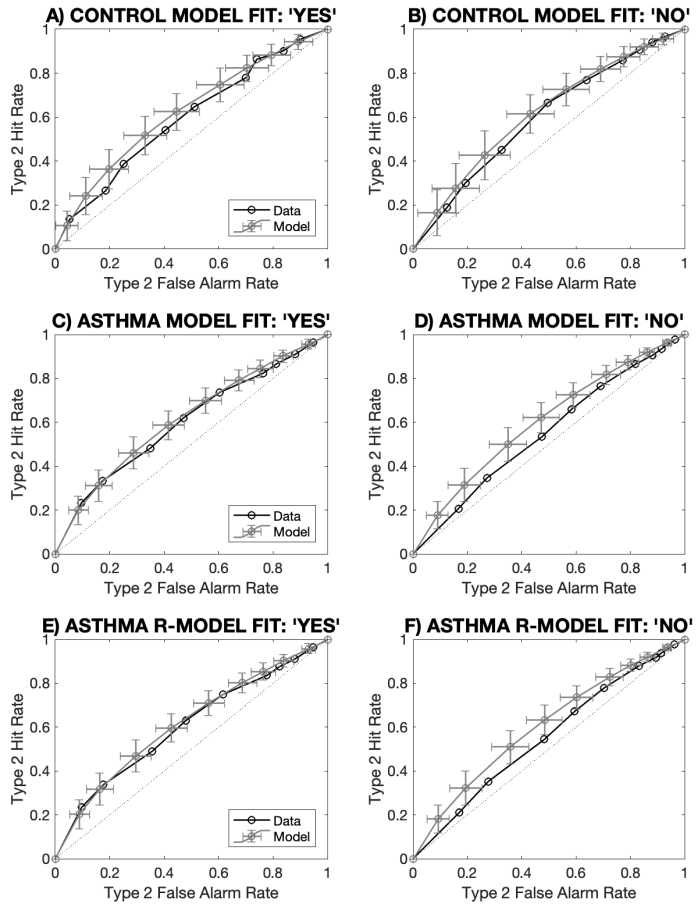
Corresponding author:
Dr Olivia Faull
Translational Neuromodeling Unit
Institute for Biomedical Engineering
University of Zurich and ETH Zurich



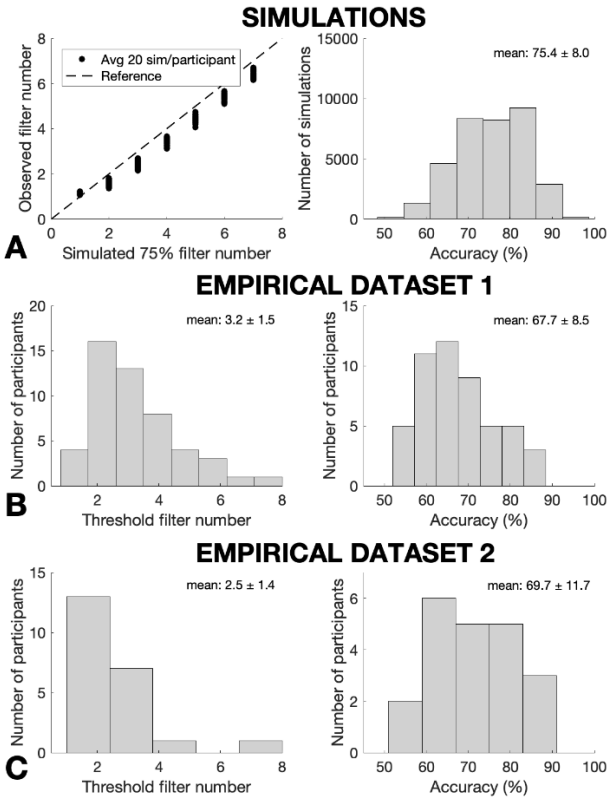
Supplementary Figure 1. Visualisation of the task performance algorithm. TOP: Practice trials, consisting of an explicit dummy (where the participants are told a sham resistance is added), followed by a large load (7 filters) where no feedback is given (this is then maintained for the rest of the experiment). If no resistance is perceived, filters are added until a resistance is reported. MIDDLE: Calibration trials, where (starting from the dummy), filters are added until two consecutive resistances are reported. Following this, one final calibration trial is performed with one less filter, to determine the starting value for the task trials. BOTTOM: Task trials, where cumulative task accuracy at each trial is transformed (using a beta distribution) into the distribution of underlying accuracies that could have produced the task performance. An upper bound (here 80%) and a lower bound (here 60%) is used to calculate the probability that the participant is performing at the targeted accuracy. If this probability falls below the error risk threshold (here 20%), a filter change is prompted – either the addition of a filter if the accuracy is too low, or the removal of a filter if the accuracy is too high. This continues until either a specified number of trials (here 60 trials) are completed at either one filter number (using the 'constant staircase' task design) or at a range of filter numbers (using the 'roving staircase' task design), the latter requiring no additional trials to be measured that will not be used in the analysis. If a constant staircase is used, the algorithm is stopped at 30 trials, and experimenter intervention can occur every 10 trials subsequently if task performance is drastically altered and no longer deemed acceptable.



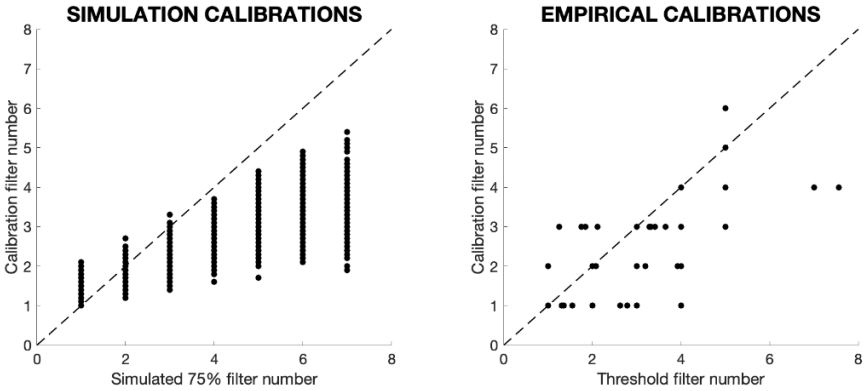
Supplementary Figure 2. Demonstration of the recovery of three group regression parameters (betas) against the logMratio using either the original HMeta-d model combined with a post-fit linear regression (HMeta-d+R), or an extended multiple regression HMeta-d (RHMeta-d) model, the latter where group linear regression parameters are simultaneously hierarchically fit alongside logMratio. Results are shown for simulations using 20 trials (A), 40 trials (B) and 60 trials (C). Grey areas denote the 95% highest density interval of the sampled estimate. Dashed lines represent ideal recovery of group beta values, and dotted lines at zero demonstrating the ability of the model fit to significantly recover group estimates (with highest density intervals not including zero).



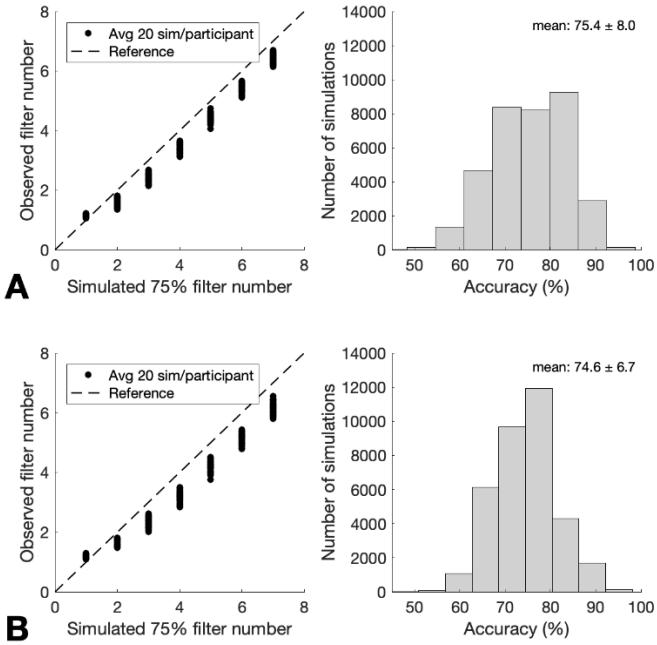
Supplementary Figure 3. Demonstrations of the model fits by comparing the observed and model estimates of the Type 2 ROC curves for both 'Yes' and 'No' responses (regarding the presence of an added inspiratory resistance). Model fits are shown for the control data fit (A and B), the asthma data fit (C and D), and for the asthma data extended regression model (E and F).



Supplementary Figure 4. Results demonstrating the use of an adapted roving staircase algorithm for targeted task difficulty over 60 trials. A) Simulation results, where data were generated from a range of logistic sigmoid functions bounded between 0.5 and 1, with 20 simulations for each sigmoid ('participant') from each of five starting points – from two filters below to two filters above the 75% threshold filter. Left: Simulated and recovered 75% filter number for each simulated 'participant'. Right: Histogram of the task accuracy scores for the 60 threshold trials for all simulations. B) Data collected using a Yes/No version of the task (with a constant staircase, but where the first 60 trials – regardless of filter number – were analysed to represent a roving staircase design), where 50 participants were measured. Left: Histogram of the measured threshold filter number for each participant. Right: Histogram of the task accuracy scores for the 60 threshold trials for the 50 measured participants. C) Data collected using a Yes/No version of the task (with a roving staircase), where 20 participants each completed 60 trials (total). Left: Histogram of the measured threshold filter number for each participant. Right: Histogram of the task accuracy scores for the 60 threshold trials for the 50 measured participants. All histograms are reported with mean ± standard deviation.



Supplementary Figure 5. Results for the calibration algorithm from both simulated and empirical data. Left: Simulated calibration filter plotted against the threshold (75%) filter, demonstrating a bias towards lower calibration values. Right: Empirical calibration results plotted against the task threshold filter number for data from both constant and roving staircase designs (roving staircase threshold filter numbers are calculated as the average filter number across the task trials).



Supplementary Figure 6. Simulated results demonstrating the use of an adapted roving staircase algorithm for targeted task difficulty over 60 trials, using two different sets of thresholds for prompting a change in filter number. In both panels, data were generated from a range of logistic sigmoid functions bounded between 0.5 and 1, with 20 simulations for each sigmoid ('participant') from each of five starting points – from two filters below to two filters above the 75% threshold filter. A) Data were generated using an upper bound of 80% and a lower bound of 65% on the beta distributions calculated from task performance scores, using 20% as a false positive threshold where probabilities below this threshold prompted a filter change. B) Data were generated using an upper bound of 75% and a lower bound of 70% on the beta distributions calculated from task performance scores, using 30% as a false positive threshold where probabilities below this threshold prompted a filter change. Both panels – Left: Simulated and recovered 75% filter number for each simulated 'participant'. Right: Histogram of the task accuracy scores for the 60 threshold trials for all simulations. Data from B show a very similar mean and reduced standard deviation for the resulting task performance accuracies generated.