Received: 10 January 2023 | Revised: 29 September 2023 | Accepted: 9 October 2023

DOI: 10.1002/ejp.2193

EJP European Journal of Pain

ORIGINAL ARTICLE

Test–retest reliability of Bayesian estimations of the effects of stimulation, prior information and individual traits on pain perception

Ariane Delgado-Sanchez[1] | **Christiana Charalambous**[1] | **Nelson J. Trujillo-Barreto**[1] | **Hannah Safi**[2,3] | **Anthony Jones**[1] | **Manoj Sivan**[4] | **Deborah Talmi**[5] | **Christopher Brown**[6]

[1]School of Health Sciences, University of Manchester, Manchester, UK

[2]Department of Medical Physics, Salford Royal Foundation Trust, Northern Care Alliance, Salford, UK

[3]Department of Electrical and Electronic Engineering, School of Engineering, University of Manchester, Manchester, UK

[4]Leeds Institute of Rheumatic and Musculoskeletal Medicine, University of Leeds, Leeds, UK

[5]Department of Psychology, University of Cambridge, Cambridge, UK

[6]Institute of Population Health, University of Liverpool and Human Pain Research Group, University of Liverpool, Liverpool, UK

Correspondence
Ariane Delgado-Sanchez, School of Health Sciences, University of Manchester, Manchester, UK.
Email: a.delgado.sanchez@mmu.ac.uk

Abstract

Background: There is inter-individual variability in the influence of different components (e.g. nociception and expectations) on pain perception. Identifying the individual effect of these components could serve for patient stratification, but only if these influences are stable in time.

Methods: In this study, 30 healthy participants underwent a cognitive pain paradigm in which they rated pain after viewing a probabilistic cue informing of forthcoming pain intensity and then receiving electrical stimulation. The trial information was then used in a Bayesian probability model to compute the relative weight each participant put on stimulation, cue, cue uncertainty and trait-like bias. The same procedure was repeated 2 weeks later. Relative and absolute test–retest reliability of all measures was assessed.

Results: Intraclass correlation results showed good reliability for the effect of the stimulation (0.83), the effect of the cue (0.75) and the trait-like bias (0.75 and 0.75), and a moderate reliability for the effect of the cue uncertainty (0.55). Absolute reliability measures also supported the temporal stability of the results and indicated that a change in parameters corresponding to a difference in pain ratings ranging between 0.47 and 1.45 (depending on the parameters) would be needed to consider differences in outcomes significant. The comparison of these measures with the closest clinical data we possess supports the reliability of our results.

Conclusions: These findings support the hypothesis that inter-individual differences in the weight placed on different pain factors are stable in time and could therefore be a possible target for patient stratification.

Significance: Our results demonstrate the temporal stability of the weight healthy individuals place on the different factors leading to the pain response. These findings give validity to the idea of using Bayesian estimations of the influence of different factors on pain as a way to stratify patients for treatment personalization.

Deborah Talmi and Christopher Brown shared senior authorship.

1 | INTRODUCTION

Pain perception is affected by multiple sensory and psychological components (Williams & Craig, 2016). Evidence shows that there is inter-individual variability in the influence of these different components, such as nociceptive focus, affecting pain perception. For instance, participants with higher levels of mindfulness have shown lower levels of perceptual biases induced by intensity informative cues (Lim et al., 2020).

In clinical settings, patients' symptoms have also been proposed to be a combination of the influence of nociception, expectations and the relative precision/weight of these components (Van den Bergh et al., 2017). Consequently, both the experimental and clinical literature suggest a potential clinical use for identifying the weight of each component on pain perception. For example, this approach could be used for patient stratification to inform individualized therapies, in which treatments could be selected based on the specific factors driving the pain response of each patient. Patient stratification is particularly relevant in chronic pain due to low treatment success and high variability in responses, both in pharmacological (Moore et al., 2013) and psychological therapies (e.g. mindfulness or cognitive behavioural therapy) (Cherkin et al., 2016; Morley, 2011; Sturgeon, 2014).

One possible approach to quantify the relative weights placed on the factors affecting an individual's pain perception is the use of Bayesian modelling (Knill & Pouget, 2004), which allows for the computation of the weight placed on each factor and the associated probability distributions. Recently, Hoskin et al. (2019) proposed an experimental paradigm with associated models that showed promise at identifying the effect of stimulation, cues and trait-like bias in the pain ratings of each participant.

The use of Bayesian modelling to quantify the influences on pain at the individual level, although promising, is only likely to be clinically valuable for treatment stratification if the estimated weight placed on each factor is relatively stable over time. Pain perception is influenced by both enduring psychological factors, such as personality and pain catastrophizing (Pulvers & Hood, 2013), and state factors, such as mood (Graham-Engeland et al., 2016; Letzen & Robinson, 2017). Furthermore, the test–retest reliability of current pain-related measures (e.g. Quantitative Sensory Testing or pain rating scales) ranges from poor to excellent (Alghadir et al., 2018; Jurth et al., 2014; Nothnagel et al., 2017), indicating that not all are stable, which further motivates the need to test reliability of new pain-related measures.

This study aimed to test the reliability of the weight placed on sensory input, expectations and trait-like bias. To do so, we used a cued pain experimental paradigm (Hoskin et al., 2019) associated with a Bayesian model that computed the weight placed on each of the three factors. Participants completed the paradigm twice with a 2-week interval between visits (Streiner et al., 2015). The test–retest reliability of the weight parameters obtained per participant was then calculated.

2 | MATERIALS AND METHODS

2.1 | Participants

We recruited 30 healthy participants (13 females and 17 males). A power analysis showed that for a conservative estimation of minimum acceptable reliability of 0.5 (threshold of moderate reliability), power of 0.8 and significance of 0.05, the necessary sample size would be 28 participants. Four participants were excluded from the final analysis: two due to failure to attend the second session; a third because in the second session, they performed the psychophysics procedure and practice trials similarly to others, but then rated every stimulus as 0, likely due to equipment malfunction; and a fourth for failing our accuracy criterion (see procedure). Our final sample consisted of 26 participants (12 females, mean age 33.65 and SD 9.91). Participants did not present substantial psychiatric history (including depression and anxiety) and were not under prescription medications that might influence their brain functioning (e.g. antidepressants, antipsychotics and analgesics for chronic pain). Furthermore, they did not have a history of drug or alcohol abuse, and/or a history of chronic pain, and they were not taking analgesic medication for acute pain or inflammation. Volunteers were reimbursed at a rate of £10 per hour. The study received ethical approval by the University of Manchester University Research Ethics Committee 2 (UREC 2) and all participants gave written informed consent.

2.2 | Experimental procedure

Electrical stimulation was delivered through a built-in house electrical stimulator (Medical Physics department, Salford Royal Hospital) that was connected to the participant by a ring electrode placed on the dorsal side of the non-dominant hand. In order to improve skin conductance and ensure homogeneous stimulation between participants, prior to the experiment, the skin was prepared with the use of Nuprep Skin Preparation Gel and Ten20 Conductive Paste.

Participants first completed a psychophysics procedure in which they were asked to rate increasing levels of electrical stimulation on a scale of 0 to 10, with anchors at 3

(pain threshold) and 7 (highest tolerable pain to be repeatedly presented). The stimuli had a duration of 2 ms, and from one stimulation to the next intensity was increased by 1.25 mA. The psychophysics procedure was repeated twice to control for habituation effects. The stimulations associated with ratings from 3 to 7 in the second run were used for the cognitive task paradigm. This is concordant with the procedure followed by previous research (Hoskin et al., 2019). The cognitive paradigm was presented with the use of Psychtoolbox in MATLAB v2020a.

A trial summary can be found in Figure 1. Each trial started with the presentation of a fixation cross for 250 ms. Afterwards, the trial continued with a choice task between a target cue (cue that participants were instructed to choose) and a lure cue. The goal of the choice task was to ensure participants' attention. Both cues (target and lure) were composed of two (spade suit) playing cards with different combinations of values ranging from 3 to 7. For example, the target may be composed of the cards depicting a '4' and a '5', and the lure the cards '7' and '5'. The number on the cards corresponded to the levels the participant rated in the psychophysics procedure. When the two cards within one cue (target or lure) depicted the same number, there was a 100% chance of getting a stimulation previously rated as the value represented in the cards. If the cue was composed of two different cards, there was a 50% chance of receiving either stimulation level. In the example, if the participant chose the target cue, they had 50% chance of getting the stimulation corresponding to 4 and 50% chance of getting the stimulation corresponding to 5. We composed the target and lure combinations so that the target cue would always be the preferred choice, in the sense that it minimized chance of high pain. This way, the lure cue was always composed of one card that depicted the same value as one of the target cards, and another card that depicted a higher value than the other card in the target cue. Consequently, choosing the lure cue would lead to either the same level of pain as the target or higher pain, making this the target cue the one with a potential gain. Participants were trained to choose the target cue. The instructions of the task explicitly explained the difference between target and lure and informed them that the goal was to pick the target. Furthermore, feedback was provided on their choice through six practice trials completed before the experiment.

Sixteen conditions were presented, each one representing a combination of target cue and associated stimulation. In Table 1, a description of all the conditions can be

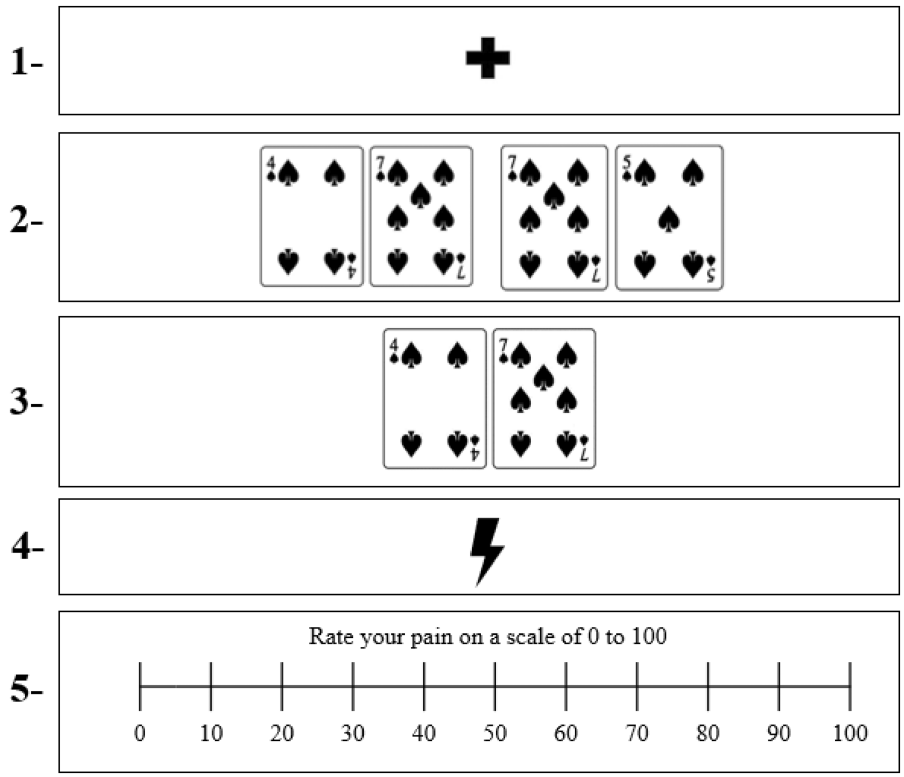

FIGURE 1 Trial Structure. 1-Participants see a fixation cross; 2-Participants are presented with two cues: target (left) and lure (right). In the example, the target cue represents a 50% probability of getting pain 4 and a 50% probability of getting pain 7. The lure represents a 50% probability of getting pain 5 and a 50% probability of getting pain 7. The target cue is always preferable to the lure cue and participants are instructed to select that one; 3-The selected cue is presented on the screen; 4-Stimulation is delivered (in this case it will be stimulation of either 4 or 7); and 5-Participants are requested to rate the pain.

TABLE 1 Conditions as a result of stimulation and cue combination.

Condition	Stimulation	Target cue	Associated lure cues
1	3	3-3	4-4, 5-5, 6-6, 7-7
2	4	4-4	5-5, 6-6, 7-7
3	5	5-5	6-6, 7-7
4	6	6-6	7-7
5	4	4-5	4-6, 4-7, 6-5, 7-5
6	5	4-5	4-6, 4-7, 6-5, 7-5
7	5	5-6	5-7, 7-6
8	6	5-6	5-7, 7-6
9	3	3-6	3-7, 4-6, 5-6, 7-6
10	6	3-6	3-7, 4-6, 5-6, 7-6
11	4	4-7	5-7, 6-7
12	7	4-7	5-7, 6-7
13	4	4-6	5-6, 7-6, 4-7
14	6	4-6	5-6, 7-6, 4-7
15	3	3-7	4-7, 5-7, 6-7
16	7	3-7	4-7, 5-7, 6-7

found. Each condition was presented five times through the experiment, resulting in 80 trials. The order of appearance of the target cues was randomized as well as the choice of the potential associated lure cues at each trial. The location of the target and lure cues on the screen (left or right) was also randomized.

After cue selection, the chosen pair of cards was displayed in the centre of the screen for 2s. Then, an electric stimulation associated with one of the depicted card values was delivered. Each card on the target cue had a 50% chance of representing the stimulation that would be delivered.

Five-hundred millisecond after the delivery of the stimulation, a 0 to 100 visual analogue scale was displayed. On each trial the starting position of the cursor on the scale was randomized to control for motor habituation effects. Participants were asked to click on this scale using the mouse to indicate the level of pain they had experienced. An inter-trial interval was randomly selected from the options of 1000, 1500, 2000, 2500 and 3000 ms. Participants were given a self-timed break every 20 trials. The same procedure was repeated 2 weeks (±2 days) after the first visit. The trials in which participants failed to choose the target cue were deleted from the final analysis since a lack of attention or understanding was assumed. If a participant chose the target cue in less than 75% of the trials, the participant was excluded from the analysis. The median number of included trials per participant in the final analysis was 79 (range 62–80) for session 1 and 80 (range 72–80) for session 2.

2.3 | Data transformation

Before the modelling took place, the data were transformed so the delivered stimuli and cues were on the same scale as the pain ratings. This way the stimulations and cues were transformed from a scale from 0 to 10, to a scale from 0 to 100 by multiplying them by 10. In the Results sections, we will refer to the cues and stimulation in the transformed scale (from 0 to 100).

A high correlation between the effect of the cue and the stimulation was encountered. This could be a phenomenological finding in which participants who have a higher somatic focus also tend to be more attentive to cues. However, it could also be a consequence of the similarity in the cue and stimulation found in some of the trials of the paradigm. To provide the most robust results possible by considering all contingencies, we progressively eliminated the trials with low standard deviations until we reached a correlation between the weight placed on stimulation and the cue below 0.70. This occurred once we eliminated the trials in which SD was 0 or 1, with the obtained correlation, $r=0.25$, falling below the value indicative of redundant constructs (Abma et al., 2016). We then repeated the full analysis on the reduced sample. The results obtained through the subsample analysis were the same as with the full sample (with slightly more conservative results for the reliability of the cue parameters) corroborating that our conclusions are not dependent on a subset of trials. A full summary of the results with the reduced sample can be found in the Supplementary Materials S1. Note that this low correlation between the weight of the stimulation and the cue is not indicative of complete independence of the parameters, in fact, results seem to point to the existence of some shared variance between two parameters which should be studied in future construct validity studies. Nevertheless, the low correlation found in the reduced sample and the consistency in reliability estimates provide reassurance of the reliability results not being dependent on a few trials.

2.4 | Specification of the models and parameters' estimation

We adopt a Bayesian observer formulation of pain perception (Petzschner et al., 2015) and use a Bayesian treatment to infer the models' parameters which are described in detail in Table 2. This is in line with the meta-Bayesian approach proposed by Daunizeau et al. (2010). The Bayesian observer models described below represent how the subjects make optimal perceptual inferences under uncertainty using a Bayesian inference framework. These models correspond to probabilistic generative models of

TABLE 2 Parameter descriptions.

Parameter	Description
β^2	The variance of the delivered stimulation. Indicator of the weight of sensory input on pain perception. Lower values are indicators of higher influence of sensory inputs.
η	A proportion of the effect of the standard deviation on the mean of the cue. Indicator of the effect of the variance of the cue on perception. A higher value indicates that when the cue variance is higher, the cue will have a lower effect on perception.
ρ^2	The variance of the cue. Indicator of the weight of the given cue mean on pain perception. Lower values are indicators of higher influence of the cue.
μ	The mean of the trait-like bias. Indicator of the prior expectations on pain perception. Higher values indicate higher expected pain independent of the given cues.
ν^2	The variance of the trait-like bias. Indicator of the variance of expectations prior to the cue. Higher values indicate a lower effect of the prior expectations.

the environmental inputs (stimulus intensity and cue), entertained by a Bayesian observer (the subject) during pain perception. Subjects' pain perception then proceeds by Bayesian inversion of these generative models to produce posterior estimates of the perceived pain intensity in the form of pain ratings. That is, uncertain representations of pain intensity (pain rating Y) correspond to posterior beliefs, which result from integrating sensory information (stimulus X and cue q) with subjective prior beliefs. The equations below represent the posterior probability density over pain ratings for each generative model considered.

2.4.1 | Model 1

This generative model assumes that the stimulus intensity (X) received by participant (i) at trial (j) is conditionally dependent on the pain rating (Y) with associated participant-specific variance parameter (β^2) to be estimated. It is assumed that the subject has no prior preference for any pain rating value (uniform (constant) prior distribution on pain ratings).

$$p\left(Y_{ij}|\,X_{ij,}\beta_i^2\right) \propto p\left(X_{ij}|\,Y_{ij},\beta_i^2\right) \quad (1)$$

2.4.2 | Model 2

This generative model assumes that the stimulus intensity (X) received by participant (i) at trial (j) and the expected value induced by the cue (q) are independent sources of input information which are both conditionally dependent on the pain rating, consequently they are modelled by independent probability densities with respective stimulation (β^2) and cue (ρ^2) participant-specific variance parameters. As in Model 1, it is assumed that the subject has no prior preference for any pain rating value

$$p\left(Y_{ij}|\,X_{ij,}\beta_i^2,q_{ij,}\rho_i^2\right) \propto p\left(X_{ij}|\,Y_{ij},\beta_i^2\right)\,p\left(q_{ij}|\,Y_{ij},\rho_i^2\right) \quad (2)$$

2.4.3 | Model 3

This generative model is similar to Model 2 but it is assumed that the cue is conditionally dependent on the pain rating minus a bias term that is proportional to the standard deviation of the cue (SD) with a proportionality constant (η). This parameterization means that the participant's pain rating inference is biased depending on the magnitude of the uncertainty associated with the cue. As in Models 1 and 2, it is assumed that the subject has no prior preference for any pain rating value

$$p\left(Y_{ij}|\,X_{ij,}\beta_i^2,q_{ij},\rho_i^2,sd_{ij},\eta_i\right) \propto p\left(X_{ij}|\,Y_{ij,}\beta_i^2\right)\,p\left(q_{ij}\,|\,Y_{ij},\eta_i,\rho_i^2\right) \quad (3)$$

2.4.4 | Model 4

This generative model is an extension of Model 3 by including a prior probability density on the pain rating with prior mean (μ_j) and variance (ν^2), which models possible trait-like bias effects of each participant (j).

$$\begin{aligned} &p\left(Y_{ij}|X_{ij,}\beta_i^2,q_{ij},\rho_i^2,sd_{ij},\eta_i,\mu_i,\nu_i^2\right) \propto p\left(X_{ij}|Y_{ij,}\beta_i^2\right) \\ &p\left(q_{ij}\,|Y_{ij},\eta_i,\rho_i^2,sd_{ij},\eta_i\right)\,p\left(Y_{ij}|\mu_{i,}\nu_i^2\right) \end{aligned} \quad (4)$$

As said above, all variance parameters, the proportionality constant of the SD of the cue, as well as the mean of the trait-like bias density, are estimated using a Bayesian inference framework. This is done by introducing additional prior probability densities on all these parameters (see Supplementary for details). In short, given stimulation, cue and pain rating data, a posterior density distribution for the parameters of each model is obtained by using each of the generative models in Formulas 1–4 as likelihood functions of the parameters to be estimated, multiplied by independent prior density distributions on each parameter. Given the factorized form of the likelihood above (one likelihood for each factor), the prior independence assumption of the parameters renders their posterior distribution factorized into three independent (posterior) distributions, one for

each of the three individual factors: stimulation, cue and trait-like bias. In other words, the inferences in the parameters in the most complex model can be carried out independently for each factor. Because of this, the estimated values of the parameters associated with each factor are not affected by adding or removing the likelihood terms associated with another factor. In simple terms, adding or removing the likelihood term of one factor does not affect the parameters of the other factors. This is particularly advantageous since it would allow progressive understanding of each factor independently. For example, future research may be interested in only one of the factors or parameters. Having a posterior distribution factorized over parameters of each factor means that in order to explore one part of the model, prior research would not need to be redone.

It should be stressed that the estimated parameters in all models are participant but not trial specific, which means they represent the average effect of each factor on each individual across all trial types. Therefore, the parameters quantify the effect of each factor on each individual across trials with different characteristics. For example, a parameter such as β^2 (effect of the stimulation) measures the uncertainty associated with the stimulation in each individual across all trial types (certain, uncertain, high or low cue).

Due to the functional form of the generative models above and the prior densities used for the parameters, a posterior density of the parameters cannot be obtained in closed form. Therefore, in all cases, parameter estimation was done using the Hamiltonian Monte Carlo (HMC) algorithm (Radford, 2011), as implemented in Rstan package (Stan Development Team, 2023).

2.5 | Further comments about the models

The factorization of the posterior density of our models over independent factors (stimulation, cue and trait-like bias) presents several benefits. First, interpretation of the factor-specific parameters is more straightforward since they can be interpreted without the need to account for interactions with other parameters. (e.g. the effect of the stimulation represents the effect of the stimulation overall across all different types of trials and not the effect of the stimulation after controlling for the effect of the cue which could shape the distribution in different ways). Second, increasing the model complexity in terms of the number of factors considered does not increase the potential for overfitting. As a result, even if the complex model was to be overfitted (which is a risk in very complex models) the results would remain consistent with those of the simpler models. That is, the parameter estimation in the complex model would be consistent with fitting a simple model for each factor separately and pulling together all the results. This is a desirable feature of models used for phenotyping where one would like to represent the effect of as many characteristics as possible.

In addition, the posterior independence of the factors allows for future research to focus on any of the individual factors without affecting the results of the others. For example, if someone is exclusively interested in the cue effects, they could focus on the likelihood and priors related to that factor without considering the other factors in the analysis.

Finally, the estimated parameters are participant specific. This means that the effects of each factor are estimated at the individual level, which will enable individual phenotyping, more so than providing average group-level estimates for the whole sample.

2.6 | Model selection

The main goal of this work was to extract individual characteristics that could be used for pain phenotyping. To fulfil this task, models that included all relevant factors that have been shown to influence the pain response (stimulation, cue and trait-like bias) were considered.

It is important to note that in most pain investigations using Bayesian methods, usually the goal is to discover an underlying mechanism that can be generalized to the wider population. Model selection, in this case, is then carried out by choosing the most generalizable model that maximizes certain information criteria (e.g. Akaike Information Criterion (AIC) and Bayesian Information Criterion (BIC)), or the model evidence among others. This is because these criteria typically penalize model complexity, hence promoting generalizability. Nevertheless, this generalizability comes at the cost of lower flexibility to adjust to more heterogeneous data (Blanchard et al., 2018), which can serve to identify individuals with specific response patterns, after the main goal in terms of phenotyping. This trade-off between model flexibility (complexity) and generalizability (parsimony) implicit in Bayesian inference has been shown to represent an automatic Occam's Razor.

However, in the case of this work, the goal is to phenotype each individual and to test the reliability of this phenotype. Consequently, unlike in other types of investigations using Bayesian methods, in the present work, a complex model would be favourable even if its generalizability was lower.

Therefore, in this study, a model with the ability to include the most parameters of interest was favoured,

as long as certain quality criteria were fulfilled. Specifically, model selection was carried out based on the ability of the model to generate reasonable outcomes while preserving flexibility by accommodating a wider range of data. This was assessed through a generative performance test, which evaluates model quality in terms of its ability to reproduce an effect of interest through data simulations (Palminteri et al., 2017). We simulated pain responses based on the estimated model parameters and trial information for each model under consideration. We first analysed the graphical representations of actual data versus simulations to test whether the outcomes obtained through models of different complexities showed reasonable estimates. Then, we used a more quantitative approach to study whether the models of higher complexity were better at capturing the variability of the data. We implemented a linear mixed-model analysis with the simulated ratings as predictors, the actual ratings as outcomes and participants as random effects, to test how much of the variance was explained by each model. The linear mixed models were then compared by looking at both marginal and conditional R^2 values.

Although our goal and procedure for model selection rely less on model generalizability, knowing which model explains the behaviour of the population better (highest generalizability) is still relevant for phenotyping. Particularly, this information can be of use in a normative approach to phenotyping, by enabling an assessment of whether an individual's behaviour is common or uncommon (out of the norm). For instance, if the most generalizable model does not include the cue, having a participant who places a lot of weight on the cue can mean that they might have an uncommon response to pain and potentially have different pain chronification trajectories or treatment responses.

To identify the most generalizable model, a model selection procedure was carried out in which models of different complexities (stimulation only, stimulation and the cue, stimulation, cue and trait-like bias) were compared based on their predictive performance, that is, the ability of a model to predict the unobserved (out-of-sample) data (Palminteri et al., 2017). In this paper predictive performance was measured via leave-one-out cross-validation as implemented in the R package 'loo'. This package evaluates predictive accuracy by using the expected log-pointwise predictive density (ELPD) of the model, which has several advantages over other model quality metrics such as BIC or AIC (Vehtari et al., 2017). In brief, this approach evaluates the ability of the model to predict the data of the general sample while adding a penalization for model complexity.

2.7 | Reliability analysis

Once the parameter values were calculated per participant and per session, the reliability of these values over time was evaluated using both relative reliability measures (the intraclass correlation coefficient (ICC)) and absolute reliability measures (coefficient of repeatability (CR) and the Bland–Altman plots).

ICC is a measure of reliability that takes into consideration both the degree of correlation and the agreement between measurements – in this case, between the two sessions. This measure provides valuable insights as to whether individuals maintain their relative position in the ranks in relation to the other members of the group; in other words, it shows whether those participants likely to obtain one of the highest scores in the first session are likely to be high scorers in the second session. Although this does not show absolute reliability (that the score obtained will be the same every time the task is run), it provides valuable information regarding the stability of each participant's response under the assumption that the circumstances in which both tests are run might present differences that induce bias on the group level. Particularly, in our case, we have to accept that in the second session, participants were more familiar with the task and that this may induce certain differences in the responses, either due to learning effects or by a reduction in any uncertainty/anxiety associated with performing an intrinsically aversive task. The model selected to conduct ICC was a two-way mixed-effects model, the type selected was single rater and the definition was absolute agreement. These specifications were selected since they are the ones that have been established as the best for reliability analysis (Koo & Li, 2016). ICC was calculated through the 'icc' function of the 'irr' package in R.

When it comes to the measures of absolute reliability, these explore whether the result obtained through the same measuring tool is the same, under the same conditions and at different time points. One drawback of this approach is that it does not provide an easy-to-interpret output in which we can assess whether the replicability is high or low as the ICC does; instead, it provides an estimate of the interval of measurements in which we can assume 95% of the scores obtained by a person will be, if no significant change has occurred, which is represented through the Bland–Altman plots (Bland & Altman, 2003). Furthermore, this is accompanied by the minimum change in the scale needed to be able to consider a change significant (and not just due to error); this is the value of the CR (Bland & Altman, 2003; Vaz et al., 2013). Whether this value is a good indicator of

reliability must then be evaluated in terms of clinical significance (Bland & Altman, 2003). In our case, the parameter values we possess are in terms of variance and means which would make comparison with pain measures difficult. Consequently, the first step in order to calculate absolute reliability was to transform the parameters to values that indicate the change in pain ratings (on a scale from 0 to 10). To do so, the following formulae were used. Note that the division by 10 was specifically to rescale the data from a 0–100 to a 0–10 scale:

- For parameters β^2, ρ^2 and ν^2, which represent variances, the pain rating difference after which a change could be considered significant was calculated using $\frac{\sqrt{\text{Parameter}}}{10}$
- For parameter μ, which represents the mean trait-like bias, the pain rating difference after which the change could be considered significant was calculated using $\frac{\text{Parameter}}{10}$
- Finally, parameter η represents the proportional effect of the standard deviation on the mean of the represented cue. Due to the nature of the parameter, we do not have clinical data to compare this to, consequently, we did not transform this parameter or evaluate its clinical significance.

After the parameters were transformed, the Bland–Altman plots were plotted and the CR was computed by multiplying the square-rooted mean of the within-subject sample variances by √2 times 1.96 (Vaz et al., 2013).

$$CR = 1.96 * \sqrt{2} * \sqrt{\sigma_w^2} \quad (5)$$

For the interpretation of the results based on clinical significance, since our study is exploring an area still in its infancy, we do not have clinical measures we can directly compare our scores. Nevertheless, in order to provide the most complete explanation of the data possible, we compared the results with existing data on the minimally clinically significant change in pain ratings studied in the chronic pain literature. This interpretation should be taken with caution since the different measures represent different constructs loosely related and in different populations. However, it is to date the only clinical data we can use to assess the repeatability coefficients.

3 | RESULTS

3.1 | Descriptive statistics

According to Table 3, pain ratings increased, on average, corresponding to an increase in the mean cue values. There appears to be more uncertainty (higher SD) around pain ratings for mean cues of 50/55 as opposed to 45. Similar results were observed about the effects of stimulation on the average pain ratings, however, the increase in the uncertainty is more prominent (shift from ~11 to ~18 in SD of pain ratings as the stimulation

TABLE 3 Descriptive statistics of pain ratings around cue and stimulation.

Session 1			Session 2		
Cue mean	**Mean pain rating**	**Standard deviation of pain rating**	**Cue mean**	**Mean pain rating**	**Standard deviation of pain rating**
30	24.82	9.94	30	23.07	10.29
40	33.37	12.05	40	32.13	13.44
45	37.84	15.18	45	36.36	15.59
50	42.39	18.23	50	40.82	18.50
55	46.87	17.06	55	45.68	17.10
60	51.91	14.84	60	51.66	15.42
Stimulation	**Mean pain rating**	**Standard deviation of pain rating**	**Stimulation**	**Mean pain rating**	**Standard deviation of pain rating**
30	25.73	11.09	30	23.92	10.77
40	35.52	12.38	40	33.27	12.67
50	42.34	13.93	50	41.75	13.71
60	49.42	15.25	60	48.93	15.62
70	58.57	17.46	70	56.74	18.32

Note: The cue and stimulation values are presented in the transformed scale (from 0 to 100).

increases from 30 to 70). These results hold for both sessions. In Figure 2, a graphical representation of the change in pain ratings as a function of stimulation and cue can be observed.

3.2 | Generative performance

The generative performance analysis showed that the estimates obtained through all models were within reasonable bounds. Nevertheless, simulated data based on model 4 explained observed pain ratings the best, whereas simulations based on model 1 performed the worst. The results of all the regression models can be seen in Table 4. The results show that the most complex models are better at capturing the individual variability of the pain ratings, which can be observed in the higher R^2 values and in Figure 3 where observed pain ratings are plotted against data simulated from each model. This renders the most complex model a good candidate for phenotyping purposes. Further graphical representations of the relationship between observed and simulated data can be found in Supplementary Materials S4. Based on these results, the most complex model was chosen for the reliability analysis.

3.3 | Predictive performance

The 'leave one out cross-validation' results showed that the simplest model (first model) was the best fit for the data, followed by the third, second and fourth models (Table 5). These results hold across both Session 1 and Session 2. This can be seen by looking at the difference in the ELPD between models, where higher ELPD values are considered indicators of better predictive accuracy. Note that when more than two models are compared, the difference is computed with respect to the model with the highest ELPD. We considered two models to be different if their ELPD difference relative to the standard error of the difference was greater than 2.

These results indicate that the simplest model is better for predicting pain perception in the whole data sample, which suggests that stimulation might be the most important factor at the population level.

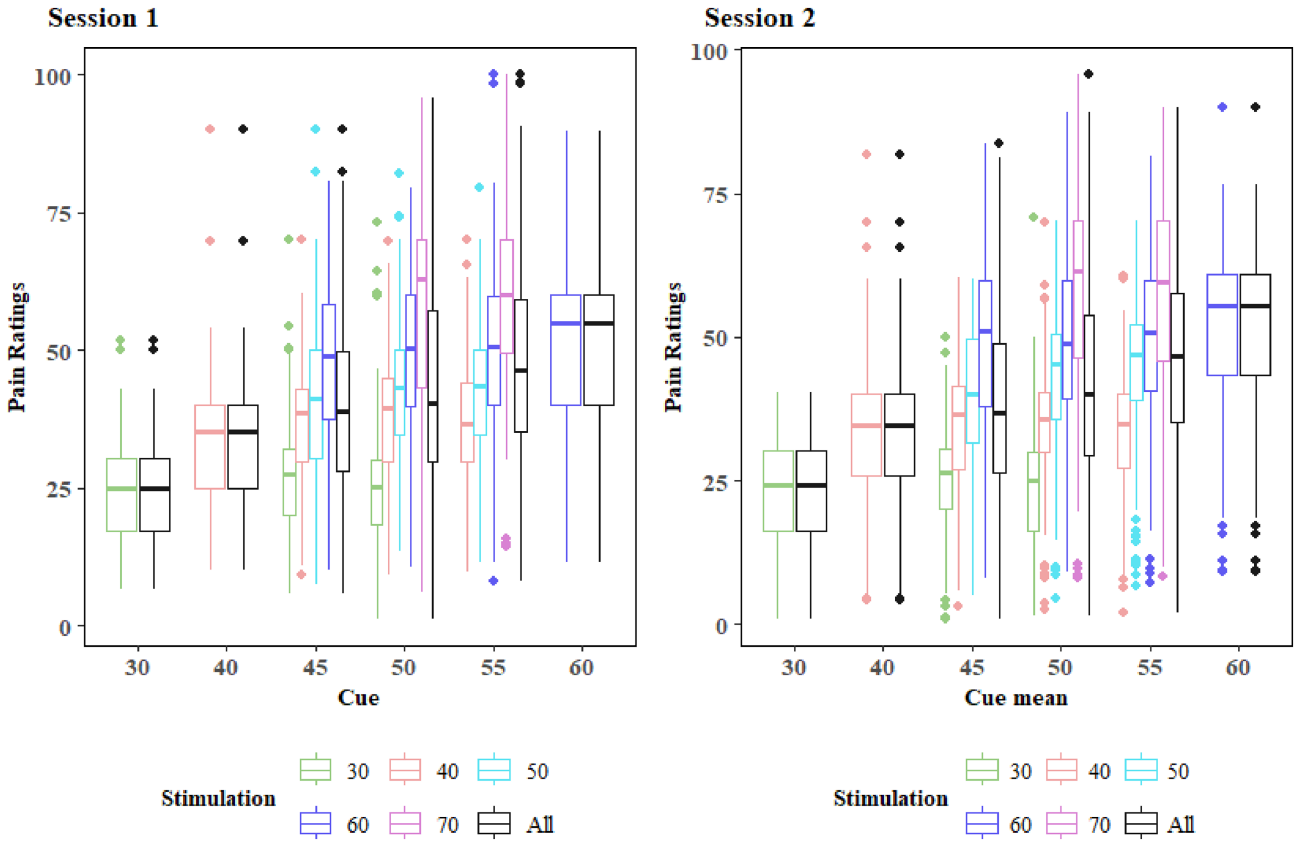

FIGURE 2 Overall pain rating change as a function of stimulation and cue means. The figure above shows the average pain ratings of all participants as the mean cue increases, divided by stimulation levels and with an overall estimate in the black boxplot that represents the average pain ratings at a certain cue over all stimulation intensities.

TABLE 4 Summary of linear mixed models.

	Fixed effect	Confidence interval 2.5%	Confidence interval 97.5%	Marginal R^2	Conditional R^2
Session 1					
Model 1	0.78	0.75	0.81	0.34	0.69
Model 2	1.07	1.02	1.11	0.35	0.70
Model 3	1.07	1.03	1.12	0.41	0.64
Model 4	1.51	1.45	1.57	0.71	0.75
Session 2					
Model 1	0.82	0.79	0.85	0.36	0.74
Model 2	1.11	1.07	1.15	0.37	0.75
Model 3	1.14	1.10	1.18	0.43	0.72
Model 4	1.54	1.49	1.60	0.77	0.82

Note: Marginal R^2: variance explained by the fixed components only; Conditional R^2: variance explained by the full model.

3.4 | Individual versus population models

The most likely reason for the different results between generative and predictive performance may rely on the fact that the stimulation alone is enough to explain the pain ratings of most participants, and the inclusion of other factors does not improve the prediction enough to warrant the increased complexity in most cases. Evidence of the log-likelihood associated with each participant per distribution supports the idea that different factors may explain the behaviour of different participants best. Furthermore, a mixed regression model with pain as an outcome, subjects as random effects and stimulation and cue as predictors show that the cue has a significant but small effect on pain (Supplementary Materials S5). This is clearly represented in Figure 3 where it can be observed that most participants present a low β^2. Nevertheless, exceptions can already be found in the sample with some participants showing higher levels of this parameter and lower levels in others. For instance, if we take participants 3 and 6 as an example, we can see that $\widehat{\beta}_6^e$ is very small compared to $\widehat{\beta}_3^2$ (where, $\widehat{\beta}_i^2$ corresponds to the posterior mean of β_i^2). This means that while participant 6's pain ratings can be explained by the stimulation delivered, in the case of participant 3, the stimulation is associated with a higher degree of uncertainty (variance). Simultaneously, we can also observe that participant 3's ρ_i^2 is small, indicating that their pain ratings are better with a model that includes the cue (Figures 4 and 5).

3.5 | Reliability analysis

The ICC for all parameters reached a level of significance below 0.05, indicating the test–retest reliability of the measures taken in this study. By the Koo and Li (2016) guidelines, the obtained ICC values (Table 6) indicate good reliability in the case of β^2, ρ^2, μ, and ν^2 and moderate reliability in the case of η (Figures 6 and 7). Furthermore, as an example of this in Figure 8, we can see the response pattern of participants 3 and 6. As mentioned previously in the manuscript, these two participants show very different β^2 and ρ^2 estimates. This way, participant 3 shows a greater reliance on the cue, whereas participant 6 relies mostly on the stimulation. In this figure, while keeping constant the value of the cue (on 50), the scatter plot has been constructed with pain rating on the y axis and stimulation on the x axis. As it can be seen, with the cue kept constant, participant 6's pain ratings increase with stimulation in a way more prominent way than participant 3's whose pain ratings stagnant around the value of 50 (the value of the cue). In Supplementary Materials S6, a gender-specific reliability analysis can be found.

When it comes to absolute reliability, in the Bland–Altman plots, 96.15%, 96.15%, 96.15%, 96.15% and 92.30% (for β^2, η, ρ^2, μ and ν^2, respectively) of the differences could be found within the upper and lower boundaries of the graph, indicating an overall good reliability (Figure 9). The computation of the upper and lower boundaries, as well as the confidence intervals for them (a recommended addition to these plots to account for the sampling error on the computation of the boundaries), has been performed by the formulae presented by Bland and Altman (Bland & Altman, 2003).

Regarding the RC values (Table 7), these indicated that a change of 0.86, 0.65, 1.45 and 0.47 (for β^2, ρ^2, μ and ν^2, respectively) would be necessary to consider a difference between measures significant.

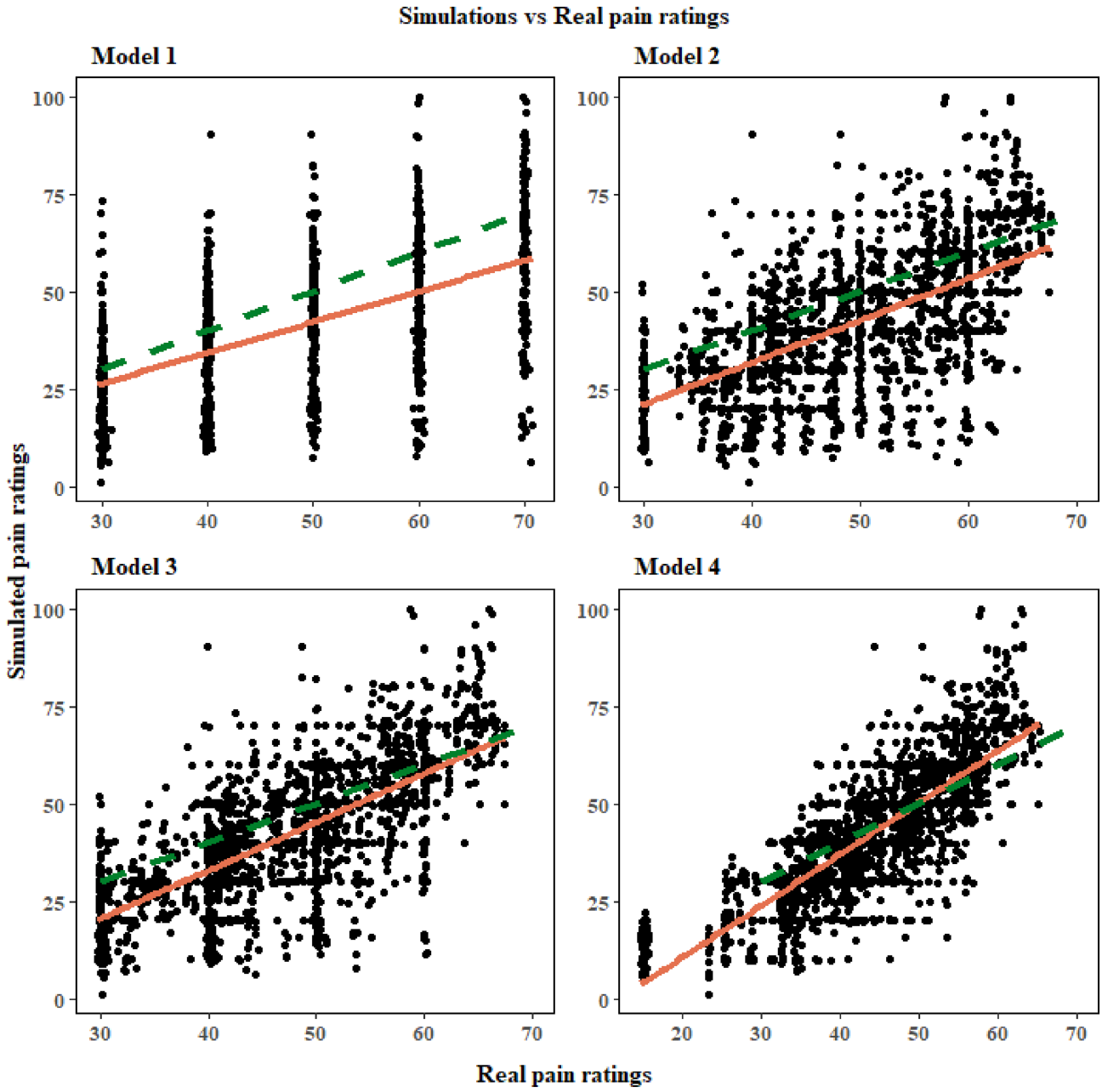

FIGURE 3 Simulations versus real data.The figure above shows the simulated ratings obtained with each one of the models (against the real pain ratings). Each dot represents the pain rating given per participant per trial and its respective simulation. The solid red line represents the regression line. The dashed green line represents the perfect fit. As it can be seen, the more complex models capture the variability of the data better than the simpler ones. The data shown are from Session 1.

4 | DISCUSSION

In this study, we aimed to investigate the temporal stability of the estimations obtained through Bayesian modelling regarding the effect of stimulation (β^2), cue ($\rho^2$and η) and trait-like bias (μ and ν^2) on pain perception. Results showed that the estimations obtained for the effect of the stimulation, cue (ρ^2) and the trait-like bias (μ and ν^2) had a good relative reliability and that the estimation for the effect of the standard deviation of the cue (η) had a moderate relative reliability. Absolute reliability measures also provided supporting evidence for the reliability of the parameters. These findings support the hypothesis that inter-individual differences in the weight placed on different pain factors is stable in time and could potentially be useful for patient stratification if the results are corroborated in patient populations.

TABLE 5 Results of predictive performance test.

Models (in order of predictive performance score)	elpd	Difference between model elpd and highest elpd	Standard error of elpd difference	elpd_difference/ standard error of the difference
Session 1				
Model 1	−8039.6	0.0	0.0	0.0
Model 3	−16307.5	−8267.9	29.4	−281.22
Model 2	−16515.9	−8476.3	29.4	−288.30
Model 4	−24321.9	−16282.3	51.4	−316.77
Session 2				
Model 1	−8056.9	0.0	0.0	0.0
Model 3	−16639.1	−8385.6	30.6	−274.0392
Model 2	−16442.5	−8582.2	30.0	−286.0733
Model 4	−24574.2	−16517.3	51.3	−321.9747

In fact, the relative reliability of the effect of the stimulation and trait-like bias is in the same range as the reliability values found for quantitative sensory testing (QST) (Felix & Widerström-Noga, 2009; Koo & Li, 2016; Lin et al., 2020; Nothnagel et al., 2017; Wylde et al., 2011). QST is a procedure widely used as a phenotyping tool in experimental pain paradigms and its results have been shown to be predictive of various clinical pain outcomes such as disability (Georgopoulos et al., 2019) and postoperative pain (Braun et al., 2021). Therefore, obtaining reliability values in the same range as this already established procedure provides support for the validity of our task and the idea of exploring the clinical application of our paradigm. In relation to the effect of the standard deviation of the cue, although the ICC values were not as high, these were still significant and in the range of some of the QST subscales such as warm and hot detection thresholds, and even the heat pain threshold of some studies (Felix & Widerström-Noga, 2009; Lin et al., 2020; Wylde et al., 2011).

When it comes to the absolute reliability results, as mentioned in the Methods section, we do not possess clinical data that would allow us to evaluate whether these values are within reasonable boundaries. However, when comparing these changes to the closest clinical data we possess (the minimum change in pain ratings needed in patient populations to consider a difference significant, e.g. as part of a therapeutic trial), we find that these show values similar to or higher than ours. In particular, the literature has shown that the minimum clinically significant change in pain ratings is around 1.5 and 2 (on a 0–10 scale) (Bahreini et al., 2020; Kovacs et al., 2008; Salaffi et al., 2004). Due to the differences in samples and methods, this comparison should be interpreted with caution. It is nonetheless an early sign that supports the reliability of the measured parameters since the changes observed between sessions are lower than these; an indicator that shows that the differences found in participants between sessions may not be clinically significant, supporting in this way the reliability of the parameters.

Although this study supported the reliability of the measured parameters, some changes did take place from one session to another. Several factors could be influencing this. To begin with, familiarity with the paradigm would unavoidably be higher in the second session. This could result from a lower pre-task anxiety/uncertainty, compared to when facing an unknown aversive task for the first time, which may have influenced their expectations, or the weight placed on each element leading to the pain response. Furthermore, it is possible that there is an influence of learning from one session to the next. In fact, the number of errors made when choosing between the target and the lure decreased between sessions (e.g. the participant excluded due to failing to pick the target cue in over 75% of the trials showed a marked improvement between sessions, with 63.75% of correct choices in the first session and 96.25% in the second session). The experimental paradigm requires some cognitive processing to be completed, which could lead to changes in response once participants gain some experience with it. In future studies, the factors that lead to different parameter values should be investigated by taking pre-task psychological measures in the sessions and by analysing learning effects, perhaps through response times or neuroimaging markers.

Moreover, it is important to discuss the results of the predictive performance test. In this study, we found that a model with only the stimulation is superior at predicting pain perception when compared to a model which includes the cue and trait-like bias. This finding is concordant with some recent findings indicating a superiority of a stimulation-only model in predicting brain

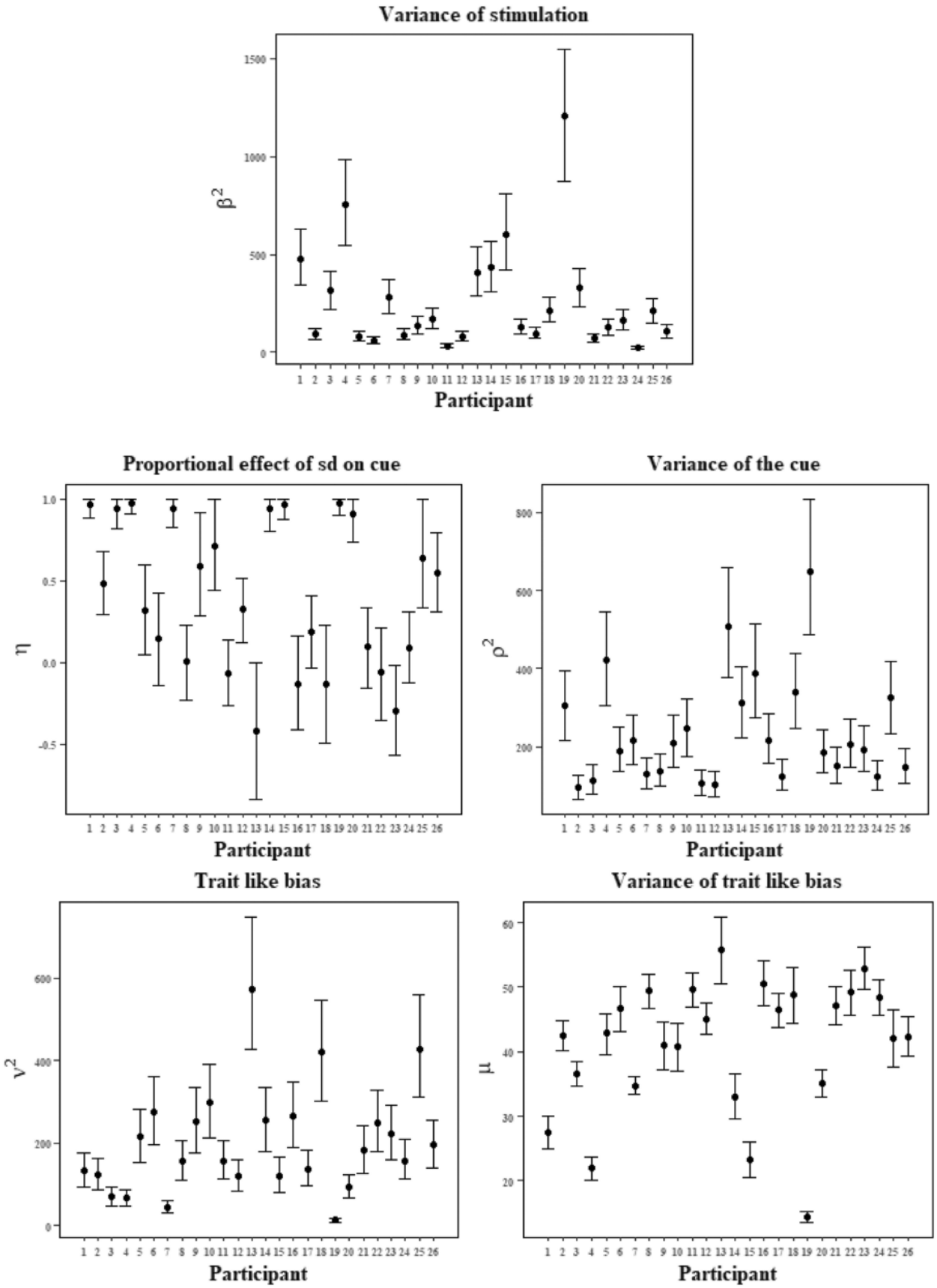

FIGURE 4 Summary of parameter calculations. The figure above shows the parameter means (dot) and credible interval (error lines) obtained by each participant in Session 1. This is a representation of the uncertainty around parameters per participant. The parameters were estimated using model 4.

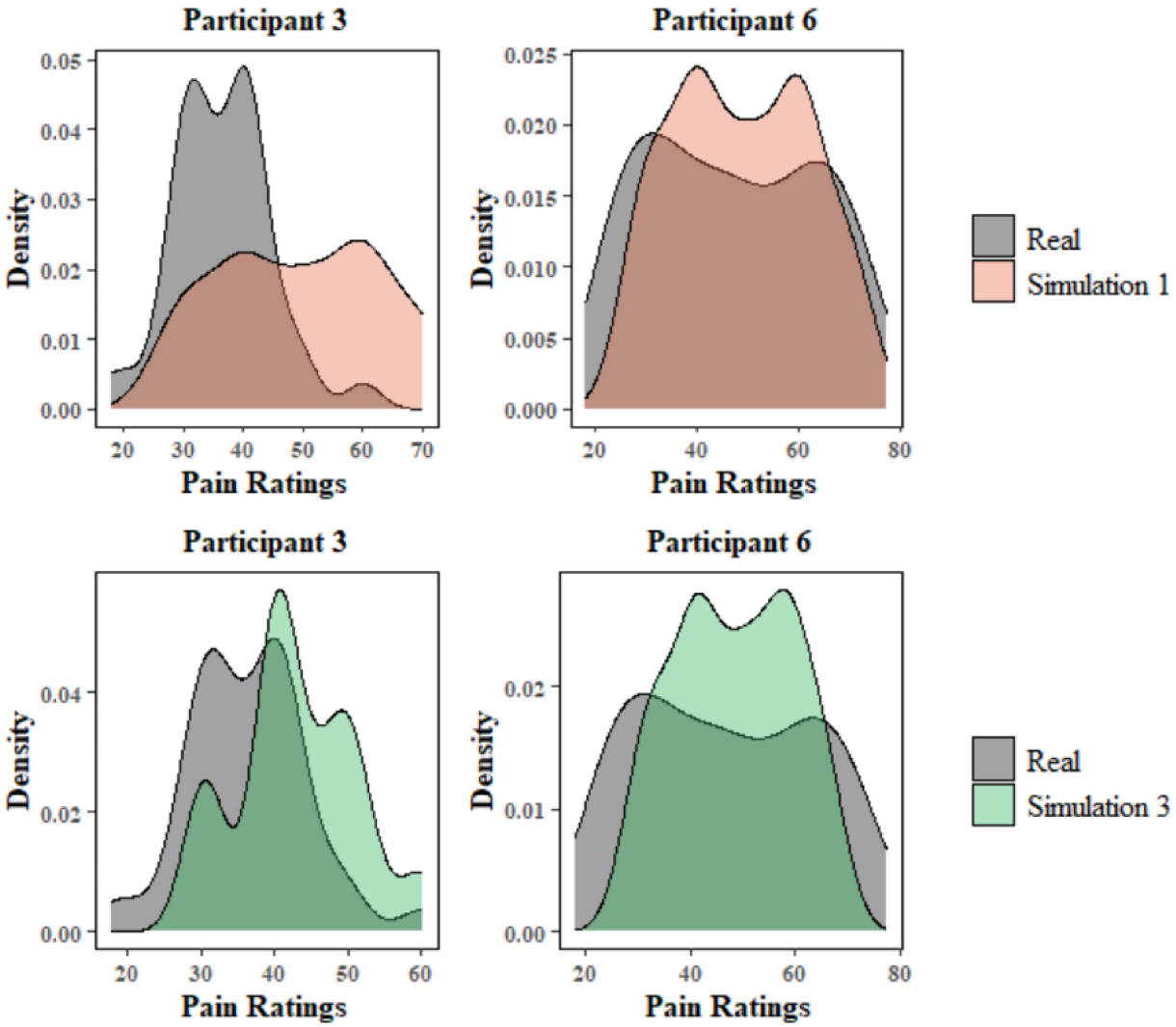

FIGURE 5 Example density plots for the effect of the stimulation. In the figure above, we compare the densities of the true pain ratings, with those obtained for the simulated pain ratings under model 1 (model with stimulation only) and model 3 (model with stimulation and cue). In the x axis, the different pain ratings are represented, and in the y axis, the density/frequency of those ratings is represented. In order to produce the two graphs on the upper row (red and grey row), the simulated data of model 1 (model with only the stimulation) were used. In order to build the two graphs in the lower row (green and grey row), the simulated data from model 3 (model with stimulation and cue) were used. As can be observed, the stimulation explains most of the pain ratings given by participant 6; however, it explains significantly less of the pain ratings given by participant 3. The pain ratings given by participant 3 are closer to the simulated data from model 3. The parameters were estimated using model 4.

TABLE 6 ICC analysis results.

Parameter	ICC	*p* value	Lower bound	Upper bound
β^2(variance of stimulation)	0.83	<0.001	0.67	0.92
η (effect of SD on cue)	0.55	0.001	0.22	0.77
ρ^2(variance of the cue)	0.75	<0.001	0.52	0.88
μ (mean of trait-like bias)	0.75	<0.001	0.52	0.88
ν^2 (variance of trait-like bias)	0.82	<0.001	0.65	0.91

responses (Nickel et al., 2022). Consequently, a potential interpretation would be that the effect of the cues presented to participants might have been overestimated in the literature perhaps due to issues such as publication bias. Nevertheless, an alternative explanation to these results may rely on sample selection. Both our study and the

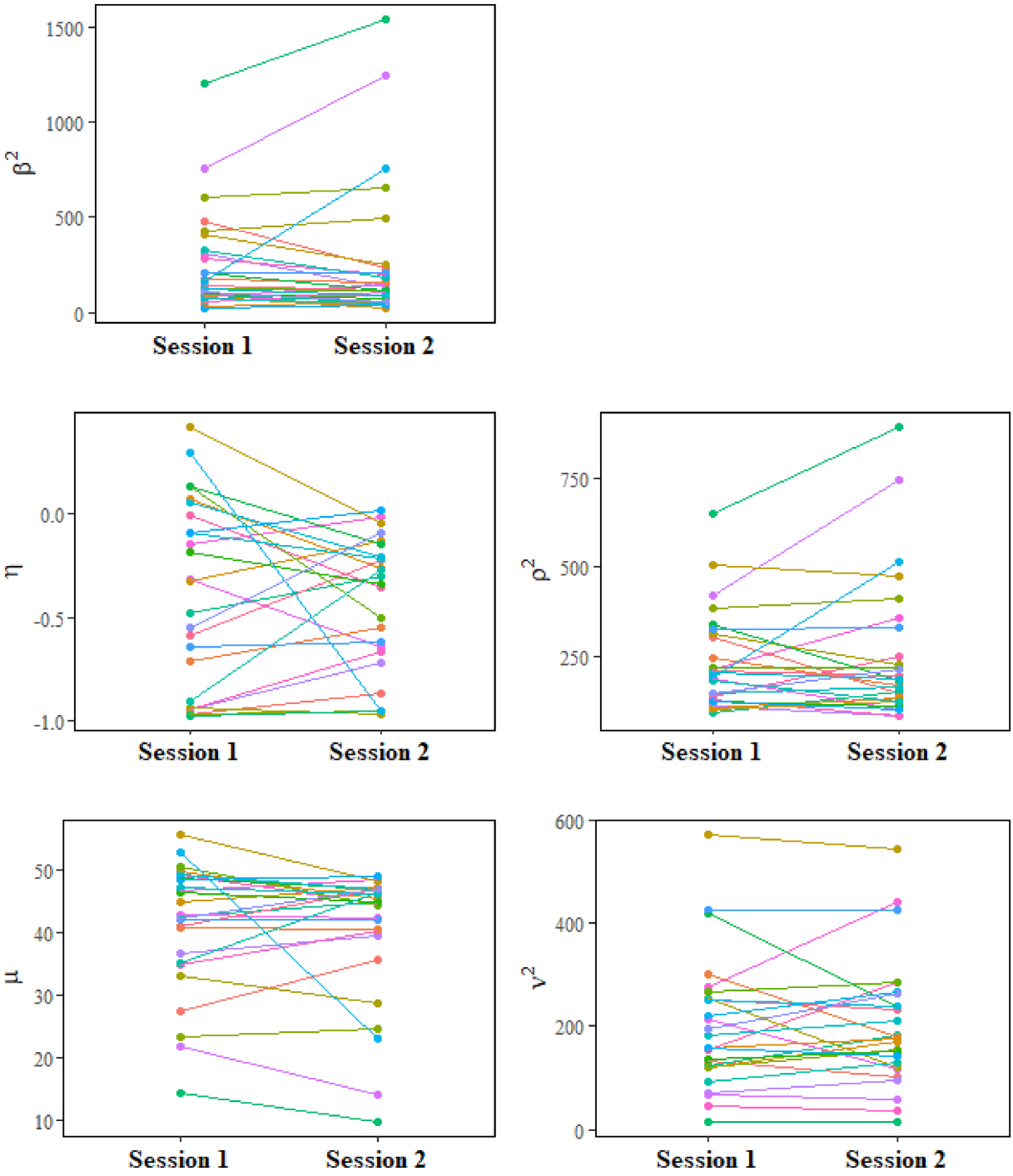

FIGURE 6 Test–retest values of all model parameters. The figure above shows the parameter means in session 1 and session 2. Each participant is represented with a different colour. The parameters were estimated using model 4.

Nickel et al. (2022) study were carried out in a healthy population. Therefore, the reliance on the cue could simply be more prominent in other populations such as chronic pain patients in which many studies have been carried out. Future research could test whether a model that includes the cue presents better predictive performance in other populations.

Nevertheless, it is still interesting to observe that even in our very homogeneous sample certain exceptions surface, with some people showing a greater reliance on the cue or their trait-like bias than on stimulation. This finding might indicate that although the majority of the population might be rating pain through a model mostly based on stimulation, and with little impact of other factors, other subsamples of the population might guide their perception through other elements such as the cue. This is a promising result for personalized medicine since it might indicate the existence of identifiable

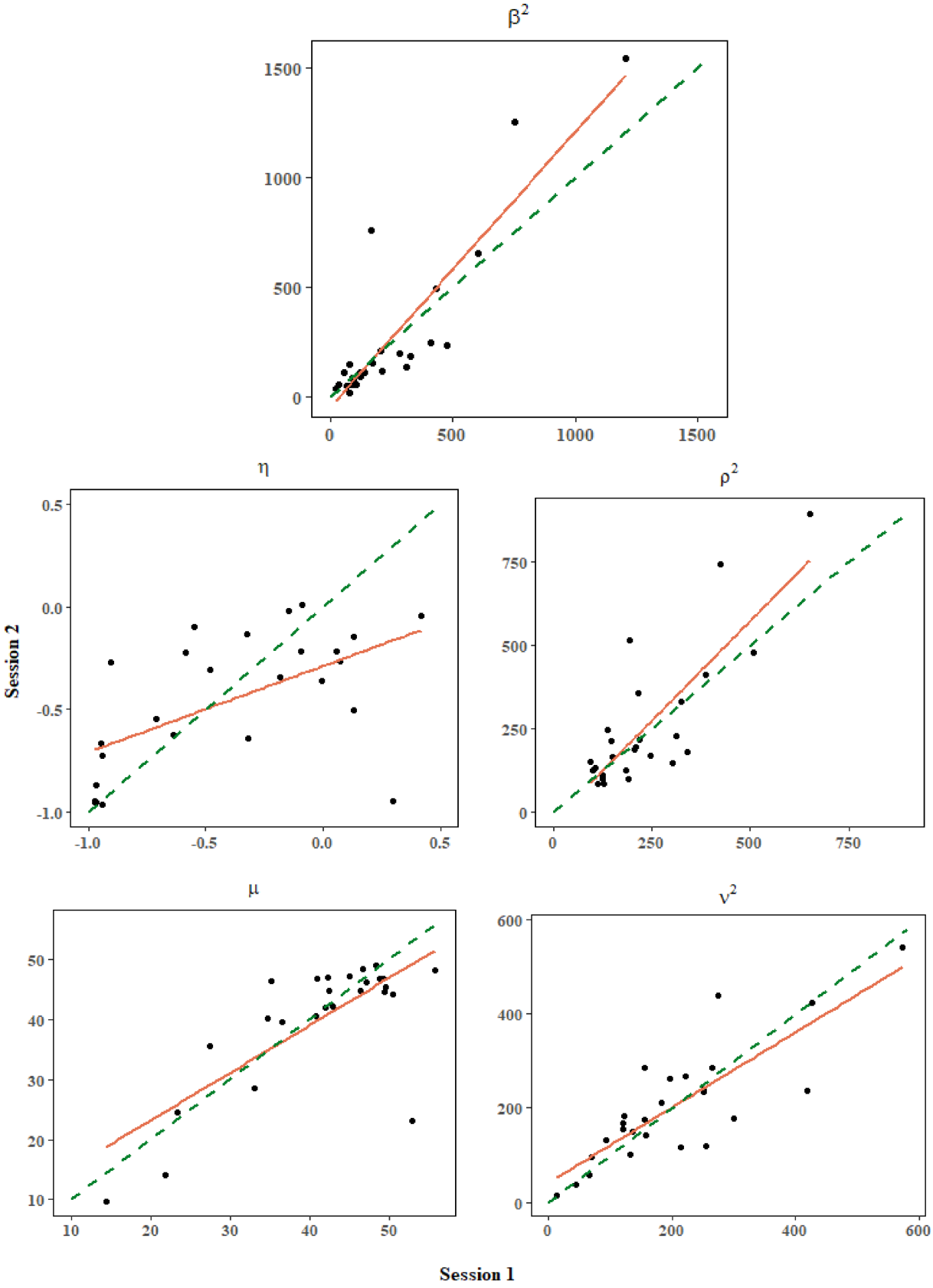

FIGURE 7 Scatter plots of the parameters in sessions 1 and 2. The figure above shows the correlation between the different parameters in session 1 and 2. The solid red line represents the regression line and the dashed green line represents the perfect fit line.

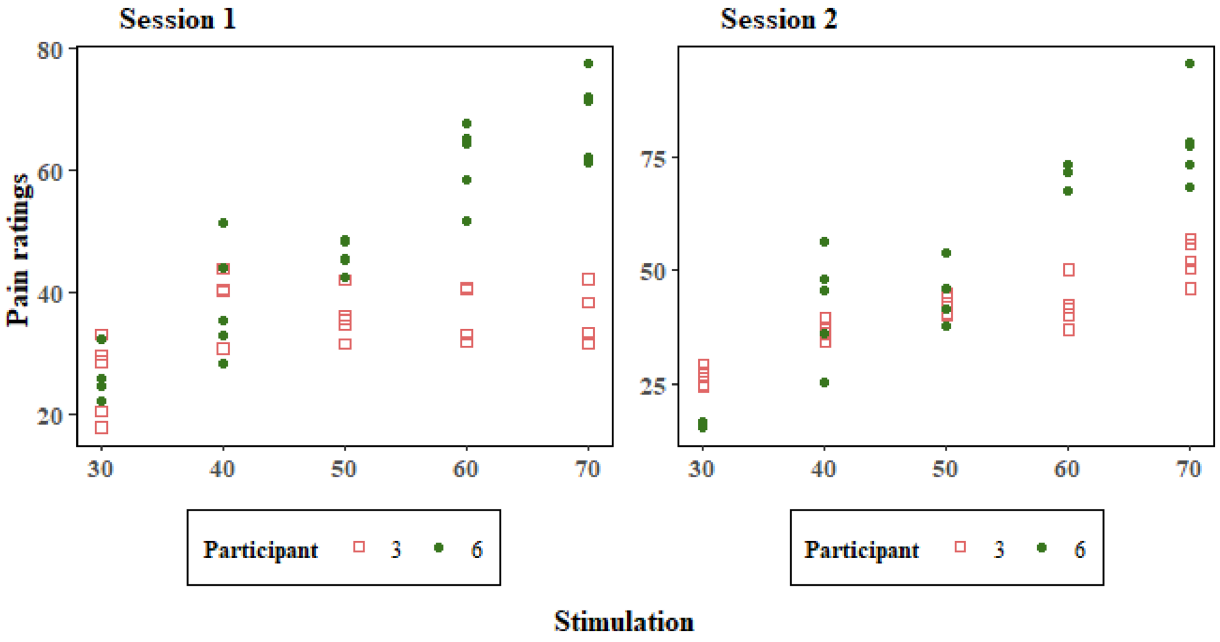

FIGURE 8 Scatter plots of pain rating stability in example participants. The figure above shows the pain ratings given by two participants to different levels of stimulation across both sessions. The stability in the response pattern can be observed through the comparison of both sessions.

pain phenotypes. Future research should aim to evaluate parameter distribution in a bigger sample to study with greater power whether participant clusters emerge from the parameter estimates.

Overall, our results carry some important implications. It has been proposed that clinical symptoms are a result of an inference made between physical causes and expectations (Van den Bergh et al., 2017), opening up the question of whether the inter-individual differences in the weight placed on the factors influencing pain perception could be a treatment target. In this work, we have shown that the weights individuals place on each factor are stable in time. This opens the possibility of using the parameters obtained through this paradigm as a patient classification tool to optimize intervention selection. This could be particularly useful where therapeutic options are potentially complex, expensive or time-consuming (e.g. cognitive and mindfulness-based therapies and more experimental therapies such as neuro-feedback and direct and indirect brain stimulation). As mentioned previously, if future research succeeded at identifying subclusters of patients based on the weight placed on different parameters, these could be used to identify the best treatment for each subgroup.

Furthermore, even if the clustering of patients was not successful, other possible approaches to classify patients based on the parameters could also be taken. For instance, it might be possible to classify patients exclusively based on the weight placed on stimulation. Perhaps, those patients who place less weight on stimulation (those with higher β^2 values) would be the ones that could benefit the most from treatments such as mindfulness that attempt to increase the focus on the actual experience. Future studies should investigate whether the values obtained through the model with regards to the effect of stimulation are predictors of treatment success. However, we should emphasize once again that this study was carried out in healthy participants and that before advancing in its application, the same approach should be tested in a chronic pain sample.

In order to validate the use of this approach in patients, first a feasibility study could be carried out followed by a reliability study like the present one. For validation purposes, it may be convenient to test patients with an associated tissue damage, such as osteoarthritis. In this group of patients, the weight placed on stimulation could be compared with the correlation each patient shows between the tissue damage (e.g. as seen in radiographies) and their pain rating. However, if the clinical validation is successful, the patients who may benefit most from this approach could be those whose condition is not associated with an identified tissue damage (e.g. Fibromyalgia). In these cases, Bayesian methods could be used to determine the relative effect of unidentified tissue damage on overall disability, since this is not possible with current techniques.

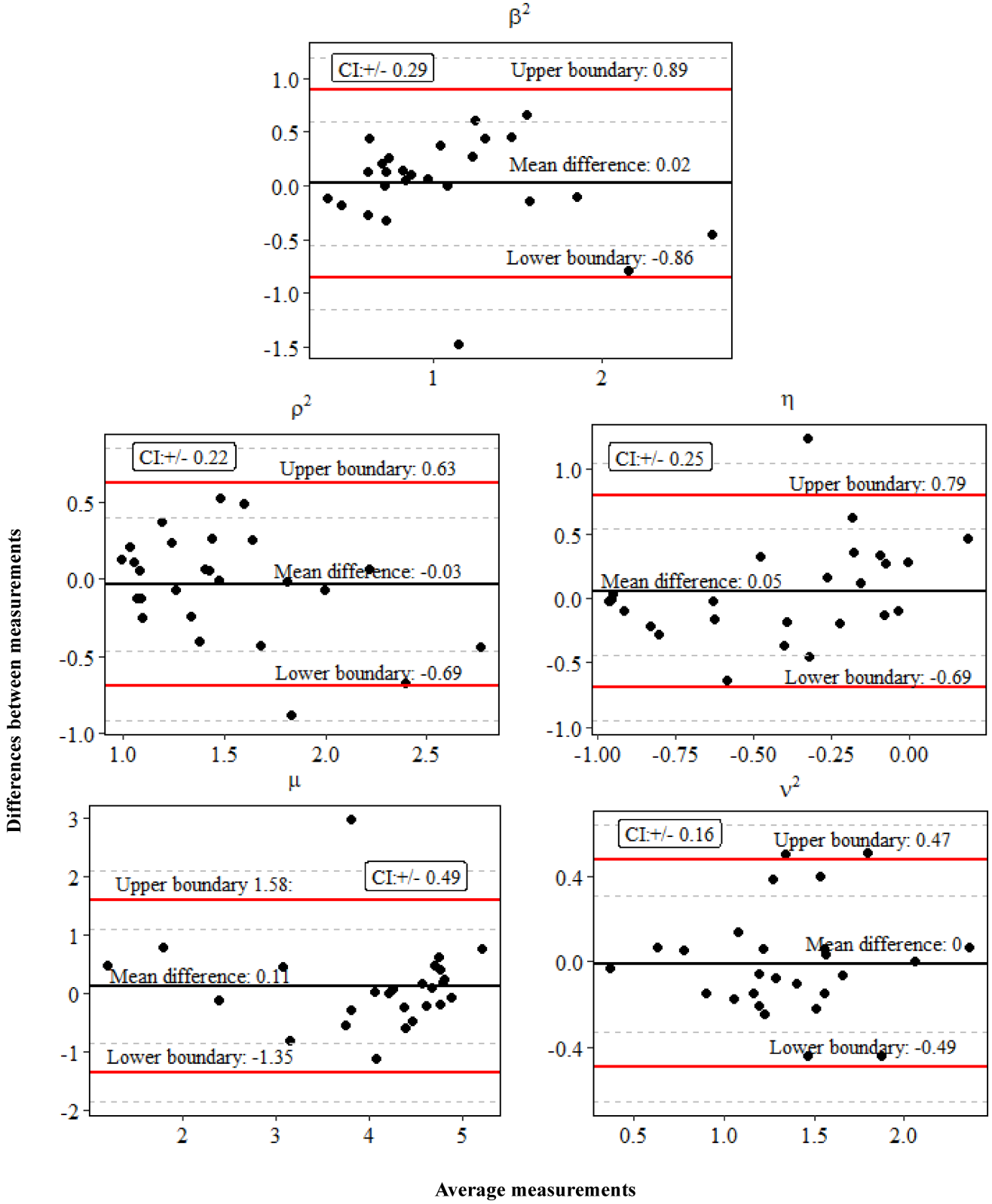

FIGURE 9 Bland-Altman plots for all parameters. The black solid line represents the mean difference in ratings between sessions. The red lines represent the upper and lower boundaries between which 95% of cases will be found and the dotted grey lines represent the confidence intervals for the boundaries.

In summary, this work has shown that the weight placed on the factors influencing pain perception is stable in time, particularly the weight placed on stimulation. Future research should study the applicability of this approach to clinical populations. If results are consistent in patients with chronic pain, this approach could open up a

TABLE 7 Repeatability coefficient per parameter.

Parameters	Repeatability coefficient
β^2	0.86
ρ^2	0.65
η	0.73
μ	1.45
ν^2	0.47

range of possibilities for patient stratification to identify patient phenotypes for targeted psychological and physiological interventions.

ACKNOWLEDGEMENTS

We would like to thank the Department of Medical Physics in Salford Royal Foundation Trust for the manufacturing and maintenance of the equipment used to carry out this study. We would also like to thank all the participants who took part in this study.

FUNDING INFORMATION

Ariane Delgado-Sanchez is funded by an MRC DTP programme (Award number: MR/N013751/1).

CONFLICT OF INTEREST STATEMENT

The authors declare no conflicts of interest.

DATA AVAILABILITY STATEMENT

The data supporting the findings reported in this paper are openly available at https://osf.io/axjfw/

ORCID

Ariane Delgado-Sanchez https://orcid.org/0000-0002-2390-6433

REFERENCES

Abma, I. L., Rovers, M., & van der Wees, P. J. (2016). Appraising convergent validity of patient-reported outcome measures in systematic reviews: Constructing hypotheses and interpreting outcomes. *BMC Research Notes*, *9*, 226.

Alghadir, A. H., Anwer, S., Iqbal, A., & Iqbal, Z. A. (2018). Test-retest reliability, validity, and minimum detectable change of visual analog, numerical rating, and verbal rating scales for measurement of osteoarthritic knee pain. *Journal of Pain Research*, *11*, 851–856.

Bahreini, M., Safaie, A., Mirfazaelian, H., & Jalili, M. (2020). How much change in pain score does really matter to patients? *The American Journal of Emergency Medicine*, *38*, 1641–1646.

Blanchard, T., Lombrozo, T., & Nichols, S. (2018). Bayesian Occam's razor is a razor of the people. *Cognitive Science*, *42*, 1345–1359.

Bland, J. M., & Altman, D. G. (2003). Applying the right statistics: Analyses of measurement studies. *Ultrasound in Obstetrics & Gynecology*, *22*, 85–93.

Braun, M., Bello, C., Riva, T., Hönemann, C., Doll, D., Urman, R. D., & Luedi, M. M. (2021). Quantitative sensory testing to predict postoperative pain. *Current Pain and Headache Reports*, *25*, 3.

Cherkin, D. C., Sherman, K. J., Balderson, B. H., Cook, A. J., Anderson, M. L., Hawkes, R. J., Hansen, K. E., & Turner, J. A. (2016). Effect of mindfulness-based stress reduction vs cognitive behavioral therapy or usual care on Back pain and functional limitations in adults with chronic low Back pain. *Jama*, *315*, 1240–1249.

Daunizeau, J., den Ouden, H. E. M., Pessiglione, M., Kiebel, S. J., Friston, K. J., & Stephan, K. E. (2010). Observing the observer (II): Deciding when to decide. *PLoS One*, *5*, e15555.

Felix, E. R., & Widerström-Noga, E. G. (2009). Reliability and validity of quantitative sensory testing in persons with spinal cord injury and neuropathic pain. *Journal of Rehabilitation Research and Development*, *46*, 69–84.

Georgopoulos, V., Akin-Akinyosoye, K., Zhang, W., McWilliams, D. F., Hendrick, P., & Walsh, D. A. (2019). Quantitative sensory testing and predicting outcomes for musculoskeletal pain, disability, and negative affect: A systematic review and meta-analysis. *Pain*, *160*, 1920–1932.

Graham-Engeland, J. E., Zawadzki, M. J., Slavish, D. C., & Smyth, J. M. (2016). Depressive symptoms and momentary mood predict momentary pain among rheumatoid arthritis patients. *Annals of Behavioral Medicine*, *50*, 12–23.

Hoskin, R., Berzuini, C., Acosta-Kane, D., El-Deredy, W., Guo, H., & Talmi, D. (2019). Sensitivity to pain expectations: A Bayesian model of individual differences. *Cognition*, *182*, 127–139.

Jurth, C., Rehberg, B., & Von Dincklage, F. (2014). Reliability of subjective pain ratings and nociceptive flexion reflex responses as measures of conditioned pain modulation. *Pain Research & Management*, *19*, 93–96.

Knill, D. C., & Pouget, A. (2004). The Bayesian brain: The role of uncertainty in neural coding and computation. *Trends in Neurosciences*, *27*, 712–719.

Koo, T. K., & Li, M. Y. (2016). A guideline of selecting and reporting intraclass correlation coefficients for reliability research. *Journal of Chiropractic Medicine*, *15*, 155–163.

Kovacs, F. M., Abraira, V., Royuela, A., Corcoll, J., Alegre, L., Tomás, M., Mir, M. A., Cano, A., Muriel, A., Zamora, J., del Real, M. T. G., Gestoso, M., Mufraggi, N., & The Spanish Back Pain Research Network. (2008). Minimum detectable and minimal clinically important changes for pain in patients with nonspecific neck pain. *BMC Musculoskeletal Disorders*, *9*, 43.

Letzen, J. E., & Robinson, M. E. (2017). Negative mood influences default mode network functional connectivity in patients with chronic low back pain: Implications for functional neuroimaging biomarkers. *Pain*, *158*, 48–57.

Lim, M., O'Grady, C., Cane, D., Goyal, A., Lynch, M., Beyea, S., & Hashmi, J. A. (2020). Threat prediction from schemas as a source of bias in pain perception. *The Journal of Neuroscience*, *40*, 1538–1548.

Lin, W., Zhou, F., Yu, L., Wan, L., Yuan, H., Wang, K., & Svensson, P. (2020). Quantitative sensory testing of periauricular skin in healthy adults. *Scientific Reports*, *10*, 1–11.

Moore, A., Derry, S., Eccleston, C., & Kalso, E. (2013). Expect analgesic failure; pursue analgesic success. *BMJ Online*, *346*, 7–9.

Morley, S. (2011). Efficacy and effectiveness of cognitive behaviour therapy for chronic pain: Progress and some challenges. *Pain*, *152*, S99–S106.

Nickel, M. M., Tiemann, L., Hohn, V. D., May, E. S., Gil Ávila, C., Eippert, F., & Ploner, M. (2022). Temporal-spectral signaling of sensory information and expectations in the cerebral processing of pain. *Proceedings of the National Academy of Sciences of the United States of America*, *119*, e2116616119.

Nothnagel, H., Puta, C., Lehmann, T., Baumbach, P., Menard, M. B., Gabriel, B., Gabriel, H. H. W., Weiss, T., & Musial, F. (2017). How stable are quantitative sensory testing measurements over time? Report on 10-week reliability and agreement of results in healthy volunteers. *Journal of Pain Research*, *10*, 2067–2078.

Palminteri, S., Wyart, V., & Koechlin, E. (2017). The importance of falsification in computational cognitive modeling. *Trends in Cognitive Sciences*, *21*, 425–433.

Petzschner, F. H., Glasauer, S., & Stephan, K. E. (2015). A Bayesian perspective on magnitude estimation. *Trends in Cognitive Sciences*, *19*, 285–293.

Pulvers, K., & Hood, A. (2013). The role of positive traits and pain catastrophizing in pain perception. *Current Pain and Headache Reports*, *17*, 330.

Radford, N. (2011). MCMC using Hamiltonian dynamics. In *Handbook of Markov chain Monte Carlo*. Chapman and Hall/CRC.

Salaffi, F., Stancati, A., Silvestri, C. A., Ciapetti, A., & Grassi, W. (2004). Minimal clinically important changes in chronic musculoskeletal pain intensity measured on a numerical rating scale. *European Journal of Pain*, *8*, 283–291.

Stan Development Team. (2023). *RStan the R interface to Stan*. R package version 2.32.3. https://mc-stan.org/

Streiner, D. L., Norman, G. R., & Cairney, J. (2015). *Health measurement scales: A practical guide to their development and use*. Oxford University Press.

Sturgeon, J. (2014). Psychological therapies for the management of chronic pain. *Psychology Research and Behavior Management*, *7*, 115.

Van den Bergh, O., Witthöft, M., Petersen, S., & Brown, R. J. (2017). Symptoms and the body: Taking the inferential leap. *Neuroscience and Biobehavioral Reviews*, *74*, 185–203.

Vaz, S., Falkmer, T., Passmore, A. E., Parsons, R., & Andreou, P. (2013). The case for using the repeatability coefficient when calculating test–retest reliability. *PLoS One*, *8*, e73990.

Vehtari, A., Gelman, A., & Gabry, J. (2017). Practical Bayesian model evaluation using leave-one-out cross-validation and WAIC. *Statistics and Computing*, *27*, 1413–1432.

Williams, A. C. D. C., & Craig, K. D. (2016). Updating the definition of pain. *Pain*, *157*, 2420–2423.

Wylde, V., Palmer, S., Learmonth, I. D., & Dieppe, P. (2011). Test-retest reliability of quantitative sensory testing in knee osteoarthritis and healthy participants. *Osteoarthritis and Cartilage*, *19*, 655–658.

SUPPORTING INFORMATION

Additional supporting information can be found online in the Supporting Information section at the end of this article.

How to cite this article: Delgado-Sanchez, A., Charalambous, C., Trujillo-Barreto, N. J., Safi, H., Jones, A., Sivan, M., Talmi, D., & Brown, C. (2024). Test–retest reliability of Bayesian estimations of the effects of stimulation, prior information and individual traits on pain perception. *European Journal of Pain*, *28*, 434–453. https://doi.org/10.1002/ejp.2193