Consistency within change: Evaluating the psychometric properties of a widely-used predictive-inference task

Alisa M. Loosen[1,2], Tricia X. F. Seow[1,2] & Tobias U. Hauser[1,2]

[1]Max Planck UCL Centre for Computational Psychiatry and Ageing Research

[2]Wellcome Centre for Human Neuroimaging, University College London

Keywords: Decision making, Learning, Predictive inference, Test-retest reliability, Internal consistency, Psychometric qualities

Abstract

Rapid adaptation to sudden changes in the environment is a hallmark of human behaviour. Many computational, neuroimaging, and even clinical investigations, which capture this ability have relied on a behavioural paradigm known as the predictive-inference task. However, the psychometric quality of this task has never been examined, leaving unanswered whether it is indeed suited to capture behavioural variation on a within- and between-subject level. Using a large-scale test-retest design (N=330), we assessed the internal (internal consistency) and temporal (test-retest reliability) stability of the task's relevant measures. We show that while the main measures capturing flexible adaptation yield good internal consistency and overall satisfying test-retest reliability, more complex markers of flexible behaviour lack convincing psychometric quality. Our findings have implications for the large corpus of previous studies using this task and provide clear guidance as to which measures should and should not be used in future studies.

Introduction

Our ability to navigate the world depends on how successfully we respond to changes in our environment. In stable environments, we should rely on past experience to guide our actions and beliefs and ignore (noisy) current deviations. However, flexibility is essential when we are in dynamic environments and exposed to sudden changes. A dynamic response to the environment is thus a hallmark of adaptive behaviour.

Research endeavouring to understand this important ability has identified specific learning mechanisms and neurocognitive substrates underpinning it. To varying degrees, human participants have been shown to engage in normative learning that involves changing rates of information integration to master navigation through changing environments (e.g.,[1–5]). The human arousal system seems to play an important role in this dynamic process[2,3,6]. Moreover, the adaptation of neural representations in cortical and subcortical regions mirroring different characteristics of the environment and the internal state[5,7–10], as well as changes in the brain's functional connectivity[11], seem to be critical.

The most common task used to study such flexible adaptation is the predictive-inference task. Pioneered by Nassar and colleagues[1], it has been widely used in a number of variations[2–4,6–8,11–19]. In this paradigm, participants are asked to predict the next position of a target that lands in a similar location for several trials. However, on some trials, the landing location will suddenly shift to a completely new position. To perform well in the task, participants must adapt flexibly to these sudden changes by altering their behaviour based on the new information while ignoring the information they received before the change. Tracking participants' actions on each trial therefore allowed researchers to characterise learning[1,3,4,12,15,20], arousal[2,6], and neural mechanisms[7,8,11,19,21] in relation to these environmental changes.

This cognitive flexibility is particularly relevant to psychiatric research, as cognitive inflexibility has been associated with several psychiatric disorders[13,18,22–32]. For example, a study using this predictive-inference task showed that schizophrenic patients were prone to extreme forms of learning (i.e., little behavioural adaptation to new evidence and complete adaptation to it)[18] while patients with obsessive-compulsive disorder (OCD) have been seen to primarily over-emphasise new information at the cost of rashly discarding previously encountered evidence[13]. These results suggest that different mechanisms may underlie cognitive inflexibility in different patient populations. However, to be able to draw such inferences about individual differences, the paradigm used to investigate them has to be sound.

Despite the great popularity of this predictive-inference paradigm[2–4,6–8,11–17,33], to our knowledge, the psychometric properties of the task measures have not yet been systematically investigated. In the light of an ongoing replicability crisis in the field[34], to which inadequate psychometric qualities are a major contributor[35,36], this is concerning, yet not surprising as it is well aligned with a general neglect of psychometric investigations of cognitive tasks[37,38]. To test the foundation of the large body of research using this predictive-inference task, we believe it is thus critical to assess the psychometric quality of its measures.

The most prominent psychometric properties that a task must satisfy are internal consistency and test-retest reliability. Internal consistency quantifies the consistency of task measures across trials within participants during a single task execution[35,39]. High internal consistency stands for minimal confounding measurement noise. In contrast, test-retest reliability characterises the stability of a task measure within participants over time. That is, the stability of the measure between one task administration and a second administration that follows after a predetermined time interval[35,40]. This is essential when making inferences about stable neurocognitive traits and comparing variability between participants, such as in psychiatric or

pharmacological studies[41]. It is therefore crucial to assess both psychometric properties when drawing inferences about intra- and inter-individual differences[35,37,39,42].

In this study, we investigated the psychometric properties of this widely used predictive-inference task by conducting a large-scale, test-retest online study in the general UK public. Participants played the task twice over several months, which allowed us to quantify internal consistency and test-retest reliability. We show that while several 'raw' measures, such as confidence and learning rate before and after environmental changes, are mostly stable and reliable, others such as learning rate at the point of environmental change or the measures' associations with Bayesian model predictions, suffer from lower psychometric properties and should be used and interpreted with caution.

Results

To examine the psychometric properties of the predictive-inference task[1–4,8,11,12,15,16,18,19] (cf. Figure 1), we conducted an online study in which 219 participants played the circle-version of the task[13,14] at two time points (T1 and T2) approximately three months apart. In this task version, participants have to catch a particle that flies from a circle centre to its edge by placing a 'bucket' on this edge. Participants then rate their confidence as to whether their placed bucket will catch the particle before they see it flying to the edge. Most of the time, the particle falls into an area that is similar to the preceding falling positions (locations are sampled from a Gaussian distribution with a standard deviation of 12 degrees; cf. Methods). However, the location of this area occasionally and unexpectedly changes to any point on this circle (i.e., the mean of the Gaussian distribution shifts). These 'change-points' (CPs) thus generally indicate a larger shift in falling position. This also means that after a CP, participants have to discard

previous experiences and update their estimate about the falling-position based on the most recent experiences only.

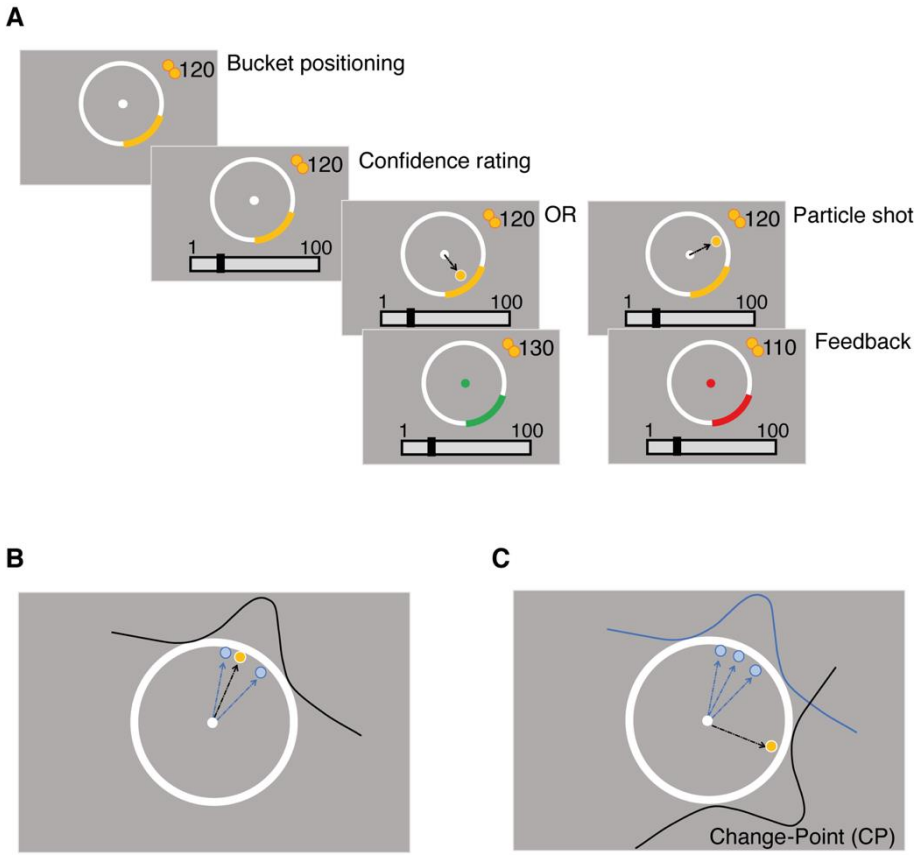


Figure 1. **Predictive-inference task.** Participants were instructed to place a bucket (yellow arc) on a circle edge to catch a flying particle and subsequently indicated their confidence in whether they would catch the particle. **(A)** Particles were fired from the centre of the circle to the edge. If the particle landed in the bucket, they won 10 points. If the particle missed the bucket, they lost 10 points. The particle landing location was sampled from a Gaussian distribution on every trial. Dashed arrow lines represent the particle trajectory in the current (black) and past (blue) trials (**B**). When a change-point (CP) occurred, the mean of the Gaussian distribution abruptly moved to another location on the circle (**C**).

To assess the psychometric properties of this task, we examined the internal consistency as well as the test-retest reliability of the most commonly used task metrics. Since the CPs are the critical experimental factor in this task, we computed all psychometric scores for metrics relative to CPs (i.e., before, at, and after CPs). Specifically, we first assessed the reliability of the behaviour-derived learning rate (LR^h) and the trial-by-trial confidence ratings. These measures are purely based on the participants' actions and ratings and are thus direct readouts of their behaviour. We subsequently probed more complex associations, such as the link between confidence and the LR^h and how these two measures are related to predictions of a quasi-optimal Bayesian learner.

Learning in a changing environment

To characterise participants' behavioural updating (positioning of the bucket) in this changing task environment, we computed trial-wise spatial learning rates for each participant (LR^h). On each trial, this LR^h was determined by the change of the bucket position from the current to the subsequent trial (i.e., action-update), divided by the prediction error they had made on this trial (distance between bucket edge and particle falling position; cf. Methods). In other words, it was defined by how much they changed their bucket position relative to the change in falling position of the particle.

We then looked at how LR^h fluctuated relative to CPs. In line with previous studies using different versions of the task[1,3,8,13], we found that CPs were associated with an increase in median LR^h (cf. Figure 2A). The LR^h then decreased and levelled off after a few trials. This shows that when participants encountered an unexpected particle location due to a CP, they reacted to the novel location information by updating the bucket position more. Thus, participants seem to appropriately react to large changes in the environment as novel

information has a higher influence on their actions while formerly acquired information becomes redundant.

Psychometric properties

We next tested the psychometric properties of the CP-related LR^h. To capture the stability of the measurement adequately relative to the CPs, we computed all psychometric estimates separately for each trial preceding CPs (4 to 1 trials before), the CP trial itself, and each trial following CPs (1 to 4 trials after). We measured internal consistency using Spearman-Brown corrected Pearson correlation (score categories: low: <0.5; moderate: ≥ 0.5 and <0.7; good: ≥ 0.7) and test-retest reliability using intraclass-correlation (ICC; score categories: low: <0.5; moderate: ≥ 0.5 and <0.75; good: ≥ 0.75; cf. Methods).

Internal Consistency

We first estimated how consistent the LR^h was within individuals at each time point and found that the internal consistency at all trials was high at T1 and T2 ($r_{SB} \geq 0.716$; cf. Figure 2B, Supplemental Table 1 and Supplemental Figure 1A). This means that the LR^h and its fluctuation relative to environmental changes were consistent within participants, supporting the notion that this task can adequately capture how humans adapt to changing environments.

Because internal consistency is critically dependent on measurement noise, which in turn is affected by the number of trials in a task, we conducted an additional analysis investigating the number of change-points that were necessary to reach a satisfying consistency. To do so, we assessed the internal consistency under a reduced number of CPs (i.e., from 6 CPs until full task length with 24 CPs; cf. Supplemental Information). We found that on trials before the CP the LR^h reached good internal stability (i.e., $r_{SB}=0.70$) after ~18 CPs, but on trials after and at the CP the LR^h did not reach good stability until the ~23 CPs (cf. Supplemental Figure 3A-B).

This suggests that with our hazard rate (H) of 0.125 (cf. below for further discussion), at least ~180 trials (totalling in 23 CPs) are required to reach a good internal consistency of the LR^h.

Test-Retest Reliability

We next estimated how much the LR^h scores from the first time point corresponded to the LR^h at the second time point approximately 3 months later, as indicated by their intra-class correlation (ICC). This analysis showed that the test-retest reliability at the trials before and after the CPs of the LR^h was predominantly good ($ICC \geq 0.761$) except for the third trial before ($ICC = 0.742$, 95% *CI* [0.677, 0.794]) and the first trial after CPs ($ICC = 0.749$, 95% *CI* [0.686, 0.799]) which were just below this threshold (cf. Figure 2C; cf. Supplemental Table 1). Thus, LR^h before and after CPs were consistent from T1 to T2, meaning that these measures seem well suited for studying individual differences.

However, the LR^h at the CPs itself had a low test-retest reliability ($ICC = 0.140$, 95% *CI* [-0.075, 0.312]). To examine this further, we looked at the total variance as decomposed for the calculation of the ICC-score and observed that the error variance was high (92%) while the between-participant variance was only 8% and between-time-point variance marginal (0.3%). Further inspection of the LR^h at the CP revealed that all LR^hs were very close to 1 for all participants, meaning that participants had very similar LR^hs at the CPs (cf. Supplemental Figure 2A). This is because all participants adapted their bucket position similarly strongly after a CP, which may be expected in this task. However, this also means that because the behaviour is very similar across participants, the CP-LR^h should not be used to assess individual differences.

We again examined how this was linked to the number of trials and found that test-retest reliability of the trials before and after the CPs reached a moderate level ($ICC = 0.50$) after ~10 CPs and most of them a good level ($ICC = 0.75$) at the ~21st CP (cf. Supplemental Figure 3C-

D), which suggests future LR^h investigations should implement a similar number of CPs as done here.

Overall, we found that the LR^h before and after CPs were consistent across time, given the implemented number of CPs, but the LR^h at CP itself was too homogeneous and therefore not capable of differentiating between individuals.

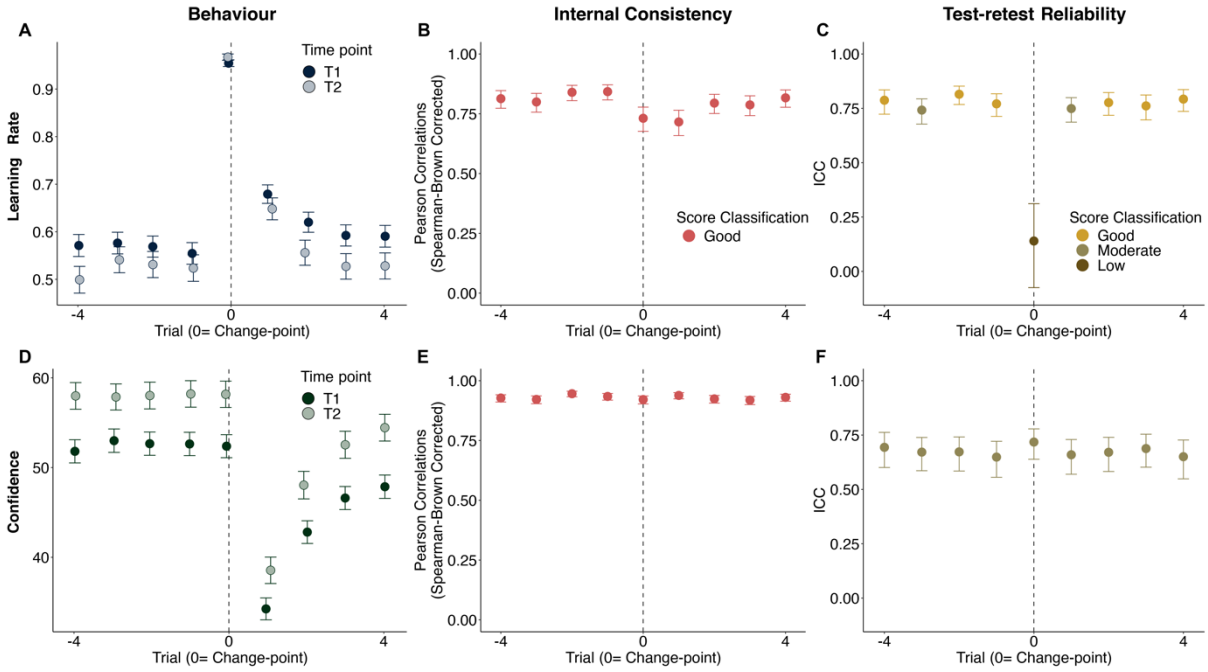


Figure 2. **Behavioural and psychometric properties of human learning rate and confidence.** Participants' learning rates (LR^h) at time point 1 (T1; N_{T1}=330) and 2 (T2; N_{T2}=219) were highest at the change-point (CPs; vertical dashed line) and decreased afterwards back to their pre-CP levels (**A**). Accordingly, confidence ratings dropped after a CP and increased back to pre-CP levels afterwards (**D**). Internal consistency (Spearman-Brown corrected Pearson correlations) for the median LR^h was good at all investigated trials before, at, and after CPs (here displayed for T1; **B**). Test-retest reliability measured by ICC-scores between median LR^hs at T1 and T2 were mostly good before and after the CP but low at the CP itself (**C**). Median confidence showed good internal consistency (**E**) and moderate test-retest reliability across all trials (**F**). Behaviour and internal consistency for both measures are here displayed for T1 only, but T2 statistics show similar results. Error bars represent standard errors in A & D. Error bars for the remaining plots represent the estimates 95% confidence interval.

Dynamic decision confidence

To measure participants' subjective uncertainty in their estimate (positioning of the bucket), we asked them to indicate at each trial how confident they were that their positioned bucket would catch the particle. Ideally, after a sudden change (CP), participants' confidence should be low as they have little evidence as to where the particle will fall next[1]. Once the environment is more stable again, and particles on subsequent trials fall into nearby positions, confidence in their estimate should increase.

Guided by this prediction, we thus investigated the development of confidence ratings relative to CPs. We indeed saw that CPs were followed by a drop in confidence, which then rose again as trials passed (cf. Figure 2D). This means, when a sudden change in the environment occurred and when participants had little information as to where the next particle would fall, they were uncertain as to whether their bucket would catch it. Subsequently, as they acquired more information regarding the new falling position with passing trials, they became more confident. This development has also been seen by previous studies using this task or adaptations of it (e.g.,[1,13]).

Psychometric properties

Internal Consistency

We analysed the internal consistency of confidence ratings in the same way as the LR^h described above. We found that the internal consistency of confidence was good across all trials, from four trials before to four trials after CPs, and at both time points (all $r_{SB} \geq 0.913$; cf. Figure 2E, Supplemental Table 1 and Supplemental Figure 1B). This level of internal consistency was, moreover, reached at all trials after only 7 CPs (cf. Supplemental Figure 3E-F), thereby only requiring approximately 56 trials with our hazard rate. This means, participants

rated their confidence consistently throughout the game, on trials before, at, as well as after the CPs, and good internal stability was reached quickly.

Test-Retest Reliability

We also investigated whether participants' confidence ratings were consistent across time, examining the ICC between T1 and T2 as done above. We found moderate test-retest reliabilities for all trials, before, at, and after CPs, ($ICC \geq 0.649$; cf. Figure 2F, Supplemental Figure 2B and Supplemental Table 1).

This level of test-retest reliability was reached after ~12 CPs for trials before and at the CPs and after ~13 CPs for trials after CPs (cf. Supplemental Figure 3G-H). However, none of these scores ever reached a good ICC-level (i.e., ≥0.75).

While this means that confidence is reliable enough to be used to compare individual differences, using more CPs than for LR^h-investigations which showed a better test-retest reliability (with the exception of the CP-trials) might be advisable.

Lower confidence is linked to a larger update of the bucket position

We went on to further characterise these opposing patterns of action and confidence around the CPs. Specifically, we assessed how the participants' action-update (i.e., the distance between the previous bucket position and chosen bucket position on a given trial; cf. Methods) was related to their confidence rating on a given trial. In line with previous studies[13,14], we chose these metrics because the above assessed LR^h is partially defined by the participants' action-update (relative to their prediction error on the preceding trial; cf. Methods). Using a summary statistics approach, we computed separate regression models for each participant, for both time points, which predicted confidence ratings on the basis of their action-update. The models showed that the size of the action-update was negatively predictive of confidence (T1: β=-

10.694, *SE*=0.314, $p<0.001$; T2: β=-11.752, *SE*=0.374, $p<0.001$; cf. Figure 3A). Thus, the more participants updated their bucket position, the less confident they were that it would catch the particle.

Internal consistency

To estimate the reliability of this negative relationship between action-update and confidence, we extracted the individual regression weights and computed their internal consistency and test-retest reliability. We found that these regression weights had a moderate internal consistency at both time points with the second time point being just below the score classification threshold (T1: r_{SB}=0.544, *95% CI* [0.463, 0.616]; T2: r_{SB}=0.495, *95% CI* [0.388, 0.589]; cf. Figure 3B). This means the link between confidence and action-updates was relatively stable within participants, within time points.

Test-retest reliability

We also computed the test-retest reliability and found that the association was also moderately reliable across time (*ICC*=0.603, *95% CI* [0.504, 0.682]; cf. Figure 3C). Thus, from one task administration to the other, the relationship between action-update and confidence was satisfactory stable, indicating that the association between the variables is still a suitable measure to investigate inter-individual differences.

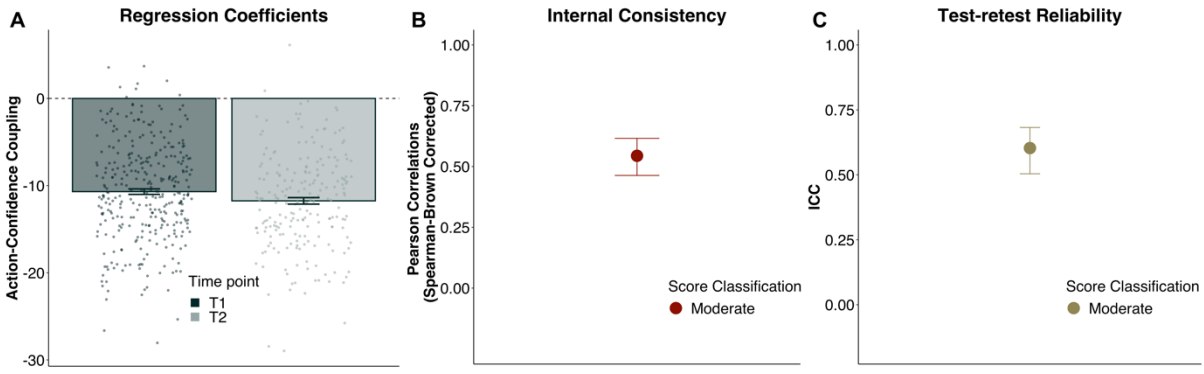


Figure 3. **Relationship between action-update and confidence.** We found a negative relationship between confidence and action-update (adaptation of the bucket position from one trial to the next) at both time points (N_{T1}=330; N_{T2}=219) showing that lower confidence was associated with larger action-updates **(A)**. This relationship showed a moderate internal consistency (Spearman-Brown corrected Pearson correlations; **B**) and test-retest reliability measured (ICC; **C**). Individual coefficients of participants are represented by circles while the bar plot represents the mean model coefficient, and their error bars represent standard errors in A. Internal consistency for all measures is here displayed for T1. Error bars for internal consistency and test-retest reliability represent the estimates 95% confidence interval.

Participants' behaviour is linked to normative factors of a Bayesian learner

To be able to relate humans' behaviour to a quasi-optimal behaviour in this task, we assessed how it was linked to a reduced quasi-optimal Bayesian learner[1,2,8] (cf. Methods and Supplemental Information). We extracted the model's normative factors, namely its prediction error PE^b (capturing the discrepancy between the model's estimation of the particle falling location and where it landed), the change-point probability (CPP; measuring the model's approximation of the probability that a CP has occurred), and the relative uncertainty (RU; the model's uncertainty about the mean of the Gaussian generating the particle landing location).

We investigated (again using a summary statistics approach) whether and how these normative factors were linked to participants' behavioural measures. Since the hazard rate and the SD of the generative distribution determining the particle landing location did not vary in our (and previous[8,13]) task implementation, PE^b and CPP were both merely driven by the mismatch

between the Bayesian learner's belief where the particle would fall and the actual particle landing location (cf. Methods and Supplemental Information)[8], and thus correlated highly (T1*:* $r(64723)=0.917$, $p<0.001$; T2: $r(42925)=0.915$, $p<0.001$). Based on this commonly observed co-linearity, we ran separate regression models for each of them (cf. Methods; similar and extended mixed-models with PE^b and CPP in a common model and demographic covariates are reported in the Supplemental Information). We additionally included RU and a 'Hit' predictor (a binary regressor indicating whether the particle was caught on the previous trial) in all models. Subsequently, we computed the internal consistency and test-retest reliability scores for participants' regression weights to examine how consistent the links between normative factors and behavioural measurements were within task runs and across time.

Confidence and the Bayesian learner

First, we conducted regressions predicting confidence on the basis of the Bayesian learner's predictions, i.e., the model's normative factors. We observed that participants' confidence was linked to all of the mentioned factors. PE^b (T1: $\beta_{PE}{}^b$=-0.215, SE=0.009, $p<0.001$; T2: $\beta_{PE}{}^b$ =-0.245, SE=0.012, $p<0.001$), CPP (β_{CPP}=-0.250, SE=0.011, $p<0.001$; T2: β_{CPP}=-0.289, SE=0.015, $p<0.001$) and RU (T1: $\beta_{PE\text{-}model}$=-0.147, SE=0.007, $p<0.001$; $\beta_{CPP\text{-}model}$=-0.187 , SE=0.009, $p<0.001$; T2: $\beta_{PE\text{-}model}$=-0.164, SE=0.010, $p<0.001$; $\beta_{CPP\text{-}model}$=-0.211, SE=0.012, $p<0.001$) negatively predicted confidence ratings, while Hit predicted it positively (T1: $\beta_{PE\text{-}model}$=0.269, SE=0.008, $p<0.001$; $\beta_{CPP\text{-}model}$=0.258, SE=0.008, $p<0.001$; T2: $\beta_{PE\text{-}model}$=0.281, SE=0.011, $p<0.001$; $\beta_{CPP\text{-}model}$=0.268, SE=0.011, $p<0.001$; cf. Figure 4A). Thus, participants' explicitly reported confidence was linked to normative factors, taking the normative magnitude of error, probability that a CP would occur, and overall uncertainty into account. Beyond that, confidence was also expectedly negatively associated with their own preceding accuracy.

Internal consistency

We then examined the internal consistency of the links between normative factors and confidence and saw that across predictors it was low to moderate (T1: PE^b: r_{SB}=0.210, *95% CI* [0.105, 0.312]; CPP: r_{SB}=0.133, *95% CI* [0.0254, 0.238]; $RU_{PE\text{-}model}$: r_{SB}=0.444, *95% CI* [0.353, 0.527]; $RU_{CPP\text{-}model}$: r_{SB}=0.452, *95% CI* [0.361, 0.534]; $Hit_{PE\text{-}model}$: r_{SB}=0.268, *95% CI* [0.165, 0.366]; $Hit_{CPP\text{-}model}$: r_{SB}=0.300, *95% CI* [0.199, 0.396]; T2: PE^b: r_{SB}=0.245, *95% CI* [0.116, 0.366], CPP: r_{SB}=0.156, *95% CI* [0.024, 0.283]; $RU_{PE\text{-}model}$: r_{SB}=0.532, *95% CI* [0.429, 0.620]; $RU_{CPP\text{-}model}$: r_{SB}=0.541, *95% CI* [0.440, 0.628]; $Hit_{PE\text{-}model}$: r_{SB}=0.539, *95% CI* [0.438, 0.627]; $Hit_{CPP\text{-}model}$: r_{SB}=0.526, *95% CI* [0.423, 0.616]; cf. Figure 4B). Thus, while the normatively estimated relative uncertainty of the task and the participants' own preceding accuracy (i.e., Hit) had a relatively stable influence on their confidence ratings throughout the task, the remaining normative factors PE^b and CPP did not seem to be associated with the confidence ratings in a consistent manner throughout the task. The latter finding could be explained by the varying size of the CPs which could potentially lead to differing associations of PE^b and CPP and confidence throughout the task.

Test-retest reliability

Next, we looked at the test-retest reliability of the regression weights of the normative factors predicting confidence. The ICC sores showed that the weights for RU and Hit were moderately reliable over time ($RU_{PE\text{-}model}$: *ICC*=0.593, *95% CI* [0.491, 0.674]; $RU_{CPP\text{-}model}$: *ICC*=0.595, *95% CI* [0.494, 0.676]; $Hit_{PE\text{-}model}$: *ICC*=0.603, *95% CI* [0.503, 0.682]; $Hit_{CPP\text{-}model}$: *ICC*=0.590, *95% CI* [0.487, 0.672]), only the ICC scores for PE^b and CPP were just below the moderate score classification threshold (CPP: *ICC*=0.488, *95% CI* [0.361, 0.590]; PE^b: *ICC*=0.406, *95% CI* [0.258, 0.525]; cf. Figure 4C). This means, from one time point to the next, the link between participants' confidence, certainty in their bucket position, and the Bayesian learner's relative

uncertainty was satisfyingly reliable. This makes these regressors suitable for investigations of inter-individual differences in the optimality of confidence fluctuations, but also shows that these links are likely to be less sensitive than the above, model-unrelated measures.

Action-update and the Bayesian learner

Having shown that participants' confidence ratings were linked to normative model predictions, we next investigated whether this was also the case for participants' action-update. Regression models predicting action-update on the basis of all model factors (with separate models for PE^b and CPP as above) showed that action-update was positively linked to PE^b (T1: β_{PE^b}=0.786, *SE*=0.008, p<0.001; T2: β_{PE^b}=0.851, *SE*=0.007, p<0.001), CPP (β_{CPP}=0.812, *SE*=0.008, p<0.001; T2: β_{CPP}=0.876, *SE*=0.007, p<0.001) and RU (T1: $\beta_{PE\text{-}model}$=0.016, *SE*=0.004, p<0.001; $\beta_{CPP\text{-}model}$=0.169; *SE*=0.004, p<0.001; T2: $\beta_{CPP\text{-}model}$=0.167, *SE*=0.005, p<0.001) although the link to RU did not prevail in the PE-model at the second time point ($\beta_{PE\text{-}model}$=-0.001, *SE*=0.005, p=0.851). Action-update was, moreover, negatively linked to Hit (T1: $\beta_{PE\text{-}model}$=-0.180, *SE*=0.004, p<0.001; $\beta_{CPP\text{-}model}$=-0.160, *SE*=0.004, p<0.001; T2: $\beta_{PE\text{-}model}$=-0.167, *SE*=0.004, p<0.001; $\beta_{CPP\text{-}model}$=-0.148, *SE*=0.004, p<0.001; cf. Figure 4D). Thus, the degree to which participants adapted the bucket position was increased when the model's prediction error (PE^b) was larger, its estimated likelihood that a CP had occurred was increased and when its relative uncertainty was higher, while the latter effect was less consistent. Participants also updated the bucket position more if their bucket position on the previous trial was wrong. This means, the participants' overall behaviour reflected the task dynamics captured by the quasi-optimal Bayesian learner.

Internal Consistency

We also examined the internal consistency of these links between action-update the normative factors and the Hit predictor. This showed that the internal consistency of all regression

weights, at both times points were low (T1: PE^b: r_{SB}=0.241, *95% CI* [0.137, 0.341]; CPP: r_{SB}=0.204, *95% CI* [0.098, 0.305]; $RU_{PE\text{-}model}$: r_{SB}=0.020, *95% CI* [-0.088, 0.128]; $RU_{CPP\text{-}model}$: r_{SB}=0.197, *95% CI* [0.091, 0.298]; $Hit_{PE\text{-}model}$: r_{SB}=0.351, *95% CI* [0.253, 0.442]; $Hit_{CPP\text{-}model}$: r_{SB}=0.373, *95% CI* [0.276, 0.463]; T2: PE^b: r_{SB}=0.247, *95% CI* [0.118, 0.368], CPP: r_{SB}=0.046, *95% CI* [-0.087, 0.177]; $RU_{PE\text{-}model}$: r_{SB}=0.205, *95% CI* [0.074, 0.328]; $RU_{CPP\text{-}model}$: r_{SB}=-0.003, *95% CI* [-0.136, 0.129]; $Hit_{PE\text{-}model}$: r_{SB}=0.183, *95% CI* [0.052, 0.309]; $Hit_{CPP\text{-}model}$: r_{SB}=0.203, *95% CI* [0.073, 0.327]; cf. Figure 4E). Thus, participants' behavioural update was not consistently linked to the normative factors or the Hit predictor throughout the task.

Test-Retest Reliability

Although just below the moderate score classification threshold, test-retest reliability of the regression weights capturing the link between normative factors and action-update were all low (PE^b: *ICC*=0.426, *95% CI* [0.276, 0.544]; CPP: *ICC*=0.438, *95% CI* [0.291, 0.553]; $RU_{PE\text{-}model}$: *ICC*=0.412, *95% CI* [0.267, 0.529]; $RU_{CPP\text{-}model}$: *ICC*=0.403, *95% CI* [0.254, 0.522]) and only the link to Hit was moderately stable (*PE-model*: *ICC*=0.608, *95% CI* [0.510, 0.686]); *CPP-model*: *ICC*=0.582, *95% CI* [0.478, 0.665]; cf. Figure 4F). This suggests that the model's normative factors' links to action-update have low reliability and are thus not appropriate for individual differences designs.

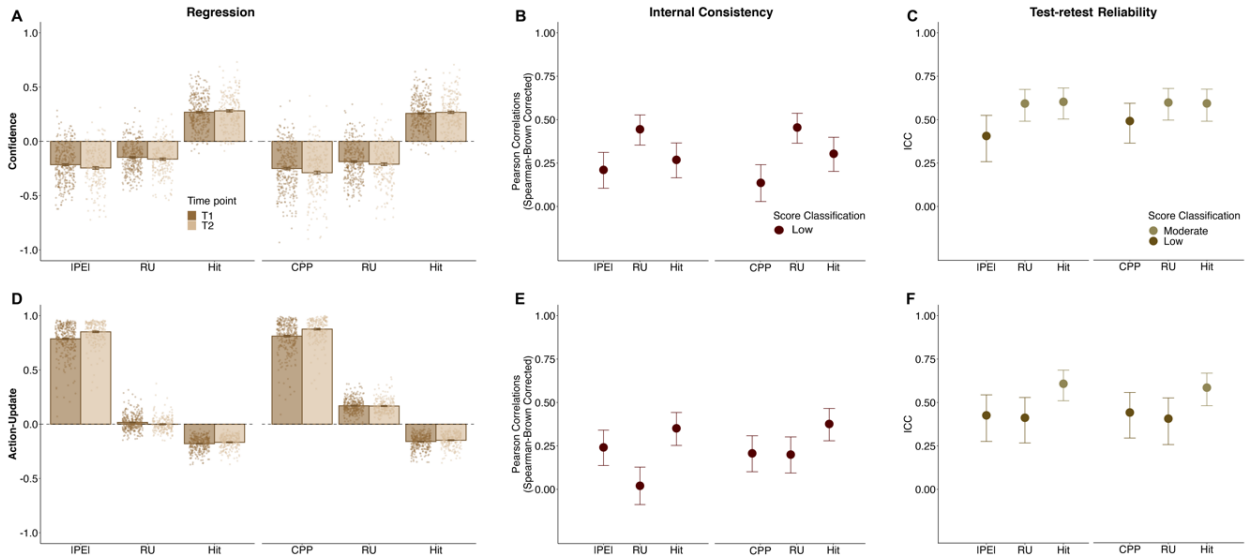


Figure 4. **Relationship between confidence, action and the Bayesian factors.** Regression models predicted trial-wise confidence ratings and action-updates based on the model-derived normative factors PE^b (model prediction error), CPP (change-point probability), RU (relative uncertainty) and Hit (accuracy on the previous trial). Separate regression models were run for PE^b (left-side of each plot) and CPP (right-side of each plot). Normative regressors negatively predicted confidence while accuracy (Hit) predicted it positively (**A**). Investigating the robustness of these associations, we saw that the internal consistency of the links between confidence and normative regression weights was low (Spearman-Brown corrected Pearson correlations; **B**). Test-retest reliability of all normative factor regression weights predicting confidence, except for PE^b and CPP, was moderate (as indicated by the ICC score; **C**). We found the reverse relationships for action-update, which was positively linked to PE^b, CPP, and RU and negatively linked to Hit (**G**). All links showed a low internal consistency (**H**). Except for the regression weight of Hit, all regression weights in the action-update model also showed low test-retest reliability (**I**). In A & D, individual coefficients of participants are represented by circles, while the bar plot represents the mean of the model coefficients, and their error bars represent standard errors. Internal consistency for all measures is here displayed for T1. Error bars for internal consistency and test-retest reliability represent the estimates 95% confidence interval in.

Associations between task measures and psychiatric dimension scores

Task reliability is of particular relevance when making inferences about inter-individual differences[42] and its lack may result in spurious or/and inconsistent findings. Previous studies have used the present predictive-inference task to investigate behavioural differences between OCD patients and controls[13] as well as along psychiatric dimensions[14]. Seow and Gillan[14]

showed that decreased action-confidence coupling (i.e., coupling of confidence and action-update) was associated with multiple psychiatric questionnaire scores such as obsessive-compulsive (OC), anxiety (before Bonferroni correction), and depressive symptoms in the general public. On a transdiagnostic level, the compulsivity dimension showed a selective negative association with the confidence-action coupling. A similar decoupling had previously been found in OCD patients[13]. Despite this apparent similarity, the driving factors were dissimilar across the studies. While the transdiagnostic dimension of compulsivity was linked to an elevated confidence[14], the OCD patient study identified higher learning rates as the underpinning factor[13]. As these findings seemed partially contradictory, we attempted to replicate them in our sample at the second time point.

In our data, we did not observe any link between any psychiatric symptom score and action-confidence coupling. None of the psychiatric scores were associated with the beta values gained from the regression linking confidence and action-update (OC symptoms: r_S=-0.004, p=0.951; anxiety: r_S=-0.010, p=0.883; depression: r_S=0.015, p=0.825).

Similarly, none of the psychiatric variables were correlated with mean confidence (OC symptoms: r_S=-0.025, p=0.716; anxiety: r_S=-0.091, p=0.182; depression: r_S=0.085, p=0.208) or action-update itself (OC symptoms: r_S=0.028, p=0.678; anxiety: r_S=0.017, p=0.798; depression: r_S=-0.024, p=0.723).

We also repeated the regression models predicting confidence and action-update based on normative factors, now additionally including the psychiatric variables as predictors. Previous literature had shown a decreased effect of CPP on confidence in individuals scoring high on the transdiagnostic compulsivity dimension[14] and OCD patients[13]. We did not find any significant correlation with the CPP beta weight (OC symptoms: r_S=0.008, p=0.906; anxiety: r_S=-0.001, p=0.983; depression: r_S=0.010, p=0.879). We also did not find any association

between OC scores and the PE^b beta weights of the action-update model (r_S=0.093, p=0.169) This means, in our sample, the psychiatric questionnaire scores tested were not associated with the link between confidence and the normative model predictions.

Finally, we investigated whether any of the psychiatric symptom scores were linked with the mean LR^h, probing previous findings of an increased LR^h in OCD patients[13] in our general public sample. However, none of the psychiatric symptoms showed an association with the mean LR^h (OC symptoms: r_S=-0.013, p=0.852; anxiety: r_S=0.037, p=0.582; depression: r_S=-0.039, p=0.566).

Given the data was collected during the COVID-19 pandemic, which had been seen to be associated with increased OC symptoms in the general public (e.g.,[43–45]), we conjectured that this general increase might have affected our replication attempt (we saw a trend-level increase in OCI-R scores compared to previous studies[14], t(463.92)=1.844, p=0.066). When repeating analysis using an OC symptom score that excluded all items that could have been influenced by the pandemic, our results did not change (cf. Supplemental Information).

Discussion

Across cognitive neuroscience and computational psychiatry, low reproducibility challenges the interpretability and generalisability of neurocognitive findings[46–48]. A particular challenge is the threat of poor psychometric properties of behavioural task measures, which can have considerable ramifications for both within- and between-subjects inference. In the present study, we have thus examined the psychometric properties, in particular the internal consistency and test-retest reliability, of a widely used predictive-inference task[2–4,6–8,11–17,33] using a large-scale, re-test online sample.

When assessing the most commonly used behavioural measures, namely learning rate (LR^h) and confidence before, at, and after environmental changes (CPs), we found a good overall internal consistency of both measures and show that current tasks could even be shortened while retaining sufficient consistency. We also found that test-retest reliability was mostly moderate to good for both measures, with the exception of a low reliability for LR^h at CPs due to low between-participant variability. This suggests that the latter should not be used in between-participant variability designs whilst the remainders are well suited.

Importantly, the psychometric properties of other common but more complex task measures were substantially worse. Especially those metrics that are often used to assess the optimality of behaviour or the coupling between behaviour and subjective experience were substantially less stable and reliable. This is specifically the case for the measures that link confidence, action-update and predictions from a quasi-optimal Bayesian model. Our findings thus demand caution when using such measures to draw conclusions about neurocognitive mechanisms.

The good internal consistency of the two main 'raw' measures, LR^h and confidence, indicates that the task captures behavioural and belief adaptations to sudden changes well[49]. This stability is an important pre-requisite when considering e.g., the effect of experimental manipulations on task performance[3] or associations within an individual such as the link between LR^h and confidence[1]. A similar picture emerges when investigating the test-retest reliability. The mostly moderate to good test-retest reliability of the 'raw' measures is particularly important for studies that showed alterations in these measures in e.g., schizophrenic[18] and OC individuals[13] or use such metrics in combination with pharmacological or other interventions (e.g.,[3]). Our results thus validate many of the previous findings using these metrics and suggest that these measures can indeed be used with great confidence.

However, we observed unsatisfactory internal consistency and re-test reliability for most of the more complex measures. Interpreting such metrics should thus be cautioned against, especially when assessing between-participants differences. Individual-difference studies would either require large(r) sample sizes or improved measures.

There are multiple factors that may have contributed to these findings. Firstly, these complex measures are interactions between multiple noisy task measures, which is known to lead to a larger overall measurement noise[50,51]. Secondly, we found that multiple model-derived predictors showed a high degree of co-linearity and thus directly affected how well the impact of these metrics could be measured when used in the same model (cf. Supplemental Information for an assessment across different modelling approaches). This co-linearity is to some extent inherent in the Bayesian model. However, if task settings, such as the hazard rate and the spread of the particle-generating Gaussian distribution (i.e., SD) varied across the experiment (e.g.,[1,11,14]), this apparent co-linearity can be reduced, and thus may be ameliorated in future studies. However, such changes do influence the task (perception) directly and may capture different (additional) cognitive processes. Future studies should investigate how changing these task settings impacts the psychometric properties of these complex task measures.

The low psychometric properties of the links between behaviour and the Bayesian learner may also be the reason why we did not find any association with OC symptoms. However, we also did not observe any links to the simpler task measures, such as mean confidence, as observed previously[13,14]. This may also be due to other differences between the studies such as the definition of the psychiatric variable (i.e., questionnaire score versus clinical diagnosis versus cross-questionnaire factors), the sample's cultural background (i.e., British versus American samples), recruiting sources or other potentially unaccounted variables.

It is important to highlight that the interval between test and re-test can substantially affect reliability estimates. Our time interval (~3 months) between testing sessions was on the timescale of many studies on behavioural change. While this underpins the robustness of the reliable measures, it should also be noted that a shorter interval might increase these estimates[40,52]. In addition, future studies should examine whether a lab-based version or one with different settings may hold different psychometric properties. However, given that they are believed to measure the same constructs, a large divergence between task types would be worrying. Lastly, as mentioned above, the ongoing Covid-19 pandemic might have altered behavioural as well as psychiatric measures. While we controlled for pandemic-related items in our psychiatric questionnaires, behavioural measures such as e.g., confidence might have been affected in subtle ways we could not account for.

In conclusion, we show that the main measures of this predictive-inference task consistently and reliably capture belief and behavioural adaptations before and after environmental changes, making them suitable for studies investigating individual-differences. We also point out that the complex links between task variables and model predictions are of mostly low psychometric quality and should only be used with caution. Our findings thus highlight the importance of a thorough assessment of the measures' psychometric quality and that such properties can differ widely even within the same task.

Methods

Procedure

We conducted a large-scale online study in which participants played the predictive-inference task[1] on two occasions. Data collection for the first time point (T1) took place at the end of April/ beginning of May 2020. At T1, participants reported their demographics and completed

a cognitive ability assessment whose total score served as a proxy for IQ[53] before playing the task. We re-contacted participants to play the task again between mid-July and mid-August 2020 (time point 2; T2). They also reported their obsessive-compulsive (OC), anxiety and, depressive symptoms (cf. Questionnaires below)[54] at T2. The mean time difference between T1 and T2 per participant was 81 days (min=69 days, max=104 days).

Sample size estimation

As there was no prior data on the estimated reliability of the predictive-inference task, we powered our sample to be sensitive to the (relatively modestly sized) associations between psychiatric measures and task indicators. We based our power analysis (conducted in G*Power[55]) on the correlation between mean task confidence and OC symptoms in the sample reported by Seow and Gillan[14] on trials that had the same hazard rate as we used here (H=0.125; r=0.272, p<0.001). This analysis suggested our study required N=190 participants to achieve 90% power at a p=0.01 significance level.

Participants

We recruited participants that were over 18 and living in the UK via the Prolific recruiting service (https://www.prolific.co/). All participants gave written informed consent online before starting the study and received £8.25/hour plus a bonus of up to £1 based on their performance in this task. The study was approved by UCL's Research Ethics Committee (15301/001).

A total of 401 participants completed the study at T1. We excluded 19 participants because they failed at least one of two attention checks (instructed questionnaire answers). We additionally applied task-based exclusion criteria in line with previous studies[13,14] to ensure good task data quality (cf. Supplemental Information). Based on these criteria, 52 additional participants were excluded. This resulted in a final T1 sample of 330 participants (187 females;

age: 34y±12.2). The T1 sample was larger than the sample size suggested by the power estimation as we expected attrition across both time points.

Participants from T1 were re-invited to take part in T2, from whom 260 completed the study a second time resulting in an overall retention rate of 79%. We again excluded participants because of failed attention checks (*N*=14) and inconsistent game completion (*N*=27). This resulted in a final T2 sample of 219 participants (119 females; 35y±12.3).

Material

We implemented all questionnaires using a web API programmed with React JS libraries (https://reactjs.org/). The predictive-inference task was implemented in JavaScript.

The predictive-inference task

Participants played the predictive-inference task (also known as 'helicopter task') pioneered by Nassar and colleagues[1–4,8,11,15,16,18,19]. We utilised a circular online version by Seow and Gillan[14], an adaptation of the in-lab version utilized by Vaghi et al.[13] (cf. Figure 1). We instructed participants to catch a particle flying from a circle centre to its edge by moving and placing a 'bucket' on this edge. Participants used the left and right arrow keys to indicate the bucket's position and the spacebar to confirm it. After placing the bucket, a confidence scale (ranging from 1-100) appeared below the circle. The confidence indicator was initially randomly placed at either 25 or 75, and participants adjusted it to report how confident they were that the particle would land in the positioned bucket. Subsequently, a particle flew from the circle centre to the edge (feedback). If the particle landed within the bucket, the bucket turned green for 500ms, and the participant gained 10 points. If the particle landed outside the bucket, it turned red, and the participant lost 10 points. Accumulated points were presented in the upper-right corner. These points were converted into a bonus payment at the end of the task

(maximum £1). Confidence ratings were not incentivized, however, to ensure active usage of the scale the task was reset if participants left their confidence ratings at the default score for more than 70% of the first 50 trials.

On each trial, the particle's landing location on the circle edge was sampled independently from a Gaussian distribution with a standard deviation of 12 degrees. The mean of this distribution remained the same until a CP trial occurred. At the CP, the mean of the generative distribution determining the landing position shifted to a new position. This new position of the mean was re-sampled from a uniform distribution U(1,360) (i.e., number of points on the circle; cf. Figure 1C). The probability of a CP occurring on each trial was determined by the hazard rate of 0.125 and was stable throughout the entire task. Participants had to learn into which part of the circle edge the particle fell after a CP had occurred and adapt their behaviour. They completed 200 trials in total (which results in approximately 24 CPs), divided into four blocks of 50 trials that were divided by self-timed short breaks.

To ensure that participants understood the task, they had to complete 10 practice trials after reading the instructions. They then had to answer five questions on the task instructions correctly before starting the actual game. They were asked to start again from instructions if they failed at least one of these questions.

Questionnaires

We assessed OC symptoms using the Obsessive-Compulsive Inventory-Revised (OCI-R)[56], as used by previous studies investigating the present task[14]. To control for potential confounds in the OCI-R score caused by the Covid-19 pandemic that was taking place during data collection[43–45], we computed an additional score controlling for items of high relevance to Covid-19 (cf. Supplemental Information). We additionally measured anxiety and depression

symptom scores using the Hospital Anxiety and Depression Scale (HADS)[54]. The HADS consists of two subscales (anxiety, depression) that are evaluated separately.

Quantitative Analysis

We pre-processed and analysed data in MATLAB 2021a (MathWorks) and R, version 3.6.2 via RStudio version 1.2.5033 (http://cran.us.r-project.org).

Computational models

Learning rate characterising human behaviour

Following approaches in the literature, we characterised participants' behaviour by computing a human prediction error (PE^h), the distance between the centre of the bucket (b) and the location where the particle landed (X) at trial t

$$PE_t^h = X_t - b_t \quad (1)$$

and a human learning rate defined as the change in bucket position from trial t to t+1 divided by the PE^h on trial t (LR^h):

$$LR_t^h = \frac{b_{t+1} - b_t}{PE_t^h} \quad (2)$$

A LR^h of 1 would here mean that newly encountered information (most recent particle falling position) would overshadow preceding information.

Normative factors characterising task fluctuations

We also characterised the task dynamics using a reduced quasi-optimal Bayesian learner as used in previous studies[8]. This model is applied to task data experienced by participants and adjusts its predictions Bayes-optimally. Normative factors derived from the learner capture

task characteristics thought to influence participants' behaviour and confidence ratings in a changing (task) environment: For each trial, the resulting factors captured (1) the discrepancy between the Bayesian learner's belief about the location of the new particle and its actual landing position (model prediction error; PE^b), (2) the Bayesian learner's approximation of the probability that a change-point has occurred (change-point probability; CPP) and (3) relative uncertainty in the learner's belief about the mean of the generative distribution determining the particle landing location (relative uncertainty; RU). The reduced quasi-optimal Bayesian learner and all normative factors are described in detail in the Supplemental Information.

Following previous studies using this task[13,14], trials where the LR^h exceeded the 99th percentile of all participants LR^h or where PE^h=0 were assumed to be unrelated to error-driven learning[15] and were thus excluded from the analyses that were based on normative factors derived from the Bayesian learner.

Psychometric Properties

As the processing of environmental changes lies at the core of this task, reliability of measurements relative to these changes (CPs) is of most importance. We, therefore, computed all psychometric scores relative to CPs to ascertain whether measures robustly captured responses to environmental changes. We also investigated how many CPs participants had to experience until the main variables reached a satisfying level on both psychometric measures (cf. Supplemental Information).

Internal consistency

We used the split-half approach[39] to estimate the internal consistency of the task. For each time point, we split the task data into two sub-datasets (odd and even) and computed all measurements for them separately. To avoid having unequal numbers of CPs in one versus the

other sub-dataset, we created the sub-datasets by splitting the full dataset based on CPs and their associated trials (i.e., all trials following a CP before the next CP occurred).

We measured internal consistency using Pearson correlation corrected with the Spearman-Brown formula[57,58]. The Spearman-Brown correction was computed using the *psych* package in R[59]. It extrapolates the reliability estimation from the length of the sub-dataset to the length of the entire dataset[60]. In line with conventions in the field[50], we categorized internal consistency coefficients below 0.5 as 'low', coefficients between 0.5, and 0.7 as 'moderate' and coefficients above 0.7 as 'good'.

When estimating the internal consistency of behavioural measurements of interest (i.e., LR^h and confidence), we assessed the internal consistency on all trials ranging from 4 trials before to 4 trials after CPs individually. For the sake of completeness, we additionally computed internal consistency across all CP relative trials. As LR^h and confidence were highly skewed, we estimated the consistency of their median score.

We computed several regression models to investigate links between behavioural measurements and psychiatric dimension scores as well as normative factors derived from the Bayesian learner. To estimate the internal consistency of all random slope regression coefficients, we computed two regression models, one entailing odd and one entailing even CPs and their associated trials. We could then estimate the extent to which these two regression models revealed the same associations.

Test-retest reliability

We estimated the task's temporal (test-retest) reliability using ICC as implemented in the R *psych* package[59]. The ICC can estimate the agreement between measurements while capturing differences in means of the compared scores (e.g., systematic shifts due to training effects),

which would not be accounted for by other measures such as Pearson correlation[61]. The ICC is the ratio of variability between participants to the total variability, including participant and error variability. It therefore also enabled us to extract the between-participant and between-time-point variance and compare it to the error variance. Measures with low between-participant relative to between-time-point (within-participant) variability may thereby be considered suitable for experimental designs, but unsuitable for correlational or individual-difference approaches[42]. As per recommendations by Shrout and Fleiss[62], we chose the ICC(2k) score (two-way random-effects model). However, in our sample, ICC(2k) scores were very similar to ICC(3k) scores, leading to the same conclusions. Following the conventional criterion[61], we categorized ICC scores below 0.5 as 'low', between 0.5 and 0.75 as 'moderate', and above 0.75 as 'good'.

To estimate the test-retest reliability of behavioural measurements (LR^h and confidence) we again examined the four trials preceding CPs, the trial of the CPs themselves and the trials right after the CPs. We also reported the ICC score across all these trials for completeness. For each trial and measurement, we looked at the ICC between the median score at T1 and the median score at T2.

We investigated the stability of the different regressions' weights within participants across time by computing the ICC between the random slope coefficients at T1 and T2. To do so, we re-run the T1-models only including participants that had also completed T2.

Regression and correlation models

To capture the association between task variables over trials, within-participants we computed the measures the repeated-measures correlation using the *rmcorr* package[63] in R. This method allows to control for repeated-measures within each participant[64].

To further characterize the link between task variables, normative factors, behavioural (i.e., action-update and confidence), and psychiatric (i.e., OC, anxiety and depressive symptoms) measures, we ran regression models, that were fit separately for each participant, with mean coefficients then tested against zero at the group level using a two-sided t-test. As all normative factor regressors used to predict action-update were linearly predictive of PE^b, they were all (except for PE^b itself) implemented as an interaction effect with PE^b in the model following previous research[1,8,13–15]. Moreover, we ran separate regression models for PE^b and CPP considering their co-linearity (cf. Supplemental Information). However, we also reported control models with both predictors combined in one model in the Supplemental Information. All task variables were z-scored across trials within participants to ensure comparability of regression coefficients.

Declaration of interest

All authors declare no conflicts of interest.

Acknowledgements

We thank Dr Vasilisa Skvortsova for implementing the predictive-inference task online and Dr Matilde Vaghi for comments on the initial draft of this manuscript. We also thank Dr Chang-Hao Kao for analyses feedback and suggestions.

A.M.L. is a pre-doctoral fellow of the International Max Planck Research School on Computational Methods in Psychiatry and Ageing Research (IMPRS COMP2PSYCH). The participating institutions are the Max Planck Institute for Human Development, Berlin, Germany, and University College London, London, UK. For more information, see: https://www.mps-ucl-centre.mpg.de/en/comp2psych. T.X.F.S. is a post-doctoral fellow at the Max Planck UCL Centre for Computational Psychiatry and Ageing Research, which is a joint initiative supported by UCL and the Max Planck Society. She is supported by a Sir Henry Wellcome Postdoctoral Fellowship (224051/Z/21/Z) from the Wellcome Trust (https://wellcome.org/grant-funding/schemes/sir-henry-wellcome-postdoctoral-fellowships). T.U.H. is supported by a Sir Henry Dale Fellowship (211155/Z/18/Z; 211155/Z/18/B; 224051/Z/21) from Wellcome & Royal Society (https://wellcome.ac.uk/funding/sir-henry-dale-fellowships), a grant from the Jacobs Foundation (2017-1261-04; https://jacobsfoundation.org/activity/jacobs-foundation-research-fellowship-program/), the Medical Research Foundation (https://www.medicalresearchfoundation.org.uk/what-we-fund), a 2018 NARSAD Young Investigator grant (27023) from the Brain & Behavior Research Foundation (https://www.bbrfoundation.org/grants-prizes/narsad-young-

investigator-grants), and a Philip Leverhulme Prize from the Leverhulme Trust (PLP-2021-040; https://www.leverhulme.ac.uk/philip-leverhulme-prizes). This project has received funding from the European Research Council (ERC) under the European Union's Horizon 2020 research and innovation programme (grant agreement No 946055; https://erc.europa.eu/funding/starting-grants). The Wellcome Centre for Human Neuroimaging is supported by core funding from the Wellcome Trust (203147/Z/16/Z). This research was funded in whole, or in part, by the Wellcome Trust (211155/Z/18/Z). For the purpose of Open Access, the author has applied a CC BY public copyright licence to any Author Accepted Manuscript version arising from this submission. The funders had no role in study design, data collection and analysis, decision to publish, or preparation of the manuscript.

Authors' contribution

T.U.H. and A.M.L. conceptualized the study. A.M.L. developed the methodology under the supervision of T.U.H. and with help from T.X.F.S.. The study was conducted by A.M.L. who also analysed the data under supervision of T.U.H. and with input from T.X.F.S. A.M.L. and T.U.H. wrote the first draft of the manuscript which was revised and edited by T.X.F.S.

Data sharing statements

Fully anonymised data and code for data analysis of this study will be available from a dedicated Github repository (https://github.com/DevComPsy) upon peer-reviewed publication.

References

1. Nassar, M. R., Wilson, R. C., Heasly, B. & Gold, J. I. An Approximately Bayesian Delta-Rule Model Explains the Dynamics of Belief Updating in a Changing Environment. *Journal of Neuroscience* **30**, 12366–12378 (2010).
2. Nassar, M. R. *et al.* Rational regulation of learning dynamics by pupil-linked arousal systems. *Nat Neurosci* **15**, 1040–1046 (2012).
3. Jepma, M. *et al.* Catecholaminergic Regulation of Learning Rate in a Dynamic Environment. *PLOS Computational Biology* **12**, e1005171 (2016).
4. Bruckner, R., Nassar, M. R., Li, S.-C. & Eppinger, B. Default beliefs guide learning under uncertainty in children and older adults. (2020) doi:10.31234/osf.io/nh9bq.
5. Behrens, T. E. J., Woolrich, M. W., Walton, M. E. & Rushworth, M. F. S. Learning the value of information in an uncertain world. *Nature Neuroscience* **10**, 1214–1221 (2007).
6. Krishnamurthy, K., Nassar, M. R., Sarode, S. & Gold, J. I. Arousal-related adjustments of perceptual biases optimize perception in dynamic environments. *Nature Human Behaviour* **1**, 1–11 (2017).
7. Nassar, M. R., McGuire, J. T., Ritz, H. & Kable, J. W. Dissociable Forms of Uncertainty-Driven Representational Change Across the Human Brain. *J. Neurosci.* **39**, 1688–1698 (2019).
8. McGuire, J. T., Nassar, M. R., Gold, J. I. & Kable, J. W. Functionally Dissociable Influences on Learning Rate in a Dynamic Environment. *Neuron* **84**, 870–881 (2014).
9. Pearson, J. M., Heilbronner, S. R., Barack, D. L., Hayden, B. Y. & Platt, M. L. Posterior cingulate cortex: adapting behavior to a changing world. *Trends in Cognitive Sciences* **15**, 143–151 (2011).

10. O'Reilly, J. X. *et al.* Dissociable effects of surprise and model update in parietal and anterior cingulate cortex. *PNAS* **110**, E3660–E3669 (2013).
11. Kao, C.-H. *et al.* Functional brain network reconfiguration during learning in a dynamic environment. *Nature Communications* **11**, 1682 (2020).
12. Razmi, N. & Nassar, M. R. Adaptive learning through temporal dynamics of state representation. *bioRxiv* 2020.08.03.231068 (2020) doi:10.1101/2020.08.03.231068.
13. Vaghi, M. *et al.* Compulsivity Reveals a Novel Dissociation between Action and Confidence. *Neuron* **96**, 348-354.e4 (2017).
14. Seow, T. X. F. & Gillan, C. M. Transdiagnostic Phenotyping Reveals a Host of Metacognitive Deficits Implicated in Compulsivity. *Scientific Reports* **10**, 1–11 (2020).
15. Nassar, M. R. *et al.* Age differences in learning emerge from an insufficient representation of uncertainty in older adults. *Nature Communications* **7**, 11609 (2016).
16. Nassar, M. R. & Troiani, V. The stability flexibility tradeoff and the dark side of detail. *Cogn Affect Behav Neurosci* (2020) doi:10.3758/s13415-020-00848-8.
17. Nassar, M. R. & Frank, M. J. Taming the beast: extracting generalizable knowledge from computational models of cognition. *Current Opinion in Behavioral Sciences* **11**, 49–54 (2016).
18. Nassar, M. R., Waltz, J. A., Albrecht, M. A., Gold, J. M. & Frank, M. J. All or nothing belief updating in patients with schizophrenia reduces precision and flexibility of beliefs. *Brain* **144**, 1013–1029 (2021).
19. Nassar, M. R., Bruckner, R. & Frank, M. J. Statistical context dictates the relationship between feedback-related EEG signals and learning. *eLife* **8**, e46975 (2019).
20. Ritz, H., Nassar, M. R., Frank, M. J. & Shenhav, A. A Control Theoretic Model of Adaptive Learning in Dynamic Environments. *Journal of Cognitive Neuroscience* **30**, 1405–1421 (2018).

21. Kao, C.-H., Lee, S., Gold, J. I. & Kable, J. W. Neural encoding of task-dependent errors during adaptive learning. *eLife* **9**, e58809 (2020).
22. Britton, J. C. *et al.* Cognitive Inflexibility and Frontal-Cortical Activation in Pediatric Obsessive-Compulsive Disorder. *Journal of the American Academy of Child & Adolescent Psychiatry* **49**, 944–953 (2010).
23. Ceaser, A. E. *et al.* Set-shifting ability and schizophrenia: a marker of clinical illness or an intermediate phenotype? *Biol. Psychiatry* **64**, 782–788 (2008).
24. Chamberlain, S. R., Fineberg, N. A., Blackwell, A. D., Robbins, T. W. & Sahakian, B. J. Motor Inhibition and Cognitive Flexibility in Obsessive-Compulsive Disorder and Trichotillomania. *AJP* **163**, 1282–1284 (2006).
25. Chamberlain, S. R. *et al.* Impaired Cognitive Flexibility and Motor Inhibition in Unaffected First-Degree Relatives of Patients with Obsessive-Compulsive Disorder. *Am J Psychiatry* **164**, 335–338 (2007).
26. Geurts, H. M., Corbett, B. & Solomon, M. The paradox of cognitive flexibility in autism. *Trends Cogn Sci* **13**, 74–82 (2009).
27. Gu, B.-M. *et al.* Neural correlates of cognitive inflexibility during task-switching in obsessive-compulsive disorder. *Brain* **131**, 155–164 (2008).
28. Hauser, T. U. *et al.* Increased fronto-striatal reward prediction errors moderate decision making in obsessive–compulsive disorder. *Psychological Medicine* **47**, 1246–1258 (2017).
29. Loosen, A. & Hauser, T. U. Towards a Computational Psychiatry of Juvenile Obsessive-Compulsive Disorder. *Neuroscience & Biobehavioral Reviews* (2020).
30. Morice, R. Cognitive inflexibility and pre-frontal dysfunction in schizophrenia and mania. *Br J Psychiatry* **157**, 50–54 (1990).

31. Zhu, C., Kwok, N. T., Chan, T. C., Chan, G. H. & So, S. H. Inflexibility in Reasoning: Comparisons of Cognitive Flexibility, Explanatory Flexibility, and Belief Flexibility Between Schizophrenia and Major Depressive Disorder. *Front Psychiatry* **11**, (2021).
32. Skvortsova, V. & Hauser, T. Distinct computational mechanisms underlying cognitive flexibility deficits in impulsivity and compulsivity. (2022) doi:10.21203/rs.3.rs-1280535/v1.
33. Kao, C.-H., Lee, S., Gold, J. I. & Kable, J. W. Neural encoding of task-dependent errors during adaptive learning. *bioRxiv* 2020.05.11.089094 (2020) doi:10.1101/2020.05.11.089094.
34. Collaboration, O. S. Estimating the reproducibility of psychological science. *Science* **349**, (2015).
35. Matheson, G. J. We need to talk about reliability: making better use of test-retest studies for study design and interpretation. *PeerJ* **7**, (2019).
36. LeBel, E. P. & Paunonen, S. V. Sexy but often unreliable: The impact of unreliability on the replicability of experimental findings with implicit measures. *Personality and Social Psychology Bulletin* **37**, 570–583 (2011).
37. Parsons, S., Kruijt, A.-W. & Fox, E. 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 (2019).
38. Zuo, X.-N., Xu, T. & Milham, M. P. Harnessing reliability for neuroscience research. *Nat Hum Behav* **3**, 768–771 (2019).
39. Green, S. B. *et al.* Use of internal consistency coefficients for estimating reliability of experimental task scores. *Psychon Bull Rev* **23**, 750–763 (2016).

40. Calamia, M., Markon, K. & Tranel, D. The robust reliability of neuropsychological measures: meta-analyses of test-retest correlations. *Clin Neuropsychol* **27**, 1077–1105 (2013).
41. Mkrtchian, A., Valton, V. & Roiser, J. P. Reliability of decision-making and reinforcement learning computational parameters. *bioRxiv* 2021.06.30.450026 (2021) doi:10.1101/2021.06.30.450026.
42. Hedge, C., Powell, G. & Sumner, P. The reliability paradox: Why robust cognitive tasks do not produce reliable individual differences. *Behav Res* **50**, 1166–1186 (2018).
43. Banerjee, D. D. The other side of COVID-19: Impact on obsessive compulsive disorder (OCD) and hoarding. *Psychiatry Research* **288**, 112966 (2020).
44. Loosen, A. M., Skvortsova, V. & Hauser, T. U. Obsessive–compulsive symptoms and information seeking during the Covid-19 pandemic. *Transl Psychiatry* **11**, 1–10 (2021).
45. Tanir, Y. *et al.* Exacerbation of obsessive compulsive disorder symptoms in children and adolescents during COVID-19 pandemic. *Psychiatry Research* **293**, 113363 (2020).
46. Button, K. S. *et al.* Power failure: why small sample size undermines the reliability of neuroscience. *Nat Rev Neurosci* **14**, 365–376 (2013).
47. Poldrack, R. A. *et al.* Scanning the horizon: towards transparent and reproducible neuroimaging research. *Nat Rev Neurosci* **18**, 115–126 (2017).
48. Szucs, D. & Ioannidis, J. P. A. Empirical assessment of published effect sizes and power in the recent cognitive neuroscience and psychology literature. *PLOS Biology* **15**, e2000797 (2017).
49. Miller, L. A. & Lovler, R. L. *Foundations of Psychological Testing: A Practical Approach*. (SAGE Publications, 2018).

50. Shahar, N. *et al.* 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**, e1006803 (2019).
51. Waltmann, M., Schlagenhauf, F. & Deserno, L. Sufficient reliability of the behavioral and computational readouts of a probabilistic reversal learning task. *Behav Res* (2022) doi:10.3758/s13428-021-01739-7.
52. McCaffrey, R. J. & Westervelt, H. J. Issues associated with repeated neuropsychological assessments. *Neuropsychol Rev* **5**, 203–221 (1995).
53. Condon, D. M. & Revelle, W. The international cognitive ability resource: Development and initial validation of a public-domain measure. *Intelligence* **43**, 52–64 (2014).
54. Zigmond, A. S. & Snaith, R. P. The hospital anxiety and depression scale. *Acta Psychiatr Scand* **67**, 361–370 (1983).
55. Faul, F., Erdfelder, E., Buchner, A. & Lang, A.-G. Statistical power analyses using G*Power 3.1: Tests for correlation and regression analyses. *Behavior Research Methods* **41**, 1149–1160 (2009).
56. Foa, E. B. *et al.* The Obsessive-Compulsive Inventory: development and validation of a short version. *Psychol Assess* **14**, 485–496 (2002).
57. Brown, W. Some experimental results in the correlation of mental abilities 1. *British Journal of Psychology, 1904-1920* **3**, 296–322 (1910).
58. Spearman, C. Correlation calculated from faulty data. *British journal of psychology* **3**, 271 (1910).
59. Revelle, W. *psych: Procedures for Psychological, Psychometric, and Personality Research*. (Northwestern University, 2020).

60. de Vet, H. C. W., Mokkink, L. B., Mosmuller, D. G. & Terwee, C. B. Spearman–Brown prophecy formula and Cronbach's alpha: different faces of reliability and opportunities for new applications. *Journal of Clinical Epidemiology* **85**, 45–49 (2017).
61. Koo, T. K. & Li, M. Y. A Guideline of Selecting and Reporting Intraclass Correlation Coefficients for Reliability Research. *J Chiropr Med* **15**, 155–163 (2016).
62. Shrout, P. E. & Fleiss, J. L. Intraclass correlations: uses in assessing rater reliability. *Psychological bulletin* **86**, 420 (1979).
63. Bakdash, J. Z. & Marusich, L. R. *rmcorr: Repeated Measures Correlation*. (2021).
64. Bakdash, J. Z. & Marusich, L. R. Repeated Measures Correlation. *Front. Psychol.* **8**, (2017).

Supplemental Information for

Consistency within change: Evaluating the psychometric properties of a widely-used predictive-inference task

Alisa M. Loosen[1,2], Tricia X.F. Seow[1,2] & Tobias U. Hauser[1,2]

[1]Max Planck UCL Centre for Computational Psychiatry and Ageing Research

[2]Wellcome Centre for Human Neuroimaging, University College London

Supplemental Methods

Exclusion criteria

At both time points (time point 1, T1; time point 2, T2), we excluded participants because of failed attention checks (T1: *N*=19; T2: *N*=14) or inconsistent game performance. Exclusion criteria for game performance were aligned with those used by Seow and Gillan[1] and as follows:

1. The confidence rating was left unchanged, thus equal to the default rating, in more than 60% of the trials (T1: *N*=14; T2: *N*=4).

2. The task was reset more than five times, whereby the reset was prompted if participants left as the default confidence rating for more than 70% of the first 50 trials. This was not the case for any of our participants.

3. If across trials, the default confidence rating correlated more than 0.5 with the confidence rating logged in by the participant (T1: *N*=17; T2: *N*=17).

4. If the mean confidence on trials that were preceded by a correct trial (i.e., caught the particle on the trial before) was lower than the mean confidence on trials that were preceded by an incorrect trial (i.e., did not catch the particle on the trial before) (T1: *N*=19; T2: *N*=6).

We additionally excluded participants who used the same confidence rating more than 90% of the time (T1: *N*=2; T2: *N*=0) to ensure usage of the full confidence scale.

Normative parameters of the reduced quasi-optimal Bayesian learner

To capture task characteristics thought to influence people's behaviour and confidence in changing (task) environments, we fitted the reduced quasi-optimal Bayesian learner that had been used by past studies using this task (e.g.,[1–3]) to the trial-wise task data. Normative parameters used in the main analyses are marked in bold.

This Bayesian learner updates its belief about the landing position distribution (i.e., point estimation of the mean of the Gaussian distribution from which the particle locations were sampled) using a delta-rule (equation 1).

$$\mathrm{B}_{t+1} = \mathrm{B}_t + {LR^b}_t \times {\mathrm{PE}^b}_t \quad (1)$$

The belief estimate is updated based on the model's learning rate (i.e., LR^b) and prediction error (i.e., $\mathbf{PE}^{\boldsymbol{b}}$ in equation 2). This prediction error $\mathbf{PE}^{\boldsymbol{b}}$ is the distance between the belief of where the particle landing position is (i.e., B_t) and the location where the particle ended up falling to (i.e., X_t).

$${\mathrm{PE}^b}_t = X_t - \mathrm{B}_t \quad (2)$$

The prediction error is weighted by the LR^b, thus, determining how much the encountered particle landing position will influence the belief on the next trial. The LR^b itself (equation 3) is determined by the change-point probability (i.e., Ω or **CPP**) and the model confidence (i.e., ν or MC).

$${LR^b}_t = \Omega_t + (1 - \Omega_t)(1 - \nu_t) \quad (3)$$

The **CPP** captures how likely it is that the mean of the generative distribution determining the particle location has shifted (equation 4). The MC captures the model's uncertainty caused by an inaccurate estimation of the mean of the generative distribution. These two parameters or,

more specifically, the last two terms of the LR^b (i.e., $(1-\Omega_t)(1-\nu_t)$) represent the normative parameter **RU**. Thus, the relative uncertainty in the belief about the mean of the generative distribution determining the particle landing location.

Unpacking the parameters further, **CPP** (equation 4) is more precisely described as the relative likelihood that the particle falling location is sampled from a new generative distribution, i.e., a change-point occurred (the distribution mean is determined by a uniform distribution U over all 360 possible locations; numerator of equation 4), or the falling location is again drawn from the same Gaussian (*N*) as the particle before, i.e., the Gaussian centred around B_t (second term in the denominator of equation 4). Both scenarios (mean of the generative distribution has changed or not) are additionally influenced by the hazard rate H, which has been pre-determined during task development (H=0.125) and is the probability that the mean of the distribution has changed.

$$\Omega_t = \frac{U(X_t|1{,}360)H}{U(X_t|1{,}360)H + N(X_t|B_t, \sigma_t^2)(1-H)} \tag{4}$$

The term σ_t^2 (equation 5) thereby represents the estimated variance of the predictive distribution of the particle falling locations. This, in turn, is determined by the variance of the overall generative Gaussian distribution σ_N^2 modulated by the model confidence, MC (ν). This means, the larger the MC, the smaller the estimated variance of the predictive distribution (or the closer to σ_N^2). As model confidence increases with trials after the change-points, the certainty about the particle falling location increases. Since our task entailed a fixed hazard rate and σ_N^2, CPP was solely driven by a mismatch between the newest particle landing location and prior expectations. This further meant, that CPP was highly correlated with PE^b (cf. Results Section of the main manuscript).

$$\sigma_t^2 = \sigma_N^2 + \frac{(1-\nu_t)\sigma_N^2}{\nu_t} \quad (5)$$

In contrast to all other parameters, the MC is computed at the end of each trial (*t*) for the subsequent trial (*t+1*). The numerator contains a weighted average of the variance of the overall generative Gaussian distribution (σ_N^2) (first term) and σ_N^2 conditional on no change-point (second term). It also includes a term capturing the variance due to the difference in means of these two conditional distributions (third term). The denominator is almost identical to the numerator but entails an additional term capturing uncertainty due to noise (σ_N^2).

$$\nu_{t+1} = \frac{\Omega_t\sigma_N^2 + (1-\Omega_t)(1-\nu_t)\sigma_N^2 + \Omega_t(1-\Omega_t)(\delta_t\nu_t)^2}{\Omega_t\sigma_N^2 + (1-\Omega_t)(1-\nu_t)\sigma_N^2 + \Omega_t(1-\Omega_t)(\delta_t\nu_t)^2 + \sigma_N^2} \quad (6)$$

As described in the manuscript, trial-wise estimates of **PE$^{\mathbf{b}}$**, **CPP** and **RU** were used to capture how uncertainty and surprise due to changes in the task influenced participants' actions and confidence.

Regression models

In addition to our summary statistics approach reported in the results section, we ran mixed-effects models implemented by previous studies using the predictive-inference task[1]. However, these models did not converge using our dataset. In an attempt to achieve convergence, we followed a step-by-step-approach lined out in the literature[4], simplifying and adapting the models. This entailed a simplification of the models and the change of the optimizer (i.e., *bobyqa*). However, further steps such as the removal of correlations between random effects via orthogonalization, were not feasible as they would have made a valid reliability analysis of the mixed-model weights impossible. All mixed models covaried for age, gender (0=female and 1=male), and IQ.

We computed the mixed-effects models using the *lme4* package[5] in R. We computed models with action-update (i.e., the change in bucket position from trial t-1 to trial t) and confidence (corresponding to trial t) as the dependent variables. We used the same predictors as in the summary statistics approach, namely the normative factors and the 'Hit' regressor and added demographic fixed-effect covariates (i.e., IQ, gender and age). Covariates were z-scored across participants, while normative factors, confidence and action-update were z-scored within participants.

In the syntax of the *lme4* package[5], the model was specified as follows:

$$Action\text{-}Update \sim PE^b + CPP + RU + Hit + Age + IQ + Gender + (1 + PE^b + CPP + RU \mid Participant)$$

When we subsequently investigated whether psychiatric questionnaire (total) scores were associated with the normative factors, we ran separate regression models for each of them. The psychiatric variables (i.e., OC symptoms, anxiety, and depression) were all z-scored across participants and included as between-subjects predictors as follows:

$$Action\text{-}Update \sim (PE^b + CPP + RU + Hit) * Psychiatric\ Variable + Age + IQ + Gender + (1 + PE^b + CPP + RU \mid Participant)$$

All normative factor regressors used to predict action-update were again implemented as an interaction effect with PE^b. Moreover, although we had run separate models for PE^b and CPP in the summary statistics approach due to their high correlation, we here left them in one mixed-model. This was for replication purposes[1], however, our main findings remained unchanged.

Supplemental Results

Internal consistency and test-retest reliability of learning rate and confidence

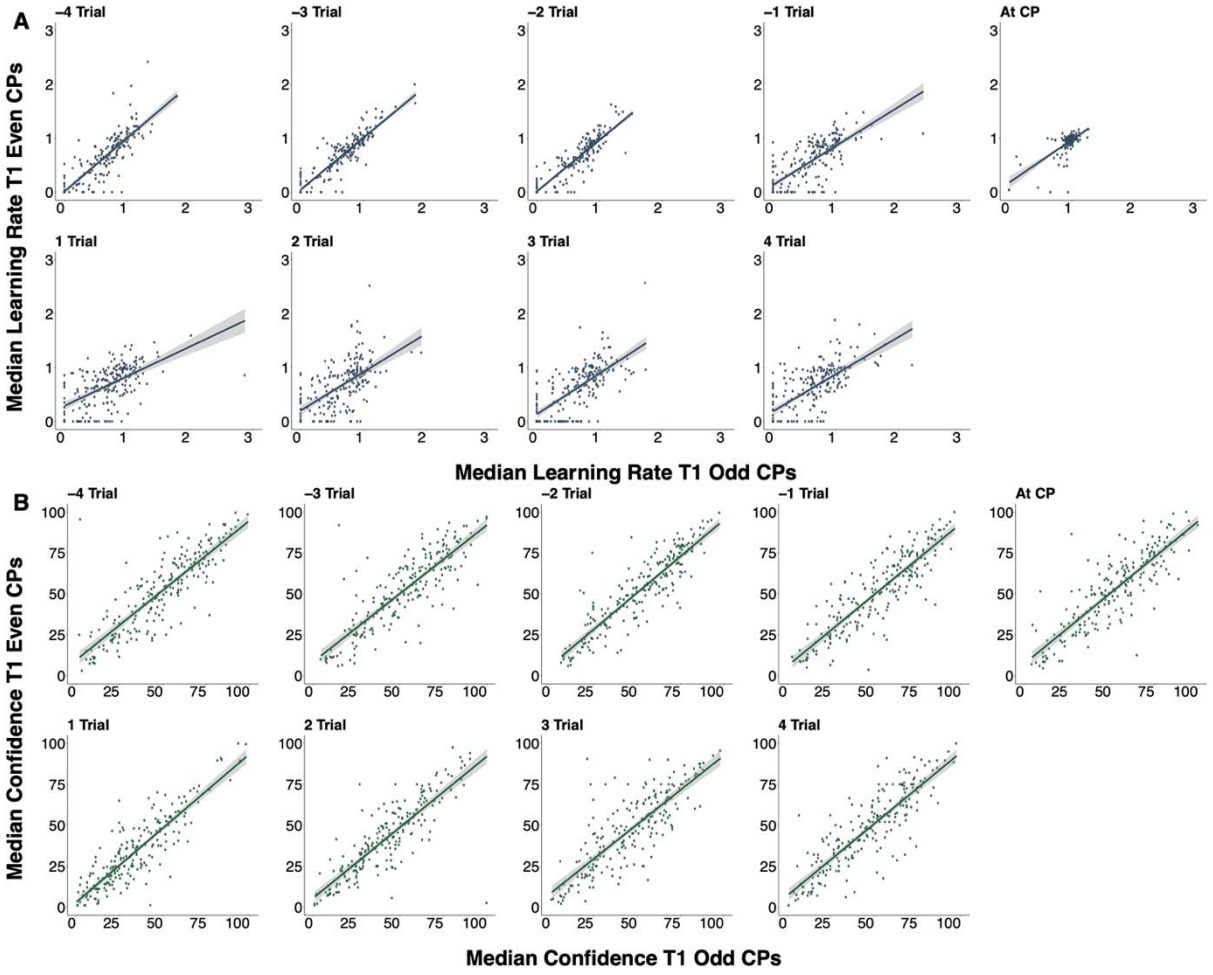
Supplemental Table 1

Internal Consistency and Test–Retest Reliability of Behavioural Measures from the Predictive-Inference Task

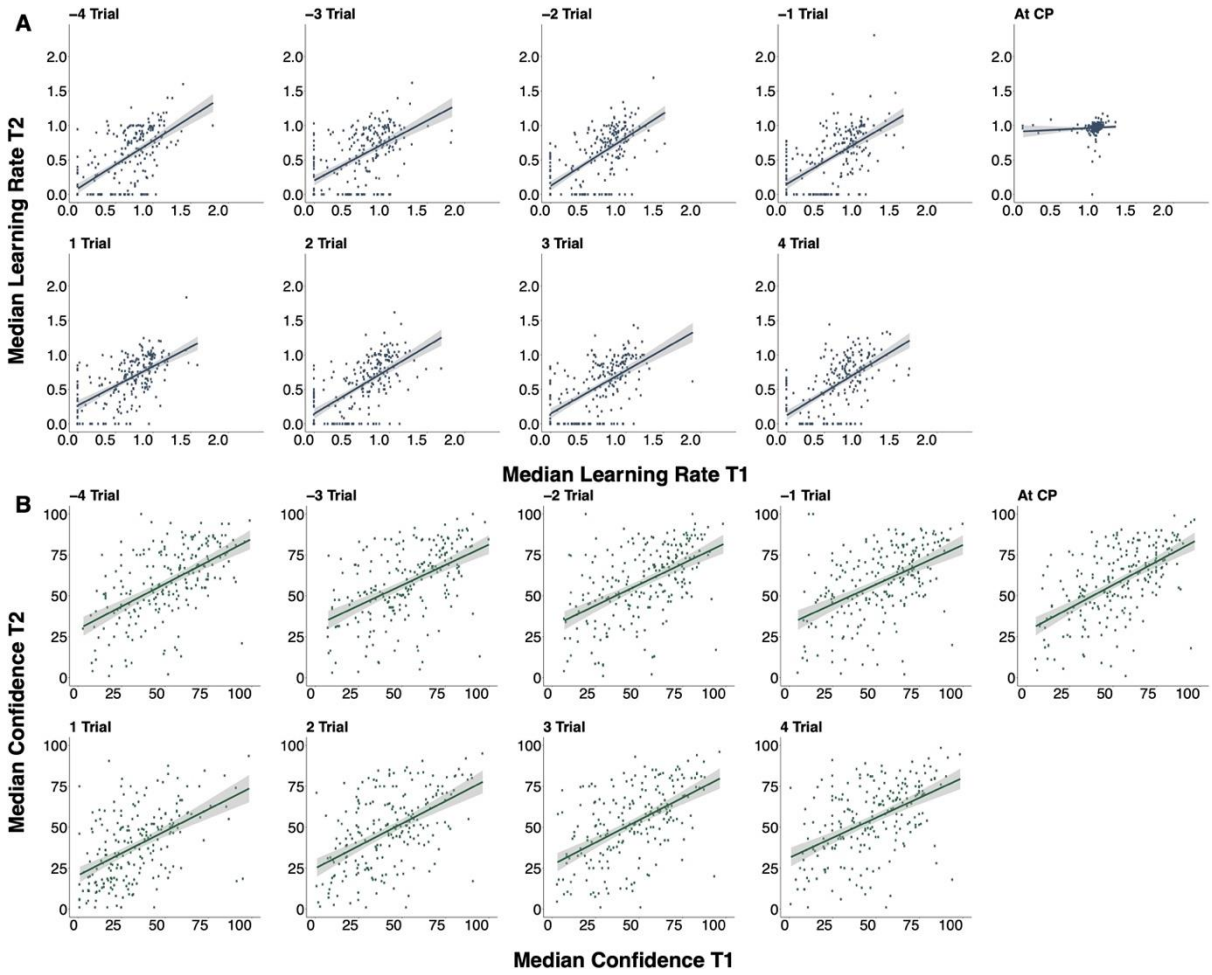
	Learning Rate (LR^h)			Confidence		
	Internal (r_{SB})	Consistency	Test-retest reliability (*ICC*)	Internal (r_{SB})	Consistency	Test-retest reliability (*ICC*)
Trial	T1	T2		T1	T2	
-4	0.813 [0.773, 0.847]	0.877 [0.842, 0.904]	0.787 [0.723, 0.835]	0.928 [0.911, 0.941]	0.914 [0.890, 0.934]	0.694 [0.601, 0.762]
-3	0.799 [0.757, 0.835]	0.817 [0.767, 0.857]	0.742 [0.678, 0.794]	0.922 [0.904, 0.937]	0.945 [0.929, 0.959]	0.672 [0.586, 0.739]
-2	0.840 [0.805, 0.869]	0.818 [0.769, 0.857]	0.815 [0.768, 0.852]	0.946, [0.933, 0.956]	0.930 [0.909, 0.946]	0.673 [0.584, 0.741]
-1	0.843 [0.808, 0.871]	0.856 [0.816, 0.888]	0.771 [0.713, 0.817]	0.934 [0.919, 0.947]	0.938 [0.920, 0.952]	0.649 [0.556, 0.721]
0 (at CP)	0.731 [0.677, 0.778]	0.764 [0.703, 0.814]	0.140 [-0.075, 0.312]	0.921 [0.903, 0.936]	0.931 [0.911, 0.947]	0.718 [0.639, 0.778]
1	0.716 [0.659, 0.765]	0.765 [0.703, 0.815]	0.749 [0.686, 0.799]	0.938 [0.924, 0.950]	0.948 [0.933, 0.960]	0.659 [0.570, 0.730]
2	0.795 [0.751, 0.831]	0.821 [0.773, 0.860]	0.777 [0.718, 0.823]	0.924 [0.906, 0.938]	0.913 [0.8889, 0.933]	0.671 [0.582, 0.739]

3	0.787 [0.742, 0.825]	0.829 [0.782, 0.866]	0.761 [0.697, 0.811]	0.919 [0.900, 0.934]	0.944 [0.927, 0.956]	0.688 [0.602, 0.754]
4	0.817 [0.777, 0.850]	0.822 [0.774, 0.861]	0.792 [0.736, 0.836]	0.930 [0.914, 0.944]	0.922 [0.899, 0.939]	0.650 [0.548, 0.727]
Entire game			0.796 [0.716, 0.849]			0.702 [0.615, 0.768]

Note. Trials refer to four trials before (-4), at (0) to four trials after (4) a change point (CP) and for completeness, we also computed the test-retest reliability over all trials of the task. *ICC*=intraclass correlation of absolute agreement; r_{SB}=Spearman-Brown corrected correlations; 95% confidence intervals are specified in square brackets.



Supplemental Figure 1. **Internal consistency of median learning rate (A) and median confidence (B) at time point 1.** Represented is the consistency of median values between odd change-points (CP; x-axes) and even CPs (y-axis) at trials ranging from 4 before (-4) to 4 after (4) a CP. Data at time point 2 showed a very similar pattern (cf. Supplemental Table 1).



Supplemental Figure 2. **Test-retest reliability of median learning rate (A) and median confidence (B).** Represented is the agreement of median values between time point 1 (T1; x-axes) and time point 2 (T2; y-axis) at trials ranging from 4 before (-4) to 4 after (4) a CP.

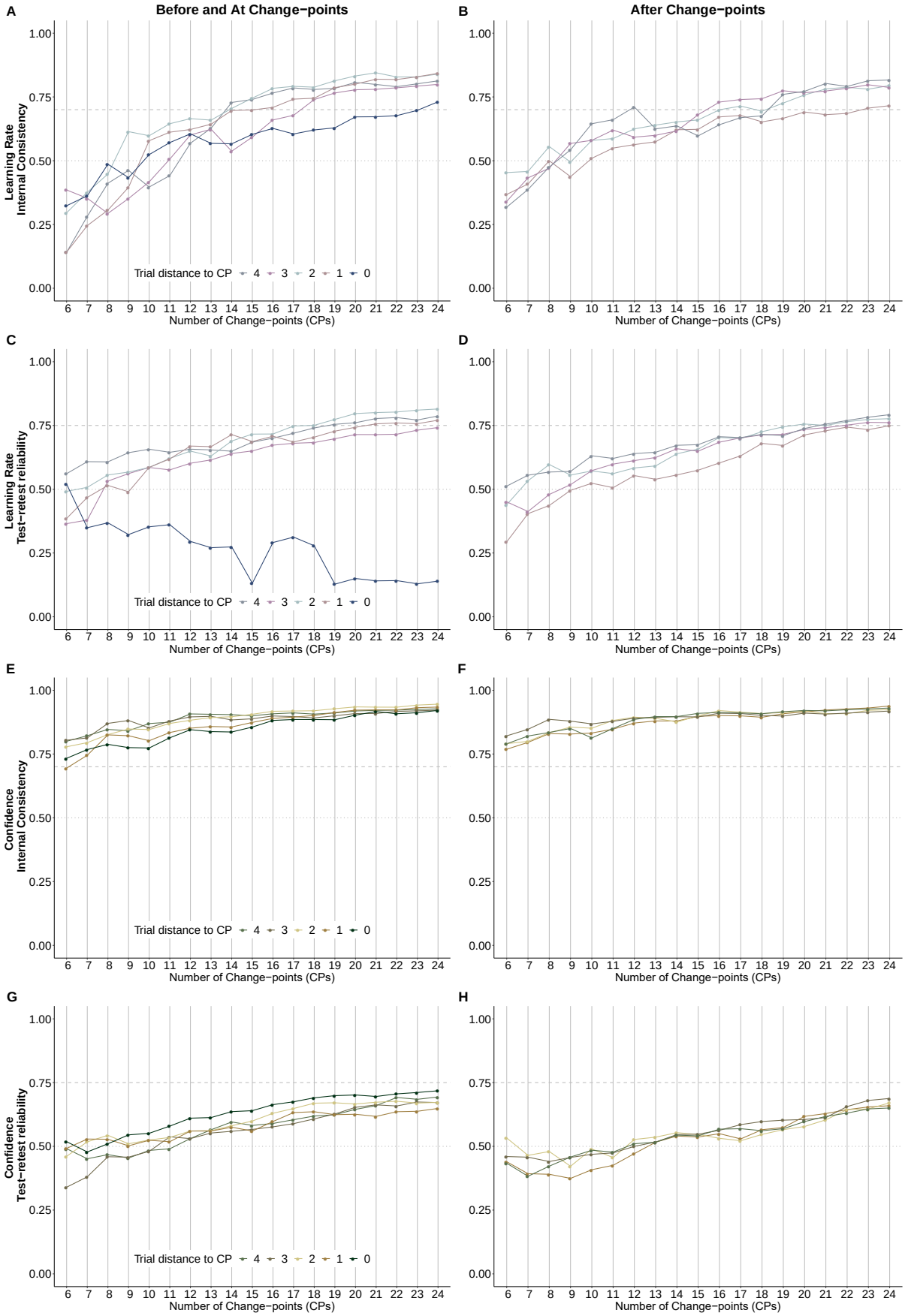
The effect of task length on stability

To assess whether the task could potentially be shortened in future studies, we additionally investigated the internal consistency and test-retest reliability of LR^h in shorter task-versions with 6 to 24 CPs (full task). To do so, we used the larger dataset from T1 (N=330) and computed psychometric scores for all CP counts separately, ranging from 6 CPs (to have a minimum of 3 odd and 3 even CPs capturing internal consistency) until the maximum in our task version, i.e., 24 CPs, was reached.

This showed, that with our hazard rate of 0.125, the LR^h at trials before the CPs reached good internal stability (r_{SB}=0.70) after ~the 18th CPs and at trials after and at the CPs LR^h did not reach stability until ~the 23rd CP (cf. Supplemental Figure 3A-B). Test-retest reliability of the LR^h at most trials before and after the CP reached a good level (ICC=0.75) at the ~21st CP (cf. Supplemental Figure 3C-D) except for the third trial before and the first trial after the CP which did not reach good reliability until full task length (i.e., 24CPs). Test-retest reliability of the LR^h at the CP itself never reached a moderate level and decreased with increasing CP count, which was explained by a decrease in between-participant variance (6 CPs: 35%; 12 CPs: 17%; 18 CPs: 16%; 24 CPs: 8 %). Thus, as participants' LR^h stabilized with time their scores also became more similar on an inter-individual level, which led to a lower reliability score. Overall, this shows that LR^h investigations should entail at least as many CP-counts as implemented here.

The same analyses for confidence showed that it reached a good internal stability at all trials after only 7 CPs (cf. Supplemental Figure 3E-F). While its test-retest reliability at the trials before and at CPs reached a moderate level after ~12 CPs, trials after the CPs reached this at the ~13th CP (cf. Supplemental Figure 3G-H). None of these scores ever reached a good reliability level. This suggests that while the maximum psychometric quality for confidence

can already be reached at ~13 CPs, more trials and/or a higher hazard rate might yield even more reliable results.


A
Before and At Change−points
B
After Change−points
C
D
E
F
G
H
Learning Rate
Internal Consistency
Learning Rate
Test−retest reliability
Confidence
Internal Consistency
Confidence
Test−retest reliability
Trial distance to CP 4 3 2 1 0
Number of Change−points (CPs)
1.00
0.75
0.50
0.25
0.00
6 7 8 9 10 11 12 13 14 15 16 17 18 19 20 21 22 23 24

Supplemental Figure 3. **Psychometric properties of human learning rate and human confidence with increasing change-points (CPs).** The internal consistency (Spearman-Brown corrected Pearson correlations) of the learning rate (LR^h) at T1 (N=330) stabilized after ~18 CPs for trials before CP (**A**) and not until ~23 CPs for trials after and at the CP (**B**). Test-retest reliability of the LR^h as captured by the intraclass correlation (*ICC*) did not reach a good level at trials before and after the CP until our full CP count of 24, while it never reached a moderate level at the CP itself (**C-D**). Internal consistency of confidence already stabilized on a good level at ~7 CPs for all trials (**E-F**). Test-retest reliability of confidence reached a moderate level at trials before CPs after ~12 CPs (**G**) and at trials after CPs at ~13 CPs (**H**). Confidence at the CP trial itself was moderately reliable throughout almost all CP-counts. Horizontal dashed lines represent the cut-off score for good (r_{SB}=0.70; ICC=0.75) and dotted lines represent the moderate cut-off score (r_{SB}=0.50; ICC=0.50).

Mixed-models investigating the link between confidence and the Bayesian learner

Confidence and the Bayesian learner

We ran a mixed-effects model on the basis of the Bayesian learner's predictions. We again observed that participants' confidence was negatively linked to all of the normative (T1: $\beta_{PE}{}^{b}$=-0.070, *SE*=0.001, p<0.001; β_{CPP}=-0.190, *SE*=0.014, p<0.001; β_{RU}=-0.184, *SE*=0.009, p<0.001; T2: $\beta_{PE}{}^{b}$ =-0.074, *SE*=0.011, p<0.001; β_{CPP}=-0.222, *SE*=0.017, p<0.001; β_{RU}=-0.208, *SE*=0.012, p<0.001), and positively to Hit (T1: β=0.254, *SE*=0.008, p<0.001; T2: β=0.263, *SE*=0.011, p<0.001; cf. Supplemental Figure 4A).

Internal consistency

We also examined the internal consistency of the links revealed by the mixed-model and again saw that across predictors it was low to moderate (T1: PE^{b}*:* r_{SB}=0.219, *95% CI* [-0.003, 0.257]; CPP: r_{SB}=-0.137, *95% CI* [-0.211, 0.003]; RU: r_{SB}=0.509, *95% CI* [0.424, 0.584]; Hit: r_{SB}=0.435, *95% CI* [0.550, 0.709], T2: PE^{b}: r_{SB}=0.219, *95% CI* [-0.003, 0.257]; CPP: r_{SB}=0.055, *95% CI* [-0.078, 0.186]; RU: r_{SB}=0.603, *95% CI* [0.521, 0.688]; Hit: r_{SB}=0.636, *95% CI* [0.550, 0.709]; cf. Supplemental Figure 4B).

Test-retest reliability

Test-retest reliability of the mixed-effects regression weights of the normative factors predicting confidence was also mostly low (PE^{b}: r_{SB}=0.364, *95% CI* [0.266, 0.454]; CPP: r_{SB}=0.125, *95% CI* [0.017, 0.230]; RU: r_{SB}=0.139, *95% CI* [0.031, 0.243]; Hit: r_{SB}=0.368, *95% CI* [0.270, 0.457]); T2: RU: r_{SB}=0.260, *95% CI* [0.132, 0.380]; PE^{b}: r_{SB}=0.318, *95% CI* [0.193, 0.4317]; CPP: r_{SB}=-0.026, *95% CI* [-0.158, 0.107]; Hit: r_{SB}=0.298, *95% CI* [0.173, 0.415]; cf. Supplemental Figure 4C).

Action-update and the Bayesian learner

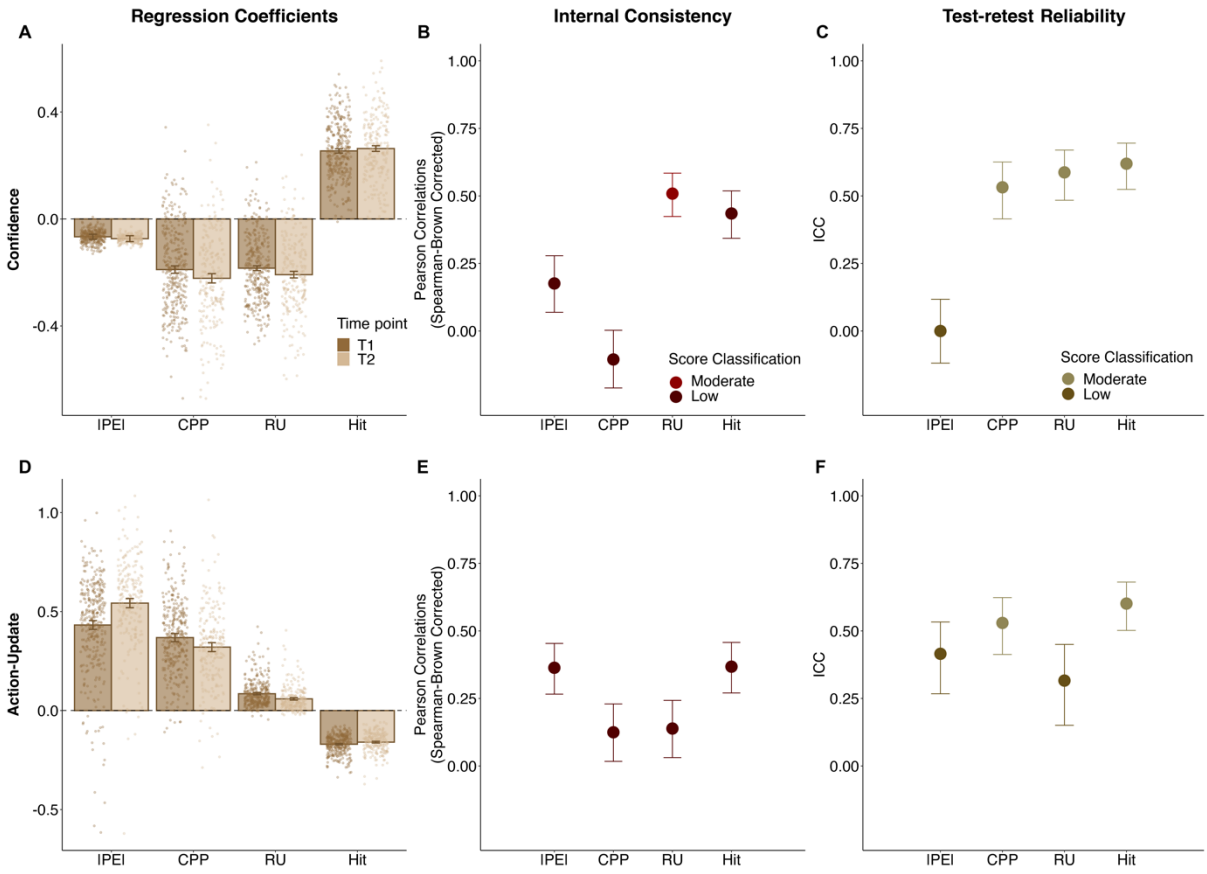
We repeated the same mixed-effects models as above now predicting action-update. This model again showed that action-update was positively linked to all normative factors at both time points (T1: β_{PE^b}=0.432, *SE*=0.022, *p*<0.001; β_{CPP}=0.368, *SE*=0.021, *p*<0.001; β_{RU}= 0.085, *SE*=0.006, *p*<0.001; T2: β_{PE^b}=0.543, *SE*=0.023, *p*<0.001; β_{CPP}=0.320, *SE*=0.022, *p*<0.001; β_{RU}=0.059, *SE*=0.006, *p*<0.001) and negatively linked to Hit (T1: β=-0.170, *SE*=0.004, *p*<0.001; T2: β=-0.160, *SE*=0.004, *p*<0.001; cf. Supplemental Figure 4D).

Internal Consistency

We also examined the internal consistency of these mixed-model links between action-update the normative factors and the Hit predictor, which were low at both time points (PE^b: r_{SB}=0.364, *95% CI* [0.266, 0.454]; CPP: r_{SB}=0.125, *95% CI* [0.017, 0.230]; RU: r_{SB}=0.139, *95% CI* [0.031, 0.243]; Hit: r_{SB}=0.368, *95% CI* [0.270, 0.457]); T2: RU: r_{SB}=0.260, *95% CI* [0.132, 0.380]; PE^b: r_{SB}=0.318, *95% CI* [0.193, 0.4317]; CPP: r_{SB}=-0.026, *95% CI* [-0.158, 0.107]; Hit: r_{SB}=0.298, *95% CI* [0.173, 0.415]; cf. Supplemental Figure 4E).

Test-Retest Reliability

Test-retest reliability of the mixed-model regression weights capturing the link between normative factors and action-update were low to moderate (PE^b: *ICC*=0.415, *95% CI* [0.267, 0.533]; CPP: *ICC*=0.530, *95% CI* [0.413, 0.623]; RU: *ICC*=0.316; *95% CI* [0.150, 0.450]; Hit: *ICC*=0.601, *95% CI* [0.502, 0.680]; cf. Supplemental Figure 4F).



Supplemental Figure 4. **Mixed-effect models results predicting confidence and action on the basis of Bayesian factors.** Mixed-effects models predicted trial-wise confidence ratings and action-updates based on the model-derived normative factors PE^b (model prediction error), CPP (change-point probability), RU (relative uncertainty) and Hit (accuracy on the previous trial). All normative factors were negatively and the Hit regressor positively linked to confidence (**A**). Investigating the robustness of these associations, we saw that the link between confidence and normative regression weights was low, except for the link with RU which showed a moderate consistency (**B**). Test-retest reliability of all normative factor regression weights predicting confidence, except for the one of PE^b, was moderate (as indicated by the ICC score; **C**). We found the reverse relationships for action-update, which was positively linked to the normative factors and negatively linked to Hit (**D**). The link between Hit and action-update showed a moderate internal consistency, while internal consistency of all remaining normative regression weights was low (**E**). Except for the regression weight of Hit and CPP all regression weights in the action-update model showed low test-retest reliability (**F**). Individual coefficients of participants are represented by circles while the bar plot represents the overall model coefficient, and their error bars represent standard errors in A & D. Internal consistency for all measures is here displayed for T1. Error bars for internal consistency and test-retest reliability represent the estimates 95% Confidence Interval.

Mixed-models investigating associations between task measures and psychiatric dimension scores

For replication purposes, we also examined mixed-effects models that had been previously implemented by Seow and Gillan[1] and which investigated the link between confidence, action-update and psychiatric dimension scores (i.e., OC symptoms, anxiety, and depression; cf. Supplemental Methods above).

We again did not observe any link between any psychiatric symptom score and action-confidence coupling. The interaction effects between the psychiatric scores and confidence in the regression models predicting action-update were non-significant for all symptom scores (OC symptoms: β=-0.076, *SE*=0.376, *p*=0.589; anxiety: β=-0.180, *SE*=0.376, *p*=0.632; depression: β=-0.197, *SE*=0.376, *p*=0.600). Thus, we still could not replicate the association between action-confidence coupling and OC or any other psychiatric symptom score when repeating the mixed-effects models.

Similarly, none of the psychiatric variables in these models were predictive of confidence (OC symptoms: β=-0.00, *SE*=0.005, *p*=1; anxiety: β=-0.00, *SE*=0.005, *p*=1; depression: β=0.00, *SE*=0.005, *p*=1) or action-update themselves (OC symptoms: β=0.262, *SE*=0.437, *p*=0.550; anxiety: β=0.035, *SE*=0.306, *p*=0.908; depression: β=-0.319, *SE*=0.275, *p*=0.247).

Moreover, none of the psychiatric variables showed a significant interaction effect with CPP on confidence (OC symptoms: β=0.001, *SE*=0.016, *p*=0.963; anxiety: β=-0.006, *SE*=0.016, *p*=0.689; depression: β=-0.001, *SE*=0.016, *p*=0.964) and we also did not find that OC scores affected the impact of PE^b on action-update (β=0.007, *SE*=0.023, *p*=0.751) using the mixed-model approach.

Covid-adapted OC symptom score and the predictive-inference task.

Since the present study was conducted during the Covid-19 pandemic, we created a new subscore for the Obsessive Compulsive Inventory-Revised (OCI-R)[6] (i.e., 'OCI-R pandemic-irrelevant items') excluding items of the OCI-R that we considered being potentially related to or influenced by the Covid-19 pandemic (cf. Supplemental Table 2).

We subsequently repeated all main analyses, investigating the link between OC symptoms and task, behavioural, and normative parameters now using the OCI-R pandemic-irrelevant items score instead of the regular OCI-R total score as a predictor.

First, we investigated the relationship between action-confidence coupling and the new OCI-R score. However, the action-confidence beta weights did not correlate with the OC pandemic-irrelevant score (r_S=0.003, p=0.970). Similarly, the OCI-R pandemic-irrelevant items score was not correlated with mean confidence (r_S=-0.013, p=0.851) or action-update (r_S=0.012, p=0.855).

We also repeated the analyses linking the CPP-weights predicting confidence and the new OCI-R score, but it was nonsignificant (r_S=0.027, p=0.693). Similarly, in the action-update model the new OCI-R score was not associated with the PE^b-weights (r_S=0.087, p=0.201).

Finally, the new OCI-R score also was not correlated with the mean LR^h (r_S=-0.018, p=0.790). Overall, this means, that even when controlling for potential inflations of the OCI-R scores due to the pandemic situation we could not replicate any of the associations between OC symptoms and task measures reported in the literature (these findings held true when adapting the mixed-model approach reported above).

Supplemental Table 2

Pandemic-relevancy of items on the OCI-R

Pandemic-relevance	Item
No	1. I have saved up so many things that they get in the way.
No	2. I check things more often than necessary.
No	3. I get upset if objects are not arranged properly.
No	4. I feel compelled to count while I am doing things.
Yes	5. I find it difficult to touch an object when I know it has been touched by strangers or certain people.
No	6. I find it difficult to control my own thoughts.
No	7. I collect things I don't need.
No	8. I repeatedly check doors, windows, drawers, etc..
No	9. I get upset if others change the way I have arranged things.
No	10. I feel I have to repeat certain numbers.
Yes	11. I sometimes have to wash or clean myself simply because I feel contaminated.
No	12. I am upset by unpleasant thoughts that come into my mind against my will.
No	13. I avoid throwing things away because I am afraid I might need them later.
No	14. I repeatedly check gas and water taps and light switches after turning them off.
No	15. I need things to be arranged in a particular order.
No	16. I feel that there are good and bad numbers.
Yes	17. I wash my hands more often and longer than necessary.
No	18. I frequently get nasty thoughts and have difficulty in getting rid of them.

Comparison of psychiatric symptom and behavioural measures' distributions

To further investigate why we did not replicate previous findings reported by Seow and Gillan[1], we carefully compared symptom and behavioural measures' distributions in our and Seow and Gillan's[1] sample in post-hoc exploratory analyses.

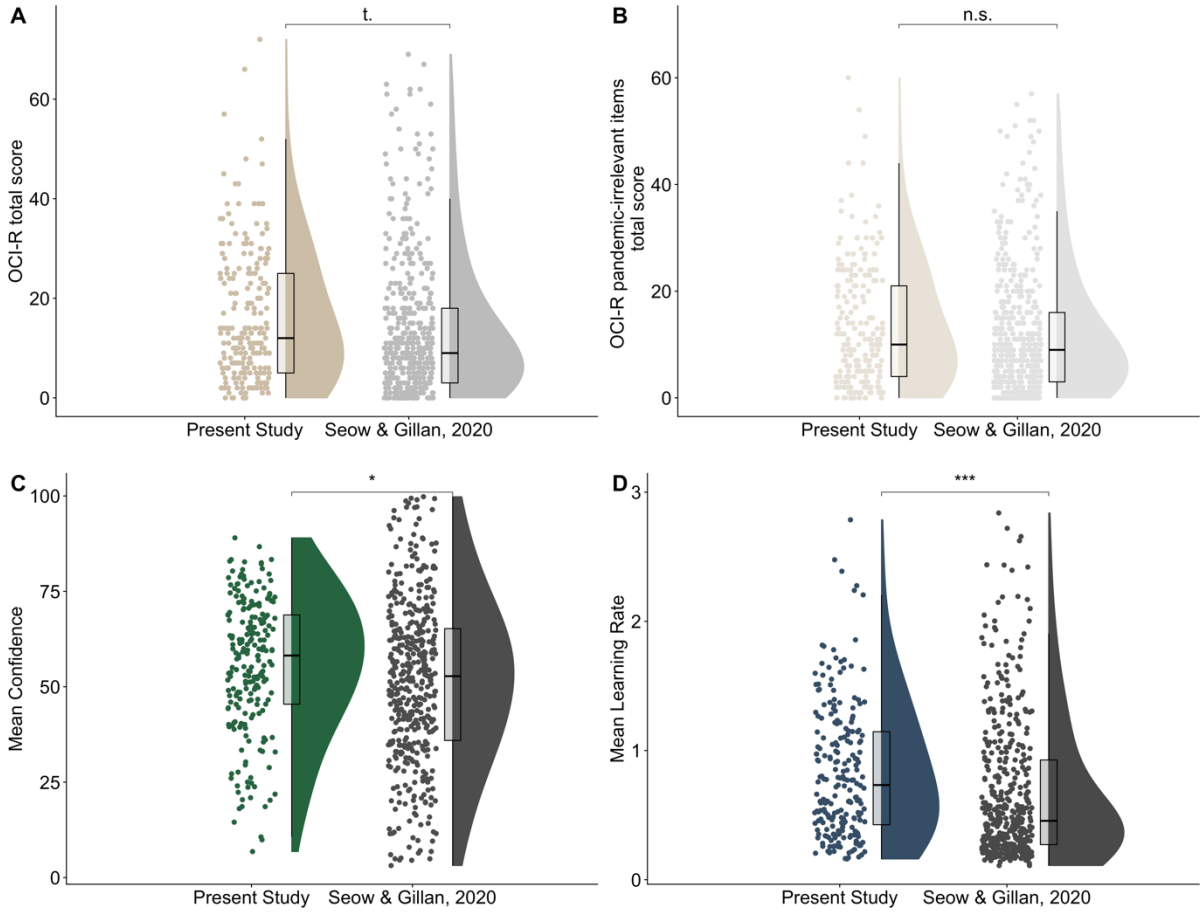
Looking at OCI-R total scores, using a simple two-sided t-test, we observed a trend for a difference between the two samples on the OCI-R total score (t(463.92)=1.844, p=0.066; cf. Supplemental Figure 5A), which vanished when using the adapted OCI-R score that controlled for items that might have been of particular relevance to the ongoing pandemic (t(456.38)=1.097, p=0.273; cf. Supplemental Figure 5B).

We also compared the mean behavioural scores of the two main variables of interest, confidence and $LR^{h,}$ across the two studies. As Seow and Gillan's study entailed two different hazard rates (H=0.025 and H=0.125), for the analyses below we only used trials in their data with the same hazard rate as ours (H=0.125). While for our T1 sample only the mean LR^h was higher than in the sample of Seow and Gillan (Confidence: t(755.67)=-0.721, p=0.471; LR^h: t(659.95)=7.668, p<0.001), both, mean confidence and mean LR^h were higher for our T2 sample (Confidence: t(518.16)=2.563, p=0.011; LR^h: t(462.81)=3.699, p<0.001; cf. Supplemental Figure 5C & D).

Approximately 50% of participants in Seow and Gillan's study had completed the trials matching our hazard rate at the beginning of the task, while the other half had first played trials with the different hazard rate. Controlling for potential biasing effects caused by the encounter of different hazard rates, we thus conducted a follow-up analysis only including Seow and Gillan's participants that had played our hazard rate at the beginning of the task. In this analysis, the difference in mean confidence and mean LR^h between their and our T2 sample turned non-significant (Confidence: t(418.66)=1.471, p=0.142); LR^h: t(412.1)=1.685,

p=0.093)). This was even though the Seow and Gillan's main findings remained true when using this adapted dataset in their analyses.

Overall, the mentioned analyses do not explain the difference in findings we observed as OCI-R scores and task measures are not strikingly different in our samples. This difference in findings across studies further stays in contrast with our main findings showing the robustness of the task measures across time and within task sessions in our sample (cf. Discussion section of the main manuscript for potential explanations for this difference in findings).



Supplemental Figure 5. **Distributions of OCI-R total (A) and new subscore (B) and behavioural measurements of confidence (C) and learning rate (D) in the sample of the present study (N=219) and the sample reported by Seow and Gillan[1] (N=437).** While regular total scores of the regular Obsessive Compulsive Inventory-Revised (OCI-R) showed a trend-level difference between samples (**A**), the newly created OCI-R subscore only including pandemic-irrelevant items did not differ across samples (**B**). Our sample showed overall higher mean confidence (**C**) and learning rate (**D**) in the predictive-inference task than the sample reported by Seow and Gillan[1]. Task measures displayed are from time point 2, while learning rate patterns were the same at both time points mean confidence between the samples was not significantly different at time point 1. Dots represent total scores of individual participants. Paired t-test (two-tailed): *t.*=trend, *n.s.*=non-significant. Plots have been created using the raincloud package in R[7].

Supplemental References

1. Seow, T. X. F. & Gillan, C. M. Transdiagnostic Phenotyping Reveals a Host of Metacognitive Deficits Implicated in Compulsivity. *Scientific Reports* **10**, 1–11 (2020).
2. McGuire, J. T., Nassar, M. R., Gold, J. I. & Kable, J. W. Functionally Dissociable Influences on Learning Rate in a Dynamic Environment. *Neuron* **84**, 870–881 (2014).
3. Vaghi, M. *et al.* Compulsivity Reveals a Novel Dissociation between Action and Confidence. *Neuron* **96**, 348-354.e4 (2017).
4. Bates, D., Kliegl, R., Vasishth, S. & Baayen, H. Parsimonious Mixed Models. *arXiv:1506.04967 [stat]* (2018).
5. Bates, D., Mächler, M., Bolker, B. & Walker, S. Fitting Linear Mixed-Effects Models Using lme4. *Journal of Statistical Software* **67**, 1–48 (2015).
6. Foa, E. B. *et al.* The Obsessive-Compulsive Inventory: development and validation of a short version. *Psychol Assess* **14**, 485–496 (2002).
7. Allen, M. *et al.* Raincloud plots: a multi-platform tool for robust data visualization. *Wellcome Open Res* **4**, 63 (2021).