Behavior Research Methods (2024) 56:6349–6362
https://doi.org/10.3758/s13428-023-02306-y

NOTES/COMMENTS/REPLY MANUSCRIPT

Temporal stability of Bayesian belief updating in perceptual decision-making

Isabella Goodwin[1] · **Robert Hester**[1] · **Marta I. Garrido**[1,2]

Accepted: 24 November 2023 / Published online: 21 December 2023
© The Author(s) 2023

Abstract

Bayesian inference suggests that perception is inferred from a weighted integration of prior contextual beliefs with current sensory evidence (likelihood) about the world around us. The perceived precision or uncertainty associated with prior and likelihood information is used to guide perceptual decision-making, such that more weight is placed on the source of information with greater precision. This provides a framework for understanding a spectrum of clinical transdiagnostic symptoms associated with aberrant perception, as well as individual differences in the general population. While behavioral paradigms are commonly used to characterize individual differences in perception as a stable characteristic, measurement reliability in these behavioral tasks is rarely assessed. To remedy this gap, we empirically evaluate the reliability of a perceptual decision-making task that quantifies individual differences in Bayesian belief updating in terms of the relative precision weighting afforded to prior and likelihood information (i.e., sensory weight). We analyzed data from participants (n = 37) who performed this task twice. We found that the precision afforded to prior and likelihood information showed high internal consistency and good test–retest reliability (ICC = 0.73, 95% CI [0.53, 0.85]) when averaged across participants, as well as at the individual level using hierarchical modeling. Our results provide support for the assumption that Bayesian belief updating operates as a stable characteristic in perceptual decision-making. We discuss the utility and applicability of reliable perceptual decision-making paradigms as a measure of individual differences in the general population, as well as a diagnostic tool in psychiatric research.

Keywords Individual differences · Perceptual decision-making · Reliability · Bayesian belief updating

Introduction

Bayesian accounts of perception suggest that the brain creates an internal model of the world to infer the cause of sensory input. This inference is generated via a weighted combination of prior expectations and incoming sensory observations which are used to estimate the state of any given environment (Hohwy, 2013, 2020; Knill & Pouget, 2004). Intrinsic uncertainty associated with prior and sensory (or likelihood) information determines how they are relatively weighted to form an internal representation of the world. Accurately incorporating this information into a veridical model of one's environment is essential for optimizing perception and effective perceptual decision-making (Friston, 2005, 2008). This framework, grounded in predictive processing, is useful for understanding complex mechanisms underlying healthy information processing. In turn, it also aids a phenomenological explanation for how aberrancies in the precision afforded to different types of information may characterize altered perceptual experiences (Hohwy, 2020). While these aberrancies in belief updating are thought to underlie symptoms of psychopathology and trait-like correlates in non-clinical populations (Fromm et al., 2023; Gibbs-Dean et al., 2023; Karvelis et al., 2023), the assumption that individual differences in belief updating is a stable characteristic is yet to be verified. This can offer valuable insight into the cognitive mechanisms underpinning behavior (Tulver et al., 2019).

Disruptions in the precision weighting afforded to prior and likelihood information have been used as a framework

✉ Isabella Goodwin
goodwin.isabella@gmail.com

[1] Melbourne School of Psychological Sciences, The University of Melbourne, Parkville Campus, Melbourne, Victoria 3010, Australia

[2] Graeme Clark Institute for Biomedical Engineering, The University of Melbourne, Melbourne, Victoria, Australia

for understanding a spectrum of clinical disorders including psychosis and schizophrenia (Adams et al., 2013; Sterzer et al., 2018), autism spectrum disorder (Palmer et al., 2015, 2017; Randeniya et al., 2021), mood disorders (Kraus et al., 2021; Putica et al., 2022) and more. This framework has also been used to investigate individual differences in non-clinical populations such as autistic traits, schizotypy, and trait anxiety (Goodwin et al., 2022; Kraus et al., 2021; Kreis et al., 2023). These individual differences in perceptual inference can be conceptualized across a continuum from stable characteristics in the general population, to more severe aberrancies in clinical disorders. This is particularly important in understanding the development and trajectory of disorders, integrating a transdiagnostic approach into understanding symptomatology (Gibbs-Dean et al., 2023; Lyndon & Corlett, 2020).

Investigating symptomatology and analogous sub-clinical characteristics of such disorders relies on the assumption that distinct perceptual phenotypes result from differences in the precision weighting of prior beliefs and sensory evidence (van Leeuwen et al., 2021). This assumes that perceptual differences in belief updating remain stable across individuals over time. While this offers important empirical utility in understanding of symptom formation, the assumption that Bayesian information integration is a stable characteristic has received little attention thus far. Furthermore, research that empirically verifies the reliability of behavioral measures to assess cognitive performance is scarce. This could largely impact the interpretation and application of such measures in our understanding of individual differences in cognitive function and its usefulness for clinical translation (Parsons et al., 2019). Therefore, it is important to verify whether measures with high variance between individuals (such as the precision weighting afforded to prior and likelihood information) can be attributed to stable individual differences in cognitive mechanisms (Parsons et al., 2019; Rouder et al., 2019).

Several studies have investigated the temporal stability of behavioral tasks that are used to measure stable characteristics in other measures of cognitive performance. For example, this has been assessed in model-based correlates of compulsivity (Brown et al., 2020) and processes underlying self-regulation (Zech et al., 2022). Behavioral and computational measures of a probabilistic reversal learning task demonstrated high reliability (Waltmann et al., 2022), validating its interpretability of individual differences in cognitive flexibility, as well as deviances across psychiatric populations in a transdiagnostic manner. This recent exploration into the psychometrics of cognitive behavioral measures have coincided with a shift in methodology to focus on hierarchical modeling that includes trial-level variation across a task, rather than traditional approaches that focus on average scores. This novel approach allows reliability estimates to vary at the individual level, meaning that researchers can directly test for homogenous within-person variance of behavioral measures (Williams et al., 2022).

Of relevance to the predictive coding framework, a recent study investigated the test–retest reliability in individuals' cross modal usage of priors in the perception of bistable visual stimuli (Pálffy et al., 2021). This paradigm adjudicated between individuals' reliance on auditory versus visual associative cues in visual perception. Importantly, substantial inter-individual variability suggested large differences in the relative use of acoustic compared to visual prior information, which researchers found to be temporally stable over two testing sessions. Despite this, it is unclear whether other aberrancies in Bayesian perceptual belief updating act as a temporally stable characteristic, particularly when the uncertainty associated with both prior *and* likelihood information are differentially altered.

This study aims to investigate whether a general reliance on likelihood relative to prior information can be considered a stable characteristic. This will be investigated with a behavioral paradigm that orthogonally manipulates prior and likelihood information and is known to yield systematic variation between individuals (Vilares et al., 2012). Along with traditional test–retest reliability measures of task performance, we will also use a hierarchical modeling approach that accounts for trial-by-trial estimates of sensory weight. This approach has consistently been shown to produce better reliability estimates in reliability, as variance in trial-by-trial data is incorporated. The behavioral paradigm will also allow us to determine whether average metrics of subjective uncertainty associated with likelihood and prior information remains stable over two testing sessions. Note however, that trial-by-trial estimates of subjective likelihood variance and subjective prior variance are not available, as these parameters are only calculated as an average metric across conditions or across the whole task.

Method

Participants

We aimed to recruit at least 30 participants, based on the number of healthy participants recruited in previous coin task paradigms (Randeniya et al., 2021; Trapp & Vilares, 2020; Vilares et al., 2012; Vilares & Kording, 2017) and based on similar test–retest studies investigating individual differences in perceptual inference (Pálffy et al., 2021). To account for potential dropout and exclusion, we initially recruited 62 participants who completed the first testing session. The final sample consisting of 37 participants who completed both testing sessions (29 female, six male, two non-binary, age range = 18–56, $M = 21.87$, $SD = 6.72$).

Participants completed the second testing session 12–17 days after the first session ($M = 14.22$, $SD = 1.80$), with the aim of averaging 2 weeks between testing sessions.

Participants were recruited through the University of Melbourne research experience program. Participants were at least 18 years of age and had corrected-to-normal vision. They were asked about their highest level of education, left or right handedness, whether English was their first language, how many years they had been speaking English if not, vision impairments, previous diagnosis of serious neurological conditions and/or emotional or psychological disorders, and any other conditions that might affect performance. All participants gave informed consent and received credit towards a university subject completion for participation. The study was approved by the University of Melbourne Human Research Ethics Committee (Ethics ID: 20592).

Experimental design

Procedure

Participants were recruited through a university online system (SONAR) that allowed completion of both testing sessions online, on their own laptops. Participants initially provided demographic details via Qualtrics (www.qualtrics.com), and were then directed to Pavlovia (www.pavlovia.com) to complete the behavioral task. The paradigm that participants completed at the second timepoint was the same as the paradigm at the first timepoint.

Coin task

Participants performed a decision-making task where they were asked to guess the position of a hidden target on a screen, requiring them to integrate both noisy sensory evidence and prior expectation of the target's location. More specifically, participants were told a coin was being thrown into a pond and were asked to guess where the coin had fallen. Likelihood and prior variance were manipulated with a two-by-two factorial design with narrow and wide variance, respectively. On each trial, five blue dots denoted "splashes" produced by the coin falling in. The variance of these splashes changed on each trial as an index of either narrow or wide likelihood conditions. The position of these splashes was drawn from a Gaussian distribution centered on the (hidden) location of the coin, with standard deviation of either 6% of the screen width ($SD = 0.096$; narrow likelihood trials) or 15% of the screen width ($SD = 0.24$; wide likelihood trials). An example trial is shown in Fig. 1. Participants were also informed that the person throwing the coin changed between blocks, and one thrower was more accurate than the other. They were told that both throwers aimed at the screen center (indicating the mean of the prior). Although they were not explicitly told which thrower was better or worse, this could be inferred through the distribution of previous coin locations from trial-to-trial. The location of the coin was drawn from a second, independent Gaussian distribution centered on the middle of the screen, with a standard deviation of either 2.5% of the screen width ($SD = 0.04$; narrow prior blocks) or 8.5% of the screen width ($SD = 0.136$; wide prior blocks). The four conditions are visually depicted in Fig. 1.

While the variance of the likelihood changed pseudorandomly from trial-to-trial (counterbalanced across all trials), the variance of the prior changed from block to block, with the order (thrower A vs thrower B) also counterbalanced across participants. Thus, there were four conditions: narrow prior and narrow likelihood (PnLn), narrow prior and wide likelihood (PnLw), wide prior and narrow likelihood (PwLn), and wide prior and wide likelihood (PwLw).

For each trial, participants were instructed to move a net (blue bar) horizontally across the screen to indicate where they thought the coin had landed. The true position of the coin (represented as a yellow dot) was then shown for 1500 ms. Scoring was tallied across each trial, with a point earned each time any part of the coin lay within the net. Participants were provided with two blocks of two practice trials before completing the main task. The main task consisted of two blocks per thrower, with each block containing 75 trials each (resulting in 300 trials total).

Likelihood only task

Prior to completing the coin task, participants completed the likelihood-only estimation task as a measure of subjective likelihood variance or sensory noise. The setup of this task was similar to the main task, without the incorporation of the prior condition. This provided an estimation of how participants perceived the center of the splashes on their own, without prior knowledge. Participants were asked to estimate where they thought the true coin location was, which was always the center of the displayed splashes, by moving the net horizontally across the screen. The true coin location (represented as a yellow dot) was revealed at the end of each trial, providing feedback on participants estimations. This task consisted of 100 trials, with an even number of wide and narrow likelihood distributions.

Behavioral analysis

Successful performance of the task required participants to move the net to the most likely location of the hidden coin. Using Bayes rule, we can determine what the optimal estimate of the position of the coin would be on each trial (Körding & Wolpert, 2004; Vilares et al., 2012):

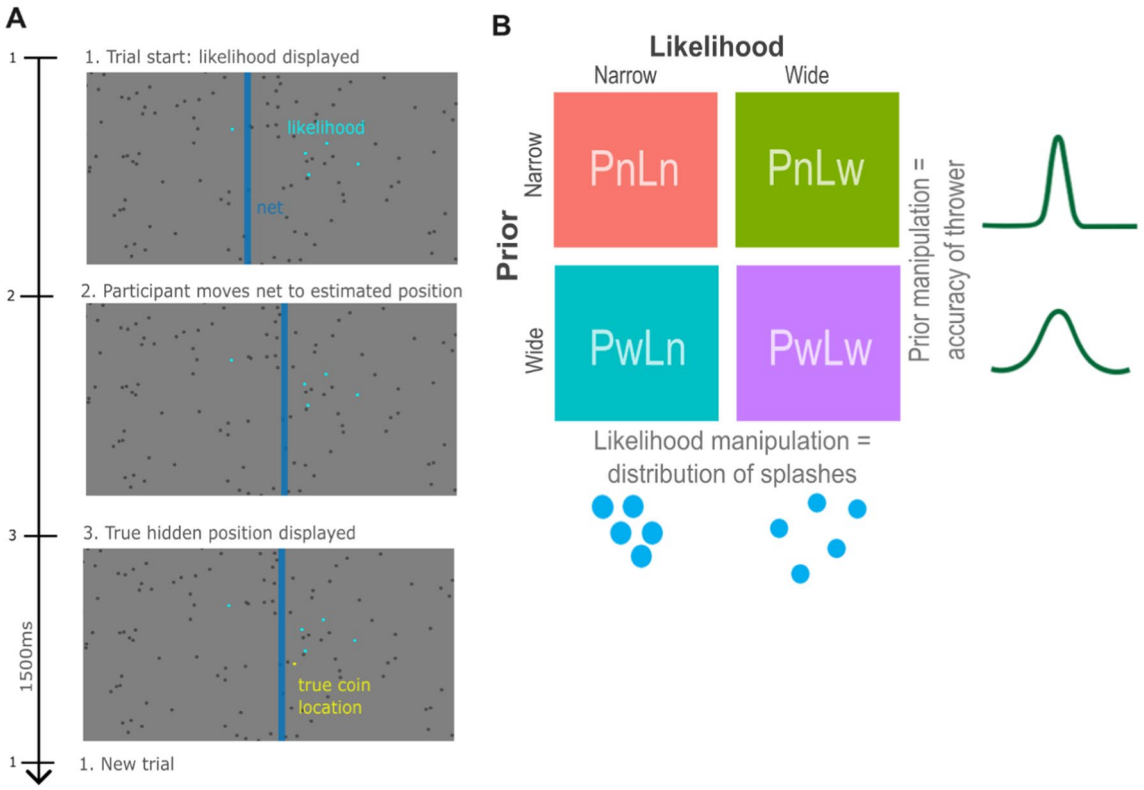

Fig. 1 Coin task paradigm as adapted from Vilares et al. (2012) demonstrating **A** the time course of a single trial and **B** the task conditions. *Note.* A) Time course of a single exemplar trial: participants are shown five blue dots to represent splashes of the location of a coin being thrown into a pond. They are then asked to move the blue bar/net to where they estimate the coin's location to be, after which the coin's true location is revealed, and they move onto the next trial. B) Task design as adapted from Vilares et al. (2012): the four conditions of the task are visually depicted including two types of likelihood as manipulated through the distribution of splashes on each trial (Ln = narrow likelihood; Lw = wide likelihood) and two types of prior as manipulated through the accuracy of the thrower on each block (Pn = narrow prior; Pw = wide prior)

$$X_{est} = \frac{\sigma_L^2}{\sigma_L^2 + \sigma_P^2}\,\mu_P + \frac{\sigma_P^2}{\sigma_L^2 + \sigma_P^2}\,\mu_L \tag{1}$$

where X_{est} is the estimated position of the coin (i.e., participants responses on each trial), (μ_P, μ_L) represent the prior and likelihood means and (σ_P^2, σ_L^2) represent the prior and likelihood variances, respectively. In our experiment, the mean of the prior was kept constant (the center of the screen, μ_P), while the mean of the likelihood was determined by the center of the five blue dots in each trial (μ_L).

Performance

Performance in the likelihood-only task was characterized by the average distance between participants' estimates of the coin location (net location) and the true center of the splashes (i.e., mean estimation error). Similarly, performance in the coin task was characterized by the average distance between participants' estimates (net location) and the true location of the coin.

Overall sensory weight (likelihood vs prior reliance)

To estimate participants reliance on likelihood relative to prior information, we fitted a linear regression to participants' estimates of the coin's position for each trial (X_{est}) as a function of the center of the splashes (i.e., the likelihood mean, μ_L):

$$sw = \frac{\sigma_P^2}{\sigma_L^2 + \sigma_P^2} \tag{2}$$

where sw is the slope of the linear regression, which indicates how much each participant relies on likelihood information. A slope closer to 1 indicates a greater reliance on the likelihood information, while a slope closer to 0 indicates greater reliance on prior information. A slope between 0 and 1 indicates that participants integrate both likelihood and prior information in their estimates. This was calculated overall, for each condition, and for each block.

Bayesian optimal sensory weights

If participants perform according to the Bayesian optimum as portrayed in Eq. (1), then the optimal values for the slopes/sensory weights should be equal to the perceived $\frac{\sigma_P^2}{\sigma_L^2+\sigma_P^2}$, where σ_P^2 is the variance associated with the prior (narrow prior $\sigma_P^2 = 0.04^2$; wide prior $\sigma_P^2 = 0.136^2$) and σ_L^2 is the variance associated with the likelihood (in this instance, narrow likelihood $\sigma_L^2 = \frac{0.096^2}{5}$; wide likelihood $\sigma_L^2 = \frac{0.24^2}{5}$). These calculations of Bayesian optimality refer to posterior computations, integrating the relative uncertainty of both prior and likelihood information.

Trial-by-trial sensory weight

Equation (1) can be rewritten to calculate an instantaneous sensory weight as an indicator of participants reliance on likelihood relative to prior information on any given trial:

$$sw_{trial} = \frac{X_{est} - \mu_P}{\mu_L - \mu_P} \tag{3}$$

where X_{est} is the participants estimated position of the coin on a given trial (net location), μ_P is the mean of the mean of the prior (assumed at the center of the screen), and μ_L is the mean of the likelihood (the center of the five blue dots for that trial). To ensure the trial-by-trial sensory weight varied from 0 to 1, we used a log-transformation for analyses.

Subjective likelihood variance

The likelihood-only task can be used to determine a proxy for participants subjective likelihood variance or sensory noise (Randeniya et al., 2021). This is determined by the variance of the participants estimates of the mean (μ_{est}) relative to the true mean of the splashes (μ_L):

$$\sigma_{LS}^2 = \frac{\Sigma\left(\mu_{est} - \mu_L\right)^2}{nTrials} \tag{4}$$

where the number of trials (nTrials) was equal to 100 in the likelihood-only task.

Subjective prior variance

To estimate participants subjective model of where each thrower would throw the coin (subjective prior variance, σ_P^2), the sensory weight from Eq. (2) can be rearranged as follows:

$$\sigma_P^2 = \frac{\sigma_L^2 * sw}{(1 - sw)} \tag{5}$$

In this equation, σ_L^2 can be assumed to be the objective likelihood variance (i.e., the variance of the splashes by design), or alternatively, this can be estimated from participants subjective likelihood variance, σ_{SL}^2 (as calculated in Eq. 4).

Statistical analysis

Firstly, mean estimation error was used as a criterion to detect poor performance or low effort, with four participants excluded in the likelihood only task, and two participants excluded in the main task (z-score greater than 3). To determine differences in key parameters across the two timepoints, *t* test and ANOVAs were calculated. A log-transformation was applied to non-parametric data to normalize the distribution. Test–retest reliability was then analyzed with Pearson correlations and intraclass correlations, ICC (2, 1), using a two-way random effects model based on absolute agreement (Koo & Li, 2016). The general formula for the traditional ICC can be calculated as follows:

$$p = \frac{\sigma_0^2}{\sigma_0^2 + \sigma_\varepsilon^2} \tag{6}$$

where σ_0^2 refers to between-person variance and σ_ε^2 refers to within-person variance. ICCs with values of < 0.5 were interpreted as poor, 0.5–0.75 as fair, 0.75–0.90 as good, and > 0.90 as excellent reliability (as suggested by Koo & Li, 2016). Additionally, split-half reliability was analyzed for trial-by-trial estimates of estimation error and sensory weight at each timepoint separately. This was done using a permutation-based split-half approach (Parsons, 2021), in which data is repeatedly split into two halves and the reliability estimate is calculated for each split, then averaged to provide a more stable estimate of reliability. We used 5000 random splits as recommended by Parsons (2021) with Spearman–Brown corrected estimates and their 95% percentile intervals reported as a measure of internal consistency. Finally, exploratory analyses of trial-by-trial sensory weight included implementing a hierarchical model that allows for individually varying intraclass correlations, to address cross-trial variability and directly test for homogenous within-person variance in test–retest reliability (Williams et al., 2022; using vICC package in R), which can be calculated as:

$$p_i = \frac{\sigma_0^2}{\sigma_0^2 + \exp\left[\eta_0 + \mu_{1i}\right]} \tag{7}$$

where i refers to the ith individual, σ_0^2 refers to between-person variance,η_0 represents the average of individual variances, and μ_{1i} represents individual departures from the fixed group effect. This model computes person-specific ICC,

allowing us to determine which (and how many) individuals belong to the common variance ICC model (as described in Eq. 6). This means that trial-by-trial sensory weight data across both testing sessions is used to inform an individual ICC metric which is unique to each participant. Like the traditional ICC approach, higher individual ICC estimates demonstrate more stability or similar scores across testing sessions, while lower individual ICC estimates demonstrate larger differences across testing sessions. If *most* participants demonstrate high individual ICC estimates, we can be more confident that this reflects *true stability* in performance across testing sessions (and vice versa for low individual ICC estimates). In other words, we are more interested in whether the majority of participants demonstrate similar individual ICC estimates, or whether the sample has a large degree of variation in individual ICC scores. Given the novelty of this methodological approach, there are no clear guidelines for the proportion of participants demonstrating low individual ICC scores to constitute *true change*. All visualizations and analyses were performed in R (version 2022.02.2).

Results

Participants

Data from a total of 37 participants was collected from participants that completed both testing sessions, with demographic information provided in Table 1. Demographic information of participants that only completed timepoint 1 and not timepoint 2 (i.e., dropped out of the study) can be found in the Supplementary (Table S1).

Performance accuracy

An analysis of performance accuracy in the likelihood-only task revealed no significant difference in overall mean estimation errors between timepoint 1 and timepoint 2 ($t(65) = 1.8$, $p = .076$, 95% CI [– 0.011, 0.0005]). When comparing performance across each condition, we found significantly greater mean estimation errors in the wide likelihood condition compared to the narrow likelihood condition at timepoint 1 ($t(32) = 5.48$, $p = 5.26 \times 10^{-8}$, 95% CI [0.0167, 0.009]) and timepoint 2 ($t(32) = 5.49$, $p = 4.83 \times 10^{-6}$, 95% CI [0.021, 0.009]). This is intuitive, given that the wide likelihood condition provides less certain information about the coin's location, replicating previous findings (Goodwin et al., 2022). When considering performance accuracy in the main task, there was no significant difference in mean estimation error between timepoint 1 and timepoint 2 ($t(139) = – 0.73$, $p = 0.467$, 95% CI [– 0.0086, 0.004]). When comparing performance across each condition, a two-way ANOVA revealed a main effect of prior (Pw>Pn; Timpoint 1: $F = 67.62$, $p = 1.41 \times 10^{-13}$; timepoint 2: $F = 30.63$, $p = 1.55 \times 10^{-7}$) and a main effect of likelihood (Lw>Ln; timepoint 1: $F = 109.26$, $p < 2 \times 10^{-16}$; timepoint 2: $F = 69.10$, $p = 8.54 \times 10^{-14}$) across both timepoints. This indicates that more estimation errors were occurring in the conditions with greater uncertainty, as expected by the task design.

We estimated the internal consistency of overall estimation error in the main task using a permutation-based split-half approach (Parsons, 2021) with 5000 random splits. The Spearman–Brown corrected split half internal consistency measure was shown to demonstrate good to excellent reliability at timepoint 1 ($r_{SB} = 0.82$, 95% HDI [0.72, 0.90]) and good to moderate reliability at timepoint 2 ($r_{SB} = 0.76$, 95% HDI [0.57, 0.88]), based on recommendations from Koo and Li (2016). Additionally, intraclass correlation coefficient of overall estimation error demonstrated moderate test–retest reliability (ICC(2, 1) = 0.47, 95% CI [0.16, 0.68]) with a positive correlation between testing sessions ($r = 0.67$, $p = 1.83 \times 10^{-5}$, bootstrapped 95% CI [0.49, 0.88]). This indicates that performance accuracy was stable within testing sessions, as well as across the two timepoints.

Table 1 Demographic information collected from participants ($n = 37$) that completed the experiment

Age (years)	M	SD	Range
	22.62	6.72	19–57
Gender	Female	Male	Other
	29	6	2
Highest level of education	Primary school	Secondary school	Tertiary education
	0	26	11
English as a first language	Yes	No	
	17	20	
Handedness	Left-handed	Right-handed	No preference
	5	31	1

Average sensory weights suggests that participants perform in a Bayesian manner (non-optimally)

Sensory weight (likelihood to prior reliance) was calculated overall and for each condition across both timepoints. We found no significant difference between the average sensory weight at timepoint 1 ($M = 0.63$, $SD = 0.17$) and timepoint 2 ($M = 0.65$, $SD = 0.18$; $t(34) = -1.15$, $p = .259$). Analysis of sensory weights across conditions, showed a main effect of prior across both timepoints (Pw>Pn; timepoint 1: $F = 25.29$, $p = 1.53 \times 10^{-6}$, timepoint 2: F =24.70, $p = 1.98 \times 10^{-6}$), as well as a main effect of likelihood (Ln>Lw; timepoint 1: $F = 8.46$, $p = .0042$, timepoint 2: $F = 7.94$, $p = .0056$). This indicates that participants relied more on likelihood information when the prior was more uncertain but relied less on likelihood information when the likelihood was more variable, as expected. Despite this, Wilcox ranked tests showed that median sensory weights across three of the four conditions were significantly different from Bayesian optimal scores at each timepoint (excepting PwLw condition; see supplementary S2). Although participants generally deviated from Bayesian optimal, their patterns of performance were verging towards optimality across each condition, as shown in Fig. 2.

Sensory weight parameter has both good internal stability and test–retest reliability

Trial-by-trial sensory weight (see Eq. (3) for calculation) demonstrated good-to-excellent internal consistency at timepoint 1 ($r_{SB} = 0.87$, 95% HDI = [0.80, 0.93]) and at timepoint 2 ($r_{SB} = 0.88$, 95% HDI = [0.81, 0.93]), as measured with a Spearman–Brown corrected split-half measure with the permutation approach. To evaluate whether individuals' average sensory weight (see Eq. 2) remained stable across the two timepoints, we performed a test–retest analysis using Pearson correlation across sessions. Further, we calculated the intraclass correlation coefficient (ICC2, 1) which reflects the absolute agreement between measurements. Overall sensory weight showed good test–retest reliability (ICC(2, 1) = 0.73, 95% CI [0.53, 0.85]) with a strong positive Pearson correlation between timepoint 1 and timepoint 2 ($r = 0.73$, $p = 6.49 \times 10^{-7}$, bootstrapped 95% CI [0.59, 0.87]) as shown in Fig. 3. When considering individuals' sensory weights across conditions, split-half reliability analyses demonstrated good to moderate internal consistency at both timepoints (supplementary S3). Similarly, measures of intraclass correlation coefficients demonstrated good to moderate test–retest reliability across the two testing sessions (supplementary S3).

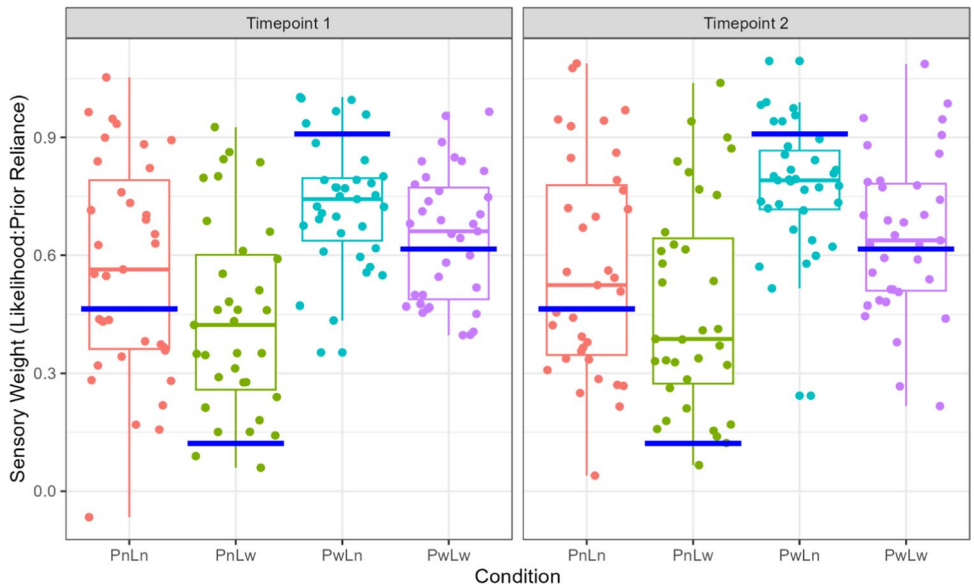

Fig. 2 Comparison of sensory weights across each condition at timepoint 1 and timepoint 2. *Note*. Sensory weight for each participant is calculated by the slope of the regression between the true center of the likelihood and participant's estimates of the coin's location for each condition. A sensory weight closer to 1 indicates greater reliance on likelihood, whilst a sensory weight closer to 0 indicates greater reliance on prior. *Blue lines* indicate the Bayesian optimal computation of the coin's location, based on the posterior integration of uncertainty in both prior and likelihood information. Conditions: PnLn = narrow prior, narrow likelihood *(red dots)*; PnLw = narrow prior, wide likelihood (*green dots*); PwLn = wide prior, narrow likelihood (*teal dots*); PwLw = wide prior, wide likelihood (*purple dots*).

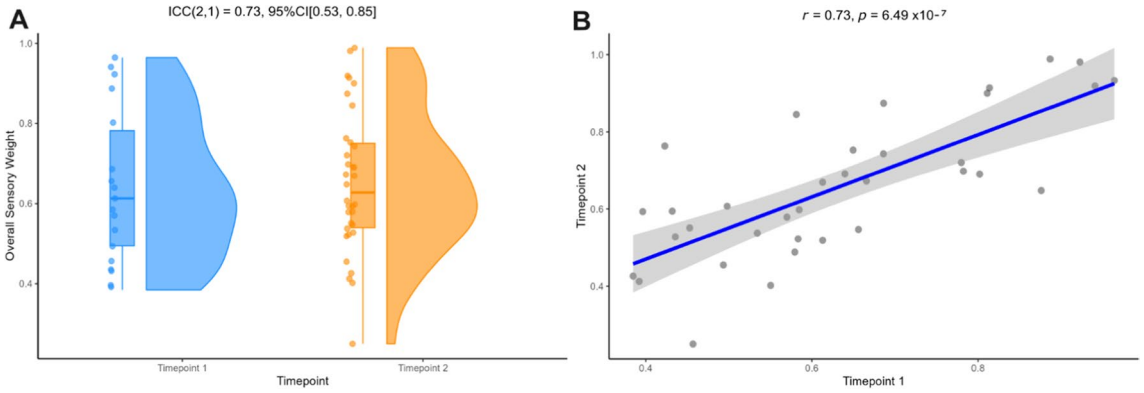

Fig. 3 Average overall sensory weight scores demonstrated **A** high test–retest reliability across two timepoints and **B** a strong positive Pearson correlation between timepoint 1 and timepoint 2

Subjective likelihood uncertainty (likelihood-only task) showed weak test–retest reliability between timepoints 1 and 2

Subjective likelihood variance was calculated in the likelihood-only task as a proxy of how much participants perceived uncertainty in likelihood information (i.e., distribution of the five blue dots) to vary across the task. Unexpectedly, there was no significant difference between average scores of subjective likelihood variance in the narrow condition, compared to the wide condition at both timepoint 1 ($t(30) = 0.165$, $p = .869$) and timepoint 2 ($t(30) = 1.99$, $p = .066$). Overall subjective likelihood variance scores were calculated for each participant, showing no significant difference between timepoint 1 ($M = 0.0048$) and timepoint 2 ($M = 0.0074$, $t(30) = 1.42$, $p = .166$). Before reliability analyses were conducted, log-transformations were applied to subjective likelihood variance scores to normalize the distribution of data (see Supplementary S4). Following this, ICC analysis of overall subjective likelihood variance scores demonstrated weak test–retest reliability between timepoint 1 and timepoint 2 (ICC(2, 1) = 0.36, 95%CI [0.013, 0.63]), with a positive correlation between timepoints ($r = 0.36$, $p = .042$, bootstrapped 95%CI [0.0079, 0.74]; see Fig. 4).

Subjective prior uncertainty showed good to moderate test–retest reliability between timepoints 1 and 2

Considering the lack of difference in subjective likelihood variance observed between narrow likelihood and wide likelihood conditions (from the likelihood-only task), the objective likelihood variance was instead used to calculate subjective prior variance (see Eq. 5). This, along with overall sensory weights, were utilized as a proxy to determine how much participants were perceiving uncertainty in the prior information (i.e., accuracy of thrower) to vary across the task. A comparison of subjective prior variance across conditions revealed no main effect of prior at timepoint 1 ($F = 3.45$, $p = .066$), but a main effect of prior at timepoint 2 (Pw>Pn; $F = 5.33$, $p = .023$). This shows that participants were more likely to perceive uncertainty in prior information

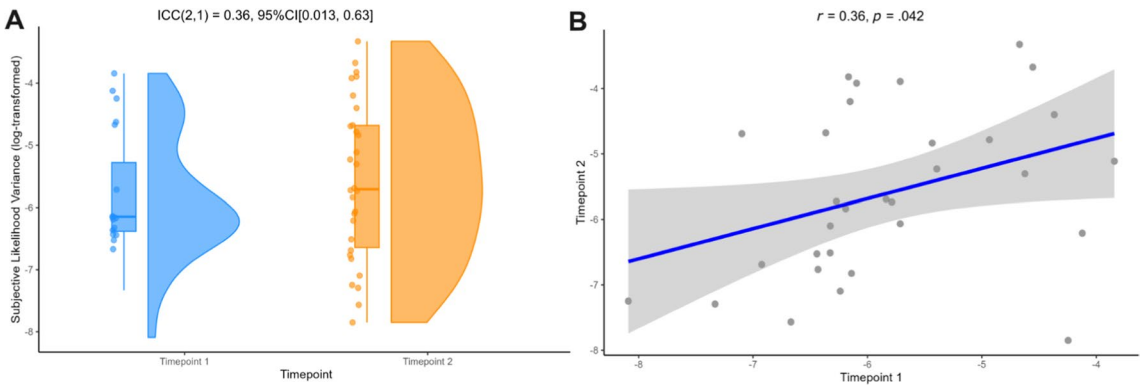

Fig. 4 Average subjective likelihood variance scores from the likelihood-only task demonstrated **A** weak test–retest reliability across two timepoints and **B** a positive correlation between timepoint 1 and timepoint 2

when it was objectively more uncertain (i.e., in the wide prior condition). However, we also unexpectedly found a main effect of likelihood (Lw>Ln; timepoint 1: $F = 13.17$, $p = 4.31 \times 10^{-4}$; timepoint 2: $F = 13.13$, $p = 4.34 \times 10^{-4}$), suggesting that the likelihood condition was also influencing perceived uncertainty in prior information. Overall subjective prior variance was also calculated for each participant, in which we found no significant difference between timepoint 1 (M = 0.018) and timepoint 2 (M = 0.020, $t(32) = -0.30$, $p = .77$). Data were normalized with a log-transformation before reliability analyses were conducted (see Supplementary S5). We found high test–retest reliability (ICC(2, 1) = 0.67, 95% CI [0.44, 0.82]), with a positive Pearson correlation between timepoint 1 and timepoint 2 ($r = 0.67$, $p = 1.72 \times 10^{-5}$, bootstrapped 95% CI [0.50, 0.86]; see Fig. 5).

Exploratory analyses: Individually varying intra-class correlation coefficients

Traditional ICC analyses rely on mean point estimates with the assumption of a common within-person variance model, suggesting that each individual is adequately described by the average within-person variance. Rouder and Haaf (2019) demonstrated that average aggregates of subject-level data can greatly attenuate measures of reliability. Instead, they suggest that modeling individual-level variability in trial-by-trial parameters can yield robust individual differences. A novel hierarchical-modeling approach developed by Williams et al. (2022) can verify whether this average-level reliability is generalizable to the individual level. This approach tests for homogeneity of within-person variance with individually varying intraclass correlation coefficients. In other words, it allows identification of which, and how many individuals belong to a common variance model (i.e., which individuals the traditional ICC is representative of), and which individuals fall outside that common variance. This allows for the possibility of individual differences in test–retest reliability, renouncing the assumption that individuals are unlikely to deviate from the average. Whilst this technique provides an excellent opportunity to deeply characterize individual differences in this research, its novelty renders these analyses exploratory at this stage. For these analyses, we computed individually varying ICCs of trial-by-trial sensory weight estimates (see Eq. 3) separated by conditions representative of trial types. The models were fitted with the R package vICC as described in Williams et al. (2022).

Across each condition, we found that individually varying ICC2 estimates of most participants belonged to the common variance model, as shown in Fig. 6. The wide prior wide likelihood condition (PwLw) demonstrated the most homogeneity in test–retest estimates, whereby all participants belonged to the common variance model (as depicted in red). Whilst the narrow prior narrow likelihood condition (PnLn) demonstrated the most heterogeneity, only nine participants fell outside what was described by the common variance model (as depicted in blue). In other words, there was not a large degree of heterogeneity in individually varying ICC calculations across each condition, suggesting that the common variance model is an accurate descriptor of test–retest reliability across the majority of individuals in this sample. Additionally, these analyses show that confidence intervals around point estimates fall within a wide range of poor to excellent test–retest reliability depending on the individual, which demonstrates variable interpretability across the sample.

Discussion

The aim of the present study was to determine whether individual differences in the precision weighting of prior to likelihood information remain temporally stable over two testing sessions. This was to empirically examine

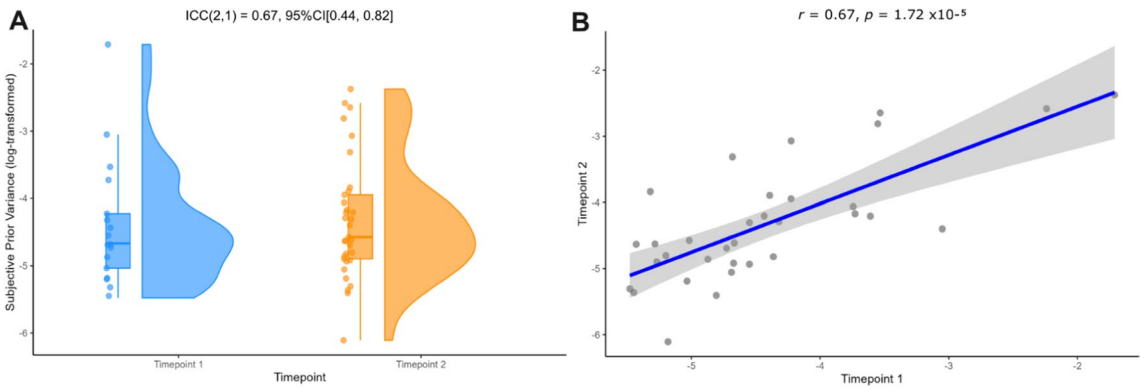

Fig. 5 Average subjective prior variance scores demonstrated **A** high test–retest reliability across two timepoints and **B** a positive Pearson correlation between timepoints 1 and 2

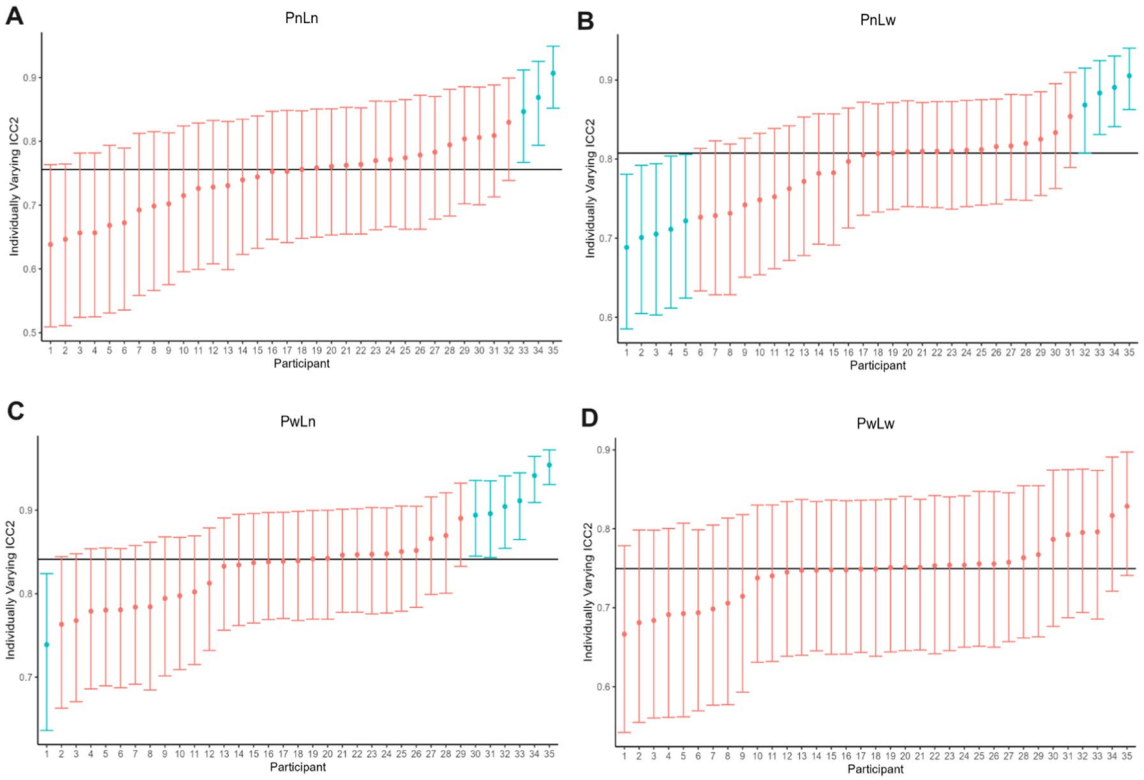

Fig. 6 Varying intraclass correlation coefficients of trial-by-trial sensory weight across each condition. *Note.* A) narrow prior narrow likelihood (PnLn), B) narrow prior wide likelihood (PnLw), C) wide prior narrow likelihood (PwLn), and D) wide prior wide likelihood (PwLw). These plots show which individuals are homogenous within in the common variance model (in *red*), and which individuals fall outside the 'traditional' or average ICC (in *blue*). The traditional mean point estimate ICC is demonstrated as the *black horizontal line* in each plot, while individually varying ICCs are demonstrated as a point estimate and confidence interval around that estimate for each participant

the assumption that this measure acts as a stable characteristic of Bayesian information integration. Precision weighting was measured with a behavioral paradigm that parameterizes the weighting afforded to prior and likelihood information via orthogonal manipulation of uncertainty across trials. Our data demonstrated high internal consistency and good test–retest reliability in individuals' average sensory weight across the task, as well as good to moderate test–retest reliability in individuals' average sensory weight across conditions. At the individual level, exploratory analyses investigating individually varying ICCs across conditions suggested high homogeneity with a common variance model, indicating good to moderate test–retest reliability of trial-by-trial sensory weights amongst individuals. Thus, sufficient temporal stability of this parameter at both the average and individual level demonstrates its suitability as a cognitive marker of individual differences in perceptual decision-making in the general population (Rouder & Haaf, 2019). This has important implications for the empirical utility of Bayesian belief updating in future clinical applications of this task (Parsons et al., 2019).

Importantly, the reliability of the precision weighting of prior to likelihood information in this study (ICC(2,1) = 0.73, 95% CI [0.53, 0.85]) is comparable to similar performance-based indicators of perceptual inference in other task designs. For example, a recent study investigated the influence of cross-modal auditory priors in the perception of bistable visual stimuli, revealing high temporal stability in individual differences of perceptual inference (median ICC = 0.83, 95% CI [0.61, 0.95]; Pálffy et al., 2021). Similarly, behavioral measures of cognitive flexibility in a probabilistic reversal learning task were also found to have excellent reliability (derived from mixed-effects models over two testing sessions; Waltmann et al., 2022). This provides evidence for the generalizability of reliable individual differences in behavioral measures of perceptual inference in healthy populations. While test-retest reliability is important to ensure a parameter acts as a stable characteristic, it is also important to verify the internal reliability of a parameter, to ensure the construct does not rapidly fluctuate within individuals. Internal consistency estimates such as split half reliability can provide an upper bound on test–retest reliability (Karvelis et al., 2023). In the current study, the internal consistency

of sensory weight was found to be good to excellent at timepoint 1 (r_{SB} = 0.87, 95% HDI = [0.80, 0.93]) and at timepoint 2 (r_{SB} = 0.88, 95% HDI = [0.81, 0.93]), meaning that individuals' precision weighting of prior to likelihood information has high reliability within the task itself.

Task-based measures of perceptual inference are often used as trait-like characteristics to assess the relationship with clinical symptoms across a spectrum of disorders such as schizophrenia (Deserno et al., 2020; Schlagenhauf et al., 2014; Weilnhammer et al., 2020) and autism spectrum disorder (Kreis et al., 2021; Randeniya et al., 2021), as well as their non-clinical trait-like correlates (Goodwin et al., 2022; Kreis et al., 2023). Mounting evidence for the reliability of distinct perceptual phenotypes across task-based measures of cognition support an important development in the translation of computational measures to clinical practice, particularly as a diagnostic tool in psychiatric research (Gibbs-Dean et al., 2023; van Leeuwen et al., 2021). For this translation to be effective, the reliability of such measures of perceptual inference should also be established in clinical populations before empirical use. The utility of computational psychiatry is growing, not only as a tool to investigate mechanisms underlying cognition and behavior, but also to inform techniques in therapy (Pott & Schilbach, 2022), predict treatment response (Hauke et al., 2023), predict transitions from 'at risk' states (Hauke et al., 2023), and the potential retraining of perceptual priors in clinical disorders (Lyndon & Corlett, 2020). This demonstrates the importance of generating robust and reliable tools to formally assess beliefs, how they change over time, and how they relate to observable behavior.

Furthermore, although sufficient temporal stability of the precision weighting of prior to likelihood information was demonstrated, the subjective uncertainty of likelihood information was found to have poor test–retest reliability in our task. Practice effects could hinder the reliability of this parameter, as exposure to the likelihood-only task (from which this parameter is calculated) in the second testing session is no longer naïve (Randeniya et al., 2021). In the first testing session, the coin task is initially presented without manipulating prior information (i.e., the likelihood-only task), in order to yield a proxy for participants' unbiased estimates of the subjective uncertainty associated with likelihood information. In the second testing session, the likelihood-only task is again presented before the main task, however participants' previous experience with the main task might be biasing their responses, due to a carryover of prior information from the first testing session to the second. Thus, following up this study with further testing sessions would provide insight into whether practice effects impact the strategy being used (indicative of individuals' subjective likelihood uncertainty) in the likelihood-only task. Alternatively, increasing the timeframe between testing sessions might provide further insight into the trade-off between practice effects and stability in responding (Karvelis et al., 2023; Zech et al., 2022).

In sum, our findings provide support for temporal stability within individuals in the precision weighting of likelihood to prior information in single task-based measures. However, it is unclear whether these findings are generalizable to other paradigms that tap into processes of Bayesian information integration. Research in this area so far has yielded contradictory findings. In attempts to understand state vs trait alterations in predictive processing, previous research has utilized different methods of prior induction to determine whether a single factor could explain performance (such as a relative reliance on priors) across multiple tasks (Andermane et al., 2020; Koblinger et al., 2021). For example, Tulver et al. (2019) investigated whether the tendency to rely on priors across multiple paradigms with noisy or ambiguous perceptual input could co-explain performance, and whether this was linked to autistic or schizotypal traits in non-clinical populations. Surprisingly, they found no single factor to explain individual differences or a common reliance on priors across four tasks with visual illusions. This might suggest that different methods of prior induction may recruit distinct neural mechanisms that operate at different levels of the visual hierarchy. Similarly, Andermane et al. (2020) investigated whether a general tendency to see the expected is general, or method specific with different facilitatory effects of perceptual priors. While they found that individual differences in expectation-based biases are closely related to attentional ability, test–retest reliability is required to decisively measure whether this operates as a consistent phenotypic difference. Whilst these computational cognitive models are often context-specific, it is yet to be verified whether measures of perceptual inference in the coin task can also predict performance across other, similar tasks.

One limitation of our task design was that the order in which participants observed narrow versus wide uncertainty in prior information was not counter balanced across testing sessions. Ensuring that each participant received a different version of the task across the two testing days could potentially improve expectation-based biases in responding and further limit practice-effects. Additionally, to ensure robust replicability of the online-version of this task, future research should verify the test–retest reliability across online and in-person testing sessions, to ensure the consistent performance of psychometrics (Zech et al., 2022). As a quality control measure for our online testing, participants with particularly high mean estimation error scores were removed, as this was deemed to be an indicator of poor adherence or engagement with the task. Additionally, participants were asked to complete practice tasks prior to completion of the main task, to ensure that they understood what was required in the main task. While

online testing has many resourceful benefits, an advantage of in-person cognitive-behavioral testing is that extraneous environmental factors can be more stringently controlled. Although cognitive processes will inevitably fluctuate across testing sessions, controlling for the time of day (i.e., circadian rhythm; Bedder et al., 2023) and assessing mood-related fluctuations (Eldar et al., 2018) could provide insight into state-related differences across testing sessions that might impact measurement reliability. Future research should include more sensitive investigations of state-like fluctuations, as these could provide insight into deviations from homogenous common variance models which may in fact be clinically meaningful (Karvelis et al., 2023; Sullivan-Toole et al., 2022). Despite this, testing perceptual inference in clinical settings might also lack this stringent stability and have inevitable fluctuations, suggesting that our research might hold better generalizability to clinical testing than a strictly controlled lab-based setting. To further aid generalizability, these findings should be replicated in a more age and demographically diverse sample, to ensure translational value to clinical populations. Similarly, the small sample size limits the generalizability of these findings, meaning that the results should be considered a promising starting point for the reliability of sensory weighting, with future replications in a larger sample to provide more conclusive evidence.

This study demonstrated good internal consistency and sufficient average and individual estimates of test–retest reliability in the precision weighting of prior to likelihood information in a perceptual decision-making task. This provides evidence that individual differences in task-based measures of Bayesian information integration perform as a stable characteristic. These results support the characterization of a latent computational measure as a means to capture potentially clinically relevant individual differences in computational psychiatry (Karvelis et al., 2023). This verification of psychometrics of cognitive-behavioral tasks should be standard practice before exploring their relationship with symptoms in clinical disorders and subclinical trait-like correlates in the general population.

Supplementary Information The online version contains supplementary material available at https://doi.org/10.3758/s13428-023-02306-y.

Open practices statement The data, analysis code, and all materials for this experiment are available at: https://osf.io/7nvsz/ and this experiment was not preregistered.

Author note We have no conflicts of interest to disclose. We gratefully acknowledge funding from the Australian Government Research Training Program Scholarship provided by the Australian Commonwealth Government and the University of Melbourne. The data and materials for this experiment are publicly available at: https://osf.io/7nvsz/

Open Access This article is licensed under a Creative Commons Attribution 4.0 International License, which permits use, sharing, adaptation, distribution and reproduction in any medium or format, as long as you give appropriate credit to the original author(s) and the source, provide a link to the Creative Commons licence, and indicate if changes were made. The images or other third party material in this article are included in the article's Creative Commons licence, unless indicated otherwise in a credit line to the material. If material is not included in the article's Creative Commons licence and your intended use is not permitted by statutory regulation or exceeds the permitted use, you will need to obtain permission directly from the copyright holder. To view a copy of this licence, visit http://creativecommons.org/licenses/by/4.0/.

References

Adams, R., Stephan, K., Brown, H., Frith, C., & Friston, K. (2013). The computational anatomy of psychosis. *Frontiers in Psychiatry, 4* https://www.frontiersin.org/article/10.3389/fpsyt.2013.00047

Andermane, N., Bosten, J. M., Seth, A. K., & Ward, J. (2020). Individual differences in the tendency to see the expected. *Consciousness and Cognition, 85*, 102989. https://doi.org/10.1016/j.concog.2020.102989

Bedder, R. L., Vaghi, M. M., Dolan, R. J., & Rutledge, R. B. (2023). Risk taking for potential losses but not gains increases with time of day. *Scientific Reports, 13*(1), 5534. https://doi.org/10.1038/s41598-023-31738-x

Brown, V. M., Chen, J., Gillan, C. M., & Price, R. B. (2020). Improving the reliability of computational analyses: Model-based planning and its relationship with compulsivity. *Biological Psychiatry: Cognitive Neuroscience and Neuroimaging, 5*(6), 601–609. https://doi.org/10.1016/j.bpsc.2019.12.019

Deserno, L., Boehme, R., Mathys, C., Katthagen, T., Kaminski, J., Stephan, K. E., Heinz, A., & Schlagenhauf, F. (2020). Volatility estimates increase choice switching and relate to prefrontal activity in schizophrenia. *Biological Psychiatry: Cognitive Neuroscience and Neuroimaging, 5*(2), 173–183. https://doi.org/10.1016/j.bpsc.2019.10.007

Eldar, E., Roth, C., Dayan, P., & Dolan, R. J. (2018). Decodability of reward learning signals predicts mood fluctuations. *Current Biology, 28*(9), 1433–1439.e7. https://doi.org/10.1016/j.cub.2018.03.038

Friston, K. (2005). A theory of cortical responses. *Philosophical Transactions of the Royal Society B: Biological Sciences, 360*(1456), 815–836. https://doi.org/10.1098/rstb.2005.1622

Friston, K. (2008). Hierarchical models in the brain. *PLOS Computational Biology, 4*(11), e1000211. https://doi.org/10.1371/journal.pcbi.1000211

Fromm, S., Katthagen, T., Deserno, L., Heinz, A., Kaminski, J., & Schlagenhauf, F. (2023). Belief updating in subclinical and clinical delusions. *Schizophrenia Bulletin Open, 4*(1), sgac074. https://doi.org/10.1093/schizbullopen/sgac074

Gibbs-Dean, T., Katthagen, T., Tsenkova, I., Ali, R., Liang, X., Spencer, T., & Diederen, K. (2023). Belief updating in psychosis, depression and anxiety disorders: A systematic review across computational modelling approaches. *Neuroscience & Biobehavioral Reviews, 147*, 105087. https://doi.org/10.1016/j.neubiorev.2023.105087

Goodwin, I., Kugel, J., Hester, R., & Garrido, M. I. (2022). Bayesian accounts of perceptual decisions in the nonclinical continuum of psychosis: Greater imprecision in both top-down and bottom-up processes. bioRxiv. https://doi.org/10.1101/2022.10.24.513606

Hauke, D. J., Charlton, C. E., Schmidt, A., Griffiths, J., Woods, S. W., Ford, J. M., Srihari, V. H., Roth, V., Diaconescu, A. O., & Mathalon, D. H. (2023). *Aberrant hierarchical prediction errors are associated with transition to psychosis: A computational single-trial analysis of the mismatch negativity*. *Cognitive Neuroscience and Neuroimaging*. https://doi.org/10.1016/j.bpsc.2023.07.011

Hohwy, J. (2013). *The Predictive Mind* (online ed.). Oxford Academic. https://doi.org/10.1093/acprof:oso/9780199682737.001.0001

Hohwy, J. (2020). New directions in predictive processing. *Mind & Language, 35*(2), 209–223. https://doi.org/10.1111/mila.12281

Karvelis, P., Paulus, M. P., & Diaconescu, A. O. (2023). Individual differences in computational psychiatry: A review of current challenges. *Neuroscience & Biobehavioral Reviews, 148*, 105137. https://doi.org/10.1016/j.neubiorev.2023.105137

Knill, D. C., & Pouget, A. (2004). The Bayesian brain: The role of uncertainty in neural coding and computation. *Trends in Neurosciences, 27*(12), 712–719. https://doi.org/10.1016/j.tins.2004.10.007

Koblinger, Á., Fiser, J., & Lengyel, M. (2021). Representations of uncertainty: Where art thou? *Current Opinion in Behavioral Sciences, 38*, 150–162. https://doi.org/10.1016/j.cobeha.2021.03.009

Koo, T. K., & Li, M. Y. (2016). A Guideline of Selecting and Reporting Intraclass Correlation Coefficients for Reliability Research. *Journal of Chiropractic Medicine, 15*(2), 155–163. https://doi.org/10.1016/j.jcm.2016.02.012

Körding, K. P., & Wolpert, D. M. (2004). Bayesian integration in sensorimotor learning. *Nature, 427*(6971), 244–247. https://doi.org/10.1038/nature02169

Kraus, N., Niedeggen, M., & Hesselmann, G. (2021). Trait anxiety is linked to increased usage of priors in a perceptual decision making task. *Cognition, 206*, 104474. https://doi.org/10.1016/j.cognition.2020.104474

Kreis, I., Biegler, R., Tjelmeland, H., Mittner, M., Reitan, S. K., & Pfuhl, G. (2021). Overestimation of volatility in schizophrenia and autism? A comparative study using a probabilistic reasoning task. *PLOS ONE, 16*(1), e0244975. https://doi.org/10.1371/journal.pone.0244975

Kreis, I., Zhang, L., Mittner, M., Syla, L., Lamm, C., & Pfuhl, G. (2023). Aberrant uncertainty processing is linked to psychotic-like experiences, autistic traits, and is reflected in pupil dilation during probabilistic learning. *Cognitive, Affective, & Behavioral Neuroscience, 23*(3), 905–919. https://doi.org/10.3758/s13415-023-01088-2

Lyndon, S., & Corlett, P. R. (2020). Hallucinations in posttraumatic stress disorder: Insights from predictive coding. *Journal of Abnormal Psychology, 129*, 534–543. https://doi.org/10.1037/abn0000531

Pálffy, Z., Farkas, K., Csukly, G., Kéri, S., & Polner, B. (2021). Cross-modal auditory priors drive the perception of bistable visual stimuli with reliable differences between individuals. *Scientific Reports, 11*(1), 16943. https://doi.org/10.1038/s41598-021-96198-7

Palmer, C. J., Lawson, R. P., & Hohwy, J. (2017). Bayesian approaches to autism: Towards volatility, action, and behavior. *Psychological Bulletin, 143*(5), 521–542. https://doi.org/10.1037/bul0000097

Palmer, C. J., Seth, A. K., & Hohwy, J. (2015). The felt presence of other minds: Predictive processing, counterfactual predictions, and mentalising in autism. *Consciousness and Cognition, 36*, 376–389. https://doi.org/10.1016/j.concog.2015.04.007

Parsons, S. (2021). splithalf: Robust estimates of split half reliability. *Journal of Open Source Software, 6*(60), 3041. 10.21105/joss.03041.

Parsons, S., Kruijt, A.-W., & Fox, E. (2019). Psychological science needs a standard practice of reporting the reliability of cognitive-behavioral measurements. *Advances in Methods and Practices in Psychological Science, 2*, 378–395. https://doi.org/10.1177/2515245919879695

Pott, J., & Schilbach, L. (2022). Tracking and changing beliefs during social interaction: Where computational psychiatry meets cognitive behavioral therapy. *Frontiers in Psychology, 13* https://www.frontiersin.org/articles/10.3389/fpsyg.2022.1010012

Putica, A., Felmingham, K. L., Garrido, M. I., O'Donnell, M. L., & Van Dam, N. T. (2022). A predictive coding account of value-based learning in PTSD: Implications for precision treatments. *Neuroscience & Biobehavioral Reviews, 138*, 104704. https://doi.org/10.1016/j.neubiorev.2022.104704

Randeniya, R., Vilares, I., Mattingley, J. B., & Garrido, M. I. (2021). Reduced context updating but intact visual priors in autism. *Computational Psychiatry, 5*(1), 140–158. https://doi.org/10.5334/cpsy.69

Rouder, J., Kumar, A., & Haaf, J. M. (2019). Why most studies of individual differences with inhibition tasks are bound to fail. *PsyArXiv.* 10.31234/osf.io/3cjr5.

Rouder, J. N., & Haaf, J. M. (2019). A psychometrics of individual differences in experimental tasks. *Psychonomic Bulletin & Review, 26*(2), 452–467. https://doi.org/10.3758/s13423-018-1558-y

Schlagenhauf, F., Huys, Q. J. M., Deserno, L., Rapp, M. A., Beck, A., Heinze, H.-J., Dolan, R., & Heinz, A. (2014). Striatal dysfunction during reversal learning in unmedicated schizophrenia patients. *NeuroImage, 89*, 171–180. https://doi.org/10.1016/j.neuroimage.2013.11.034

Sterzer, P., Adams, R. A., Fletcher, P., Frith, C., Lawrie, S. M., Muckli, L., Petrovic, P., Uhlhaas, P., Voss, M., & Corlett, P. R. (2018). The predictive coding account of psychosis. *Biological Psychiatry, 84*(9), 634–643. https://doi.org/10.1016/j.biopsych.2018.05.015

Sullivan-Toole, H., Haines, N., Dale, K., & Olino, T. M. (2022). Enhancing the psychometric properties of the iowa gambling task using full generative modeling. *Computational Psychiatry, 6*(1), 189–212. https://doi.org/10.5334/cpsy.89

Trapp, S., & Vilares, I. (2020). Bayesian decision-making under stress-preserved weighting of prior and likelihood information. *Scientific Reports, 10*. https://doi.org/10.1038/s41598-020-76493-5

Tulver, K., Aru, J., Rutiku, R., & Bachmann, T. (2019). Individual differences in the effects of priors on perception: A multi-paradigm approach. *Cognition, 187*, 167–177. https://doi.org/10.1016/j.cognition.2019.03.008

van Leeuwen, T. M., Sauer, A., Jurjut, A.-M., Wibral, M., Uhlhaas, P. J., Singer, W., & Melloni, L. (2021). perceptual gains and losses in synesthesia and schizophrenia. *Schizophrenia Bulletin, 47*(3), 722–730. https://doi.org/10.1093/schbul/sbaa162

Vilares, I., Howard, J. D., Fernandes, H. L., Gottfried, J. A., & Kording, K. P. (2012). Differential representations of prior and likelihood uncertainty in the human brain. *Current Biology: CB, 22*(18), 1641–1648. https://doi.org/10.1016/j.cub.2012.07.010

Vilares, I., & Kording, K. P. (2017). Dopaminergic medication increases reliance on current information in Parkinson's disease. *Nature Human Behaviour, 1*(8), 1–7. https://doi.org/10.1038/s41562-017-0129

Waltmann, M., Schlagenhauf, F., & Deserno, L. (2022). Sufficient reliability of the behavioral and computational readouts of a probabilistic reversal learning task. *Behavior Research Methods, 54*(6), 2993–3014. https://doi.org/10.3758/s13428-021-01739-7

Weilnhammer, V., Röd, L., Eckert, A.-L., Stuke, H., Heinz, A., & Sterzer, P. (2020). Psychotic Experiences in Schizophrenia and Sensitivity to Sensory Evidence. *Schizophrenia Bulletin, 46*(4), 927–936. https://doi.org/10.1093/schbul/sbaa003

Williams, D. R., Martin, S. R., & Rast, P. (2022). Putting the individual into reliability: Bayesian testing of homogeneous within-person variance in hierarchical models. *Behavior Research Methods, 54*(3), 1272–1290. https://doi.org/10.3758/s13428-021-01646-x

Zech, H., Waltmann, M., Lee, Y., Reichert, M., Bedder, R. L., Rutledge, R. B., Deeken, F., Wenzel, J., Wedemeyer, F., Aguilera, A., Aslan, A., Bach, P., Bahr, N. S., Ebrahimi, C., Fischbach, P. C., Ganz, M., Garbusow, M., Großkopf, C. M., Heigert, M., et al. (2022). Measuring self-regulation in everyday life: Reliability and validity of smartphone-based experiments in alcohol use disorder. *Behavior Research Methods.* https://doi.org/10.3758/s13428-022-02019-8

Publisher's note Springer Nature remains neutral with regard to jurisdictional claims in published maps and institutional affiliations.