Test-retest reliability of behavioral and computational measures of advice taking under volatility

Povilas Karvelis[*,1], Daniel J. Hauke[2], Michelle Wobmann[3], Christina Andreou[4], Amatya Mackintosh[3], Renate de Bock[3], Stefan Borgwardt[4], and Andreea O. Diaconescu[1,5,6,7]

[1]Krembil Centre for Neuroinformatics, Centre for Addiction and Mental Health (CAMH), Toronto, ON, Canada
[2]Centre for Medical Image Computing, Department of Computer Science, University College London, London, United Kingdom
[3]Department of Psychiatry (UPK), University of Basel, Basel, Switzerland
[4]Department of Psychiatry and Psychotherapy, Translational Psychiatry, University of Lubeck, Lubeck, Germany
[5]Department of Psychiatry, University of Toronto, Toronto, ON, Canada
[6]Institute of Medical Sciences, University of Toronto, Toronto, ON, Canada
[7]Department of Psychology, University of Toronto, Toronto, ON, Canada
[*]corresponding author: povilas.karvelis@camh.ca

Abstract

The development of computational models for studying mental disorders is on the rise. However, their psychometric properties remain understudied, posing a risk to undermine their use in empirical research and clinical translation. Here we investigated test-retest reliability (with a 2-week interval) of a computational assay probing advice-taking under volatility with a Hierarchical Gaussian Filter (HGF) model. In a sample of 39 healthy participants, we found the computational measures to have largely poor reliability (intra-class correlation coefficient or $ICC < 0.5$), on par with the behavioral measures of task performance. Further analysis revealed that reliability was substantially impacted by intrinsic measurement noise (indicated by parameter recovery analysis) and to a smaller extent by practice effects. However, a large portion of within-subject variance remained unexplained and may be attributable to state-like fluctuations. Despite the poor test-retest reliability, we found the assay to have face validity at the group level. Overall, our work highlights that the different sources of variance affecting test-retest reliability need to be studied in greater detail. A better understanding of these sources would facilitate the design of more psychometrically sound assays, which would improve the quality of future research and increase the probability of clinical translation.

Keywords: computational psychiatry, test-retest reliability, Hierarchical Gaussian Filter, advice-taking, individual differences

1 Introduction

Computational modelling is a powerful tool for studying complex systems such as the brain. Combining computational models with cognitive tasks allows one to measure various latent neurocomputational variables (e.g., learning rate, reward sensitivity, sensory precision, decision noise) that would otherwise be unobservable. There is plenty of enthusiasm that such computational measures may become a new basis for psychiatric nosology and treatment personalization (Friston, 2023; Yip et al., 2022; Huys et al., 2021; Hauser et al., 2022; Patzelt et al., 2018; Karvelis et al., 2022; Karvelis and Diaconescu, 2022).

However, for the computational measures to be sufficiently informative at the individual level - for studying individual differences in the lab (e.g., correlational analyses with symptoms) or for personalized care in the clinic (e.g., treatment response prediction) - they must have excellent psychometric

properties (reliability and construct validity). Up until recently, the psychometric properties of computational measures have received little attention (Browning et al., 2020). Emerging empirical evidence indicates that both reliability and construct validity are often poor (Karvelis et al., 2023), but the psychometric properties of many currently used computational assays remain unknown and require further investigation.

Here we investigate test-retest reliability of a computational assay probing advice-taking under volatility. The assay is comprised of a social advice-taking task and a Hierarchical Gaussian Filter (HGF) model (Diaconescu et al., 2014). The task has been designed to probe how people infer the intentions of others (helpful vs. unhelpful) in the context where the intentions themselves are changing over time. This allows studying the dynamics of how people learn to trust or distrust others. Combining the task with the HGF model further helps to infer various cognitive variables of interest, in particular as it relates to participants' estimation of how volatile the adviser is and whether they are likely to be helpful or unhelpful at any given moment. Variants of this assay have been used to study multiple disorders that exhibit difficulties in correctly inferring the intentions of others, including schizophrenia (Hauke et al., 2024; Henco et al., 2020), autism (Sevgi et al., 2020), and borderline personality disorder (Henco et al., 2020), as well as aging (Reiter et al., 2021), but the reliability of the measures afforded by the assay has not yet been studied.

A more general aim of this paper is to examine the different sources of variance that underlie poor test-retest reliability in more detail - something that is rarely done when test-retest reliability of computational measures is assessed (Karvelis et al., 2023). Understanding what drives low test-retest reliability is the first step towards designing more psychometrically sound assays.

2 Methods

2.1 Participants

A total of 43 participants were recruited from the general population via online advertisements and advertisements in public places in Basel, Switzerland. Four participants completed only a single task session, leaving 39 participants for the analysis presented in this paper. The remaining sample included 20 female and 19 male participants, with an average age of 28.85 (SD = 4.03). All participants provided informed written consent. The study was approved by the local ethics committee (Ethikkommission Nordwest- und Zentralschweiz, no. 2017-01149) and conducted in accordance with the latest version of the Declaration of Helsinki. All data collection was done between November 1st 2017 and July 31st 2022.

2.2 Study design

The data analyzed here was collected as part of a larger study investigating first-episode psychosis patients and individuals at clinical high-risk for psychosis (Hauke et al., 2024). Each participant completed the cognitive task twice, 2 weeks apart. One of the sessions was combined with fMRI and one with EEG measurements (the order of fMRI and EEG was randomized across participants). In this paper, however, we will focus solely on the behavioral data and only on the 39 healthy controls.

2.3 The task

Participants engaged in a probabilistic learning task (**Fig 1**) where they had to infer on the changing intentions of an adviser and decide how much to rely on their advice (for previous studies using this task see Diaconescu et al., 2014, 2017; Hauke et al., 2024). The task had two phases: a stable phase with consistent helpful advice and a volatile phase where the adviser's intentions changed rapidly (**Fig 1B**). On each trial, participants predicted the outcome of a binary lottery using information from a non-social cue (displaying true winning probabilities) and a prerecorded video recommendation from

a human adviser who had previously attempted to help or deceive a player in a human-human interaction. The participants were informed that the adviser had privileged information about the outcome, but not complete information, and that inaccurate advice could be due to mistakes of the adviser or the adviser having a different incentive structure that could motivate them to provide misleading advice during the experiment. In total, there were 153 trials.

In addition to making choices on every trial, participants were also prompted to explicitly report how helpful the adviser was at multiple times throughout the task. They were asked "How do you feel about the advice you are receiving right now?" on trials 14, 49, 73, 99, and 134, and responded by choosing among "helpful", "random", and "misleading". To probe their prior expectations, before starting the task (on trial 1) participants were also asked "How do you think the advice you are about to receive will be?" and had to respond by choosing among the same three options.

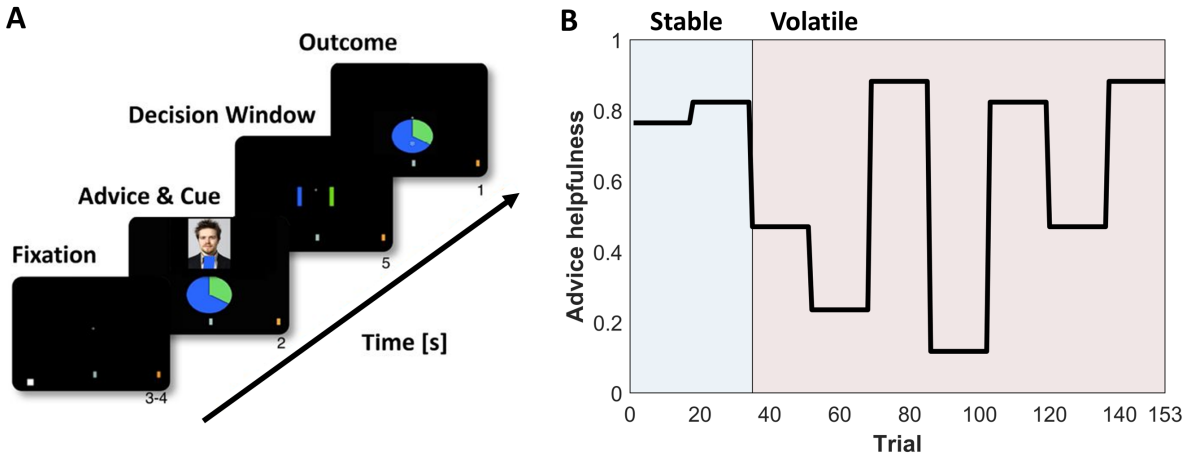


Figure 1: **The task.** (**A**) The trial structure: On each trial, participants are presented with two pieces of information: a social cue (a recording of an adviser holding up one of the colors), a non-social cue (a pie chart) indicating true winning probabilities of the two options. (**B**) Probability of adviser's fidelity (helpfulness) across the task.

2.4 Behavioral measures

For the behavioral analysis we computed standard behavioral measures. This included: 1) Total accuracy, which was defined as the fraction of trials in which participants chose the option with the higher reward probability. This measure captures the overall level of performance. 2) Advice taking probability, which was defined as the fraction of trials in which participants acted in accordance with the social cue. This measure captures the tendency to rely on advice, and we computed it separately for stable and volatile phases of the task, as well as for the whole task. 3) Win-stay probability, which was defined as the fraction of trials in which participants kept the same strategy (trusting the advice or not) following a winning trial. 4) Lose-switch probability, which was defined as the fraction of trials in which participants switched their strategy (trusting the advice or not) following a losing trial. Both win-stay and lose-switch measures capture how quickly participants learn from positive and negative feedback, respectively. However, it is a rather rough proxy for the learning rate, because we expect it to dynamically change depending on how volatile the adviser's strategy is perceived to be at any given point in the task – to capture these dynamics more explicitly we next turn to computational modelling.

2.5 Computational modelling

2.5.1 Hierarchical Gaussian Filter

To model participants' behavior in the task we used the same analysis pipeline as in the previous work (Hauke et al., 2024), using a 3-level Hierarchical Gaussian Filter (HGF) (Mathys et al., 2011,

2014). This modelling framework assumes that participants infer a hierarchy of hidden states (x_1, x_2, and x_3) that characterize outcome probabilities and their temporal dynamics. The participants' beliefs about the hidden states at each level i on trial k are denoted as $\mu_i^{(k)}$. In this task, the states that participants need to infer based on the inputs are structured as follows: the lowest level state represents the accuracy of the advice, which can be either accurate ($x_1^{(k)} = 1$) or inaccurate ($x_1^{(k)} = 0$). The relationship between this state and the second level state ($x_2^{(k)}$), representing the adviser's fidelity or tendency towards helpful advice, is described by a Bernoulli distribution linked to the level above with a unit sigmoid transformation:

$$p(x_1^{(k)}|x_2^{(k)}) = s(x_2^{(k)})^{x_1^{(k)}}(1 - s(x_2^{(k)}))^{1-x_1^{(k)}} \sim \text{Bernoulli}(x_1^{(k)}; s(x_2^{(k)})), \tag{1}$$

with s being a sigmoid function:

$$s(z) = \frac{1}{1 + e^{-z}}. \tag{2}$$

The second level state ($x_2^{(k)}$) is modeled as a normal distribution with a mean ($x_2^{(k-1)}$) and variance determined by the volatility of the adviser's intentions ($x_3^{(k)}$):

$$p(x_2^{(k)}|x_2^{(k-1)}, x_3^{(k)}, \kappa_2, \omega_2) \sim \mathcal{N}(x_2^{(k)}; x_2^{(k-1)}, \exp(\kappa_2 x_3^{(k)} + \omega_2)) \tag{3}$$

where ω_2 is the evolution rate that shapes tonic learning, and κ_2 is the coupling strength between the second and third levels shaping phasic learning. The third level state ($x_3^{(k)}$) represents the volatility of the adviser's intentions and is also modeled as a normal distribution:

$$p(x_3^{(k)}|x_3^{(k-1)}, \vartheta) \sim \mathcal{N}(x_3^{(k)}; x_3^{(k-1)}, \vartheta) \tag{4}$$

where ϑ is meta-volatility at the third level.

Following previous work (Hauke et al., 2024; Cole et al., 2020), this part of the model was augmented with an additional tendency to revert back to equilibrium beliefs about environmental volatility, akin to an Ornstein-Uhlenbeck process in discrete time (Uhlenbeck and Ornstein, 1930):

$$p(x_3^{(k)}|x_3^{(k-1)}, \vartheta, \phi_3, m_3) \sim \mathcal{N}(x_3^{(k)}; x_3^{(k-1)} + \phi_3(m_3 - x_3^{(k-1)}), \vartheta), \tag{5}$$

where ϕ_3 is the drift rate and m_3 is the equilibrium point towards which the state moves over time.

After each observation, the model updates inferences about each of the states. The HGF update equations can be derived using a variational approximation, where the belief update at each level depends on the precision from the level below and the prediction error:

$$\Delta\mu_i^{(k)} \propto \frac{\hat{\pi}_{i-1}^{(k)}}{\pi_i^{(k)}}\delta_{i-1}^{(k)}, \tag{6}$$

where $\mu_i^{(k)}$ is the expectation or the belief at trial k and level i of the hierarchy, $\hat{\pi}_{i-1}^{(k)}$ is the precision (inverse of the variance) from the level below, $\pi_i^{(k)}$ is the updated precision at the current level, and $\delta_{i-1}^{(k)}$ is the prediction error from the level below (the difference between the expected and the observed outcome).

Inferred states about the world are translated into decisions (i.e., to go with or against the advice) via the response model:

$$b^{(k)} = \zeta\hat{\mu}_1^{(k)} + (1-\zeta)c^{(k)}, \tag{7}$$

where ζ captures the relative weighing of the advice compared to the non-social cue, $c^{(k)}$ is the outcome probability indicated by the pie chart, transformed to reflect the colour probability indicated by the adviser, while $\hat{\mu}_1^{(k)}$ is the belief that the adviser is providing accurate advice.

The probability that a participant follows the advice ($y = 1$) can then be described by a sigmoid transformation of the integrated belief b:

$$p(y=1|b) = \frac{b^\beta}{b^\beta + (1-b)^\beta}, \tag{8}$$

with

$$\beta = \exp(-\hat{\mu}_3^{(k)} + \nu). \tag{9}$$

where $\hat{\mu}_3^{(k)}$ is the current estimate of the volatility of the advisers' intentions, and ν is decision noise that is independent of the perception of volatility and is constant across trials.

In summary, the HGF models participants' behavior in the social learning task by inferring hidden states and updating them based on the outcome of each trial. The mean-reverting HGF additionally incorporates a drift in volatility estimates towards a subject-specific equilibrium (**Fig 2**). The response model integrates beliefs and cues to make decisions, with the noise level determined by the perception of volatility.

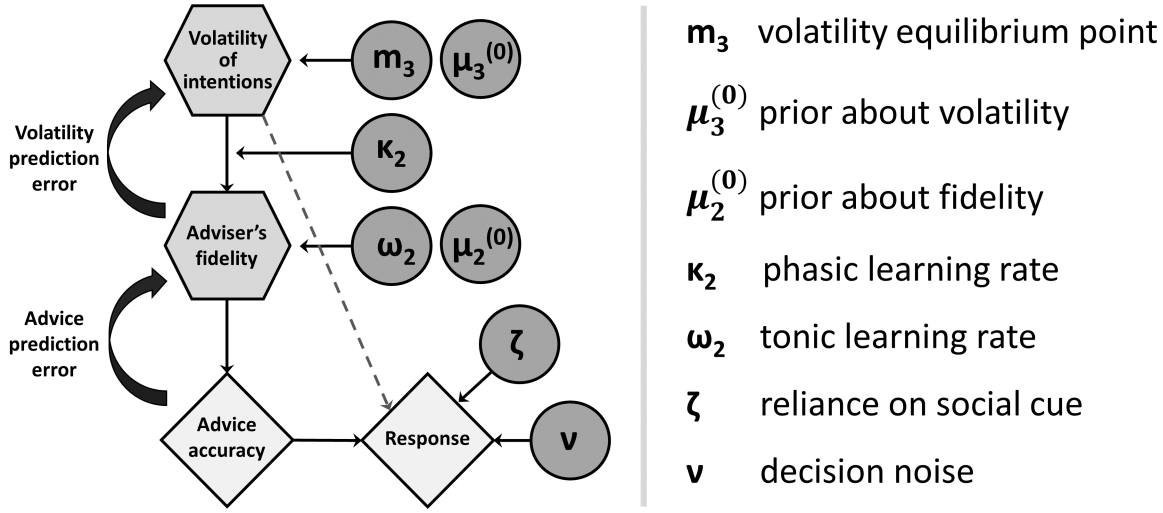


Figure 2: **Graphical representation of the mean-reverting 3-level HGF.** The highlighted parameters are free parameters. m_3 is volatility equilibrium point towards which the agent slowly drifts irrespective of empirical observations. μ_3 and μ_2 are the priors about advice volatility and fidelity, respectively, that the agent has before starting the task. κ_2 is the phasic learning rate which regulates how much the inferences about the volatility (at the third level) modulate learning about adviser's fidelity (at the second level). ω_2 is the baseline leaning rate (evolution rate), which regulates the learning rate at the second level independent from the estimates at the third level. ζ parameter captures how much participants rely on the social cue relative to the non-social cue. Finally, ν captures intrinsic noise at the stage of decision-making that is unrelated to inferred volatility of the adviser.

2.5.2 Model space

In total, we have tested 2 variants of the HGF: standard HGF, (**Eq. 4**) (model 1), and HGF with the mean-reverting process at the third level, (**Eq. 5**) (model 2). For each of these variants, we included an additional control model (model 1b and model 2b), in which all parameters of the perceptual model were fixed to parameter values of an ideal Bayesian observer optimised based on the inputs alone.

The models were implemented in Matlab R2020a using HGF 3.0 open-source toolbox included in Translational Algorithms for Psychiatry-Advancing Science (TAPAS) (Frässle et al., 2021) software collection `https://github.com/translationalneuromodeling/tapas/releases/tag/v3.0.0`). The code implementing the models can be found in 'tapas_hgf_binary.m' function for model 1 and 'tapas_hgf_ar1_binary.m function for model 2.

2.5.3 Model fitting

Following the modelling pipeline in Hauke et al. (2024), meta-volatility ϑ was fixed to 0.5, and the drift rate parameter ϕ_3 was fixed to 0.1, which was done to reduce the number of free parameters to help with model convergence. The remaining parameters were free, including: the evolution rate at the second level ω_2, the coupling strength between the second and third level κ_2, prior expectations before seeing any input about the adviser's fidelity $\mu_2^{(0)}$ and the volatility of the adviser's intentions $\mu_3^{(0)}$, the relative weighing of the advice compared to the non-social cue ζ, decision noise ν, and volatility equilibrium m_3.

Model fitting (i.e., finding parameter values with the highest posterior probability) was done using quasi-Newton optimization algorithm. The priors for all model parameters were based on previous work, the details of which can be found in Table 1 in Hauke et al. (2024). We briefly summarize them here in **Table 1**.

	κ_2	ω_2	$\mu_2^{(0)}$	$\mu_3^{(0)}$	ζ	ν	m_3
model 1	$(0,1),1$	$(-2,4)$	$(0,1)$	$(1,1)$	$(0,1),1$	$(\log(48),1)$	
model 2	$(0,1),1$	$(-2,4)$	$(0,1)$	$(1,1)$	$(0,1),1$	$(\log(48),1)$	$(1,1)$

Table 1: **Priors on free model parameters**. Prior means and their respective variances are denoted in brackets, and for some parameters followed by an upper bound: (mean, variance), upper bound. See **Fig 2** for a detailed explanation of each parameter.

2.5.4 Model selection

To arbitrate among the models, we performed random-effects Bayesian model selection (Rigoux et al., 2014; Stephan et al., 2009), using VBA toolbox (Daunizeau et al., 2014) (`https://mbb-team.github.io/VBA-toolbox/`). We present protected exceedance probabilities (PEP), which quantify the likelihood of a model being more plausible than any other model in the model space, while accounting for the possibility that model differences could arise by chance alone (Stephan et al., 2009; Rigoux et al., 2014). Additionally, we report relative model frequencies or f as an effect size measure, representing the probability that a randomly selected participant would be best accounted for by a specific model.

2.5.5 Parameter and model recovery

To better understand test-retest reliability results, we also performed parameter recovery. Parameter recovery involves generating synthetic data of task performance using a range of model parameter values and then fitting the same model to that data (Wilson and Collins, 2019; Palminteri et al., 2017). This procedure allows us to assess how uniquely different sets of parameter values map onto specific behavioral patterns, and vice versa. The similarity between the 'true' and the 'recovered' parameters indicates how reliable the estimates would be under ideal conditions: when task performance stays exactly the same across instances. As such, it is constrained only by the intrinsic measurement error

and can be considered to provide an upper bound on test-retest reliability, because the latter would be further affected by within-subject changes in task performance (Karvelis et al., 2023).

For simulating task data, we simply used the same parameter estimates obtained from empirical data. This means that we had 39 synthetic agents and each of them completed 153 task trials. Due to the stochasticity in decision making, we reran this procedure 20 times with different random seeds. We report parameter recovery results for one of the runs as well as averaged cross the 20 runs.

To make sure that our models are recoverable, we repeated this procedure for all other models, while also fitting each model to the data generated by the other models. We then performed random-effects Bayesian model selection (based on the free energy approximation of the Bayesian model evidence) on each of the datasets and averaged the resulting protected exceedance probabilities across the 20 runs to produce a confusion matrix to assess model recovery.

2.6 Statistical analysis

Test-retest reliability of the behavioral and computational measures was assessed using intra-class correlation coefficients (ICCs). More specifically, we computed ICCs(A,1): two-way mixed, single-measure, absolute agreement ICC (McGraw and Wong, 1996; Qin et al., 2019). Note that in the Fleiss (2011) notation, this would be denoted as ICC(2,1). Calculations were performed using *icc* function from *irr* package (version 0.84.1) in R (version 4.2.2). For referring to different ranges of ICC values, we chose Koo and Li (2016) labelling guidelines: poor < 0.5, moderate 0.5–0.75, good 0.75–0.9, and excellent > 0.9. Although note that these labels are somewhat arbitrary (Karvelis et al., 2023; Weir, 2005).

Alongside ICCs, we also report Pearson's correlation coefficients in order to highlight when and how much these metrics diverge. Pearson's r is not sensitive to offset and scaling errors, which makes it a less suitable metric for assessing reliability (Karvelis et al., 2023). However, as we will see, in practice this might be rarely a problem. Furthermore, we will use the comparison between Pearson's r and ICC to gauge the impact of practice effects on test-retest reliability.

Practice effects were investigated using paired t-tests, comparing the obtained measures between the two time points. Note that this addresses only fixed effects (e.g., people getting better at the task or less engaged with the task on average), but does not account for potential idiosyncratic practice effects. Given the limitations of frequentist statistics, wherever appropriate we computed Bayesian Factors (Kass and Raftery, 1995) to test the strength of evidence for the null hypothesis (BF_{01}).

To make our findings more robust we performed the above tests by excluding extreme outliers, which were based on the conventional Tukey's definition: as being below $Q1 - 3 * IQR$ or above $Q3 + 3 * IQR$.

3 Results

3.1 Test-retest reliability of behavioral measures

First, we analyzed test-retest reliability of standard behavioral measures (**Fig 3**). We found most measures to be in the poor reliability range ($ICC < 0.5$), except for accuracy ($ICC = 0.61$) and win-stay ($ICC = 0.52$). Comparing both absolute ICC and Pearson's correlation for each of the measures also revealed that both indices provide similar results (with the largest difference being 0.03).

3.2 Model recovery and Bayesian model selection

In line with previous work (Hauke et al., 2024), we found that models 1 and 2 are well recoverable (**Fig 1 in S1 Appendix**). We also found model 2 (mean-reverting HGF) to be the winning model (**Fig 2 in S1 Appendix**). This was consistent across the two task sessions. Protected exceedance

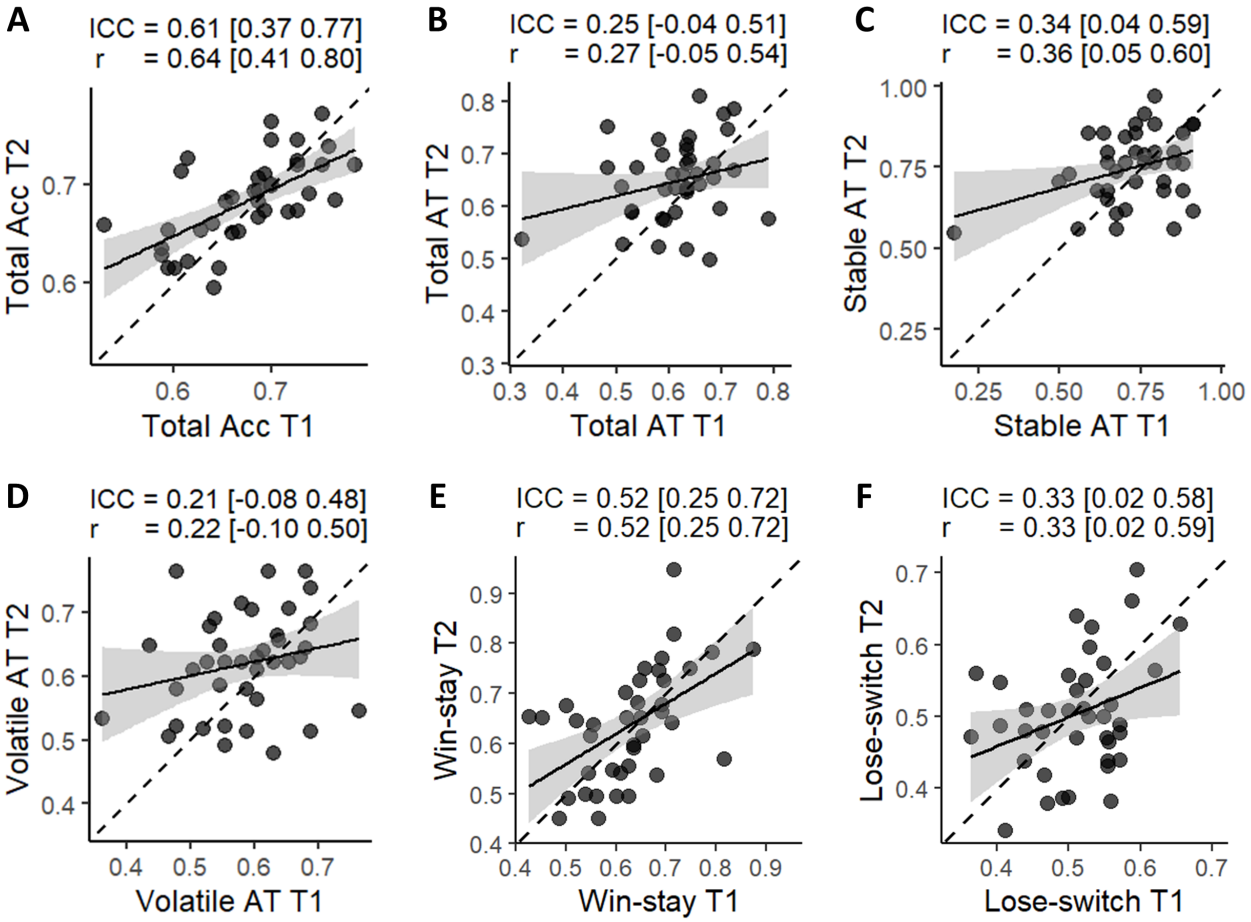


Figure 3: **Test-retest reliability of behavioral measures.** (**A**) Total accuracy. (**B**) Total advice taking. (**C**) Advice taking during the stable phase. (**D**) Advice taking during the volatile phase. (**E**) the probability of staying with the same advice taking strategy (trusting advice or not) after a win in a previous trial (**F**) the probability of changing the advice taking strategy (trusting advice or not) after a loss in a previous trial. ICC(A,1) and Pearson's correlation coefficients are shown above each panel. The square brackets indicate 95% confidence intervals.

probability for this model was ∼0.7 for session 1 and ∼1 for session 2. Note that < 0.8 might be considered inconclusive, but our results do not hinge on this because we found similar test-retest reliability for both models.

3.3 Test-retest reliability of computational measures

Next, we assessed test-retest reliability of all model parameter estimates of the winning model (**Fig 4**). Similar to the behavioral measures, we found most parameter estimates to be in the poor reliability range ($ICC < 0.5$), except for decision noise, which was slightly higher ($ICC = 0.56$). The lowest reliability was exhibited by κ_2 at $ICC = 0.04$. However, further analysis showed that removing data points with outsized effect on ICC as based on Cook's distance drastically improved the reliability of this parameter to $ICC = 0.40$ [0.09 0.64] (see **Fig 3 in S1 Appendix** for more details). Note, the results for in **Fig 4G** exclude one extreme outlier at that had a value of ~ 100 at T1 (see **Fig 3G in S1 Appendix** for the outlier).

3.4 Parameter recovery vs test-retest reliability

To investigate to what extent poor test-retest reliability of parameter estimates was due to intrinsic measurement error (as opposed to within-subject variability), we compared test-retest reliability re-

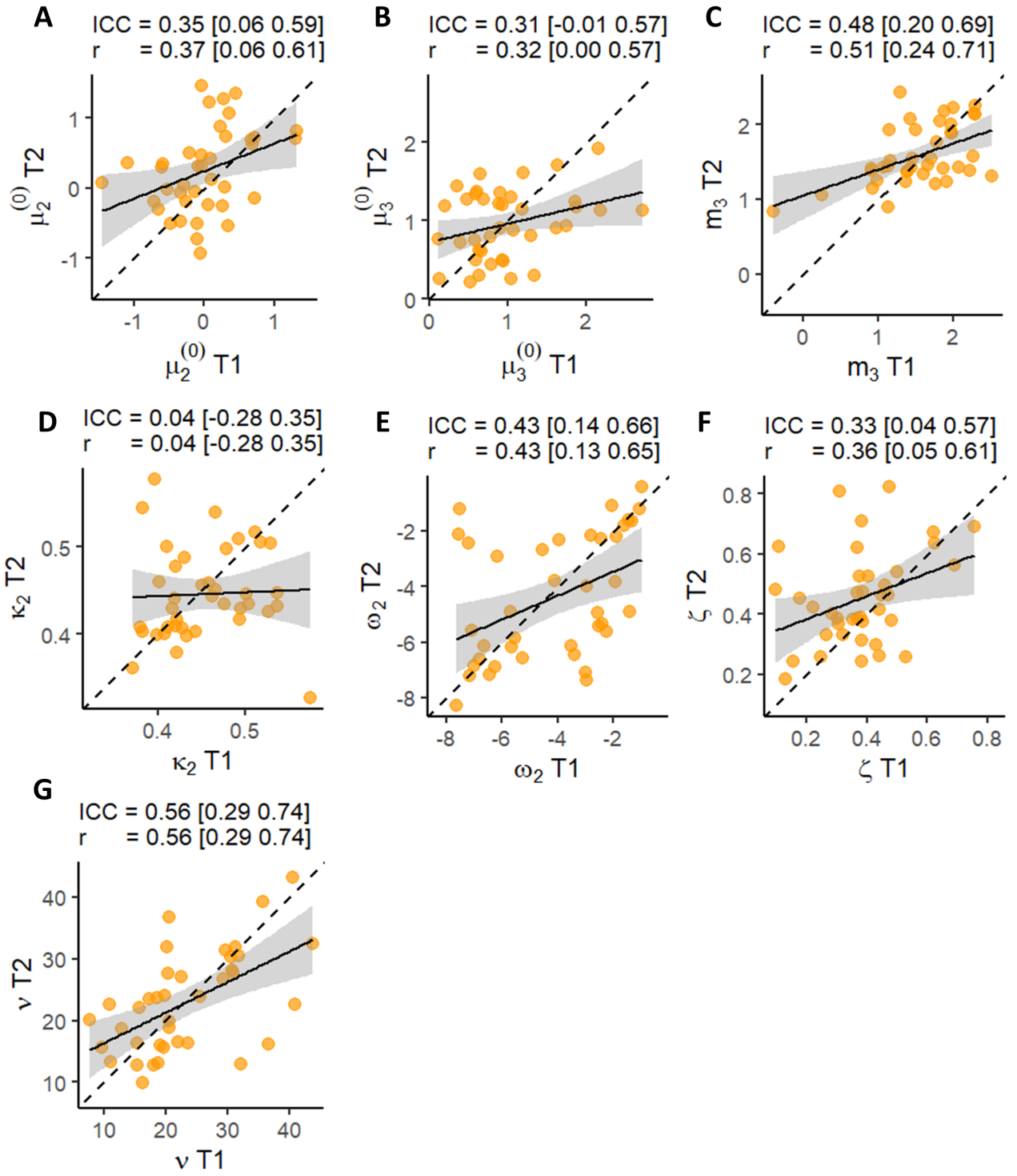


Figure 4: **Test-retest reliability of computational measures.** (**A**) Prior expectations about the adviser's fidelity before starting the task (**B**) Prior expectations about the adviser's volatility before starting the task. (**C**) Volatility equilibrium point (**D**) Phasic learning rate about the adviser's fidelity (**E**) Baseline learning rate about the adviser's fidelity (**F**). The relative weighing of the advice compared to the non-social cue. (**G**) Decision noise. ICC(A,1) and Pearson's correlation coefficients are shown above each panel. The square brackets indicate 95% confidence intervals.

sults to parameter recovery results. Parameter recovery was variable across the model parameters, varying from poor to excellent (**Fig 5**), with three of the model parameters being in the poor recoverability range $ICC \approx 0.5$, and four parameters being within moderate-to-excellent range ($ICC > 0.7$). The most recoverable parameter was ω_2 ($ICC = 0.96$), and the least recoverable parameter was κ_2 ($ICC = 0.44$).

Test-retest reliability was substantially lower than parameter recoverability for all parameters, but especially for ζ, ω_2, m_3, and κ_2 (**Fig 5H**). This indicates that for these parameters a major source of variance underlying low test-retest reliability was within-subject variability. These results were very similar when analyzing model 1 as well (**Fig 4 in S1 Appendix**).

We also investigated whether the low parameter recoverability for some parameters ($\mu_2^{(0)}$, $\mu_3^{(0)}$ and κ_2) was due to collinearity among parameters. To test this, we fixed these parameters to their prior values one by one and examined if it improved the recovery of the remaining parameters. We found that it did not, suggesting that collinearity was not the reason behind low recoverability (see **Fig 5 in S1 Appendix** for details).

3.5 Practice effects

Next, we investigated to what extent poor reliability of behavioral and computational measures was due to practice effects (i.e., participants getting better at the task with more practice). We used a paired t-test to test for differences between the two sessions, and we computed Bayesian factors to assess the null hypothesis.

Among the behavioral measures we found significant increases during the second session in advice-taking during the stable phase, close to significant increases in advice-taking during the volatile phase, significant increases in advice-taking overall, and increased win-stay probability (**Fig 6**). Notably, these changes in task performance did not result in improved performance accuracy ($BF_{01} = 5.6$).

Among the computational measures, only $\mu_2^{(0)}$ and ζ estimates were significantly different between the sessions (**Fig 7**).). Both of these parameters were found to be increased in the second session, indicating that participants expected the adviser to be more helpful before starting the second session ($\mu_2^{(0)}$), and tended to rely on the social cue more throughout the task (ζ) For all other parameter estimates, there was substantial evidence ($BF_{01} > 3$) for there being no practice effects.

It is interesting to note that while for both $\mu_2^{(0)}$ and ζ, practice effects were of moderate size (Cohen's $d = 0.37$ and $d = 0.42$, respectively), its contribution to low test-retest reliability of these parameters was minimal, reducing it only by 0.02 and 0.03, respectively, as indicated by the differences between the absolute agreement ICC and Pearson's correlation (**Fig 4**).

To make sure that different neuroimaging environments (EEG vs. fMRI) did not confound practice effects, or test-retest reliability results more generally, we compared parameter estimates between EEG and fMRI sessions. We found substantial evidence ($BF_{01} > 3$) that none of the computational measures were different between EEG and fMRI sessions (**Fig 6 in S1 Appendix**).

To further investigate how practice effects found in behavioral measures and parameter estimates relate to one another, we performed a correlational analysis (**Fig 8**). Changes in ζ were found to be associated with changes in all three behavioral measures, while changes in $\mu_2^{(0)}$ were associated only with changes in advice-taking during the stable phase. These results contribute to face validity of the model (see the following section for further results on this): $\mu_2^{(0)}$ is a prior of a dynamic variable that gets updated throughout the task and would be expected to have a detectable effect only at the beginning of the task, which is what we find (with the stable phase being the first 35 trials). ζ on the other hand, remains fixed throughout the task, and thus we see its effects on all three measures.

3.6 Face validity of the task and the model

Given the rather poor test-retest reliability results, we performed further analyses to assess to what extent the assay was measuring what it was intended to measure. First, we compared explicit reports of adviser's helpfulness with the actual helpfulness (**Fig 9A,C**), and found that participants consistently reported the adviser to be more helpful when the adviser was indeed more helpful. This

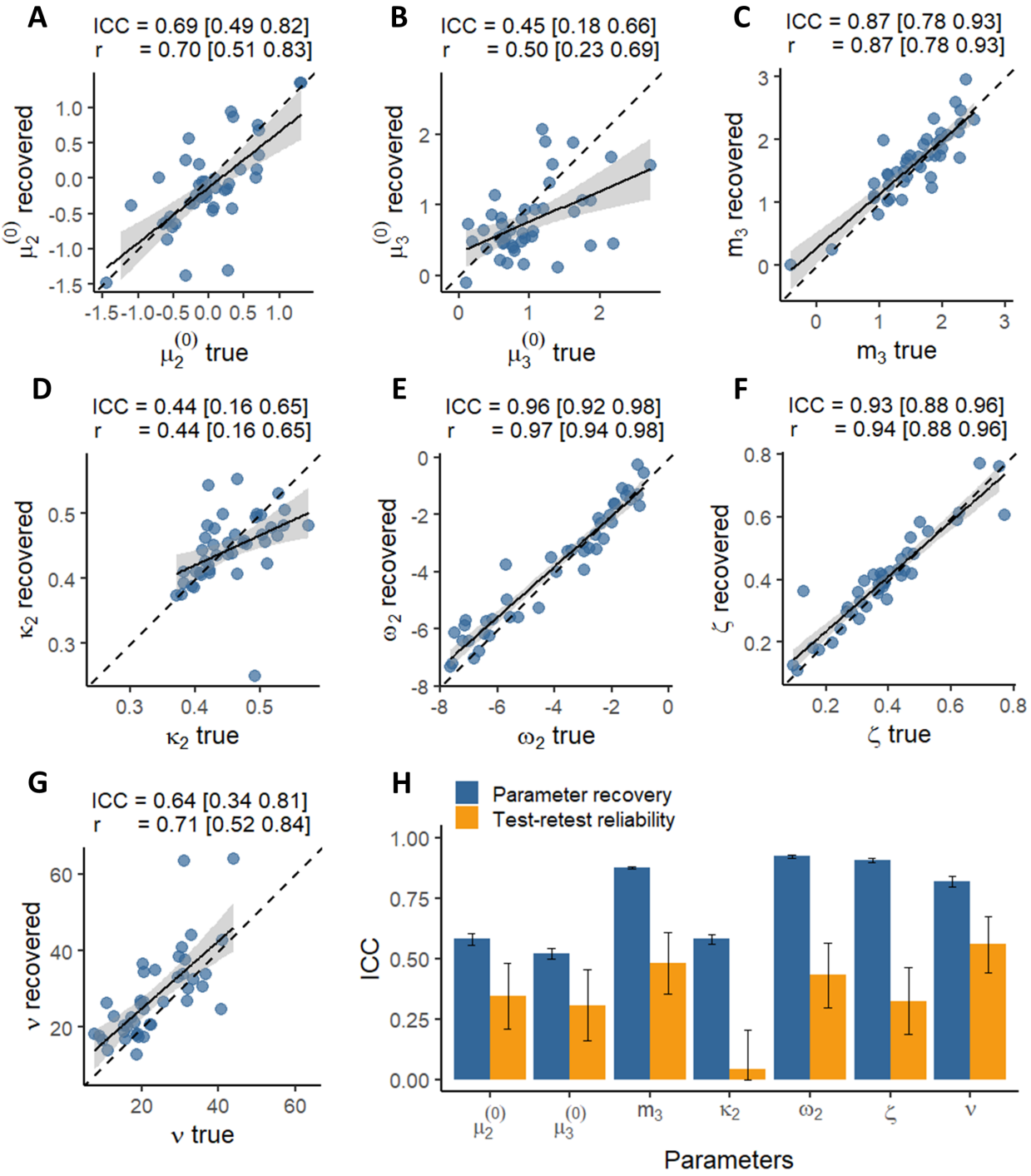


Figure 5: **Parameter recovery and test-retest reliability.** (**A**) Prior expectations about adviser's fidelity before starting the task (**B**) Prior expectations about the adviser's volatility before starting the task. (**C**) Volatility equilibrium point (**D**) Phasic learning rate about the adviser's fidelity (**E**) Baseline learning rate about the adviser's fidelity (**F**). The relative weighing of the advice compared to the non-social cue. (**G**) Decision noise. ICC(A,1) and Pearson's correlation coefficients are shown above each panel. The square brackets indicate 95% confidence intervals. (**H**) Comparison of parameter recovery vs. test-retest reliability for each model parameter. Note, (**A-G**) results are based on one particular random number generator seed (used to generate synthetic data), while (**H**) shows results averaged across 20 simulations with different seeds. The error bars denote standard error.

indicates face validity of the task - i.e., that the intended manipulation worked.

Next, we investigated whether model-derived estimates of inferred adviser's fidelity ($\hat{\mu}_2$) agreed

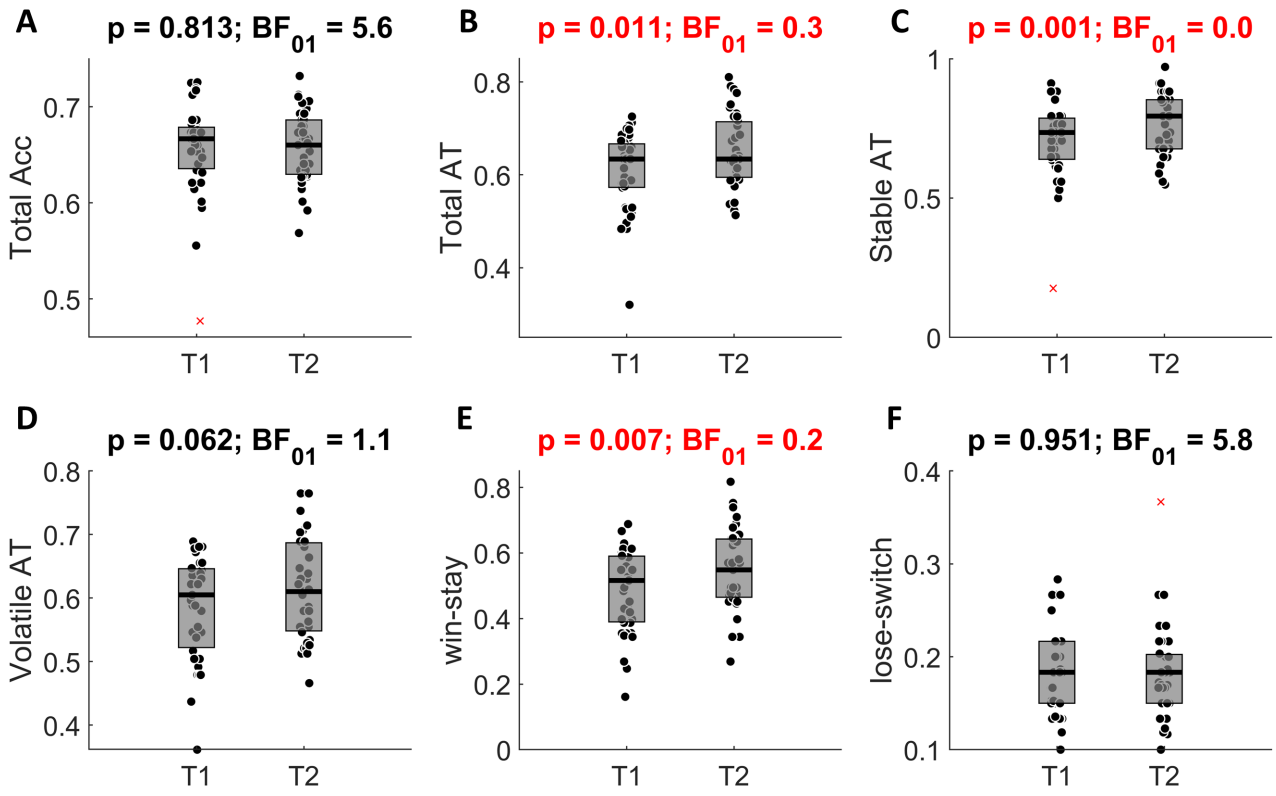


Figure 6: **Practice effects in the behavioral measures.** (**A**) Total accuracy. (**B**) Total advice taking. (**C**) Advice taking during the stable phase. (**D**) Advice taking during the volatile phase. (**E**) the probability of staying with the same advice taking strategy (trusting advice or not) after a win in a previous trial (**F**) the probability of changing the advice taking strategy (trusting advice or not) after a loss in a previous trial. Red crosses indicate outliers. p-values and Bayesian factors for the null hypothesis (BF_{01}) of paired t-tests are presented above each plot.

with explicitly reported inferred fidelity at different points in the task (**Fig 9B,D**). We found that $\hat{\mu}_2$ was significantly higher when advice was reported to be helpful vs when it was reported misleading ($p < 0.01$ in both testing sessions). This suggests that $\hat{\mu}_2$ captures what it is assumed to capture, which contributes to the face validity of the model.

4 Discussion

Here, we assessed psychometric properties of a computational assay designed to probe how people infer intentions in the context of receiving volatile advice (Diaconescu et al., 2014). We found this assay to have mostly poor test-retest reliability: this was true for both behavioral and computational measures. Our work is one of the first one to investigate test-retest reliability of a HGF model (see Mikus et al., 2024), and it adds to the emerging body of research showing that many computational assays used to study mental disorders suffer from poor psychometric properties, which poses considerable challenges for their use in research and for clinical translation efforts (Karvelis et al., 2023).

4.1 Why is test-retest reliability low?

While many studies report poor test-retest reliability of computational measures (Karvelis et al., 2023), post hoc analysis of different sources of variance affecting reliability is seldom performed. Here, we sought to tease apart the underlying sources of variance by directly comparing parameter recoverability with test-retest reliability for each parameter. Parameter recovery is constrained only by measurement error (and is not affected by within-subject variability across testing time points) providing an upper bound on test-retest reliability (Karvelis et al., 2023). In our study, measurement error proved to be

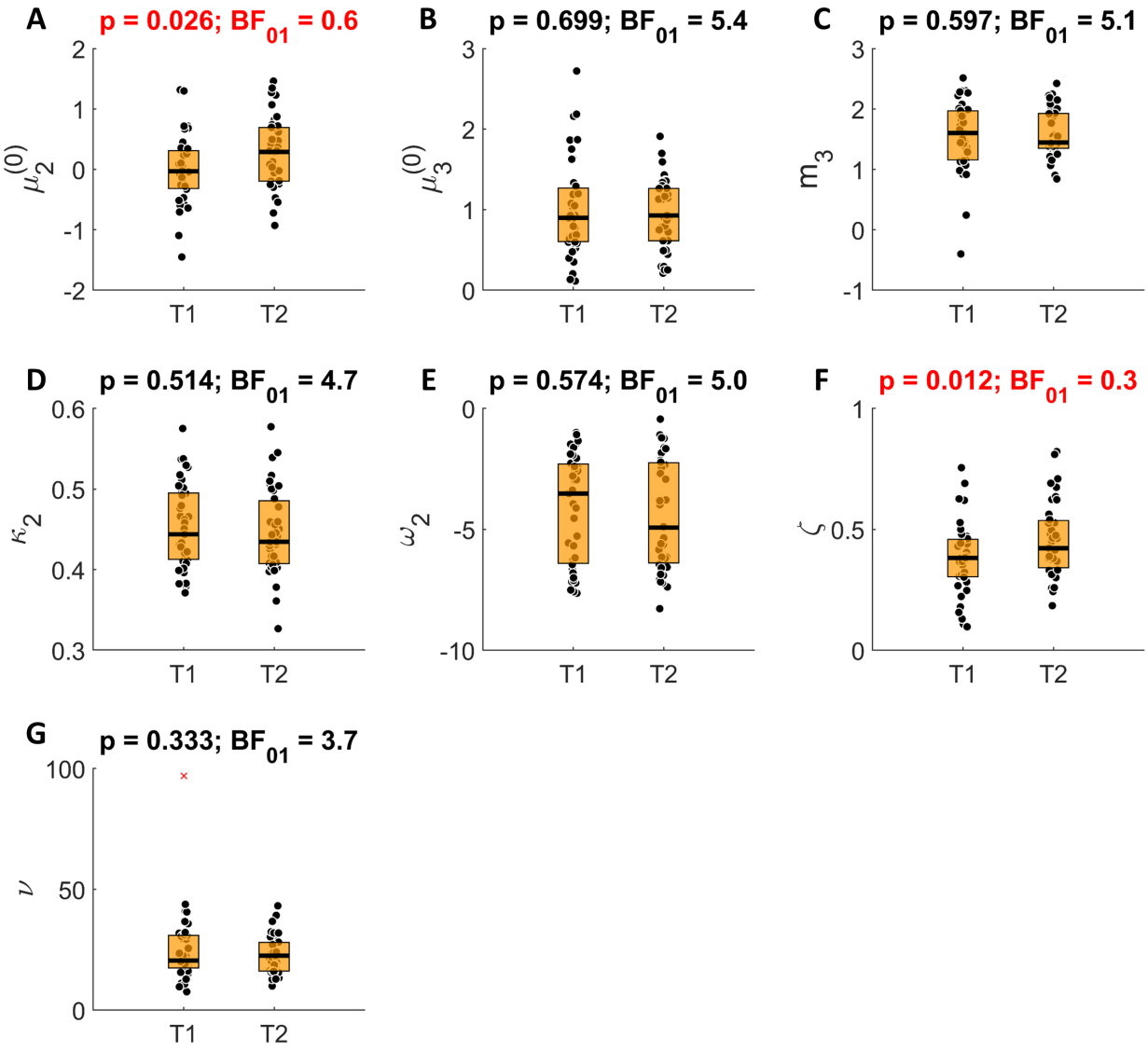


Figure 7: **Practice effects in the computational measures.**(**A**) Prior expectations about adviser's fidelity before starting the task (**B**) Prior expectations about the adviser's volatility before starting the task. (**C**) Volatility equilibrium point (**D**) Phasic learning rate about the adviser's fidelity (**E**) Baseline learning rate about the adviser's fidelity (**F**). The relative weighing of the advice compared to the non-social cue. (**G**) Decision noise. Red crosses indicate outliers. p-values and Bayesian factors for the null hypothesis (BF_{01}) of paired t-tests are presented above each plot.

an important factor behind low test-retest reliability: three parameters had poor recoverability (in the range of $ICC \approx 0.5$), indicating high measurement error for these parameters.

Still, for most parameters test-retest reliability was substantially lower than parameter recoverability suggesting that within-subject variance also played a major role in reducing test-retest reliability. We considered three sources of within-subject variance: practice effects, state-like fluctuations, or trait-like changes. We did find significant practice effects in some parameters, and while these effects were of moderate size (Cohen's $d = 0.37$ - 0.42), they accounted only for a small fraction ($ICC \approx 0.02 - 0.03$) of the reduction in test-retest reliability. Together with other studies reporting small to medium practice effects (Smith et al., 2022; Mkrtchian et al., 2023), it suggests that fixed practice effects might be a fairly minor factor behind low test-retest reliabilities of computational assays (although see Schurr et al. (2023) for a more extensive analysis). Nonetheless, practice effects still remain understudied and under-reported (Karvelis et al., 2023).

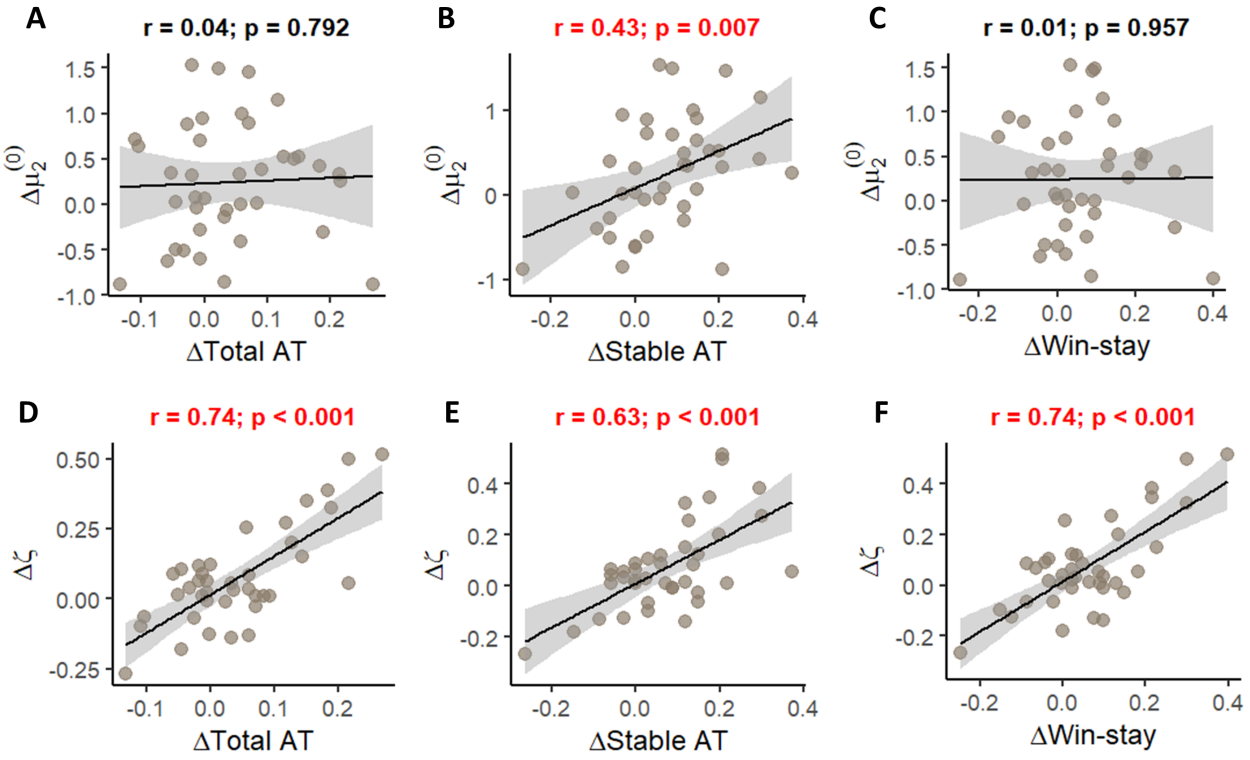


Figure 8: **Correlation between the changes in behavioral and computational measures.** (**A-C**) Prior expectations about adviser's fidelity before starting the task vs. total advice taking (**A**), vs. advice taking during the stable phase (**B**), and vs. win-stay frequency (**C**). (**D-F**) The relative weighing of the advice compared to the non-social cue vs. total advice taking (**D**), vs. advice taking during the stable phase (**E**), and vs. win-stay frequency (**F**). Pearson's correlation coefficients and corresponding p-values are presented above each panel.

Our test-retest interval was short (2 weeks), which rules out the possibility that trait-like changes (which by definition happen at longer timescales) affected the observed test-retest reliability. This means that the majority of within-subject variance might be explained by state-like fluctuations (mood, motivation, energy levels, etc.). However, since we did not collect such data, we are unable to demonstrate this definitively and to examine which state-like factors may be driving the effects.

In general, the degree to which state-like fluctuations impact computational measures remains understudied, even though previous work indicates its importance (Otto et al., 2013; Eldar and Niv, 2015; van Dooren et al., 2021). Studies investigating these effects in the context of test-retest reliability are only starting to emerge: Sullivan-Toole et al. (2022) found current mood intensity (positive or negative) to be associated with learning rate estimates in the Iowa gambling task, Schaaf et al. (2023) found stress and happiness levels to be associated with the loss learning rate parameter in Two-armed Bandit task, while Schurr et al. (2023) found affective valence to be associated with the inverse temperature parameter in Intertemporal Choice task, the lapse rate in Go/No-Go task, and the weight of relative uncertainty in Two-armed Bandit task. However, the effect sizes in all of these studies were rather small. Furthermore, Schurr et al. (2023) explicitly showed that valence explained only a small fraction of within-subject variance in parameter estimates while most of the variance was explained by subject-specific noise terms (the noise term simply captured any remaining unexplained variance). This indicates there could be many more state-like factors that might be driving these effects and need to be further examined.

Another recent study found that sleep deprivation affected perceptual stability in random dot kinematogram task, meanwhile psychosis proneness (a schizotypal trait) did not have any effect on task performance (Balter et al., 2022). While this study looked at behavioral rather than computational

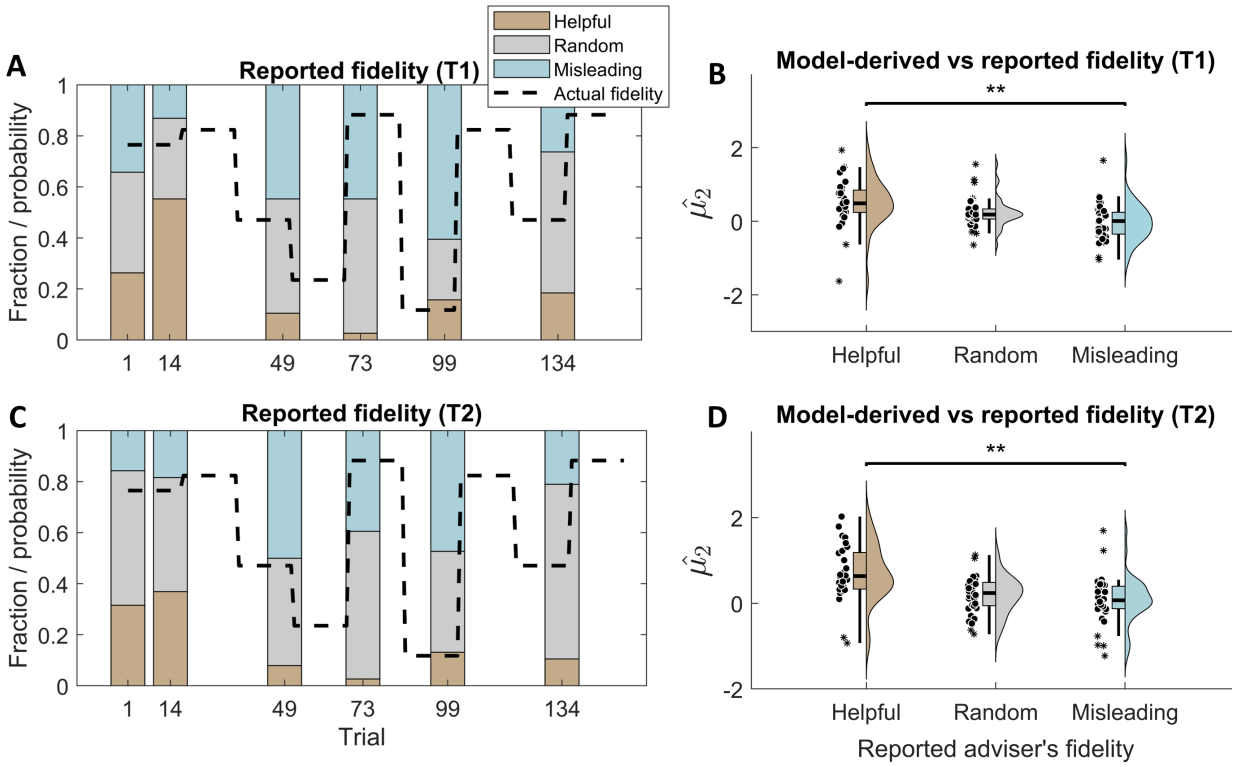


Figure 9: **Face validity: the task and the model.** Reported adviser's fidelity vs actual fidelity during test (**A**) and retest (**C**) at different points in the task. Model-derived estimates of adviser's fidelity ($\hat{\mu}_2$) broken down according to the explicitly reported fidelity at the corresponding points in the task during test (**B**) and retest (**D**).

measures, it nicely highlights that in some cases, state-like factors may have stronger effects than the primary variables of interest (e.g., traits, symptoms, diagnostic categories, etc.).

Overall, our work highlights that by directly comparing recoverability with reliability one can begin teasing apart which sources of variance are responsible for low test-retest reliability. While a handful of studies investigating test-retest reliability of computational measures have reported parameter recovery results (Waltmann et al., 2022; Mkrtchian et al., 2023; Smith et al., 2021; Sullivan-Toole et al., 2022; Hitchcock et al., 2022; Moutoussis et al., 2018; Smith et al., 2022; Shahar et al., 2019), such direct comparison has not been performed. Pooling the data from these studies does show that test-retest reliability tends to be substantially lower than parameter recoverability, highlighting the importance of accounting for within-subject variability (**Fig 7 in S1 Appendix**).

It must be noted, however, that our attempt to tease apart the different sources of variance has some caveats. First, the parameter recovery analysis only indicates the intrinsic measurement error, while the actual measurement error is likely to be higher. That is, given that we cannot assume the model to perfectly match the underlying cognitive mechanisms (which is true by construction, because modelling in its nature is an attempt to simplify reality), the mismatch will introduce additional measurement error. Second, our practice effects analysis accounts only for fixed effects. Any kind of idiosyncratic practice effects would be statistically indistinguishable from measurement error. Thus, it is possible that in addition to the state-like factors, that we emphasize in our interpretation, the remaining within-subject variance might also be driven by idiosyncratic task strategies and heuristics. However, how such strategies are selected may still be affected by state-like factors - ultimately blurring the distinction between state-like effects and idiosyncratic practice effects. Our ability to explore these and other more nuanced questions was limited by our experimental design, but should be investigated in the future using more intentional experimental designs (e.g., see Schurr et al., 2023).

Additionally, the social nature of the task might have led participants to employ higher-level strategies during their first session based on prior expectations of potential deception. Although the advice-taking task did not involve deception and participants were truthfully informed that the adviser had a different incentive structure, the lack of explicit details about this structure until debriefing might have fostered skepticism towards the task and experimenter's instructions. Potential concerns of deception may have influenced participants' behavior uniquely during the first session. By the second session, participants would have become familiar with the procedure, potentially reducing their concerns about deception. Consequently, the first session may differ substantially from subsequent sessions: we already saw this reflected in practice effects at the group level, but the individual differences in these effects would be adding to error variance. Therefore, to obtain more accurate insights into test-retest reliability, future studies should consider examining the reliability between the second and third sessions, after participants' initial concerns about deception have been alleviated.

4.2 What are the implications of our findings for the previous work using this assay?

The only previous work that used the exact version of the assay, the same modelling pipeline, and some of the same data was done by Hauke et al. (2024), which examined first-episode psychosis patients, individuals at clinical high-risk for psychosis, and matched healthy controls (note that this study had slightly less trials, 136 instead of 153 analyzed here; however, this had a negligible effect on reliability - see **Fig 8 in S1 Appendix**). Hauke et al. (2024) found that computational measures (m_3 and κ_2) to be different across the groups as well as showing weak correlations with clinical symptoms - although these results did not survive correction for multiple comparisons. Our finding of κ_2 having a near-zero test-retest reliability makes it even less likely that the group differences in κ_2 reported in Hauke et al. (2024) are true positives. However, our further analysis suggested that κ_2 reliability may have been slightly underestimated due to an outsized effect of 4 participants, with the ICC being $\sim$ 0.4 after removing them (**Fig 3D in S1 Appendix**). In such case, as well as for m_3, which also showed similar reliability of ICC ≈ 0.5, the group differences are less likely to be in question - while this level of reliability significantly reduces the observable effect size (e.g., Karvelis and Diaconescu, 2024; Zuo et al., 2019), if the true effect is very large (Cohen's $d > 1$) it would still be detectable with the sample size that the study had.

When it comes to other studies using this assay, our findings are less straightforward to generalize because each study used a slightly different version of the assay. For example, some previous studies used different volatility manipulations, used a different version of HGF and had either fewer (120) (Reiter et al., 2021) or more (210) trials (Diaconescu et al., 2020; Wellstein et al., 2020). On top of that, some other studies used a more difficult version of the task where instead of being shown explicit reward probabilities in a form of pie chart on every trial participants had to infer these reward probabilities from feedback (Henco et al., 2020; Sevgi et al., 2020). All of these differences among different versions of the task can lead to varying degrees of test-retest reliability (Karvelis et al., 2023; Zorowitz and Niv, 2023). This stresses the importance of making test-retest reliability analysis a routine practice (Karvelis et al., 2023; Parsons et al., 2019).

4.3 How to improve test-retest reliability?

Recent work has demonstrated that modelling task performance jointly across the two testing sessions can improve reliability (Brown et al., 2020; Sullivan-Toole et al., 2022; Waltmann et al., 2022). In this approach, each parameter is estimated in a form of a bivariate distribution, from which a correlation coefficient between the sessions can be derived. A simulation analysis by Waltmann et al. (2022) has demonstrated that estimating reliability this way is more accurate compared to modelling the sessions separately. However, computing ICCs using parameter estimates obtained from the joint modelling of the sessions resulted in inflated reliability (Waltmann et al., 2022). It means that using these point estimates in further analyses without accounting for their uncertainty and their non-independence would also introduce biases. To avoid these issues, future analyses should utilize full posterior distributions rather than relying solely on point estimates. Nonetheless, here we used a simpler modelling

approach that relies on fixed priors (see Methods for details), as this is the only established optimization pipeline available for fitting HGF models (Mathys et al., 2014). Joint modelling of the sessions - or even modelling the sessions separately but with empirical priors – poses significant challenges for the HGF. Without fixed priors, there is less control over regularization, increasing the risk of parameter values falling into regions that violate model assumptions and causing the optimization to fail. These limitations will be addressed in future work.

While hierarchical modelling approaches may provide some improvement in the reliability of parameter estimates, focusing on task design optimization seems to be a much more promising avenue as it can address the root causes of low reliability. Most cognitive tasks have been optimized for producing strong group effects, which often means low between-subject variation and as a result, low test-retest reliability (Hedge et al., 2018). A new wave of research is needed that places more emphasis on high test-retest reliability: see Zorowitz and Niv (2023) for some of the main task design factors to consider.

The first thing to ensure in task optimization is high parameter recoverability. Here we found that some parameters were recovered poorly, making it impossible to have high test-retest reliability for these parameters. Note that parameter recoverability of ~ 0.5 would not have been considered too problematic in previous research, as it would be interpreted as a 'large' effect size. However, considering parameter recoverability as the upper limit for test-retest reliability (Karvelis et al., 2023) imposes much stricter criteria for what recoverability can be considered satisfactory.

How the parameter recoverability itself could be improved does not have a straightforward answer. Sometimes low recoverability of parameters stems from collinearity – when parameters trade-off among each other to provide an equally good fit to the data. Here we ruled out this possibility by fixing poorly recovered parameters one by one and showing that it did not improve recoverability of the remaining parameters (**Fig 5 in S1 Appendix**). Another potential way to improve parameter recovery is to increase the number of trials. However, because our task depends on trial-by-trial dynamics, it immediately reverts us back to a more general question of what advice probability structure should be chosen (e.g., how should the volatility of the advisor change over time). It is likely that a different structure than was used in this task could constrain the parameter estimates better – however, discovering such structure in a principled way is a rather challenging problem. Finally, one more easily addressable factor that may be limiting parameter recoverability is the binary nature of responses in this task (Mikus et al., 2024). Future studies should consider incorporating continuous response measures, which contain more information and can therefore constrain parameter estimates better.

4.4 Construct validity

Despite the low test-retest reliability, we were able to show that the task has face validity at the group level: (1) participants reported the adviser to be more helpful when the adviser was indeed more helpful (task face validity) and (2) model-derived estimates of inferred adviser's fidelity ($\hat{\mu}_2$) covaried with explicitly reported inferred adviser's fidelity (model face validity). This adds to previous validation work that also found HGF parameters in this task to be associated with interpersonal reactivity index (IRI) scores (Diaconescu et al., 2014).

A more thorough validation of the assay would require demonstrating convergent and divergent validity of each model parameter across a battery of tasks; see Eckstein et al. (2022) for an example in the context of reinforcement learning models and Weigard et al. (2021) for an example in the context of drift-diffusion models. Ideally, these tasks would vary in their similarity among multiple important dimensions (e.g., instructions, stimulus types, volatility protocols, adviser qualities, etc.), so that a sufficient variety of tasks is achieved for assessing convergent and divergent validity. This work could be carried out in parallel with testing the predictive and longitudinal validity of these measures (Karvelis et al., 2023).

4.5 Limitations

Our sample size in this study was relatively small (N = 39), which resulted in rather wide confidence intervals around the estimated mean ICC values. While there is no consensus for the optimal sample size for reliability studies - the suggested optimal sample sizes range from ~ 60 (Williams et al., 2023), to 100 (Kennedy, 2022), to 400 subjects (Charter, 1999) - larger samples can provide more precise reliability estimates, which is especially important when reliability itself is poor. Another limitation in our study is the fact that our test and retest sessions were done during EEG and fMRI scanning. Even though we showed that there were no fixed effects on task performance between the two experimental environments, it may have nonetheless contributed to the error variance.

5 Conclusion

Our findings of the largely poor test-retest reliability of behavioral and computational measures of advice taking under volatility warrant more caution when interpreting previous findings of this (Hauke et al., 2024) and potentially other versions of the assay (Henco et al., 2020; Sevgi et al., 2020; Reiter et al., 2021). More generally, our findings add to the growing body of research showing that many behavioral and computational task measures suffer from poor test-retest reliability (Karvelis et al., 2023; Enkavi et al., 2019; Hedge et al., 2018). We sought to illustrate how using simple methods, one can start to disentangle the different sources of variance underlying low test-retest reliability of computational measures, however, more intentional experimental designs are needed for a more in-depth analysis of these sources (e.g., Schurr et al., 2023). A better understanding of what factors drive low test-retest reliability would facilitate the design of more psychometrically sound assays, which would increase the quality of future research and the possibility of clinical translation.

6 Supporting information

S1 Appendix. All supplementary materials.

7 Data availability

The data is available at https://osf.io/a6v8w.
The code is available at https://github.com/povilaskarvelis/compi_ioio_reliability.

8 Credit authorship contribution statement

Povilas Karvelis performed the data analysis and wrote up the original draft of the manuscript. Daniel Hauke contributed to the development of the analysis pipelines and provided feedback on the manuscript. Andreea Diaconescu conceptualized and supervised the study. Michelle Wobmann recruited participants, collected data, performed analyses on behavioral measures and provided feedback. Christina Andreou provided supervision and rater training and feedback on the manuscript. Amatya Mackintosh and Renate de Bock recruited participants, collected data and provided feedback. Stefan Borgwardt provided infrastructure, supervision, and feedback.

9 Competing interests

The authors have no competing interests to declare.

References

Balter, L. J., Matheson, G. J., Sundelin, T., Sterzer, P., Petrovic, P., and Axelsson, J. (2022). Experimental sleep deprivation results in diminished perceptual stability independently of psychosis proneness. *Brain Sciences*, 12(10):1338.

Brown, V. M., Chen, J., Gillan, C. M., and 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.

Browning, M., Carter, C. S., Chatham, C., Den Ouden, H., Gillan, C. M., Baker, J. T., Chekroud, A. M., Cools, R., Dayan, P., Gold, J., et al. (2020). Realizing the clinical potential of computational psychiatry: report from the banbury center meeting, february 2019. *Biological Psychiatry*, 88(2):e5–e10.

Charter, R. A. (1999). Sample size requirements for precise estimates of reliability, generalizability, and validity coefficients. *Journal of Clinical and Experimental Neuropsychology*, 21(4):559–566.

Cole, D. M., Diaconescu, A. O., Pfeiffer, U. J., Brodersen, K. H., Mathys, C. D., Julkowski, D., Ruhrmann, S., Schilbach, L., Tittgemeyer, M., Vogeley, K., et al. (2020). Atypical processing of uncertainty in individuals at risk for psychosis. *NeuroImage: Clinical*, 26:102239.

Daunizeau, J., Adam, V., and Rigoux, L. (2014). Vba: a probabilistic treatment of nonlinear models for neurobiological and behavioural data. *PLoS computational biology*, 10(1):e1003441.

Diaconescu, A. O., Mathys, C., Weber, L. A., Daunizeau, J., Kasper, L., Lomakina, E. I., Fehr, E., and Stephan, K. E. (2014). Inferring on the intentions of others by hierarchical bayesian learning. *PLoS computational biology*, 10(9):e1003810.

Diaconescu, A. O., Mathys, C., Weber, L. A., Kasper, L., Mauer, J., and Stephan, K. E. (2017). Hierarchical prediction errors in midbrain and septum during social learning. *Social cognitive and affective neuroscience*, 12(4):618–634.

Diaconescu, A. O., Wellstein, K. V., Kasper, L., Mathys, C., and Stephan, K. E. (2020). Hierarchical bayesian models of social inference for probing persecutory delusional ideation. *Journal of Abnormal Psychology*, 129(6):556.

Eckstein, M. K., Master, S. L., Xia, L., Dahl, R. E., Wilbrecht, L., and Collins, A. G. (2022). The interpretation of computational model parameters depends on the context. *Elife*, 11:e75474.

Eldar, E. and Niv, Y. (2015). Interaction between emotional state and learning underlies mood instability. *Nature communications*, 6(1):6149.

Enkavi, A. Z., Eisenberg, I. W., Bissett, P. G., Mazza, G. L., MacKinnon, D. P., Marsch, L. A., and Poldrack, R. A. (2019). Large-scale analysis of test–retest reliabilities of self-regulation measures. *Proceedings of the National Academy of Sciences*, 116(12):5472–5477.

Fleiss, J. L. (2011). *Design and analysis of clinical experiments*. John Wiley & Sons.

Frässle, S., Aponte, E. A., Bollmann, S., Brodersen, K. H., Do, C. T., Harrison, O. K., Harrison, S. J., Heinzle, J., Iglesias, S., Kasper, L., et al. (2021). Tapas: an open-source software package for translational neuromodeling and computational psychiatry. *Frontiers in Psychiatry*, 12:680811.

Friston, K. (2023). Computational psychiatry: from synapses to sentience. *Molecular psychiatry*, 28(1):256–268.

Hauke, D. J., Wobmann, M., Andreou, C., Mackintosh, A. J., de Bock, R., Karvelis, P., Adams, R. A., Sterzer, P., Borgwardt, S., Roth, V., et al. (2024). Altered perception of environmental volatility during social learning in emerging psychosis. *Computational Psychiatry*, 8(1):1.

Hauser, T. U., Skvortsova, V., De Choudhury, M., and Koutsouleris, N. (2022). The promise of a model-based psychiatry: building computational models of mental ill health. *The Lancet Digital Health*.

Hedge, C., Powell, G., and Sumner, P. (2018). The reliability paradox: Why robust cognitive tasks do not produce reliable individual differences. *Behavior research methods*, 50(3):1166–1186.

Henco, L., Diaconescu, A. O., Lahnakoski, J. M., Brandi, M.-L., Hörmann, S., Hennings, J., Hasan, A., Papazova, I., Strube, W., Bolis, D., et al. (2020). Aberrant computational mechanisms of social learning and decision-making in schizophrenia and borderline personality disorder. *PLoS computational biology*, 16(9):e1008162.

Hitchcock, P. F., Britton, W. B., Mehta, K. P., and Frank, M. J. (2022). Self-judgment dissected: A computational modeling analysis of self-referential processing and its relationship to trait mindfulness facets and depression symptoms. *Cognitive, Affective, & Behavioral Neuroscience*, pages 1–19.

Huys, Q. J., Browning, M., Paulus, M. P., and Frank, M. J. (2021). Advances in the computational understanding of mental illness. *Neuropsychopharmacology*, 46(1):3–19.

Karvelis, P., Charlton, C. E., Allohverdi, S. G., Bedford, P., Hauke, D. J., and Diaconescu, A. O. (2022). Computational approaches to treatment response prediction in major depression using brain activity and behavioral data: A systematic review. *Network Neuroscience*, pages 1–52.

Karvelis, P. and Diaconescu, A. O. (2022). A computational model of hopelessness and active-escape bias in suicidality. *Computational Psychiatry*, 6(1).

Karvelis, P. and Diaconescu, A. O. (2024). Clarifying the reliability paradox: poor test-retest reliability attenuates group differences.

Karvelis, P., Paulus, M. P., and Diaconescu, A. O. (2023). Individual differences in computational psychiatry: a review of current challenges. *Neuroscience & Biobehavioral Reviews*, page 105137.

Kass, R. E. and Raftery, A. E. (1995). Bayes factors. *Journal of the american statistical association*, 90(430):773–795.

Kennedy, I. (2022). Sample size determination in test-retest and cronbach alpha reliability estimates. *British Journal of Contemporary Education*, 2(1):17–29.

Koo, T. K. and Li, M. Y. (2016). A guideline of selecting and reporting intraclass correlation coefficients for reliability research. *Journal of chiropractic medicine*, 15(2):155–163.

Mathys, C., Daunizeau, J., Friston, K. J., and Stephan, K. E. (2011). A bayesian foundation for individual learning under uncertainty. *Frontiers in human neuroscience*, 5:39.

Mathys, C. D., Lomakina, E. I., Daunizeau, J., Iglesias, S., Brodersen, K. H., Friston, K. J., and Stephan, K. E. (2014). Uncertainty in perception and the hierarchical gaussian filter. *Frontiers in human neuroscience*, 8:825.

McGraw, K. O. and Wong, S. P. (1996). Forming inferences about some intraclass correlation coefficients. *Psychological methods*, 1(1):30.

Mikus, N., Lamm, C., and Mathys, C. (2024). Computational phenotyping of aberrant belief updating in individuals with schizotypal traits and schizophrenia. *Biological Psychiatry*.

Mkrtchian, A., Valton, V., and Roiser, J. P. (2023). Reliability of decision-making and reinforcement learning computational parameters. *Computational Psychiatry*, 7(1):30–46.

Moutoussis, M., Bullmore, E. T., Goodyer, I. M., Fonagy, P., Jones, P. B., Dolan, R. J., Dayan, P., and in Psychiatry Network Research Consortium, N. (2018). Change, stability, and instability in the pavlovian guidance of behaviour from adolescence to young adulthood. *PLoS computational biology*, 14(12):e1006679.

Otto, A. R., Raio, C. M., Chiang, A., Phelps, E. A., and Daw, N. D. (2013). Working-memory capacity protects model-based learning from stress. *Proceedings of the National Academy of Sciences*, 110(52):20941–20946.

Palminteri, S., Wyart, V., and Koechlin, E. (2017). The importance of falsification in computational cognitive modeling. *Trends in cognitive sciences*, 21(6):425–433.

Parsons, S., Kruijt, A.-W., and 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(4):378–395.

Patzelt, E. H., Hartley, C. A., and Gershman, S. J. (2018). Computational phenotyping: using models to understand individual differences in personality, development, and mental illness. *Personality Neuroscience*, 1.

Qin, S., Nelson, L., McLeod, L., Eremenco, S., and Coons, S. J. (2019). Assessing test–retest reliability of patient-reported outcome measures using intraclass correlation coefficients: recommendations for selecting and documenting the analytical formula. *Quality of Life Research*, 28:1029–1033.

Reiter, A. M., Diaconescu, A. O., Eppinger, B., and Li, S.-C. (2021). Human aging alters social inference about others' changing intentions. *Neurobiology of Aging*, 103:98–108.

Rigoux, L., Stephan, K., Friston, K., and Daunizeau, J. (2014). Bayesian model selection for group studies — revisited. *NeuroImage*, 84:971–985.

Schaaf, J. V., Weidinger, L., Molleman, L., and van den Bos, W. (2023). Test-retest reliability of reinforcement learning parameters.

Schurr, R., Reznik, D., Hillman, H., Bhui, R., and Gershman, S. J. (2023). Dynamic computational phenotyping of human cognition.

Sevgi, M., Diaconescu, A. O., Henco, L., Tittgemeyer, M., and Schilbach, L. (2020). Social bayes: using bayesian modeling to study autistic trait–related differences in social cognition. *Biological Psychiatry*, 87(2):185–193.

Shahar, N., Hauser, T. U., Moutoussis, M., Moran, R., Keramati, M., Consortium, N., and Dolan, R. J. (2019). Improving the reliability of model-based decision-making estimates in the two-stage decision task with reaction-times and drift-diffusion modeling. *PLoS computational biology*, 15(2):e1006803.

Smith, R., Kirlic, N., Stewart, J. L., Touthang, J., Kuplicki, R., McDermott, T. J., Taylor, S., Khalsa, S. S., Paulus, M. P., and Aupperle, R. L. (2021). Long-term stability of computational parameters during approach-avoidance conflict in a transdiagnostic psychiatric patient sample. *Scientific reports*, 11(1):1–13.

Smith, R., Taylor, S., Stewart, J. L., Guinjoan, S. M., Ironside, M., Kirlic, N., Ekhtiari, H., White, E. J., Zheng, H., Kuplicki, R., et al. (2022). Slower learning rates from negative outcomes in substance use disorder over a 1-year period and their potential predictive utility. *Computational Psychiatry*, 6(1).

Stephan, K. E., Penny, W. D., Daunizeau, J., Moran, R. J., and Friston, K. J. (2009). Bayesian model selection for group studies. *NeuroImage*, 46(4):1004–1017.

Sullivan-Toole, H., Haines, N., Dale, K., Olino, T., et al. (2022). Enhancing the psychometric properties of the iowa gambling task using full generative modeling. *Computational Psychiatry*, 6(1):12=89–212.

Uhlenbeck, G. E. and Ornstein, L. S. (1930). On the theory of the brownian motion. *Physical review*, 36(5):823.

van Dooren, R., de Kleijn, R., Hommel, B., and Sjoerds, Z. (2021). The exploration-exploitation trade-off in a foraging task is affected by mood-related arousal and valence. *Cognitive, Affective, & Behavioral Neuroscience*, 21(3):549–560.

Waltmann, M., Schlagenhauf, F., and Deserno, L. (2022). Sufficient reliability of the behavioral and computational readouts of a probabilistic reversal learning task. *Behavior Research Methods*, pages 1–22.

Weigard, A., Clark, D. A., and Sripada, C. (2021). Cognitive efficiency beats top-down control as a reliable individual difference dimension relevant to self-control. *Cognition*, 215:104818.

Weir, J. P. (2005). Quantifying test-retest reliability using the intraclass correlation coefficient and the sem. *The Journal of Strength & Conditioning Research*, 19(1):231–240.

Wellstein, K. V., Diaconescu, A. O., Bischof, M., Rüesch, A., Paolini, G., Aponte, E. A., Ullrich, J., and Stephan, K. E. (2020). Inflexible social inference in individuals with subclinical persecutory delusional tendencies. *Schizophrenia Research*, 215:344–351.

Williams, B., FitzGibbon, L., Brady, D., and Christakou, A. (2023). Sample size matters when estimating variability of flexible, goal directed behaviour.

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

Yip, S. W., Barch, D. M., Chase, H. W., Flagel, S., Huys, Q. J., Konova, A. B., Montague, R., and Paulus, M. (2022). From computation to clinic. *Biological Psychiatry Global Open Science*.

Zorowitz, S. and Niv, Y. (2023). Improving the reliability of cognitive task measures: A narrative review. *Biological Psychiatry: Cognitive Neuroscience and Neuroimaging*.

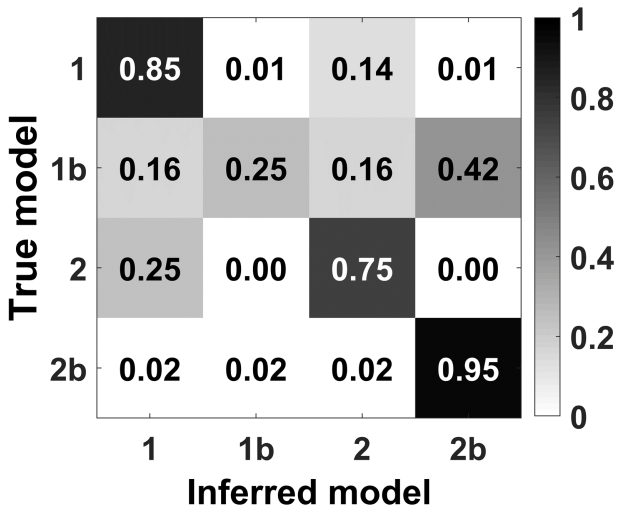
Zuo, X.-N., Xu, T., and Milham, M. P. (2019). Harnessing reliability for neuroscience research. *Nature human behaviour*, 3(8):768–771.

Supplementary materials: Test-retest reliability of behavioral and computational measures of advice taking under volatility

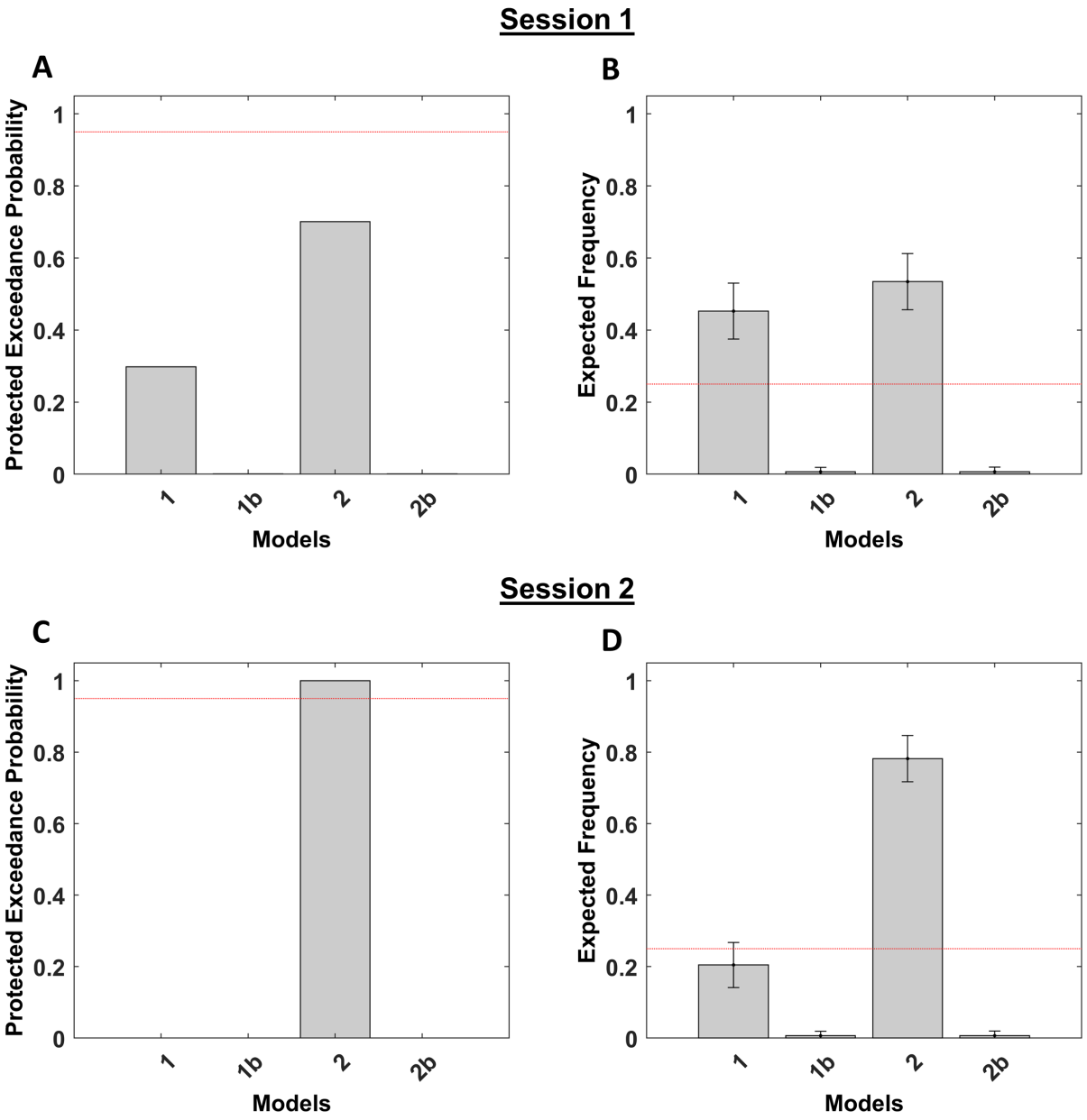
1 Model recovery and model selection

Just like Hauke et al. (2024), we found that that the control model 1b could not be well recovered. All other models - and most importantly, the primary models of interest (model 1 and model 2) - were recoverable with high probability (**Supplementary Fig 1**).

Random-effects Bayesian model selection revealed that model 2 was the winning model for both session 1 and session 2 (**Supplementary Fig 2**).



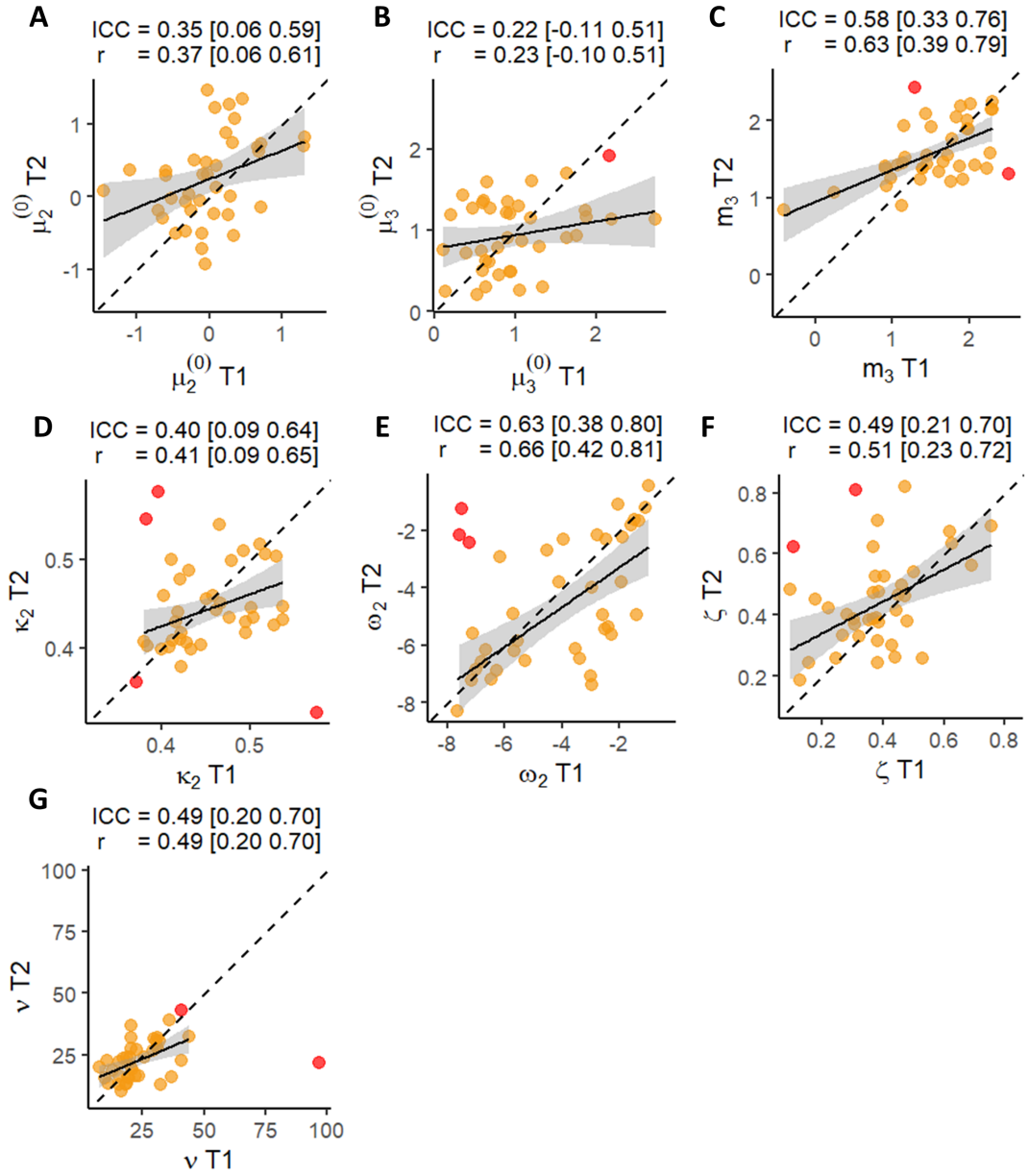
Supplementary Figure 1: **Model recovery.** The confusion matrix shows average protected exceedance probability: averaged across 20 data sets simulated with the same parameters but with different random seeds to account for the stochasticity effects.



Supplementary Figure 2: **Random-effects Bayesian Model Selection results.** Models: standard HGF with free parameters (1) or Bayes Optimal parameters (1b) and mean-reverting HGF with free parameters (2) or Bayes Optimal parameters (2b). In both testing sessions model 2 is the winning model.

2 Test-retest reliability results after removing influential data points

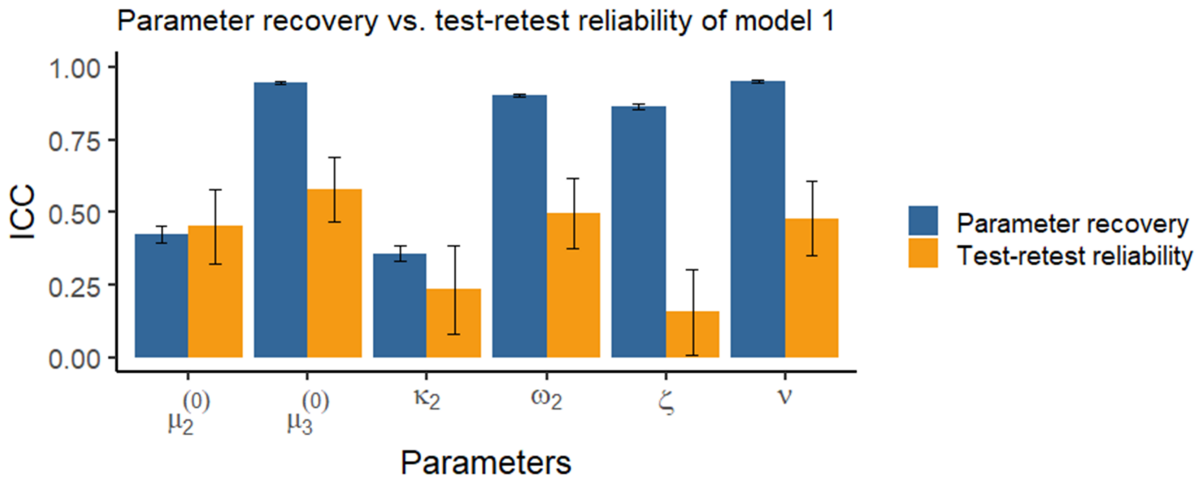
The visual inspection of the scatter plot of κ_2 test-retest reliability suggested that some data points, while not being extreme univariate outliers may nonetheless have exerted outsized influence on the estimated reliability. To verify this, we used Cook's distance (Cook, 1977) to remove the most influential data points affecting a linear regression model ($x_{T1} \sim x_{T2}$). We used the standard threshold of 4/N, where N is the number of data points, which in our case was 39. This indeed resulted in much better test-retest reliability results for κ_2 (**Supplementary Fig 3D**), improving it from 0.04 to 0.40. For completeness we also report how this analysis affected other parameters, although the changes were less noteworthy (**Supplementary Fig 3**).



Supplementary Figure 3: **Test-retest reliability of model 2 parameters after removing data points with outsized influence, as determined using Cook's distance.**.

3 Parameter recovery and test-retest reliability of model 1

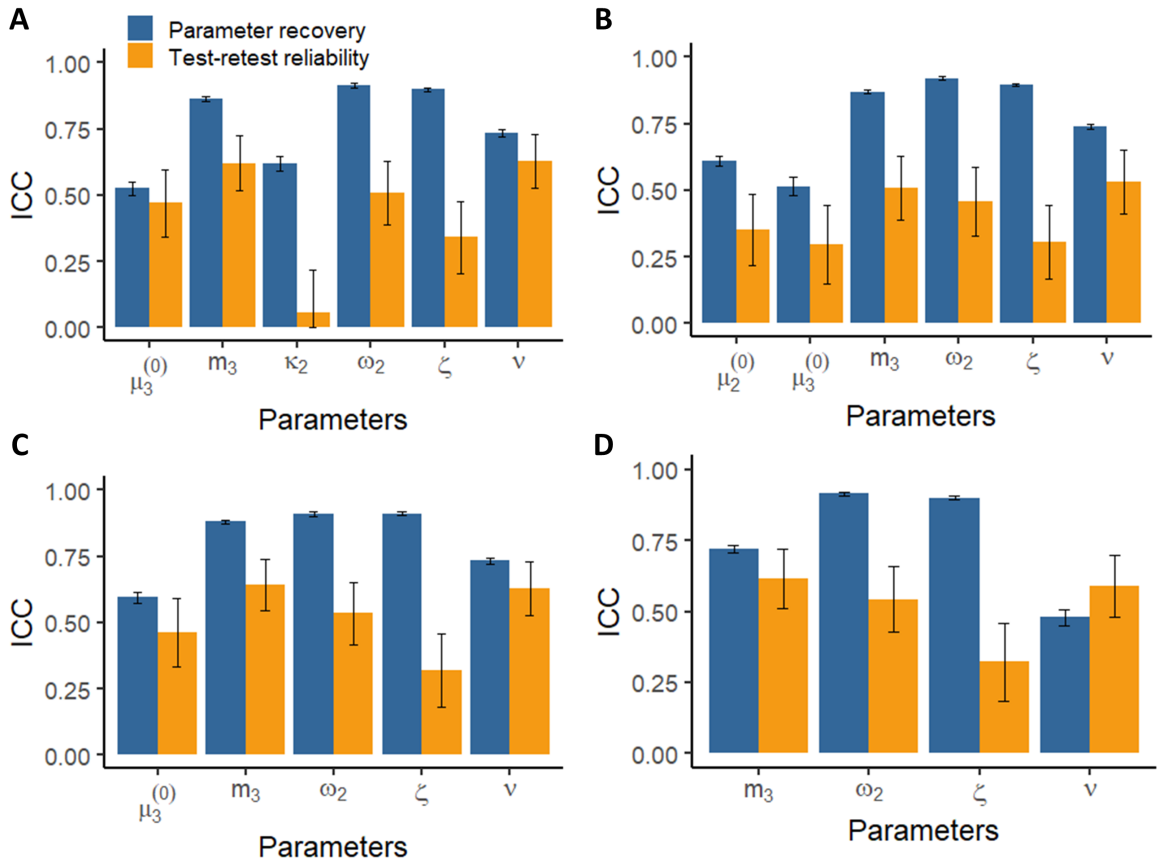
To make sure that our results about the reliability of computational measures did not hinge on the winning model only, we also assessed model 1. Given that model 1 is less complex (it does not involve the drift at the 3rd level and has 6 instead of 7 free parameters), we expected it might fare better in terms of reliability. However, the results were rather similar: test-retest reliability was poor ($ICC < 0.5$) for most parameters, even when recoverability was very high (**Supplementary Fig 4**)



Supplementary Figure 4: **Parameter recovery vs. test-retest reliability of model 1 parameters.**.

4 Fixing parameters to test for collinearity

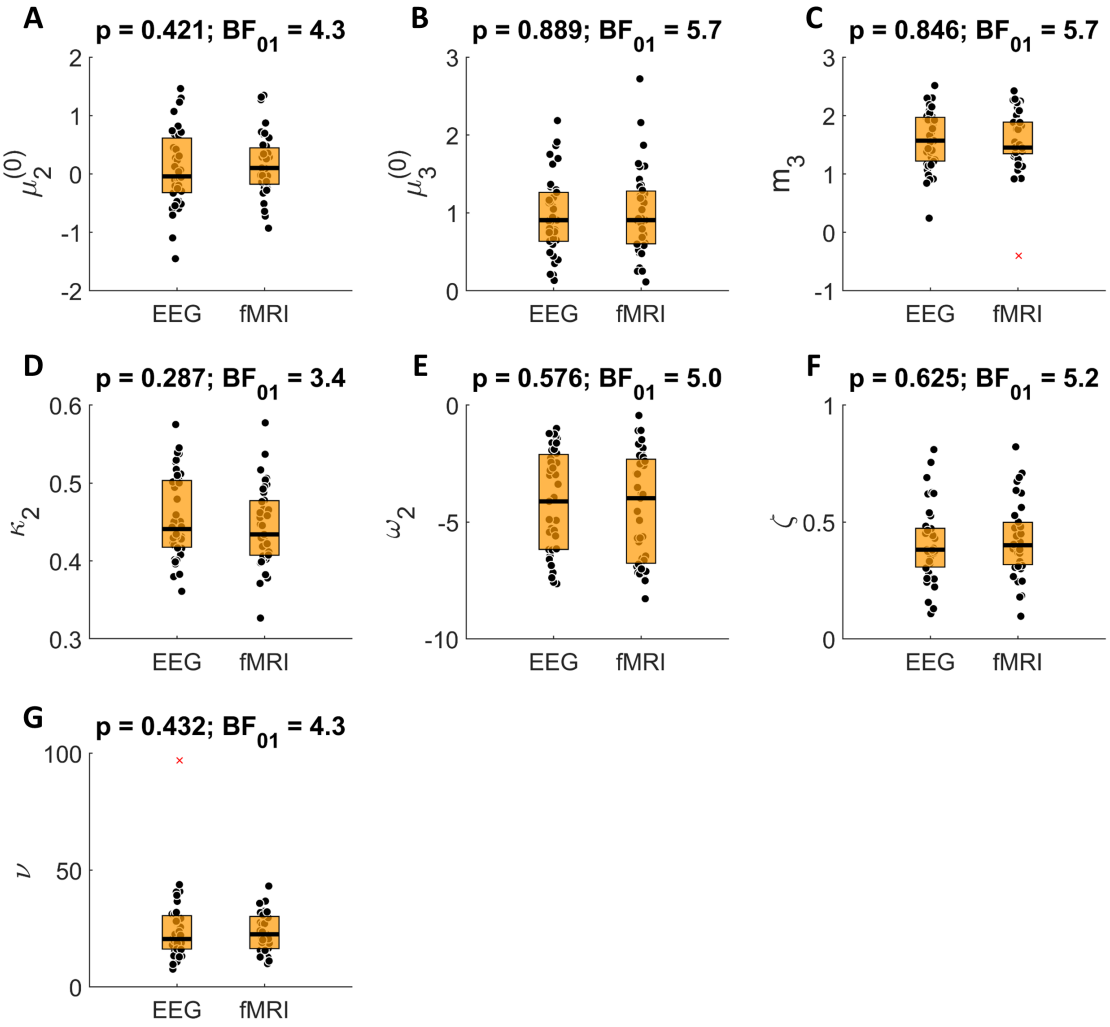
To investigate if collinearity was the reason why parameters $\mu_2^{(0)}$, $\mu_3^{(0)}$, and κ_2 in model 2 had low recoverability, we refit the model to the data by fixing each of these parameters at a time and observing how it affected the recoverability of the remaining parameters - if collinearity was the issue, fixing one of the parameters would improve the recovery of the remaining ones. We followed the same fitting procedure as described in the main text and fixed the parameters to the corresponding priors as detailed in **Table 1** in the main text. We did not observe any major improvements after fixing $\mu_2^{(0)}$ (**Supplementary Fig 5A**), or fixing κ_2 (**Supplementary Fig 5B**), or even fixing both of these parameters at once (**Supplementary Fig 5C**). Interestingly, fixing all three parameters worsened the recoverability of parameters ν and m_3 (**Supplementary Fig 5D**).



Supplementary Figure 5: **Parameter recovery vs. test-retest reliability for reduced versions of model 2, to test for potential collinearity.** (**A**) Fixed $\mu_2^{(0)}$. (**B**) A model with fixed κ_2. (**C**) Fixed $\mu_2^{(0)}$ and κ_2. (**D**) Fixed $\mu_2^{(0)}$, κ_2, and $\mu_3^{(0)}$.

5 EEG vs fMRI sessions

To assess if test-retest reliability might have been affected by different environments, we compared parameter estimates between EEG and fMRI sessions (the order of which were randomized across participants). We found substantial evidence ($BF_{01} > 3$) that neither of the parameters were systematically affected by different neuroimaging environments (**Supplementary Fig 6**).

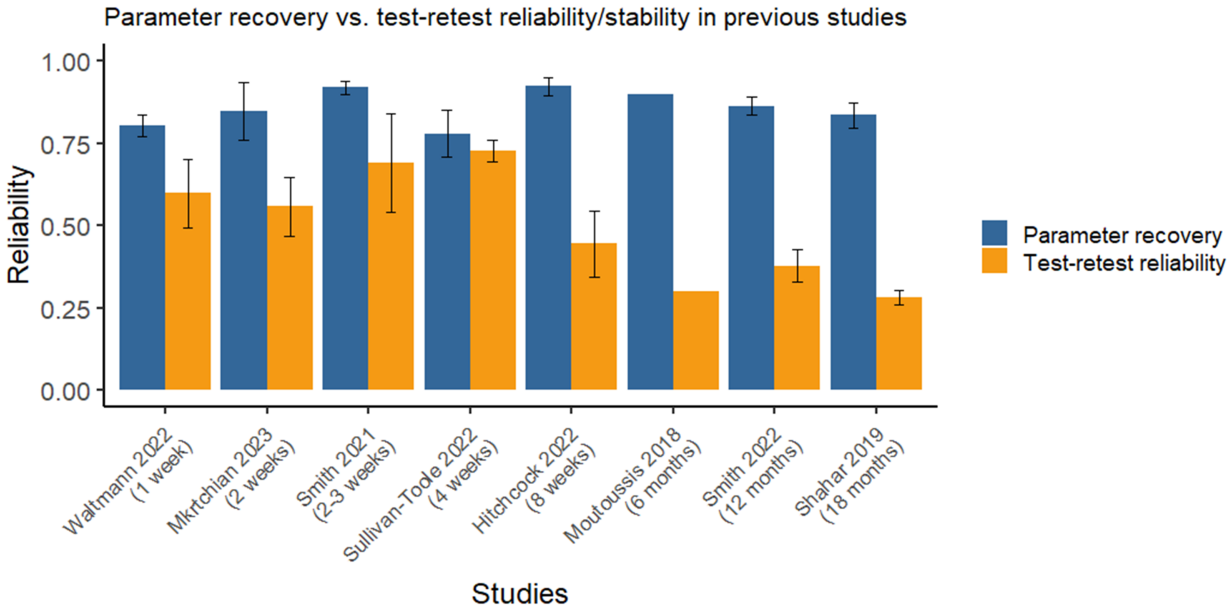


Supplementary Figure 6: **Parameter estimates from EEG and fMRI sessions.** (**A**) Prior expectations about adviser's fidelity before starting the task (**B**) Prior expectations about adviser's volatility before starting the task. (**C**) Volatility equilibrium point (**D**) Phasic learning rate about adviser's fidelity (**E**) Tonic learning rate about adviser's fidelity (**F**). The relative weighing of the advice compared to the non-social cue. (**G**) Decision noise. Red crosses indicate outliers. p-values and Bayesian factors for the null hypothesis (BF_{01}) of paired t-tests are presented above each plot.

6 Parameter recovery vs test-retest reliability in previous studies

Not many studies assessing test-retest reliability also report parameter recovery results (Karvelis et al., 2023). In such cases it is difficult to know which sources of variance are primarily responsible for low test-retest reliability (e.g., measurement error or within-subject variability). Even when parameter recovery results are reported, they tend to be not fully utilized to disambiguate the different sources of variance. We pooled the data from such studies and put them together in a single plot to gain more insight (**Supplementary Fig 7**).

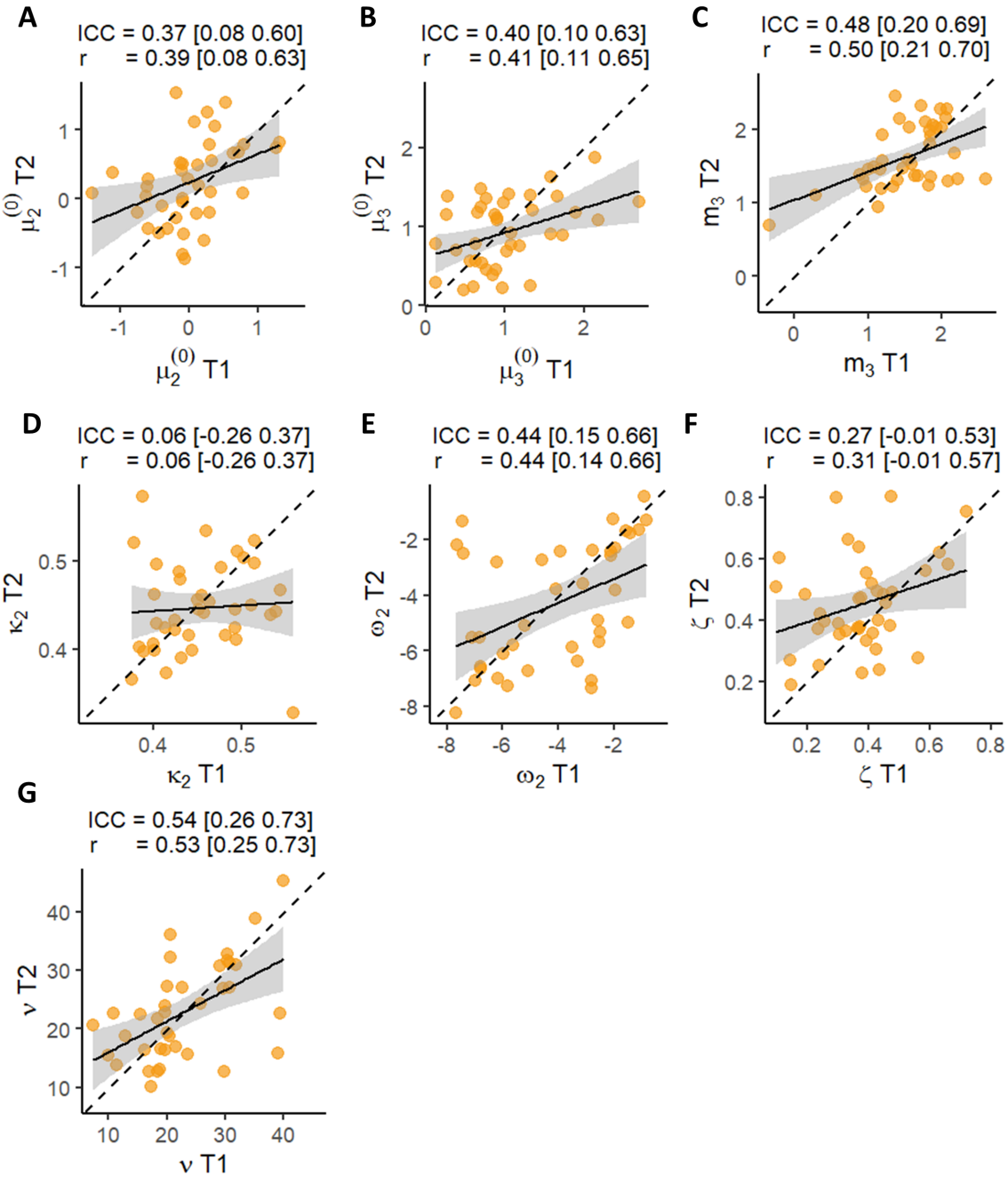
First, we find that all studies report considerably high parameter recoverability (> 0.75). Second, test-retest reliability tends to always be considerably lower than parameter recoverability. This difference also tends to increase with longer intervals between test and retest.



Supplementary Figure 7: **The differences in parameter recovery and test-retest reliability of parameter estimates in previous studies.** The studies are sorted by the length of the interval between test and retest (indicated within the parentheses): Waltmann et al. (2022); Mkrtchian et al. (2023); Smith et al. (2021); Sullivan-Toole et al. (2022); Hitchcock et al. (2022); Moutoussis et al. (2018); Smith et al. (2022); Shahar et al. (2019). Each column represents reliability averaged across all parameter estimates; the error bars indicate standard error. Note that some studies used ICC, some used Pearson's correlation coefficient, and some used Spearman's correlation coefficient to assess test-retest reliability and parameter recoverability - here we ignore the differences between these metrics and plot them along the same Y-axis.

7 Test-retest reliability of model 2 parameters when using the same number of trials as in Hauke et al. (2024)

To be able to more directly relate our results to the findings of Hauke et al. (2024), we reran the test-retest analysis using the same number of trials (136 instead of 153). Perhaps unsurprisingly, we found that test-retest reliability of model parameters was very similar to that reported in the main text (**Supplementary Fig 8**).



Supplementary Figure 8: **Test-retest reliability of model 2 using the same number of trials (136) as in Hauke et al. (2024)** (**A**) Prior expectations about adviser's fidelity before starting the task (**B**) Prior expectations about adviser's volatility before starting the task. (**C**) Volatility equilibrium point (**D**) Phasic learning rate about adviser's fidelity (**E**) Tonic learning rate about adviser's fidelity (**F**). The relative weighing of the advice compared to the non-social cue. (**G**) Decision noise. ICC(A,1) and Pearson's correlation coefficients are shown above each panel. The square brackets indicate 95% confidence intervals.

References

Cook, R. D. (1977). Detection of influential observation in linear regression. *Technometrics*, 19(1):15–18.

Hauke, D. J., Wobmann, M., Andreou, C., Mackintosh, A. J., de Bock, R., Karvelis, P., Adams, R. A., Sterzer, P., Borgwardt, S., Roth, V., et al. (2024). Altered perception of environmental volatility during social learning in emerging psychosis. *Computational Psychiatry*, 8(1):1.

Hitchcock, P. F., Britton, W. B., Mehta, K. P., and Frank, M. J. (2022). Self-judgment dissected: A computational modeling analysis of self-referential processing and its relationship to trait mindfulness facets and depression symptoms. *Cognitive, Affective, & Behavioral Neuroscience*, pages 1–19.

Karvelis, P., Paulus, M. P., and Diaconescu, A. O. (2023). Individual differences in computational psychiatry: a review of current challenges. *Neuroscience & Biobehavioral Reviews*, page 105137.

Mkrtchian, A., Valton, V., and Roiser, J. P. (2023). Reliability of decision-making and reinforcement learning computational parameters. *Computational Psychiatry*, 7(1):30–46.

Moutoussis, M., Bullmore, E. T., Goodyer, I. M., Fonagy, P., Jones, P. B., Dolan, R. J., Dayan, P., and in Psychiatry Network Research Consortium, N. (2018). Change, stability, and instability in the pavlovian guidance of behaviour from adolescence to young adulthood. *PLoS computational biology*, 14(12):e1006679.

Shahar, N., Hauser, T. U., Moutoussis, M., Moran, R., Keramati, M., Consortium, N., and Dolan, R. J. (2019). Improving the reliability of model-based decision-making estimates in the two-stage decision task with reaction-times and drift-diffusion modeling. *PLoS computational biology*, 15(2):e1006803.

Smith, R., Kirlic, N., Stewart, J. L., Touthang, J., Kuplicki, R., McDermott, T. J., Taylor, S., Khalsa, S. S., Paulus, M. P., and Aupperle, R. L. (2021). Long-term stability of computational parameters during approach-avoidance conflict in a transdiagnostic psychiatric patient sample. *Scientific reports*, 11(1):1–13.

Smith, R., Taylor, S., Stewart, J. L., Guinjoan, S. M., Ironside, M., Kirlic, N., Ekhtiari, H., White, E. J., Zheng, H., Kuplicki, R., et al. (2022). Slower learning rates from negative outcomes in substance use disorder over a 1-year period and their potential predictive utility. *Computational Psychiatry*, 6(1).

Sullivan-Toole, H., Haines, N., Dale, K., Olino, T., et al. (2022). Enhancing the psychometric properties of the iowa gambling task using full generative modeling. *Computational Psychiatry*, 6(1):12=89–212.

Waltmann, M., Schlagenhauf, F., and Deserno, L. (2022). Sufficient reliability of the behavioral and computational readouts of a probabilistic reversal learning task. *Behavior Research Methods*, pages 1–22.