Further perceptions of probability: in defence of trial-by-trial updating models

Mattias Forsgren[1], Peter Juslin[1] and Ronald van den Berg[1,2]

[1]Department of Psychology, Uppsala University, Uppsala, Sweden
[2]Department of Psychology, Stockholm University, Stockholm, Sweden

Running head: Further perceptions of probability

Corresponding author
Ronald van den Berg
Department of Psychology,
Frescati Hagväg 9A, 11419, Stockholm, Sweden
Tel: +46707929895
Email: ronald.van-den-berg@psychology.su.se

ABSTRACT

Extensive research in the behavioural sciences has addressed people's ability to learn probabilities of stochastic events, typically assuming them to be stationary (i.e., constant over time). Only recently have there been attempts to model the cognitive processes whereby people learn – and track – *non-stationary* probabilities, reviving the old debate on whether learning occurs trial-by-trial or by occasional shifts between discrete hypotheses. Trial-by-trial updating models – such as the delta-rule model – have been popular in describing human learning in various contexts, but it has been argued that they are inadequate for explaining how humans update beliefs about non-stationary probabilities. Specifically, it has been claimed that these models cannot account for the discrete, stepwise updating that characterises response patterns in experiments where participants tracked a non-stationary probability based on observed outcomes. Here, we demonstrate that the rejection of trial-by-trial models was premature for two reasons. First, our experimental data suggest that the stepwise behaviour depends on details of the experimental paradigm. Hence, discreteness in response data does not necessarily imply discreteness in internal belief updating. Second, previous studies have dismissed trial-by-trial models mainly based on qualitative arguments rather than quantitative model comparison. To evaluate the models more rigorously, we performed a likelihood-based model comparison between stepwise and trial-by-trial updating models. Across eight datasets collected in three different labs, human behaviour is consistently best described by trial-by-trial updating models. Our results suggest that trial-by-trial updating plays a prominent role in the cognitive processes underlying learning of non-stationary probabilities.

KEYWORDS

Probability learning; change-point model; delta rule; belief updating; hypothesis testing

INTRODUCTION

When making decisions, we often rely on subjective estimates of the probability that certain events will occur. Not surprisingly, the issue of how people assess – and should assess – probabilities has been pivotal to the behavioural sciences since at least the Enlightenment. How people learn, estimate, and reason with probability has thus been studied extensively, especially in psychology and behavioural economics. Typically, this has occurred in the context of assuming *stationary probabilities* in the environment (i.e., probabilities that stay constant over time). This research shows that people are good at learning stationary probabilities from experience with relative frequencies (e.g. Edwards, 1961; Estes, 1976; Fiedler, 2000; Peterson & Beach, 1967), and it has been suggested that frequencies are among the few properties of the environment that are encoded automatically (Zacks & Hasher, 2002). At the same time, the research on heuristics-and-biases shows that probability assessments are sometimes also swayed by subjective ("intentional") aspects, like prototype-similarity (representativeness) or ease of retrieval, leading to biased judgements (Kahneman & Frederick, 2005). People also appear to over-weight extreme probabilities in their decisions when encountering them in numeric form (Tversky & Kahneman, 1992), but under-weight them when they are learned inductively from trial-by-trial experience (Hertwig & Erev, 2009). People frequently have problems with reasoning according to probability theory, leading to phenomena like base-rate neglect and conjunction fallacies (Kahneman & Frederick, 2005; Tversky & Kahneman, 1983), at least if they cannot benefit from natural frequency formats (Gigerenzer & Hoffrage, 1995) that highlight the set-relations between the events (Barbey & Sloman, 2007).

However, not all probabilities are stationary, as when, for example, the risks of default in a mortgage market fluctuate over time or the risk of hurricanes changes with a changing global climate. A small and mostly recent literature has started to model the cognitive processes by which people learn – and track – *non-stationary probabilities* (Gallistel, Krishan, Liu, Miller, & Latham, 2014; Khaw, Stevens, & Woodford, 2017; Ricci & Gallistel, 2017; Robinson, 1964). Because this research addresses changes in people's beliefs about probability it has (once again) highlighted the classical issue of learning by trial-by-trial updating or occasional shifts between discrete hypotheses (Bruner, Goodnow, & Austin, 1956), with the initial studies reporting support for processes of explicit hypothesis testing. In this article, we complement the existing literature in two ways. First, we report an experiment that investigates the robustness of the stepwise learning patterns that have been taken as evidence for hypothesis testing models over trial-by-trial updating models in the previous studies. Second, for the first time, we report a formal comparison between the competing models, applied to our own data as well as data from two other laboratories.

Tracking Probabilities in Non-Stationary Environments

Several previous studies have started to address how people learn and reason with non-stationary probabilities. They used tasks in which participants were presented with outcomes from a Bernoulli distribution that changed over time. Participants were asked to estimate the hidden Bernoulli parameter, by having them adjust a physical lever (Robinson, 1964) or a slider on a computer screen (Gallistel et al., 2014; Khaw et al., 2017; Ricci & Gallistel, 2017), with the option to change their estimate after each new observation.

Most versions of this paradigm have asked participants to estimate the proportion of items of a certain colour in a hypothetical box visualised on a computer screen (Gallistel et al., 2014; Khaw et al., 2017; Ricci & Gallistel, 2017) (Figure 1A). The participants drag a slider to indicate a value between 0 and 100 percent to indicate their current estimate, before locking in their guess, which initiates another draw of an item from the box. The participant may then choose to revise their estimate or leave it unchanged. This procedure is repeated for many trials. The data of interest are the realised outcomes, the underlying true probabilities of the outcomes, and the participant's estimates of these probabilities (Figure 1B). Most participants in previous studies exhibited stepwise updating behaviour: for long periods they did not adjust their estimates, at other times more often, but never on every trial.

As in many areas of the psychology of learning, there are two different ways of explaining how people infer probabilities from experience: models with their origin in the associationist traditions of behaviourism, reinforcement learning, and connectionist models emphasise the continuous updating of beliefs "trial-by-trial", while models with their origin in cognitive psychology emphasise the testing of discrete shifting between hypotheses.

A defining feature of trial-by-trial models is that the internal beliefs are updated each time a new data point is observed. They can be further separated into at least two kinds: delta-rule and memory-based models. The delta learning rule was introduced by Widrow and Hoff (1960) as an algorithm for updating the weights of nodes in a connectionist network (see Widrow & Lehr, 1993, for a review). In psychology, the most famous model based on this rule is the Rescorla-Wagner model of classical conditioning (Rescorla & Wagner, 1972), but it has also been adopted in many other domains (Behrens, Woolrich, Walton, & Rushworth, 2007; Busemeyer & Myung, 1988; Neal & Dayan, 1997; Verguts & Van Opstal, 2014).

In the context of probability estimation, delta-rule learning can be implemented as

$$\hat{p}_t = (1-\gamma)\,\hat{p}_{t-1} + \gamma\delta_{t-1} \tag{1}$$

where $\hat{p}_t$ is the probability estimate at time t, $\hat{p}_{t-1}$ the previous estimate, δ_{t-1} the prediction error at time $t-1$, and γ the learning rate. This rule has the advantage of being recursive: it can operate without access to memories going back any further than the latest observation.

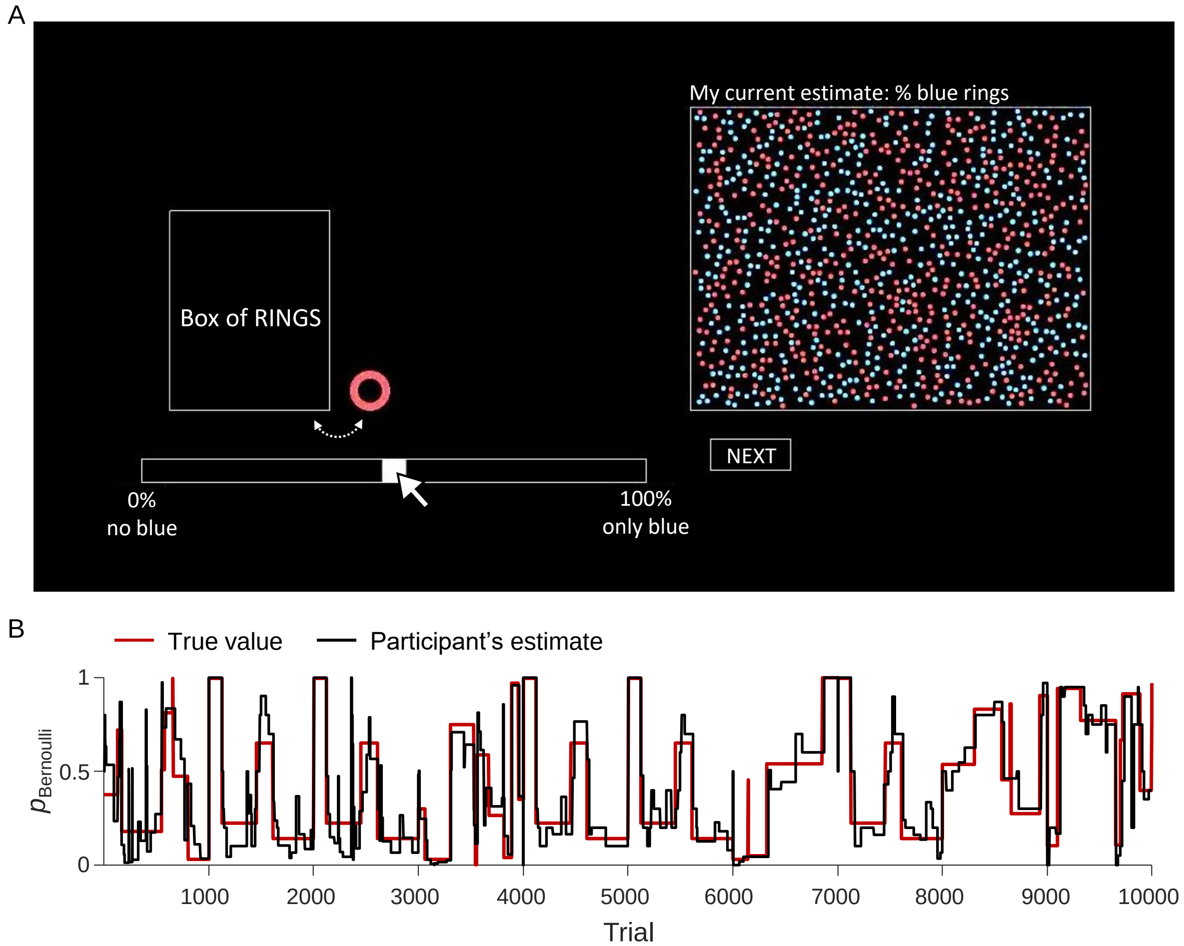

Figure 1 | Experimental paradigm. (A) Screenshot of our replication of the visual design of the experiments by Gallistel et al. (2014), Khaw, Stevens and Woodford (2017), and Ricci and Gallistel (2017). All text translated from Swedish to English and slightly enlarged for readability. (B) Example of response data (black) in an experiment where the hidden Bernoulli probability (red) was changing in a stepwise fashion (Participant 1 in Gallistel et al., 2014).

Memory-based models, on the other hand, rely on the memory of previously observed outcomes. They encode and then retrieve memories of events, often in the form of recency-constrained samples, to calculate beliefs on-line. These models have been applied to a variety of domains, including perceptual classification (Nosofsky & Palmeri, 1997), decision making (Lebiere, Stewart, & West, 2009), probability judgments (Costello & Watts, 2014; Juslin & Persson, 2002; Juslin, Winman, & Hansson, 2007), speech recognition (Gemmeke, Virtanen, & Hurmalainen, 2011), and consumption decisions (Mullainathan, 2002). Memory-based models have the advantage that, although they potentially draw on an extensive long-term memory, they are flexible in the sense that nothing needs to be pre-computed, but the computations are primarily performed at the time of judgement.

By contrast, hypothesis-testing models assume that people learn about the world by testing between explicit hypotheses about the state of the world based on the confirming or disconfirming feedback (Brehmer, 1974; Bruner et al., 1956). Hypothesis testing models have been applied to, for example, research on reasoning (e.g. Klayman & Ha, 1987; Oaksford & Chater, 1994; Wason & Johnson-Laird, 1970), categorisation (Ashby & Valentin, 2017; Bruner et al., 1956), and function learning (Brehmer, 1974, 1980). Because a single data point typically provides little evidence about a hypothesis, these models predict that the beliefs may sometimes stay unchanged over many trials.

According to current theory, trial-by-trial models are unable to account for the stepwise patterns found in experiments where participants track non-stationary probabilities (Gallistel et al., 2014; Ricci & Gallistel, 2017) (Figure 1B). Instead, it has been proposed that the stepwise response pattern is caused by discreteness in how the participants update their beliefs, which Gallistel et al. (2014) formalised in a hypothesis-testing model that they named the "If it ain't broke, don't fix it" (IIAB) model. According to this model, participants assess whether their current belief is "broke" after each new observation and only update their belief if the answer is in the affirmative. The suggestion is that humans do not estimate probabilities directly: they estimate changes in the hidden Bernoulli parameter and infer probabilities from this.

Purpose of this study

In the present work, we address three potential weaknesses in previous studies. The first one is related to the available data. Four previous studies (Gallistel et al., 2014; Khaw et al., 2017; Ricci & Gallistel, 2017; Robinson, 1964) have reported stepwise response updating in probability learning experiments with non-stationary probabilities. In three of those experiments (Gallistel et al., 2014; Khaw et al., 2017; Robinson, 1964), the underlying probability changed discretely. As noted by Ricci and Gallistel (2017), this is problematic, because it could mean that the discreteness in response patterns simply reflects the discreteness in the true underlying function, rather than discreteness in belief updating. Therefore, competing models of probability learning should primarily be tested using data from experiments in which the Bernoulli parameter changes in a *continuous* fashion. To the best of our knowledge, the study by Ricci and Gallistel (2017) is the only one so far that has performed such an experiment. However, for three[1] of their nine participants, the Bernoulli processes consisted of long periods of no change followed by a quite abrupt change, thus closely resembling a discretely changing parameter. Altogether, this means that current theories about human learning of non-stationary probabilities rely heavily on data from only six participants. The first purpose of the present study is to study the robustness of previous findings by using a larger participant sample.

A second potential weakness of previous studies is that the experimental design may unintentionally have invited stepwise behaviour. In all previous studies, participants were informed that the distribution they were inferring would change over the course of the experiment. If participants had reason to believe that the changes in the probability that they were tracking were discrete (e.g., because they were told that the box would be replaced "*from time to time*"), then this may have invited stepwise response behaviour. In addition to this, the bodily effort required to change one's estimate was in all previous studies substantially greater than that needed to maintain it. Robinson (1964) had the participants adjust a lever while Gallistel et al. (2014), Ricci and Gallistel (2017) and Khaw et al. (2017) required them to move the computer mouse, adjust a slider and move the mouse back again before clicking "Next". In contrast, maintaining one's previous guess merely required pressing the left mouse button once (Gallistel et al., 2014; Khaw et al., 2017; Ricci & Gallistel, 2017) or no action at all (Robinson, 1964). The asymmetry between the effort required to maintain or change the estimate may have affected the rate of re-estimations, especially when considering that participants performed 10,000 trials.[2] In Gallistel et al. (2014) and Ricci and Gallistel (2017) a further asymmetry existed in that a participant could move the slider by clicking right or left of its current position, which would make it jump a set distance. This made it easier to move it in large steps than in small ones. The second purpose of our study is to examine whether experimental design choices regarding instructions and response mode affect the degree of discreteness in response patterns.

A third and perhaps the most important weakness of previous work is that competing models have never been tested against each other using formal quantitative model comparison methods. Gallistel et al. (2014) compared models mainly based on visual comparisons of

[1] Subjects S1, S3, and S4 in the "aperiodic" condition.
[2] We do not know the exact number of trials in Robinson (1964) but each of his subjects performed the task for about 15 hours, which is a substantial amount of time.

summary statistics in the participant data with those produced by the models. Khaw et al. (2017) performed model comparison with the Bayesian Information Criterion (Schwarz, 1978) but only between trial-by-trial models from the economic literature. The third purpose of this study is to perform a comprehensive, formal comparison of competing models.

To summarise, the main contributions of the present article are as follows. First, we substantially increase the participant sample of data from learning experiments with continuously changing probabilities. Second, we investigate whether response effort and instructions affect the degree of discreteness in people's response patterns. Third, we perform a rigorous, likelihood-based comparison of hypothesis-testing and trial-by-trial updating models on all available data, which has not been attempted before.

EXPERIMENT

Previous studies on human learning and tracking of a non-stationary probabilities interpreted stepwise response behaviour as evidence that participants update their internal beliefs in a discrete manner (Gallistel et al., 2014; Ricci & Gallistel, 2017). This interpretation rests on the assumption that the discrete learning pattern constitutes a fairly stable and robust phenomenon that derives from the participant's mental shift between discrete hypotheses. In the present experiment we investigate the extent to which these results are sensitive to superficial specifics of the task, by experimentally varying two factors that we believe may affect the rate of re-estimations in the observed response behaviour. The first factor is the amount of information provided in the instructions to the participants about the non-stationarity of the probability they are asked to estimate. The second factor is the amount of effort required to make an update to the response slider.

Method

Participants. Sixty-two participants were recruited using posters advertising the study at several university campuses in Uppsala. Data from two participants were excluded from the analysis since they chose to terminate early. The mean age of those who completed the experiment was 24.7 (SD = 6.3). Forty-seven of these participants identified as female, eleven as male, and two as other. Participants were rewarded with gift vouchers for a major Swedish book shop chain (Akademibokhandeln). The total reward value depended on a participant's task accuracy, with a minimum fixed to the approximate equivalent of USD 11 and the maximum being approximately equivalent to USD 28.[3] Two participants in Condition 1, six in Condition 2, six in Condition 3 and five in Condition 4 received a signature on a participation form instead of gift cards. The study was approved by the Regional Ethical Review Board in Uppsala and conducted according to the Declaration of Helsinki Principles.

Stimulus and task. We replicated the visual design of the experiment described by Gallistel et al. (2014) to the best of our ability. The stimulus consisted of a screen showing a box labelled "Box of RINGS", a bar with a slider, and a rectangle filled with red and blue dots (Figure 1A). At the beginning of each trial, a ring would move out of the box and then stay beside it until the end of the trial. The task of the participant was to estimate the proportion of blue rings in the box by changing the value indicated by a slider on a bar that was labelled with "0% - No blue" and "100% - Only blue" on the left and right ends, respectively. Adjusting the slider caused the proportion of red and blue dots in the square labelled "My current estimate: % blue rings" to change to reflect the new proportion indicated by the slider position, which was intended as a visual aid to help participants "see" their currently chosen estimate.

Table 1.

Overview of Experimental Conditions as Combinations of the Response Mode and the Instruction Mode.

Condition	Effort mode	Response mode
1	High effort	Uninformed
2	High effort	Informed
3	Low effort	Uninformed
4	Low effort	Informed

Conditions. The experiment followed a two-by-two factorial design, with "Response Mode" and "Instruction Mode" as the independent variables (see Table 1). The first variable had two levels: "Low Effort" and "High Effort". In the High Effort response mode, participants revised their estimate by first clicking on the slider and then dragging it to adjust its value. When they were finished, they would click a "next" button to the right of the slider to initiate the next trial. In the Low Effort response mode of our experiment, no cursor or "next" button was visible, and the slider value would change whenever the mouse was moved. Participants initiated the next trial by a mouse click. The second independent variable also had two levels.

[3] Calculated using 2017 OECD purchasing power parity estimates.

In the "Informed" Instruction Mode, participants were explicitly informed about the non-stationarity of the generative process: they were told that the contents of the box might change after each draw and that these changes would occur throughout the task. They were also told that the changes could be fast or slow and that their task was to track the proportion as it changed. Participants in the "Uninformed" Instruction Mode were not provided with this information. In all four conditions, the hidden Bernoulli parameter was a sinusoidal with a minimum of 0, a maximum of 1, and a period of 500. Its value at the very first trial was 0.50. Condition 2 is almost identical to the design described in Ricci and Gallistel (2017). To the best of our knowledge, the only difference is that in the original study, the slider would jump a set distance when the participant clicked to the left or right of it.[4]

Procedure. At the start of the experiment, participants read a paper detailing that they were allowed to discontinue their participation at any stage; that the experiment would be divided into two sessions with a break in between; that the average difference between each of their guesses and the correct answer would determine their reward; and what the highest possible reward was. Meanwhile, a Swedish translation of the instructions found in Appendix A in Gallistel et al. (2014) was displayed on the screen, but without the passages relating to reporting that the box had changed. In the Low effort conditions, the relevant parts of the instructions were altered to explain how to answer using the Low Effort response mechanism. In the Informed conditions, paragraphs were added to explain that the box could be swapped every time a ring was put back into it, that these changes could be large or small, and that their task was to estimate the proportion of blue rings in the box and track it as it changed throughout the task (see the online materials at https://osf.io/zhv2r/ for English translations of the instructions). Participants were not told anything about how often they were supposed to make a change to the slider.

When the participant indicated that they had read everything, the experimenter would approach them to ask if they had understood all that they had read and if they had any further questions. If asked a question regarding anything not revealed in the instructions, the experimenter would respond that he was unable to provide that information. Any question pertaining to practicalities of how to carry out the task would be clarified upon request. The participants then completed 1,000 trials before a pause screen was displayed, inviting them to take a break. At their leisure, participants were allowed to commence the second session of

[4] This subtlety was not mentioned in the methods of the original study and we only became aware of it when scrutinising the methods of Khaw et al. (2017) who mention it in relation to their own experiment.

1,000 trials. The length of the break varied strongly across participants, ranging from 12 seconds to 17 minutes, with a mean of 3 minutes and 6 seconds.

After finishing the experiment, the participants filled out post-test questionnaires with questions concerning their beliefs about the generative function, self-assessed statistics proficiency, age, gender and education. Finally, they were asked to draw the probability of drawing a blue ring as a function of trial count into a graph. The questionnaires were administered on paper and filled in with pen. However, we found little use for the questionnaire data and did not analyse them.[5]

Analysis. All statistical analyses are performed using the JASP software package with default settings (JASP Team, 2019) and R (R Core Team, 2014).

Results

Accuracy. A visual inspection of the mean estimations (Figure 2A) shows that, on average, the participants tracked the wave-like pattern of the underlying probability reasonably well in all four conditions of the experiment. However, average accuracy is clearly highest in the condition where the participants were informed about the non-stationary generative function and making changes to the slider involved more effort (Figure 2B). We next perform statistical tests to determine if there is evidence for effects of Information Mode and Effort Mode on the root mean squared error (RMSE) between the generating probability and the participant's estimate.

Since the data violate the normality assumption of standard ANOVA analyses (Kolmogorov-Smirnov test, $p<10^{-13}$), we apply a Kruskal-Wallis and a Friedman test, with the two between-participant conditions as fixed factors and repeated measurement across blocks of 500 trials each. An initial main effects analysis suggests a main effect of Information Mode ($H(1) = 8.919$, $p = 0.003$) but not of Effort Mode ($H(1) = 0.685$, $p = 0.408$) or Block of Trials ($\chi^2(3) = 1.043$, $p = .791$). However, Dunn's post hoc test between the four between-participant cells indicates that this main effect is secondary to the interaction between Information Mode and Effort Mode presented in Figure 2C, with significantly lower median RMSE (approximately 0.13) in the Informed, High Effort condition than in the other three conditions (median RMSE > 0.30; $p_{\text{holm}} < .020$; see Appendix A for details on the Dunn's post hoc test).

To get an indication of how well participants performed in an absolute sense, we compare their accuracy to that of fictive observers who always responds 0.50 (Figure 2C, dashed lines)

[5] All questionnaire data are available in the online materials at https://osf.io/zhv2r/.

or randomly (Figure 2C, dotted lines). It is clear that despite that the average estimates track the functions in all conditions in Figure 2A, in three of the conditions the trial-by-trial accuracy in terms of RMSE is no better than what is expected from a participant who always responds with the probability 0.50. In sum: participants did not improve with training and although the average estimates tracked the underlying function, the trial-by-trial accuracy was poor in all conditions, except when the participants were informed about the nonstationary process and used the more effortful response method.

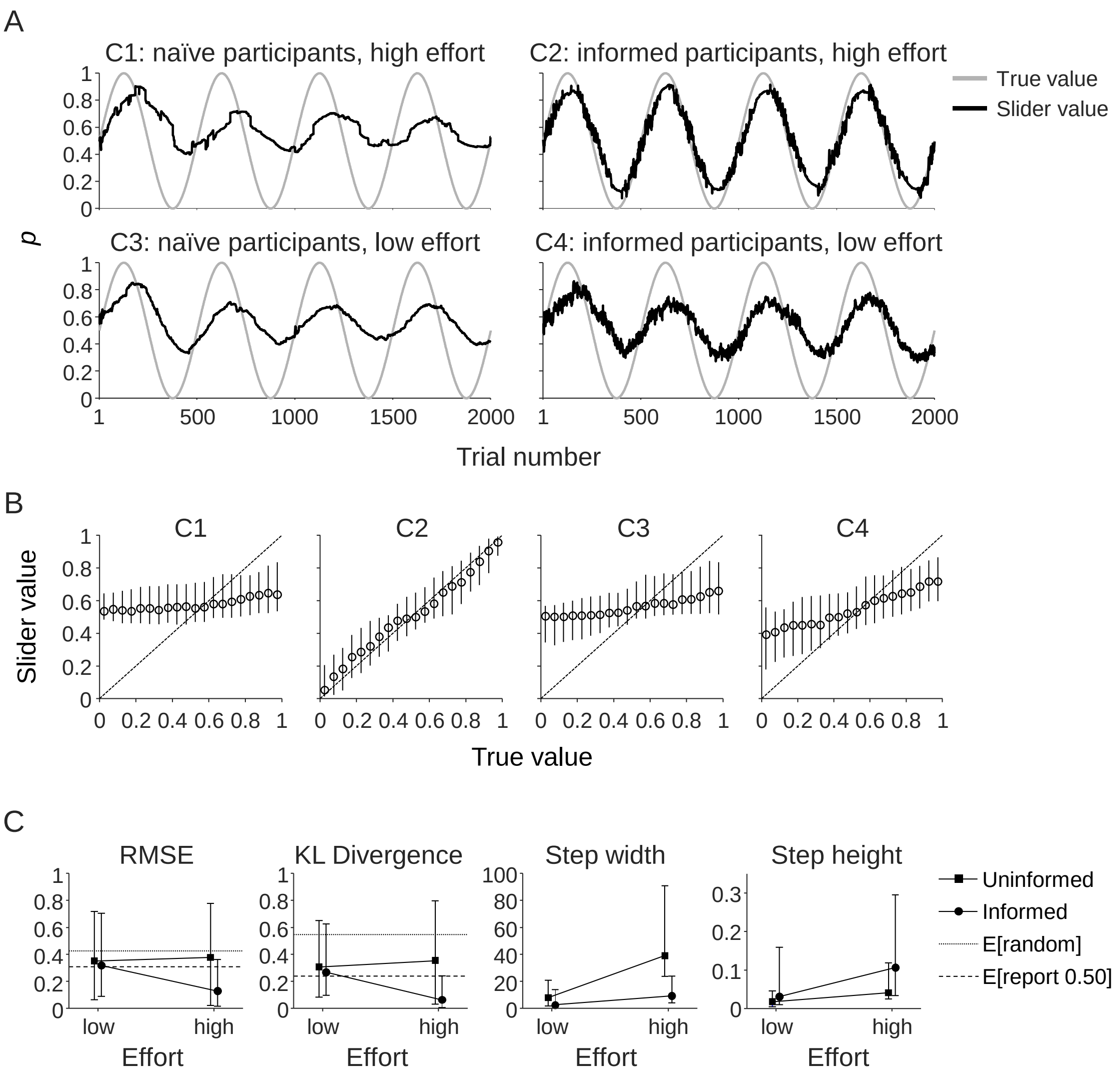

Figure 2 | Experimental results. (A) Average response in the four experimental conditions. (B) Median slider value plotted as a function of the true value of the tracked probability. The error bars indicate the interquartile range. (C) Median values of four summary statistics, split by condition. The error bars indicate the 25% and 75% quantiles. The title of each plot specifies the quantity on the y-axis. RMSE stands for root mean square error and KL stands for Kullback-Leibler. The dashed lines indicate the expected value of the summary statistic for an observer who responds randomly and an observer who always responds 0.50.

Following earlier work (Gallistel et al., 2014; Ricci & Gallistel, 2017), we consider the Kullback-Leibler (KL) divergence as an alternative measure of accuracy. We perform the same analyses with the KL divergence as the dependent variable and find an initial main effect of Information Mode ($H(1) = 8.656$, $p = 0.003$) but not of Effort Mode ($H(1) = 0.367$, $p = 0.544$) or Block of Trials ($\chi^2(3) = 2.187$, $p = 0.534$). Dunn's post hoc test shows that it is secondary to the interaction between Information Mode and Effort Mode (Figure 2C). The median KL divergence in the informed high effort condition (approximately 0.064) is significantly lower than in the other three conditions (median KL divergence > 0.267; $p_{\text{holm}} \leq .030$; see Appendix A for details on the Dunn's post hoc test). Hence, the results are consistent between the RMSE and KL divergence.

Step width. We next examine whether the experimental manipulations affect the average number of trials between slider updates, in previous studies referred to as "step width" (Gallistel et al., 2014; Ricci & Gallistel, 2017). The initial main effects analyses, with the same non-parametric tests as we applied to the RMSE, suggest significant main effects of Information Mode ($H(1) = 9.46$, $p = 0.002$), Effort Mode ($H(1) = 15.12$, $p < 0.001$) and Block of Trials ($\chi^2(3) = 69.33$, $p < 0.001$). The main effect of Block of Trials is an increasing step width, and thus decreasing rate of re-estimation, with additional training. The main effects of Information Mode and Effort Mode are qualified by the interaction illustrated in Figure 2C. Dunn's post hoc test shows that the median step width is significantly higher in the condition with no information about the non-stationarity of the process and a High Effort response mode (approximately 39) as compared to the other three conditions (medians between approximately 2 and 9: $p_{\text{holm}} < 0.020$, see Appendix A for details on Dunn's post hoc test). In sum: with more training the step width increased somewhat, and it was much larger in the condition without information about nonstationary and a high-effort response mode. In other words, when the participants were uninformed that the probability would change over time and the response required more effort, they were more reluctant to change their estimate.

Step height. Finally, we test if Information Mode and Effort Mode affected the average magnitude of the slider adjustments on trials when the estimate was updated, referred to as the "step height" in Gallistel et al. (2014) and Ricci and Gallistel (2017). Applying the same statistical tests as above, the results suggest main effects of Information Mode ($H(1) = 14.633$, $p < 0.001$) and Effort Mode ($H(1) = 11.363$, $p < 0.001$), but not of Block of Trials ($\chi^2(3) = 6.766$, $p = 0.080$). Dunn's post hoc test supports both a main effect of Information Mode and an interaction between Information Mode and Effort Mode, as illustrated in Figure 2C. The median step height was significantly greater with information about the non-stationarity than without, both with the Low Effort response mode (medians 0.0312 *vs.* 0.0177; $p_{\text{holm}} = 0.043$) and the High Effort response mode (medians .107 *vs.* 0.0445; $p_{\text{holm}} = 0.003$), suggesting a main effect of information regardless of the amount of effort required to update the response. In addition, the Informed, High Effort condition had a higher median than all of the other three conditions, suggesting a (catalytic) interaction for this specific condition (see Appendix A for the full results of Dunn's post hoc test). In sum, Block of Trials had no effect on the step height, but information about non-stationarity of the process increased it, especially when the high-effort response mode was used. Thus, when the participants were told that the underlying probability could change over time, the changes they made were larger, and this was especially the case if the response mode required more effort.

Discussion

Although the average estimates track the sinusoid function in all conditions (Figure 2A), in absolute terms the trial-by-trial accuracy was poor in three of the four conditions, in the sense that the deviation from the true probability on a given trial was no smaller than expected from a participant who responds with 0.50 on each trial (median RMSE approximately 0.35, see Figure 2C). In part, of course, this reflects the relative complexity of the task the participants are faced with. It takes at least a few observations to get a reliable estimate of the underlying probability. When this probability changes on each trial – as in our experiment – the observer's estimate will always lag behind the generating value. Optimal performance would require participants to infer the abstract function that relates the trial number to the true probability and to use this function to *predict the true probability on the next trial*. To induce this function from the "foggy" output of a constantly changing Bernoulli distribution is difficult, especially so if the observer is provided with only minimum information about the generative process. For this reason, some previous studies have assessed participant performance by comparing their responses to those of an optimal observer rather than to the true generating value (Gallistel et al., 2014; Khaw et al., 2017; Ricci & Gallistel, 2017). These analyses are helpful when investigating the degree of optimality of participants. However, here we are primarily interested in the relative performance between groups, for which any measure of accuracy seems suitable.

The high accuracy and distinctly stepwise re-estimation behaviour observed in Ricci and Gallistel (2017) and the other previous studies were only replicated when the participants were informed about the non-stationarity of the process beforehand and used the more effortful response mode, which are the conditions under which it has previously been observed. Better performance with more accurate prior information about the task is obviously no surprise. But this effect interacted with the effort required by the response mode in an interesting way. With a low effort response mode, there are frequent but small adjustments (median step width of approximately 5, suggesting about 100 re-estimations per block of 500 trials, of a median size of .03), and this holds regardless of whether participants are informed about non-stationarity or not. With the High Effort response mode, the pattern with relatively rare, large re-estimations only occurred with prior information that the process is non-stationary. The behavioural differences are indeed large. Participants without information about the non-stationarity and with the more effortful response mode rarely re-estimate and make rather small adjustments when they do (median step width of 39 trials, suggesting approximately 13 re-estimations per block of 500 trials, with a median size of .04). The participants with information about the non-stationarity and with the more effortful response mode often change their estimates (median step width of 9 trials suggesting approximately 56 re-estimations per block of 500 trials) and usually by quite a lot (median step height of .11) The characteristic stepwise patterns of the predictions of the IIAB-model (Gallistel et al., 2014) were thus observed in only one cell and appear to arise under specific conditions, suggesting that rare but large re-estimations are not necessarily intrinsic to the cognitive process.

An alternative explanation of the effects of the independent variables on step width and step height is that they merely reflect the fact that the Low Effort response mode results in an increase in the number of small, accidental adjustments. When the slider is “stuck” to the mouse cursor, participants might occasionally produce unintended adjustments. When the slider has to be dragged, this is less likely to occur. This kind of “shaky hand” error would decrease both the average step width and step height. There are relatively small negative main effects of having a low effort response mode on both of those dependent variables. Since we cannot rule out that the shaky hand effect exists, these should be interpreted with some caution. However, the substantial interaction between High Effort and Information Mode is not possible to attribute to such error. If unintentional adjustments as a result of the low effort response mechanism is a pervasive phenomenon, it should affect the results equally regardless of what information is provided. We would therefore argue that the main result of our experiment – that the previously observed stepwise updating arises as a result of particular combinations of circumstances – holds regardless of whether the Low Effort response mode increases the number of accidental adjustments.

A tentative interpretation of the results is that people spontaneously tend to be “myopic”, only considering small samples of the most recent observations, which they project onto the next trial as an estimate of the probability. This estimate can, in principle, change from trial to trial, as is consistent with the small and frequent adjustments produced by the participants in several conditions, and their overt expression of the estimate is affected by the effort required to produce the response, as is consistent with the significant effects of Response Mode. Intriguingly, the effortful response mode seems to have invited participants to consider larger sample sizes, allowing them to better track changes in the underlying probability.

To conclude, a key implication of these results is that the discreteness of the response data seems sensitive to external factors, which calls into question whether it should be thought of as inherent to human probability inference as has been done in previous literature. Instead, the pattern may reflect adaptations to the particulars of the task at hand. In other words, it is possible that the internal belief updating is continuous and only the slider adjustments occur discretely.

MODELLING

According to the currently leading theory, human behaviour in probability estimation tasks is consistent with hypothesis-testing models and cannot be explained by any trial-by-trial updating model (Gallistel et al., 2014; Ricci & Gallistel, 2017). Above, we presented experimental evidence that calls the first part of this claim into question; the remainder of this paper is dedicated to evaluating the plausibility of the second part, by using formal model comparison techniques. Our approach makes four important methodological improvements on previous studies. First, instead of setting parameters manually, we use maximum-likelihood fitting to determine parameter values. Second, instead of fitting models to summary statistics, we fit them to the raw data. This way, we use all available information and avoid having to decide which statistics to look at and how to weight them against each other. Third, instead of evaluating goodness of fit through visual inspection of plots, we use formal model comparison techniques. Fourth, instead of evaluating the models only against our own data, we also include all available data from other studies in our analyses.

Factorial model design

When models differ from each other in multiple ways, it is hard to identify which factor explains the success of one model over another. To circumvent such identifiability problems, we apply a method known as *factorial model comparison* (van den Berg, Awh, & Ma, 2014). Just as in factorial experimental designs and factorial ANOVAs, this means that we pair every choice in one factor with every possible choice in the other factors. The goal is not only to identify the model that best captures the underlying process, but also to quantify evidence for each factor level, much as an ANOVA quantifies the evidence for each of the main effects. We deconstruct the models that we consider here into two factors: the updating mechanism and the threshold mechanism. For convenience, Table 2 provides an overview of the most important mathematical terms and symbols appearing in the model specifications.

Table 2. *Overview of Mathematical Terms Used in the Model Specifications.*

Term	Description
p_{true}	True value of the Bernoulli parameter that participants are trying to estimate (p_g in Gallistel et al., 2014)
p_{slider}	The current estimate of p_{true} as represented by the slider ($\hat{p}_{\text{g}}$ in Gallistel et al., 2014)
p_{observed}	The current estimate of p_{true} based on the (latest) outcome observations (p_{o} in Gallistel et al., 2014)
O_t	The observed outcome (0 or 1) on trial t
E	Discrepancy between p_{slider} and p_{observed}, measured as the absolute difference or KL divergence (comparable to E in Gallistel et al., 2014)
N	Number of trials since the last slider update took place
T_1	Threshold on ε, determining whether a slider update is performed (T_1 in Gallistel et al., 2014); this parameter appears in all models
T_2	Threshold on the posterior odds of a change, determining whether the observer beliefs that a change point was missed (T_2 in Gallistel et al., 2014); this parameter only appears in IIAB models
Λ	Learning rate; this parameter only appears in delta-rule models
A	Memory weight; this parameter only appears in memory-based averaging models
$\sigma_{\text{unexplained}}$	Standard deviation of the normally distributed error term, which takes care of unexplained variance
μ_{T1}, σ_{T1}	Mean and standard deviation of the distribution of T_1

Factor 1: Updating mechanism. This factor determines how and when the observer updates their belief about the hidden Bernoulli probability, p_{true}. We consider three options: the IIAB mechanism, a delta-rule mechanism, and a memory-based averaging mechanism. The essence of the IIAB mechanism (Gallistel et al., 2014) is that it maintains a list of "change points" that is updated through hypothesis testing. The change points summarise at which earlier time points there was, according to the model, a change in p_{true} and how large each

supposed change was. After making a new observation, the mechanism tests the hypothesis that "something is broke". It does so by computing how much the currently held belief about p_{true} – as encoded in the most recently registered change point – deviates from the estimate based on all observations since the last change point. When this discrepancy exceeds a threshold T_1, it is concluded that "something is broke" and that it "needs fixing." The updating mechanism then proceeds to a second stage, where three further hypotheses are tested about what might be wrong: the last registered change point was incorrect and must be expunged, it was at the wrong point and should be moved, or there has been a new change point after the last one encoded, which now needs to be registered. Once a decision has been made on this, the mechanism updates the list of change points accordingly and adjusts the slider value, p_{slider}, to make it consistent with what is now the latest estimated change point. For a detailed description of the mechanism, see Gallistel et al. (2014). Importantly, since it can take many observations before it is detected that "something is broke", slider updates in this type of model tend to happen in a discrete fashion.

The second updating mechanism that we consider is the delta rule, which we abbreviate as "Delta". Unlike the IIAB mechanism, the delta rule has no notion of hypothesis testing and, therefore, has no threshold on its belief updating. Instead, it updates its estimate of the hidden Bernoulli parameter after each new observation. It does so by computing a weighted average of the previous estimate, $p_{\text{observed},t-1}$, and latest observation, O_t, through

$$p_{\text{observed},t} = (1-\lambda)\, p_{\text{observed},t-1} + \lambda O_t, \tag{2}$$

where parameter λ is the learning rate. Another difference to the IIAB mechanism is that since an update is made on each trial, the magnitude of the updates will often be very small. However, considering that it is effortful in both time and energy to adjust the slider value, it seems reasonable to assume that observers only do so when the discrepancy between slider and belief has grown sufficiently large. Therefore, we impose a response threshold T_1 on this discrepancy, such that a slider update is only made when it is considered to be worth the effort.

The third and final updating mechanism that we consider is a memory-based weighted average, which we abbreviate as "M-Avg". In this mechanism, the probability estimate is computed as

$$p_{\text{observed},t} = \sum_{i=1}^{t} w_i O_i, \tag{3}$$

where the weights decrease exponentially in history, $w_i = \frac{\alpha^{t-i}}{\sum_{j=1}^{t} \alpha^{t-j}}$. Parameter α is constrained to the range [0,1] and can be thought of as a history weight: the larger its value, the more weight is given to observations further back in time. If $\alpha = 0$, then p_{observed} is equal to the last observation; if $\alpha = 1$, then p_{observed} equals a plain average of all observations; if $0 < \alpha < 1$, then p_{observed} is a weighted average of all observations, with higher weight given to more recent observations. Just as in the Delta mechanism, we include a response threshold such that slider updates are made only when the discrepancy between belief and current slider value is sufficiently large.

Factor 2: Threshold mechanism. All three updating mechanisms described above involve a threshold, denoted as T_1: the IIAB mechanism has an "is it broke" threshold that prevents hypothesis updating when there is too little evidence that something is wrong and the other two updating mechanisms have a response threshold that prevents slider updating when it is not worth the effort. In the original formulation of the IIAB model, the "is it broke" discrepancy is measured as KL divergence, ε=KL($p_{\text{observed}} \parallel p_{\text{slider}}$)×$n$, where p_{observed} is an estimate of p_{blue} based on the outcomes observed since the last change point, p_{slider} is the currently held belief and n is the number of trials since the last update. For the response threshold in the other two mechanisms, however, a more obvious measure of discrepancy is the absolute difference, $\varepsilon=|p_{\text{observed}} - p_{\text{slider}}|$. This is indeed what Gallistel et al. (2014) used in their implementations of delta-rule models. These two proposals differ from each other in two ways: the discrepancy is either measured as KL divergence (ε=KL(Δ)) or as an absolute difference ($\varepsilon=|\Delta|$) and it is either multiplied by n (ε=KL(Δ)×n; $\varepsilon=|\Delta|\times n$) or not. To dissociate the effects of threshold choice from effects of updating mechanism on goodness of fit, we cross these options factorially, which gives rise to four different threshold mechanisms. Combining each updating mechanism with each threshold mechanism results in a total of 12 models (see Table 3).

Threshold variability. Since cognitive processes are generally noisy, it seems plausible that threshold T_1 varies from trial to trial. Therefore, following the proposal by Gallistel et al. (2014), we draw the value of T_1 on each trial from a normal distribution with a mean μ_{T1} and standard deviation σ_{T1}, both of which are fitted as free parameters.

Table 3. *Overview of Factors and Factor Levels in the Factorial Model Design. The First Factor Specifies the Updating Mechanism, of Which Three are Considered: IIAB, the Delta Dule, and Memory-based Averaging. The Second Factor Specifies the Threshold Mechanism,*

of Which Four are Considered: Absolute Error, Absolute Error Multiplied by the Number of Trials Since the Last Slider Update, KL Divergence, and KL Divergence Multiplied by the Number of Trials Since the Last Slider Update.

Factor name	Level name	Level-related parameters
Updating mechanism	IIAB	T_2
	Delta	λ
	M-Avg	α
Threshold mechanism	$\varepsilon = \lvert p_{\text{observed}} - p_{\text{slider}} \rvert$	μ_{T1}, σ_{T1}
	$\varepsilon = \lvert p_{\text{observed}} - p_{\text{slider}} \rvert n$	μ_{T1}, σ_{T1}
	$\varepsilon = \mathrm{KL}\left(p_{\text{observed}} \parallel p_{\text{slider}} \right)$	μ_{T1}, σ_{T1}
	$\varepsilon = \mathrm{KL}\left(p_{\text{observed}} \parallel p_{\text{slider}} \right) n$	μ_{T1}, σ_{T1}

Model fitting methods

Due to the existence of latent variables in the IIAB models and the presence of trial dependencies, the proper likelihood function is intractable for some of the models. Therefore, we use a simplified, "custom" likelihood function for model fitting (Appendix B). We use the Bayesian Adaptive Direct Search (BADS) method (Acerbi & Ma, 2017) to find the parameters that maximise this function. In order to reduce the risk of terminating in local maxima, we run BADS thirty times with different initial parameter values. Prior to each run, we evaluate the likelihood function for five hundred randomly drawn parameter vectors and choose the vector that gives the highest outcome as the initial parameter vector for BADS. Results from a model recovery analysis confirm that these methods allow for reliable model comparison (see Appendix C).

Benchmark dataset

To get the most out of the model comparison, we fit the models to both our own data and the data from three previous studies, which were made available to us by the respective authors (Gallistel et al., 2014; Khaw et al., 2017; Ricci & Gallistel, 2017; see Table 4).[6] The number of trials per participant varied from 2,000 to 10,000 across experiments, with a grand total of

[6] There is one other study using the same paradigm (Robinson, 1964), but it has no preserved record of the data known to us.

408,000 trials. To the best of our knowledge, all experiments were conducted in sessions of 1,000 trials each, with breaks between consecutive sessions. Because of these breaks, we suspect that parameter values might not be stable across sessions. Therefore, we fit the models separately to each session, of which we have 408 in total (Table 4). All data are available online as a benchmark data set at https://osf.io/zhv2r/.

Table 4. *Overview of Datasets Used to Evaluate the Models.*

Exp. ID	Study	Underlying function	Number of participants	Number of trials per participant	Number of trials per session	Total number of sessions
E1	Gallistel et al. (2014)	Stepwise	10	10,000	1,000	100
E2	Ricci & Gallistel (2017)	Continuous (aperiodic)	5	10,000	1,000	50
E3	Ricci & Gallistel (2017)	Continuous (periodic)	3[7]	9,000 (2x) 10,000 (1x)	1,000	28
E4	Khaw et al. (2017)	Stepwise	11	10,000	1,000	110
E5	Present study	Continuous (Condition 1)	15	2,000	1,000	30
E6	Present study	Continuous (Condition 2)	15	2,000	1,000	30
E7	Present study	Continuous (Condition 3)	15	2,000	1,000	30
E8	Present study	Continuous (Condition 4)	15	2,000	1,000	30

Model comparison

We fit the twelve models (Table 3) separately to each of the 408 datasets (Table 4) for a total of 4,896 fits. In doing so, we include only the first 750 trials from each dataset, so that we can use the remaining 250 trials for cross validation.

Model comparison based on AIC values shows a large heterogeneity between participants (Figure 3A, left): there is not a single model that provides a good fit to all datasets and every model seems to perform well on at least one dataset. Despite this heterogeneity, it is clear that

[7] This experiment had 4 subjects, but we suspect that for one of them the responses were flipped between two sessions. We excluded this subject from our analyses.

some models perform better overall than others. In particular, the IIAB models generally fit worse than the Delta and M-Avg models. When averaging the relative AIC values across datasets (Figure 3A, right), the most successful model is the one with a memory-based updating mechanism and a threshold mechanism based on the absolute difference (M-Avg with *ε*=|Δ|). All other models have an average AIC value of at least 50 points larger, which would even under a very conservative criterion be reason to reject them all. However, given the heterogeneity at the individual level, it seems unwarranted to rule out individual models at this stage.

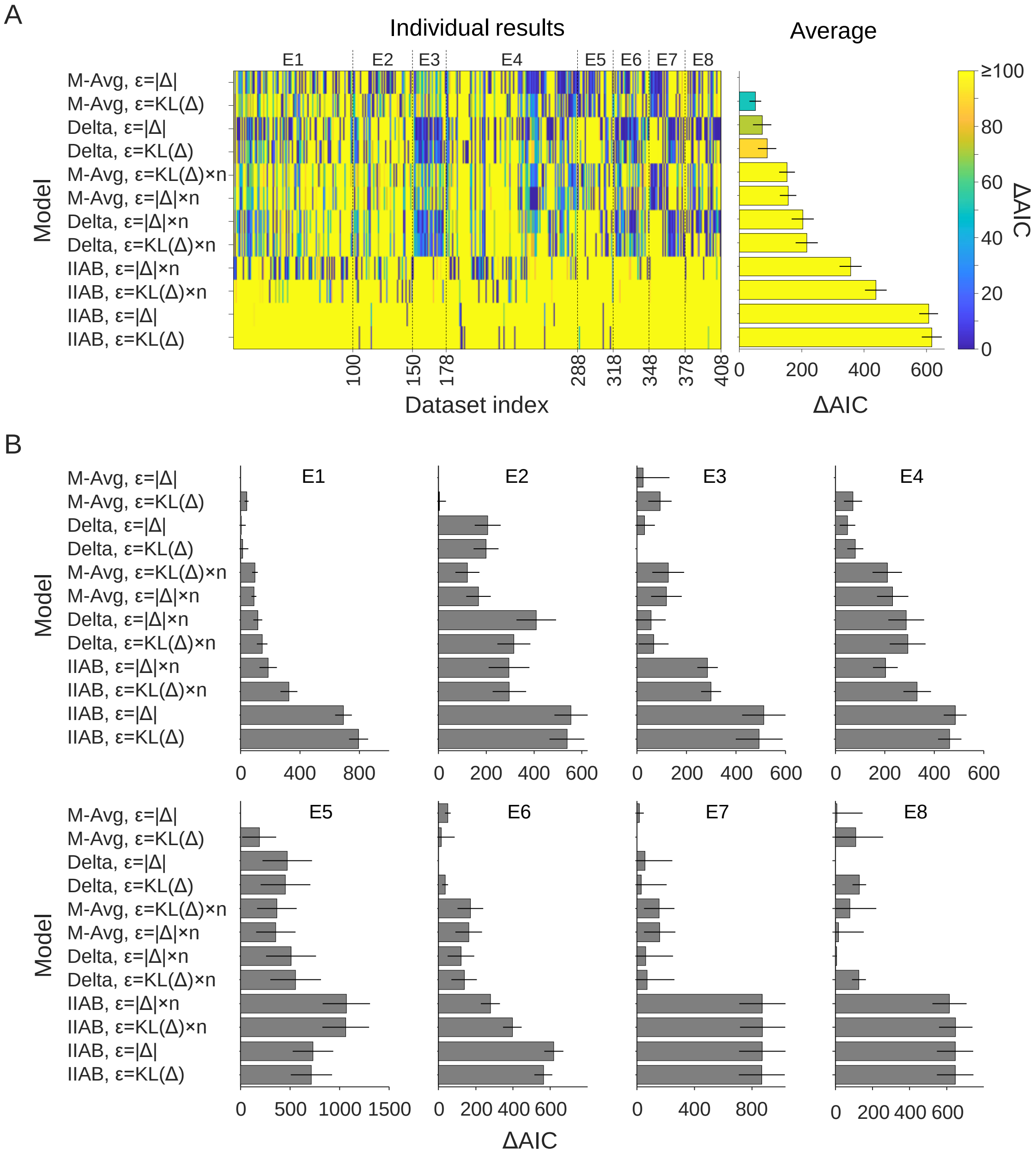

Figure 3 | Model comparison based on AIC scores. (A) AIC-based comparison of the twelve main models fitted to 408 datasets. Left: AIC values relative to the best-fitting model for individual datasets. Right: Relative AIC values averaged across all datasets. (B) Model comparison split by experiment, with the models ordered in the same way as in panel A.

Instead of looking at individual models, it may be more informative to look at the success of each factor level. To this end, we compute the *log factor likelihood* as proposed by Shen and Ma (2019) to quantify the evidence for each factor level (Figure 4). Consistently across experiments, the results reveal strong evidence against the IIAB updating mechanism, while the two trial-by-trial mechanisms perform approximately equally well in most experiments. In terms of threshold mechanisms, we observe that there is evidence against models that incorporate the number of trials since the last slider update, while there is approximately equal evidence for mechanisms based on the absolute difference and mechanisms based on KL divergence.

While AIC is widely used as a measure of *fit*, it is not necessarily a good measure of *prediction* due to possible overfit. Therefore, we next compare models based on the log likelihood of the last 250 trials of each session, which were not included during model fitting. The results of this cross-validation analysis (Appendix D) show a pattern that is largely similar to the AIC-based results: there is large heterogeneity at the level of individual datasets, models with an IIAB updating mechanism generally perform poorly, and there is no strong evidence in favour or against specific threshold mechanisms. However, the evidence is now more even between the Delta and M-Avg mechanisms and it is harder to distinguish between the threshold mechanisms.

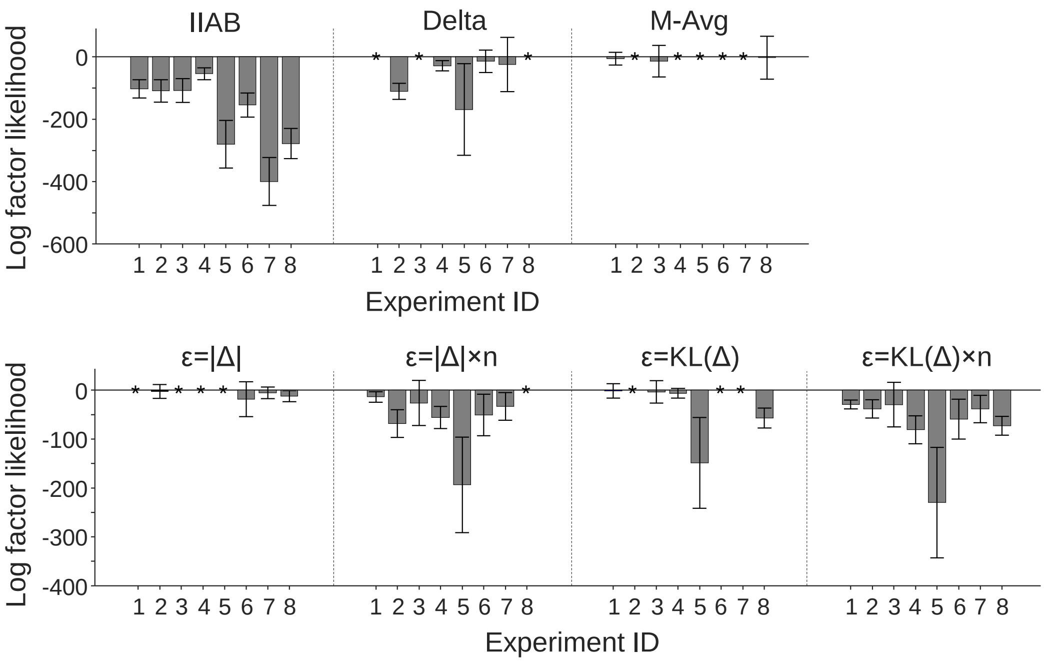

Figure 4 | Factor level comparison. Top: Evidence for each level in the first factor relative to the most successful level, combined across all models. Bottom: Evidence for each level in the second factor relative to the most successful level, combined across all models. The most successful levels in each experiment are indicated by asterisks.

Model fits

The model comparison results provide insight into how well the models perform in relation to each other. However, those results would be of little value if all models were extremely poor descriptions of the data. Visual inspection of the fits indicates that the best model overall (M-Avg with ε=|Δ|) generally does a good job in describing the participant responses (see Figure 5 for a few examples; an overview of all fits can be found online at https://osf.io/zhv2r/). Across all 408 datasets, the average RMSE between the maximum-likelihood fit of this model and the participant data is 0.139 ± 0.004. Consistent with the results of the formal model comparison, we find that the RMSE is higher for the best-fitting Delta model (0.142 ± 0.004) and the best-fitting IIAB model (0.153 ± 0.004).

Parameter estimates

An overview of the maximum-likelihood parameter estimates for each model is found in Appendix E. The estimate of $\sigma_{\text{unexplained}}$ is on average smaller in the M-Avg and Delta models than in the IIAB models, suggesting that the latter kind of model leaves more variance unexplained than the former two, which is consistent with the model comparison results. In the best-fitting model (M-Avg with ε=|Δ|), the median value of this parameter is 5.64×10^{-2}. This is rather small in relation to the response scale (0 to 1), which corroborates our earlier conclusion that the model provides a reasonably good account of the data. For parameters μ_{T1} and σ_{T1} we find median values equal to 0.470 and 0.207, respectively. These values indicate a relatively high response threshold with quite a high degree of trial-by-trial variability. We speculate that the variance captured by these parameters also includes other sources of variability in response behaviour (e.g., noise in the calculation of ε and variability in the applied learning rate or memory weight) which are not specified in the models.

Finally, we estimate how much outcome history the winning M-Avg takes into account in its trial-by-trial estimates of p_{true}. The memory weight in this model drops exponentially with history length, with a rate that is determined by parameter α. We quantify the history length as the number of trials that cover 95% of the total weight mass. Based on the maximum-likelihood estimates of α, we find a median length of 33 trials (25% quantile: 19 trials; 75% quantile: 97 trials).

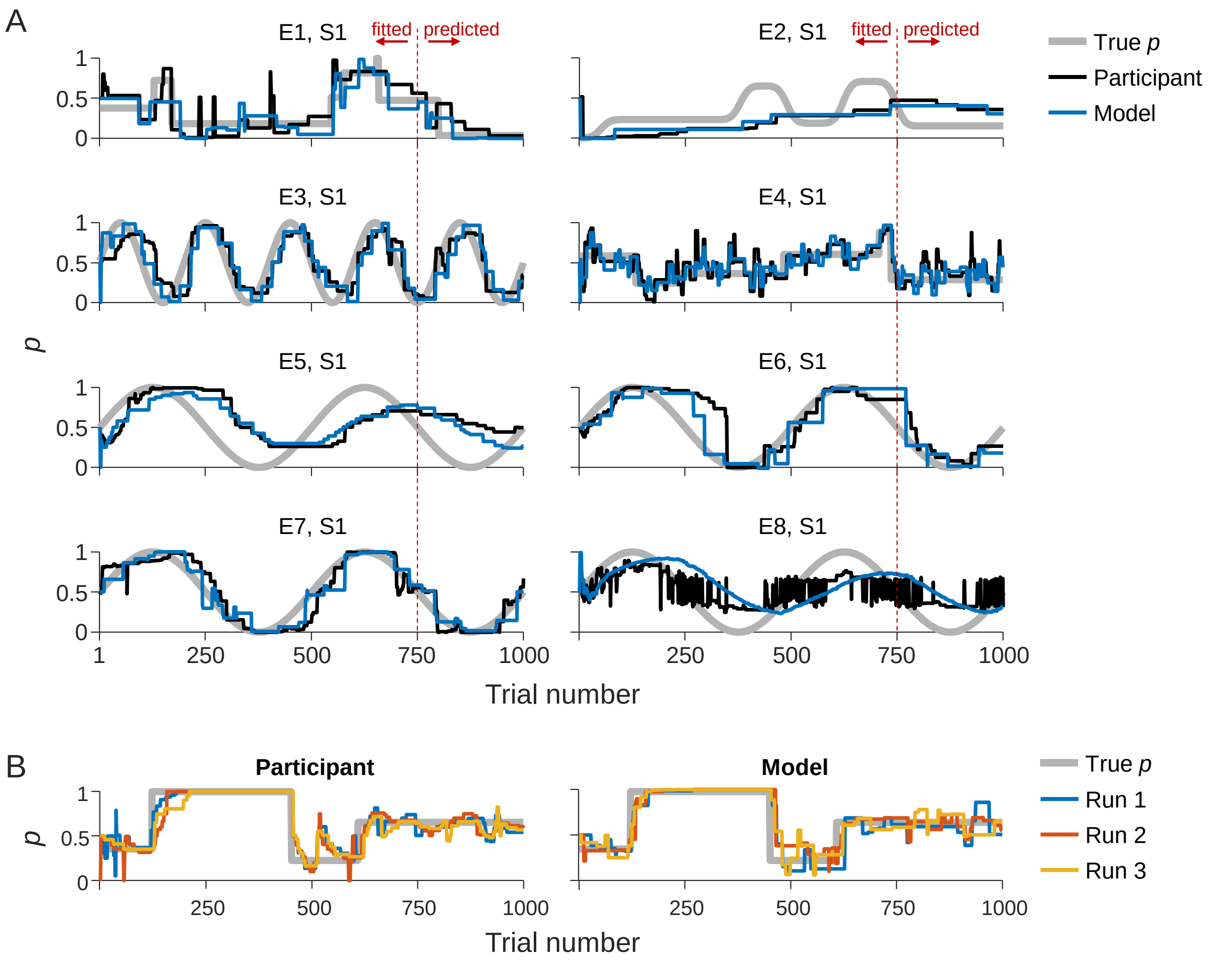

Figure 5 | Fits of the best-fitting model (M-Avg with ε=|Δ|) to the raw response data. (A) Data and model fit for the first session of the first participant in each of the eight experiments. The model fit was computed using a forward simulation using the maximum-likelihood parameter estimates. (B) Left: Responses of Participant 1 in experiment E4 to sessions 2 (blue), 6 (red), and 9 (yellow). The value of $p_{\text{Bernoulli}}$ (grey) as well as the observed outcomes presented to the participant were identical in those sessions. Right: three runs of the model with parameters fixed to the maximum likelihood estimates obtained from fitting the data of session 2. Note that the variability across runs is of similar magnitude between participant and model.

Model comparison with fixed thresholds

All models that we have tested so far had a variable threshold. We next address two questions regarding this variability. First, how much do the fits suffer if the variable threshold is replaced by a fixed one? Second, do the conclusions that we draw from the model comparison depend on the existence of threshold variability? To answer these questions, we re-fit the twelve models with σ_{T1} fixed to 0. While the AIC value worsens for each of the twelve models – by a minimum of 728±38 points – the model order is near-identical to the order we found with the models with variable thresholds (Figure 6A). Hence, while the assumption of variability in thresholds contributes strongly to the success of all tested models, our main conclusions do not critically depend on it.

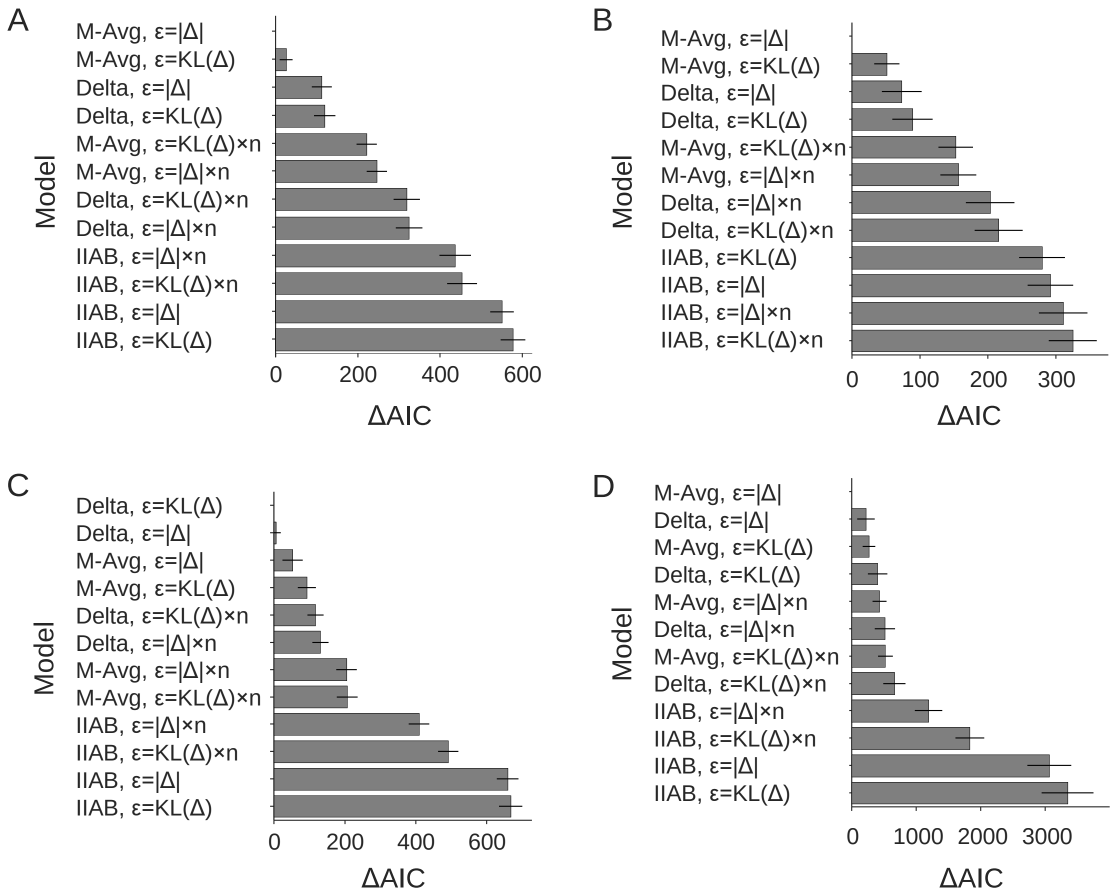

Figure 6 | Results from additional model comparisons. (A) Model comparion results after removing threshold variability. (B) Model comparison results after adding a response threshold to the IIAB models. (C) Model comparison results after adding a second kernel to the Delta models. (D) Model comparison results based on fitting models to full datasets instead of sessions.

IIAB with a response threshold

The IIAB models have a threshold at the belief updating stage, while the trial-by-trial updating models have a threshold at the response stage. This creates a potential interpretation problem regarding the model comparison results: is the relatively poor performance of the IIAB models due to its belief updating mechanism or due to it lacking a threshold at the response stage? Or, put differently: can the IIAB model be salvaged by adding a response threshold? To answer this question, we add a response threshold to the IIAB models and fit them again to all 408 datasets. We find that this modification improves the average AIC values of the IIAB models by 200±6 points. However, despite this substantial improvement, the models still perform poorly compared to the trial-by-trial models (Figure 6B).

Two-kernel delta-rule model

Under conditions where there are large and infrequent changes, as in much of the experimental data considered in this study, the standard versions of the delta-rule and memory-averaging models face a problem. If a lot of weight is put on the most recent history (by having a high learning rate in the delta model or a low memory weight in the memory-averaging model), the model will quickly catch on to changes but exhibit excessive volatility during the long periods where the true probability is unchanged. If, on the other hand, it is only given a little weight, excessive volatility will be avoided but the model will be slow to catch on to sudden changes. As a potential solution, Gallistel et al. (2014) considered a two-kernel variant that keeps track of two running averages. One kernel has a fast learning rate and the other a slow one. When there is a sudden change, the discrepancy between the two estimates is large, which is used as a signal that there has been a change and that the fast kernel should be trusted. After some observations, the slow kernel will catch up and the discrepancy will decrease, signalling that the fast kernel is no longer relevant. The model will then revert to reporting the slow kernel's estimate. A similar extension is conceivable for the memory-averaging model, by using two memory weights, but we limit our present analysis to the Delta model.

We next test whether a two-kernel delta-rule model is a serious contender to the other models we have considered so far. The model keeps two estimates of the Bernoulli probability, $p_{\text{slow},t} = (1-\lambda_{\text{slow}})p_{\text{slow},t-1} + \lambda_{\text{slow}}O_t$ and $p_{\text{fast},t} = (1-\lambda_{\text{fast}})p_{\text{fast},t-1} + \lambda_{\text{fast}}O_t$. On trials where the absolute difference between the two estimates is larger than a threshold Δ_c, the model takes p_{fast} as its estimate of the Bernoulli probability; otherwise it uses p_{slow} as its estimate. The model thus has two additional parameters compared to the standard delta-rule model tested above. As in the main analysis, we combine this updating mechanism with all four thresholding mechanisms (Table 3). We find that across all 1,632 fits, the additional kernel improves the AIC value of the delta-rule models on average by 133±5 points. In terms of model comparison, the two-kernel delta-rule model with ε=KL(Δ) outperforms all other tested models (Figure 6C).

Fits to full datasets

In the analyses presented above, we have been fitting models to sessions of 1,000 trials each to allow for the possibility that parameters can vary between sessions. To verify that our conclusions do not critically depend on this choice, we next fit the models to the full datasets, that is, with only one set of parameters per participant. Although there are small differences in the model order (Figure 6D), the overall findings are the same as before: the M-Avg model with ε=|Δ| comes out as the overall best model and the four IIAB models perform poorly. Hence, the general conclusions of our model comparison do not seem to critically depend on whether we fit the models to single sessions or to full datasets.

Fits to summary statistics

So far, we have been comparing models based on log likelihoods computed from fitting raw data. One might argue, however, that it is also important that a model captures key summary statistics derived from the raw data. In the context of probability estimation, Gallistel et al., (2014) argued that two important summary statistics are the step widths and step heights. While we agree with this, we are not convinced by their conclusion that it is impossible for *any* trial-by-trial updating model to account for the empirical joint distributions of these statistics. The problem is that this conclusion was based on visual inspection of model behaviour for a supposedly small number of manually picked parameter settings, rather than on a systematic exploration of the parameter space.

To investigate more formally how well the models are able to account for the empirical joint distributions of step widths and step heights, we use an optimisation algorithm to find the parameters that minimise the Jensen-Shannon divergence[8] (JSD) between the empirical and the predicted distributions. Since repeated computation of joint distributions makes this optimisation very time-consuming, we fit the models with only one threshold variant in the second model factor. To make it unlikely that our choice biases the results in favour of the trial-by-trial models, we choose $\varepsilon=|\Delta|\times n$ for all three models, which was the most successful variant for the IIAB model in the main analysis (Figure 3). We fit these models to full datasets, because joint distributions for session-based data often contain too few data points for reliable fitting.

The left panel of Figure 7A presents the empirical data that led Gallistel et al. (2014) to conclude that there are serious discrepancies between the kind of patterns generated by participants and those generated by trial-by-trial models. In contrast to their conclusion, however, we find that the three models perform approximately equally well, both visually (Figure 7A) and in terms of JSD (IIAB: 0.22±0.03; Delta: 0.22±0.04; M-Avg: 0.19±0.04). Also at the individual level, visual inspection of the fits does not indicate an advantage of the IIAB model over the M-Avg and Delta models in any of the experiments (Figure 7B). In fact, when averaging the JSD across all 89 participants (Figure 8A), the IIAB model accounts for the distributions substantially worse than the M-Avg and Delta models (IIAB: 0.28±0.017; Delta: 0.17±0.013; M-Avg: 0.17±0.011).

[8] The Jensen-Shannon divergence is a symmetric variant of the Kullback-Leibler divergence and has the advantage that it is always finite, even when one of the inputs is zero.

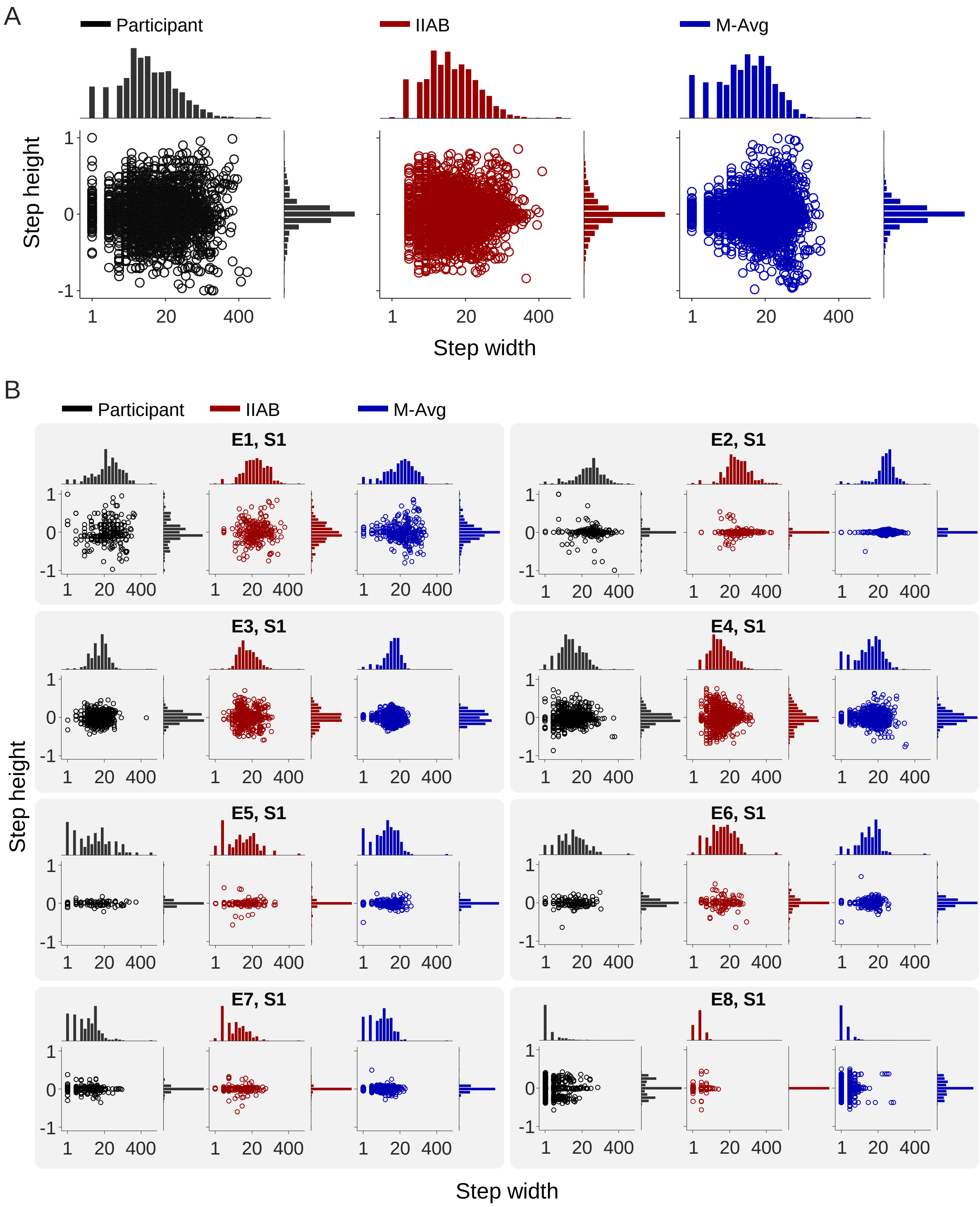

Figure 7 | Model fits to summary statistics. (A) Left: Joint distribution of step widths and step heights of all participants in E1 pooled together (cf. Figure 15 in Gallistel et al., 2014). Center: pooled fits of the IIAB model. Right: pooled fits of the M-Avg model. (B) Subect-level joint distributions of step widths and step heights and fits of the IIAB and M-Avg models. The first participant of each experiment is shown. Fits of the Delta model look very similar to those of the M-Avg model (see Supplementary Materials).

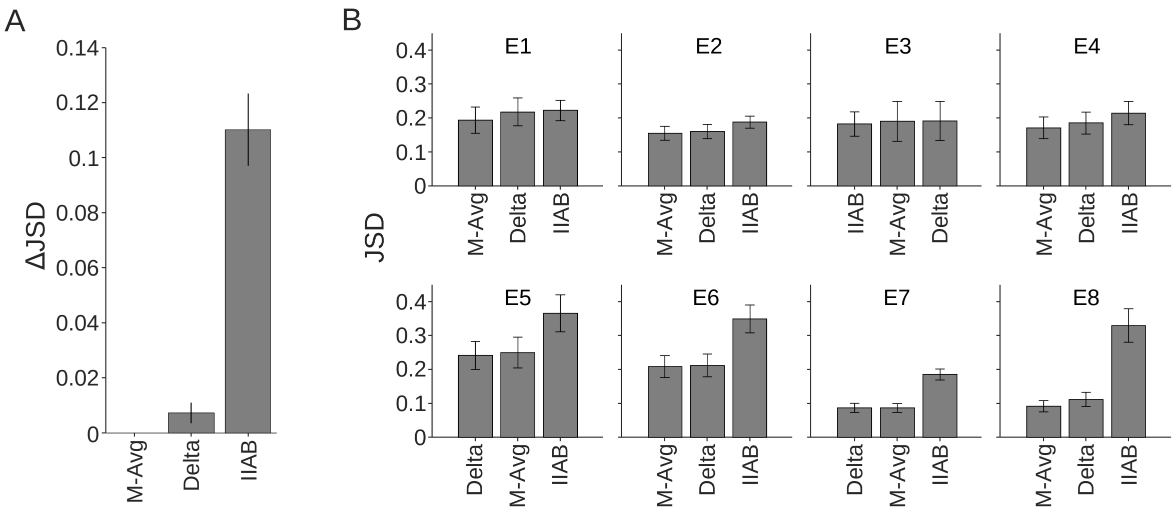

Figure 8 | Model comparison based on fits to summary statistics. (A) Jensen-Shannon divergence (JSD) between data and fit, averaged across all participants and expressed relative to the M-Avg model. A larger values indicates a worse fit. (B) JSD values averaged across participants and split by experiment.

At the level of individual experiments, the IIAB model has the worst JSD in seven of the eight cases (Figure 8B); the only exception is E3, where all models have approximately equal JSD, probably because it consists of only three participants. Overall, these results are consistent with our main analysis in the sense that the Delta and M-Avg mechanisms perform roughly equally well and better than the IIAB mechanism. However, it has to be noted that the JSD differences are very small in comparison to the AIC differences (Figure 3). This is because a summary statistic can never contain more information than the raw data from which it is derived, which follows from a theorem known as the data processing inequality (Cover & Thomas, 2005). We quantified this difference in a previous study (albeit in a different context), where we found that the summary statistics contained only 0.15% of the evidence present in the raw data (van den Berg & Ma, 2014). In light of this, we prefer to give more weight to likelihood-based comparisons than comparisons based on summary statistics.

In conclusion, even if one considers the joint distribution of step widths and step heights as the sole criterion to evaluate models on, there seems to be no ground for ruling out trial-by-trial models. If anything, the trial-by-trial models explain the data better than the hypothesis-testing model.

Slider updating consistency

The three updating mechanisms considered in this study (IIAB, Delta, M-Avg) have in common that belief updates are always consistent with the most recent observation: observing a blue increases the estimate of p_{blue} and observing a ring of the other colour decreases it. However, we find that across all 89 participants in our dataset, on average only 75.8±1.8% of the updates were consistent with the most recent observation (range: 68.9% to 80.3%). Hence, about one in every four updates was made in the direction opposite to the most recent observed outcome. Threshold variability may be one source of these inconsistencies. To see why this is the case, suppose that a participant observes three blue rings followed by a red one. If the updating threshold happened to be high in the first three trials and low in the last trial, it can happen that a slider update is made only in the fourth trial.

In agreement with our intuitions, we find that updating behaviour in the fits (to full datasets) is 100% for all M-Avg and Delta models without threshold variability. However, somewhat to our surprise, for the IIAB model we find that a small proportion of the updates (1.4±0.3% across all 89 participants) is inconsistent with the last observation. We suspect that this may have to do with the ability of the model to have "second thoughts", that is, to take back an earlier made update. In any case, models without threshold variation predict much higher updating consistency than what is observed in the data.

For models with threshold variation, we find substantially lower consistency values in the fits: 91.6±0.8% (IIAB with ε=|Δ|), 83.7±1.6% (Delta with ε=|Δ|), and 83.6±1.0% (M-Avg with ε=|Δ|). These results show that threshold variance may be one explanation for participants' updating consistency rates. However, since they are still somewhat overestimated by these models, it is likely that there are other sources too. Participants could, for example, be inferring local sequential dependencies in the data. This would lead to beliefs of the form "the next ring will surely be red since I have just drawn three blue ones" as opposed to "there is a high chance of drawing a blue ring given that I have just drawn several of them", and thus inconsistent updating.

Discussion

The most important point to take away from the modelling analyses is that – contrary to previous claims – we find no compelling evidence against trial-by-trial updating in human estimation of non-stationary probabilities. In fact, we find this class of models to be more successful at explaining behaviour than the hypothesis-testing models, with very high consistency: it holds across all eight available datasets; it holds for models with and without threshold variability; it is independent of whether model comparison is based on AIC values or on cross-validation; it is independent of whether model comparison is based on raw data or summary statistics; it is independent of whether we fit the models to full data sets or per session; and it still holds if we add a second variable threshold to the IIAB model.

It is difficult to say which of the two types of trial-by-trial models is the more successful one. When applied to data from probability estimation tasks, M-Avg models have a slight advantage over Delta models in AIC-based model comparison. However, the results are reversed in model comparison based on cross validation and in the results from the binary prediction task. Altogether, these results suggest to us that the two classes of models make very similar predictions, but that M-Avg models may be more susceptible to overfitting.

Allowing the threshold to vary is important for any model to describe the participants' behaviour well. This kind of variance could have multiple origins. For example, it could be that the neural representation of the threshold varies due to neural noise. Another possibility is that the revisions of the threshold depend on the participant's level of attention, which may fluctuate over time, especially in long experiments of the type considered here. Similarly, the threshold as such can be interpreted in several ways. Gallistel et al., (2014) assumed any threshold to be an integral part of the estimation procedure, while Khaw et al., (2017) suggest that it arises from rational adaptation to the cognitive costs of updating. Yet others may envisage it as the result of motor "laziness", which could be an equally rational outcome of a trade-off between motor cost and expected reward. All in all, the psychological interpretation of the updating threshold requires further study.

Our finding that the two-kernel delta-rule model outperformed all other models on the probability estimation task suggests that participants may have been keeping track of both slow and fast changes in the probability that they were estimating. Another possible explanation is that they were in fact behaving as described by a single-kernel model that updates its learning rate as a function of the prediction errors, as suggested by Behrens et al. (2007). Intuitively, this mechanism should be able to solve the problem which a regular trial-by-trial model will face when tracking a function with large but infrequent changes: that the estimate sometimes needs to be highly sensitive to new observations and at other times less sensitive in order to track it well. This is an interesting question for future work.

Lastly, we made an interesting observation which to the best of our knowledge has not been reported before: a rather large proportion of the slider updates was inconsistent with the most recent draw from the Bernoulli distribution. While threshold variability may be part of the explanation, we suspect that there are other sources too. Since the origin of these inconsistencies could be informative about the underlying belief updating mechanism, further investigation of this issue could lead to important improvements of the theories.

GENERAL DISCUSSION

While there is an extensive literature on human estimation of stationary probabilities (Edwards, 1961; Estes, 1976; Fiedler, 2000; Peterson & Beach, 1967), research on estimation of non-stationary probabilities has only just begun. An important observation made by the studies that have been pioneering this area is that humans tend to report their probability updates in a stepwise manner (Gallistel et al., 2014; Khaw et al., 2017; Ricci & Gallistel, 2017; Robinson, 1964). Ricci and Gallistel (2017) posited that explaining this kind of behaviour is the number one challenge for any model based on trial-by-trial updating. In this article, we took up this challenge and scrutinised the claim in two ways. First, we reported empirical data which investigated the malleability of these observed stepwise behaviours, and which expanded the empirical data base for distinguishing between the different models considerably. Second, we evaluated the different models using more rigorous likelihood-based model comparisons, applying them both to our new data and to the data sets from three previously published studies.

In the experiment, using two novel manipulations, we found evidence that particulars of the experimental design affect the discreteness in the response patterns, in turn suggesting that the stepwise behaviours need not exclusively or mainly be a signature of hypothesis testing. In particular, the finding that the extent of stepwise behaviours is strongly affected by the effort required to produce the response indicates that there are covert changes in beliefs that are not disclosed when there are asymmetric costs of maintaining vs. changing the response. The rate of stepwise behaviour was also affected by instructions about the non-stationarity of the process, indicating that there are a priori adaptations of the process that are responsive to instructions (e.g., changes in the priors across a hypothesis space or changes in the sampling window effectively used for estimation). The characteristic patterns of rare and large changes observed in the previous studies were not general, but mainly observed in one of the four experimental cells.

Furthermore, using rigorous model comparison methods, we found that not only our own data, but also all previous data sets are better accounted for by models based on trial-by-trial updating than by models based on hypothesis testing. This conclusion held across eight data sets and across a variety of different criteria for evaluating the fit of the models. However, we should immediately point out that the ambition of this article is not to proclaim the death of hypothesis testing models, but rather to suggest that the reports of the death of trial-by-trial learning models have been greatly exaggerated. Ultimately, we would expect that – as is true in most areas of cognitive science – the mind is able to draw on several different cognitive processes for learning about a property as fundamental to adaptation as probability.

More challenges

While the modelling results presented above may appear conclusive, Ricci and Gallistel (2017) raised several additional challenges for trial-by-trial models in excess of the question of how to explain stepwise updating. Here, we briefly address these. The first one is to explain that "participants perceive the changes themselves" when there are abrupt and large changes. The authors considered the possibility of a trial-by-trial model with both a slow and fast kernel, the latter of which should be able to detect abrupt changes. However, they rejected that model because they were unable to find parameter settings that produced summary statistics matching the patterns in participant data. Here, we performed a rigorous model comparison and found that the two-kernel delta-rule model actually beats all other models that we tested. Based on this finding, we believe that it would be interesting for future work to examine to what extent perceptions of abrupt changes in a two-kernel Delta-rule model coincide with those perceived by participants.

Another challenge posited by Ricci and Gallistel (2017) is to explain that participants sometimes have "second thoughts about previously perceived changes in the hidden parameter". An elegant property of the IIAB model is that the prediction of second thoughts is integral to its updating mechanism. However, we believe that it would be wrong to reject trial-by-trial model based on the fact that they need additional assumptions to account for second thoughts, because they might very well be governed by a separate process. A circumstance (in this case a button) which explicitly invites people to re-evaluate their previous beliefs might induce them to do so, but that is not to say that such behaviour must be integral to the iterated online estimation which the present paradigm investigates.

A final challenge posited by Ricci and Gallistel (2017) is to explain that participants are able to extract abstract information about the function that guides the true value of the probability that they are tracking. In line with their findings, we observed in the post-experiment questionnaires that many participants produced something that resembled a sinusoidal function when asked to draw the function they believed they had been tracking. An appealing feature of the IIAB is that the higher-order structure of the generative function may be derived from its record of change points. However, the same is true for the M-Avg models, which keep a history of previous outcomes. As was the case with the issue of second thoughts, we argue that inference of the underlying function may be governed by a mechanism that is separate from the updating mechanism. We agree with Ricci and Gallistel (2017) that such a mechanism should rely on some sort of sequence memory, but that does not imply that the updating must too. To shed more light on this, more data are required about the relation between sequences of observed outcomes and the kind of abstract structures that participants infer from these sequences.

Heterogeneity in updating strategies

Our model comparison results were unambiguous when considered at the group level: the M-Avg mechanism accounted best for the data, followed by first the Delta mechanism and then the IIAB mechanism (Figures 3 and 4). However, at the level of individual participants, we observed substantial heterogeneity in the results (Figure 3A). There are two possible explanations for this. First, there may be true heterogeneity in the underlying cognition, in which case it would be misleading to consider only group-level results. Second, the heterogeneity could be an artefact caused by limitations of the analysis, such as the finite size of the dataset, the use of a custom likelihood function, and the lack of guarantee that the optimisation algorithm always converged to the maximum of this function. Indeed, the model recovery analysis (Appendix C) showed some misclassifications even when the true model was in the set of fitted models, although never between updating mechanisms. We can, at present, neither rule out nor confirm that different individuals used different updating strategies.

Limitations

A first limitation of the present study is that we did not test hybrid models. Since the main goal was to scrutinise previous conclusions drawn about the viability of trial-by-trial models, we considered the testing of hybrid models outside the scope of the present work. However, since hypothesis-testing and trail-by-trial updating are not necessarily mutually exclusive, the most promising models might be ones that combine the two processes.

We also mentioned above that there remain unexplained differences between the observed consistency rates and those predicted by the models. Intuitively, one possible cause is that participants infer sequential dependencies within random processes (Ayton & Fischer, 2004). A participant who is under the impression that, say, three blues in a row indicate that the next ring is most likely going to be red should update inconsistently after observing that sequence. This has not been addressed in our experimentation or modelling, but experimental data exists from a paradigm similar to our own. Toda (1958) rigged the Bernoulli sequence in his probability estimation task in such a way that there were sequential patterns in the outcomes, allowing him to study if these were inferred through observing the participants' subjective probabilities. He inferred from the data that participants estimate probabilities in a way that is approximately the Bayesian solution of a higher order Markov process – a non-trivial trial-by-trial model. We are, however, reluctant to accept this conclusion. The problem is that the probability estimates in Toda's task were derived indirectly from decisions in an ultimatum bargaining game and thus likely to have been affected by first-mover advantage and people's fairness concerns (Güth, 1995; Güth & Van Damme, 1998; Slembeck, 1999; Thaler & Camerer, 1995). This may have biased his estimates. Future studies could adapt the present task with Toda's (1958) rigged sequences to see if this increases the inconsistency rates beyond those in a non-rigged control condition.

Another limitation is that we performed model comparisons based on a custom likelihood function, because the proper likelihood function was intractable. Even though model recovery analysis confirmed that the chosen function allowed for reliable model comparison, better choices might have been possible and could have led to more conclusive results in terms of distinguishing the four threshold mechanisms in the second model factor. We constructed the custom likelihood function mainly based on "educated guesses" of what aspects are important to consider. An alternative and probably better way would have been to *derive* a likelihood function by starting with the proper one and then make simplifications until it becomes tractable.

Lastly, during our debriefings, some participants reported that they counted or chunked the observations. This could possibly imply a trivial dual-strategy hypothesis: some people attempt to solve the task by counting, a strategy which is highly inefficient in the chaotic world outside of the laboratory. When they update intuitively, they use a different system which does not require working memory retention of observations. Manipulating working memory capacity may confirm or reject this hypothesis and inform future studies which want to use similar tasks – since most scientists presumably will be more interested in the second, intuitive system we must know if we need to control for counting.

Relation to behavioural economics

In their seminal work "Theory of Games and Economic Behavior", originally published in 1944, von Neumann and Morgenstern (2007) begin by recognising the fact that a "universal system" of economic theory is not achievable in the foreseeable future, largely due to the lack of a sufficient body of empirical observations. In anticipation of that, they make-do with "some commonplace experience of human behavior" to demonstrate the mathematical framework we today recognise as game theory. These behavioural assumptions have been criticised by behavioural economists and cognitive psychologists (e.g. Mullainathan & Thaler, 2015; Schoemaker, 1982; Tversky, 1975). Some studies have introduced modifications (e.g. Caplin

& Leahy, 2001; O'Donoghue & Rabin, 1999), but there have been few comprehensive replacements. A well-validated, robust theory of probability perception would be an important step towards such an end. We believe that the present work is a contribution to the construction of such a theory.

Concluding remarks

To the best of our knowledge, the first study that investigated human estimation of non-stationary probabilities directly was performed in 1964 (Robinson, 1964). After that, it took another 50 years before a serious modelling attempt was initiated to obtain an understanding of the mechanism behind this important cognitive function (Gallistel et al., 2014). That attempt culminated in a rejection of the entire class of trial-by-trial models and the proposal that humans instead use hypothesis testing to track non-stationary probabilities. Here, we scrutinised that proposal and found that there is actually much stronger evidence for trial-by-trial updating than for hypothesis testing. Hence, the rejection of trial-by-trial models seems to have been premature. However, considering the juvenility of this field of research, we believe that it would be equally wrong to use these results to rule out hypothesis-testing models. In the end, it may turn out that humans use a mix of strategies. Therefore, future studies might benefit from starting to look into hybrid models instead of continuing to restrict themselves to one particular class. In doing so, they should strive to bring all the findings – from function learning through binary choice to probability inference – under one umbrella. That way, applied researchers such as economists may find important uses for the work.

ACKNOWLEDGEMENTS

We thank Randy Gallistel and Matthew Ricci for sharing their data with us and for several helpful discussions about their experiments. The research was funded by the Swedish Research Council (Grant 2018-01947) and the Marcus and Amalia Wallenberg Foundation (MAW 2016.0132).

CREDIT AUTHOR STATEMENT

Mattias Forsgren: Conceptualization, Methodology, Validation, Formal Analysis, Investigation, Data Curation, Writing – Original Draft, Writing – Reviewing & Editing. **Peter Juslin:** Conceptualization, Methodology, Formal Analysis, Writing – Original Draft, Writing – Reviewing & Editing, Supervision, Project Administration, Funding Acquisition. **Ronald van den Berg:** Conceptualization, Methodology, Software, Validation, Formal Analysis, Data

Curation, Writing – Original Draft, Writing – Reviewing & Editing, Visualization, Supervision, Project Administration, Funding Acquisition.

REFERENCES

Acerbi, L., & Ma, W. J. (2017). Practical Bayesian Optimization for Model Fitting with Bayesian Adaptive Direct Search. *Advances in Neural Information Processing Systems 30*, 1836–1846. https://doi.org/https://doi.org/10.1101/150052

Akaike, H. (1974). A new look at the statistical model identification. *IEEE Transactions on Automatic Control*, *19*(6), 716–723. https://doi.org/10.1109/TAC.1974.1100705

Ashby, F. G., & Valentin, V. V. (2017). Multiple systems of perceptual category learning: Theory and cognitive tests. In *Handbook of categorization in cognitive science, 2nd ed.* (pp. 157–188). https://doi.org/10.1016/B978-0-08-101107-2.00007-5

Ayton, P., & Fischer, I. (2004). The hot hand fallacy and the gambler's fallacy: Two faces of subjective randomness? *Memory and Cognition*. https://doi.org/10.3758/BF03206327

Barbey, A. K., & Sloman, S. A. (2007). Base-rate respect: From ecological rationality to dual processes. *Behavioral and Brain Sciences*. https://doi.org/10.1017/S0140525X07001653

Behrens, T. E. J., Woolrich, M. W., Walton, M. E., & Rushworth, M. F. S. (2007). Learning the value of information in an uncertain world. *Nature Neuroscience*. https://doi.org/10.1038/nn1954

Brehmer, B. (1974). Hypotheses about relations between scaled variables in the learning of probabilistic inference tasks. *Organizational Behavior and Human Performance*. https://doi.org/10.1016/0030-5073(74)90002-6

Brehmer, B. (1980). In one word: Not from experience. *Acta Psychologica*. https://doi.org/10.1016/0001-6918(80)90034-7

Bruner, J. S., Goodnow, J. J., & Austin, G. A. (1956). A study of thinking. In *A study of thinking.* Oxford, England: John Wiley and Sons.

Busemeyer, J. R., & Myung, I. J. (1988). A New Method for Investigating Prototype Learning. *Journal of Experimental Psychology: Learning, Memory, and Cognition*. https://doi.org/10.1037/0278-7393.14.1.3

Caplin, A., & Leahy, J. (2001). Psychological expected utility theory and anticipatory feelings. *Quarterly Journal of Economics*. https://doi.org/10.1162/003355301556347

Costello, F., & Watts, P. (2014). Surprisingly rational: Probability theory plus noise explains biases in judgment. *Psychological Review*. https://doi.org/10.1037/a0037010

Cover, T. M., & Thomas, J. A. (2005). Elements of Information Theory. In *Elements of*

Information Theory. https://doi.org/10.1002/047174882X

Edwards, W. (1961). Probability learning in 1000 trials. *Journal of Experimental Psychology*. https://doi.org/10.1037/h0041970

Estes, W. K. (1976). The cognitive side of probability learning. *Psychological Review*. https://doi.org/10.1037/0033-295X.83.1.37

Fiedler, K. (2000). Beware of samples! A cognitive-ecological sampling approach to judgment biases. *Psychological Review*. https://doi.org/10.1037/0033-295X.107.4.659

Gallistel, C. R., Krishan, M., Liu, Y., Miller, R., & Latham, P. E. (2014). The perception of probability. *Psychological Review*. https://doi.org/10.1037/a0035232

Gemmeke, J. F., Virtanen, T., & Hurmalainen, A. (2011). Exemplar-based sparse representations for noise robust automatic speech recognition. *IEEE Transactions on Audio, Speech and Language Processing*. https://doi.org/10.1109/TASL.2011.2112350

Gigerenzer, G., & Hoffrage, U. (1995). How to improve Bayesian reasoning without instruction: Frequency formats. *Psychological Review*. https://doi.org/10.1037/0033-295X.102.4.684

Grosse, I., Bernaola-Galván, P., Carpena, P., Román-Roldán, R., Oliver, J., & Stanley, H. E. (2002). Analysis of symbolic sequences using the Jensen-Shannon divergence. *Physical Review E - Statistical Physics, Plasmas, Fluids, and Related Interdisciplinary Topics*. https://doi.org/10.1103/PhysRevE.65.041905

Güth, W. (1995). On ultimatum bargaining experiments - A personal review. *Journal of Economic Behavior and Organization*. https://doi.org/10.1016/0167-2681(94)00071-L

Güth, W., & Van Damme, E. (1998). Information, Strategic Behavior, and Fairness in Ultimatum Bargaining: An Experimental Study. *Journal of Mathematical Psychology*. https://doi.org/10.1006/jmps.1998.1212

Hertwig, R., & Erev, I. (2009). The description-experience gap in risky choice. *Trends in Cognitive Sciences*. https://doi.org/10.1016/j.tics.2009.09.004

JASP Team. (2019). JASP (Version 0.10.2). *[Computer Software]*.

Juslin, P., & Persson, M. (2002). PROBabilities from EXemplars (PROBEX): A “lazy” algorithm for probabilistic inference from generic knowledge. *Cognitive Science*. https://doi.org/10.1016/S0364-0213(02)00083-6

Juslin, P., Winman, A., & Hansson, P. (2007). The Naïve Intuitive Statistician: A Naïve Sampling Model of Intuitive Confidence Intervals. *Psychological Review*. https://doi.org/10.1037/0033-295X.114.3.678

Kahneman, D., & Frederick, S. (2005). A model of heuristic judgment. *The Cambridge*

Handbook of Thinking and Reasoning.

Khaw, M. W., Stevens, L., & Woodford, M. (2017). Discrete adjustment to a changing environment: Experimental evidence. *Journal of Monetary Economics*. https://doi.org/10.1016/j.jmoneco.2017.09.001

Klayman, J., & Ha, Y. W. (1987). Confirmation, Disconfirmation, and Information in Hypothesis Testing. *Psychological Review*. https://doi.org/10.1037/0033-295X.94.2.211

Lebiere, C., Stewart, T., & West, R. (2009). Applying cognitive architectures to decision-making: How cognitive theory and the equivalence measure triumphed in the Technion Prediction Tournament. *Proceedings of the Annual Meeting of the Cognitive Science Society*, 31(31).

Mullainathan, S. (2002). A memory-based model of bounded rationality. *Quarterly Journal of Economics*. https://doi.org/10.1162/003355302760193887

Mullainathan, S., & Thaler, R. H. (2015). Behavioral Economics. In *International Encyclopedia of the Social & Behavioral Sciences: Second Edition*. https://doi.org/10.1016/B978-0-08-097086-8.71007-5

Neal, R. M., & Dayan, P. (1997). Factor Analysis Using Delta-Rule Wake-Sleep Learning. *Neural Computation*. https://doi.org/10.1162/neco.1997.9.8.1781

Nosofsky, R. M., & Palmeri, T. J. (1997). An Exemplar-Based Random Walk Model of Speeded Classification. *Psychological Review*. https://doi.org/10.1037/0033-295X.104.2.266

O'Donoghue, T., & Rabin, M. (1999). Doing it now or later. *American Economic Review*. https://doi.org/10.1257/aer.89.1.103

Oaksford, M., & Chater, N. (1994). A Rational Analysis of the Selection Task as Optimal Data Selection. *Psychological Review*. https://doi.org/10.1037/0033-295X.101.4.608

Peterson, C. R., & Beach, L. R. (1967). Man as an intuitive statistician. *Psychological Bulletin*. https://doi.org/10.1037/h0024722

R Core Team. (2014). R Core Team (2014). R: A language and environment for statistical computing. *R Foundation for Statistical Computing, Vienna, Austria. URL Http://Www.R-Project.Org/*.

Rescorla, R. A., & Wagner, A. R. (1972). A Theory of Pavlovian Conditioning: Variations in the Effectiveness of Reinforcement and Nonreinforcement BT - Clasical conditioning II: current research and theory. In *Clasical conditioning II: current research and theory*.

Ricci, M., & Gallistel, R. (2017). Accurate step-hold tracking of smoothly varying periodic and aperiodic probability. *Attention, Perception, and Psychophysics*.

https://doi.org/10.3758/s13414-017-1310-0

Robinson, G. H. (1964). Continuous Estimation Of A Time-Varying Probability. *Ergonomics, 7*(1), 7–21. https://doi.org/10.1080/00140136408930721

Schoemaker, P. J. H. (1982). The Expected Utility Model: Its Variants, Purposes, Evidence and Limitations. *Journal of Economic Literature*.

Schwarz, G. (1978). Estimating the Dimension of a Model. *The Annals of Statistics*. https://doi.org/10.1214/aos/1176344136

Shen, S., & Ma, W. J. (2019). Variable precision in visual perception. *Psychological Review*. https://doi.org/10.1037/rev0000128

Slembeck, T. (1999). Reputations and Fairness in Bargaining - Experimental Evidence from a Repeated Ultimatum Game With Fixed Opponents. *Universität St. Gallen Discussion Paper*.

Thaler, R. H., & Camerer, C. F. (1995). Ultimatums, Dictators and Manners. *Journal of Economic Perspectives 9(2):209-19*.

Toda, M. (1958). Subjective inference vs. objective inference of sequential dependencies. *Japanese Psychological Research*. https://doi.org/10.4992/psycholres1954.1958.1

Tversky, A. (1975). A critique of expected utility theory: Descriptive and normative considerations. *Erkenntnis*. https://doi.org/10.1007/BF00226380

Tversky, A., & Kahneman, D. (1983). Extensional versus intuitive reasoning: The conjunction fallacy in probability judgment. *Psychological Review*. https://doi.org/10.1037/0033-295X.90.4.293

Tversky, A., & Kahneman, D. (1992). Advances in prospect theory: Cumulative representation of uncertainty. *Journal of Risk and Uncertainty*. https://doi.org/10.1007/BF00122574

van den Berg, R., Awh, E., & Ma, W. J. (2014). Factorial comparison of working memory models. *Psychological Review*, *121*(1), 124–149. https://doi.org/10.1037/a0035234

van den Berg, R., & Ma, W. J. (2014). “Plateau”-related summary statistics are uninformative for comparing working memory models. *Attention, Perception & Psychophysics*. https://doi.org/10.3758/s13414-013-0618-7

Verguts, T., & Van Opstal, F. (2014). A delta-rule model of numerical and non-numerical order processing. *Journal of Experimental Psychology: Human Perception and Performance*. https://doi.org/10.1037/a0035114

von Neumann, J., & Morgenstern, O. (2007). Theory of games and economic behavior. In *Theory of Games and Economic Behavior*. https://doi.org/10.2307/3610940

Wason, P. C., & Johnson-Laird, P. N. (1970). A conflict between selecting and evaluating information in an inferential task. *British Journal of Psychology*. https://doi.org/10.1111/j.2044-8295.1970.tb01270.x

Widrow, B., & Hoff, M. E. (1960). Adaptive Switching Circuits. In *Technical report no. 1553-1*.

Widrow, B., & Lehr, M. A. (1993). ARTIFICIAL NEURAL NETWORKS OF THE PERCEPTRON, MADALINE, AND BACKPROPAGATION FAMILY. In *Neurobionics*. https://doi.org/10.1016/b978-0-444-89958-3.50013-9

Zacks, R. T., & Hasher, L. (2002). Frequency processing: a twenty-five year perspective. In *Etc. Frequency Processing and Cognition*. https://doi.org/10.1093/acprof:oso/9780198508632.003.0002

APPENDIX A – Dunn's post hoc comparisons

Table A1. *Dunn's Post Hoc Comparisons of RMSE Between Conditions.*

Condition		z-score	W_{left}	W_{right}	p	$p_{bonferroni}$	p_{holm}
HE-UI	HE-IN	4.297	42.333	14.933	< .001	< .001	< .001
	LE-UI	1.599	42.333	32.133	0.055	0.329	0.165
	LE-IN	1.526	42.333	32.600	0.063	0.381	0.165
HE-IN	LE-UI	-2.697	14.933	32.133	0.003	0.021	0.014
	LE-IN	-2.770	14.933	32.600	0.003	0.017	0.014
LE-UI	LE-IN	-0.073	32.133	32.600	0.471	1.000	0.471

Table A2. *Dunn's Post Hoc Comparisons of Kullback-Leibler Divergence Between Conditions.*

Condition		z-score	W_{left}	W_{right}	p	$p_{bonferroni}$	p_{holm}
HE-UI	HE-IN	4.098	42.200	16.067	< .001	< .001	< .001
	LE-UI	1.589	42.200	32.067	0.056	0.336	0.148
	LE-IN	1.652	42.200	31.667	0.049	0.296	0.148
HE-IN	LE-UI	-2.509	16.067	32.067	0.006	0.036	0.030
	LE-IN	-2.446	16.067	31.667	0.007	0.043	0.030
LE-UI	LE-IN	0.063	32.067	31.667	0.475	1.000	0.475

Table A3. *Dunn's Post Hoc Comparisons of Step Width Between Conditions.*

Condition		z-score	W_{left}	W_{right}	p	$p_{bonferroni}$	p_{holm}
HE-UI	HE-IN	2.718	47.933	30.600	0.003	0.020	0.013
	LE-UI	3.293	47.933	26.933	< .001	0.003	0.002
	LE-IN	4.924	47.933	16.533	< .001	< .001	< .001
HE-IN	LE-UI	0.575	30.600	26.933	0.283	1.000	0.283
	LE-IN	2.206	30.600	16.533	0.014	0.082	0.041
LE-UI	LE-IN	1.631	26.933	16.533	0.051	0.309	0.103

Table A4. *Dunn's Post Hoc Comparisons of Step Height Between Conditions.*

Condition		z-score	W_{left}	W_{right}	p	$p_{bonferroni}$	p_{holm}
HE-UI	HE-IN	-3.230	27.800	48.400	< .001	0.004	0.003

Table A4. *Dunn's Post Hoc Comparisons of Step Height Between Conditions.*

Condition		z-score	W_{left}	W_{right}	p	$p_{bonferroni}$	p_{holm}
	LE-UI	1.861	27.800	15.933	0.031	0.188	0.063
	LE-IN	-0.324	27.800	29.867	0.373	1.000	0.373
HE-IN	LE-UI	5.091	48.400	15.933	< .001	< .001	< .001
	LE-IN	2.906	48.400	29.867	0.002	0.011	0.007
LE-UI	LE-IN	-2.185	15.933	29.867	0.014	0.087	0.043

Legend: HE is High Effort, LE is Low Effort, UN is Uninformed, and IN is Informed. W_{left} and W_{right} are the summed ranks of the condition in the leftmost and second to leftmost column, respectively. Non-integer values are due to rank ties.

APPENDIX B – Custom likelihood function

In its most general form, the log likelihood function for the models considered in this study takes the form

$$\log p\left(\mathbf{R} \mid \boldsymbol{\theta}, \boldsymbol{\psi}, \mathbf{O}\right) = \sum_{t=1}^{n} \log p\left(R_t \mid \boldsymbol{\theta}, \boldsymbol{\psi}_{1,\ldots,t-1}, R_{1,\ldots,t-1}, O_{1,\ldots,t-1}\right), \tag{4}$$

where $\mathbf{R}=\{R_1, R_2, \ldots, R_n\}$ is a vector with subject responses for all n trials, $\boldsymbol{\theta}$ is a vector with parameter values, $\boldsymbol{\psi}$ is a matrix with latent variables, and $\mathbf{O}=\{O_1, O_2, \ldots, O_m\}$ is a vector with all Bernoulli outcomes observed by the subject. The IIAB model has multiple time-varying latent variables, including a list of change points and parameters of a beta distribution representing the observer's prior belief that any given trial is a change point (see Table 1 in Gallistel et al., 2014). The existence of these latent variables in combination with the fact that the model predictions are not independent across trials makes evaluation of the likelihood function computationally prohibitive.

To circumvent this problem, we construct a "custom" likelihood function that captures the main aspects of the likelihood function proper in a computationally tractable way, yet still allows for reliable model comparison, which will be verified by a model recovery analysis (Appendix C).

We believe that there are two important aspects that the likelihood function should cover in order to allow it for reliable model fitting and comparison. First, obviously, it should punish models for discrepancies between the predicted slider value and the slider value chosen by the subject. Second, since one of the main differences between the models is when they predict slider updates, it is probably also important that the likelihood function punishes models that predict slider updates on trials where the subject made no update and vice versa. With this in mind, we choose to compute the likelihood of parameters $\boldsymbol{\theta}$ for model M as follows. Let $\mathbf{R}_{\text{subject}}$ denote the vector with subject responses and $\mathbf{O}$ the vector with observed Bernoulli outcomes. First, we compute the model's predicted response vector $\mathbf{R}_{\text{M}}$. Assuming for the moment that there is no threshold noise, $\mathbf{R}_{\text{M}}$ is a deterministic function of $\boldsymbol{\theta}$ and $\mathbf{O}$ for all models that we consider here. We can obtain $\mathbf{R}_{\text{M}}$ efficiently using a forward simulation of the model, feeding it with $\mathbf{O}$ while fixing the parameters to $\boldsymbol{\theta}$. After obtaining $\mathbf{R}_{\text{M}}$, we compute the probability of the subject response on each trial t as follows,

$$p\left(R_{\text{subject},t} \mid R_{\text{M},t}\right) \equiv \begin{cases} 0 & R_{\text{subject},t} - R_{\text{subject},t-1} = 0,\ R_{\text{M},t} - R_{\text{M},t-1} \neq 0 \\ 0 & R_{\text{subject},t} - R_{\text{subject},t-1} \neq 0,\ R_{\text{M},t} - R_{\text{M},t-1} = 0 \\ N\left(R_{\text{subject},t}; R_{\text{M},t}, \sigma_{\text{unexplained}}\right) & \text{otherwise,} \end{cases} \quad (5)$$

where $N(x; \mu, \sigma)$ is a normal distribution with mean μ and standard deviation σ, evaluated at point x. This function strongly punishes models that predict an update when the subject did not make an update (first line of last expression in Eq. (5)) or vice versa (second line). If, on the other hand, the updating behaviour is consistent between model and subject (third line), the probability of the subject response is measured as a draw from a normal distribution centred on the response predicted by the model. This normal distribution can be thought of as a way to capture variance in the data that is left unexplained by the model: the better the model, the smaller the estimate of $\sigma_{\text{unexplained}}$. Part of this variance could be due to variability in motor responses, but there may be other sources too. To avoid log likelihoods equal to negative infinity, we assume in each model that the observer sometimes produces a random response drawn from a uniform distribution on [0,1]. We fix the rate of such random responses to 1 in 1,000 trials.

So far, we have assumed fixed thresholds in our construction of the likelihood function. However, all models that we consider here have a variable threshold, which makes the predictions non-deterministic: for a fixed set of parameters $\boldsymbol{\theta}$ and input vector $\mathbf{O}$, prediction $\mathbf{R}_{\text{M}}$ varies from run to run. To approximate the probability of the subject's response under a variable response threshold, we average the model prediction over 100 runs. We thus obtain the following custom log likelihood function:

$$L(\boldsymbol{\theta}) = \sum_{t=1}^{n} \log\left(\frac{1}{100}\sum_{i=1}^{100} p\left(R_{\text{subject},t} \mid R_{\text{M},t}\right)\right), \quad (6)$$

where $p(R_{\text{subject},t} \mid \mathrm{R}_{M,t})$ is as specified in Eq. (5).

APPENDIX C – Model recovery

We created a group of five synthetic data sets from each of the twelve models with threshold noise, giving a total of sixty synthetic datasets. Next, we used maximum-likelihood estimation to fit the twelve main models twenty times to all datasets. For each fit, we computed the Akaike Information Criterion (AIC; Akaike, 1974). At the level of individual data sets, AIC-based model comparison picks out the correct model in forty-six of the sixty cases (Figure C, Panel A). In the remaining fourteen cases, a mistake was made with respect to the second modelling factor, that is, the threshold mechanism. This indicates that at the individual level, our methods are adequate for selecting the right updating mechanism (IIAB, Delta or M-Avg), but it has some difficulties in selecting the right threshold mechanism. At the group level, on the other hand, the correct model was selected in all cases (Figure C, panel B). These results also indicated that the quality of fit improved very little after about ten runs of the optimizer (Figure C, panel C).

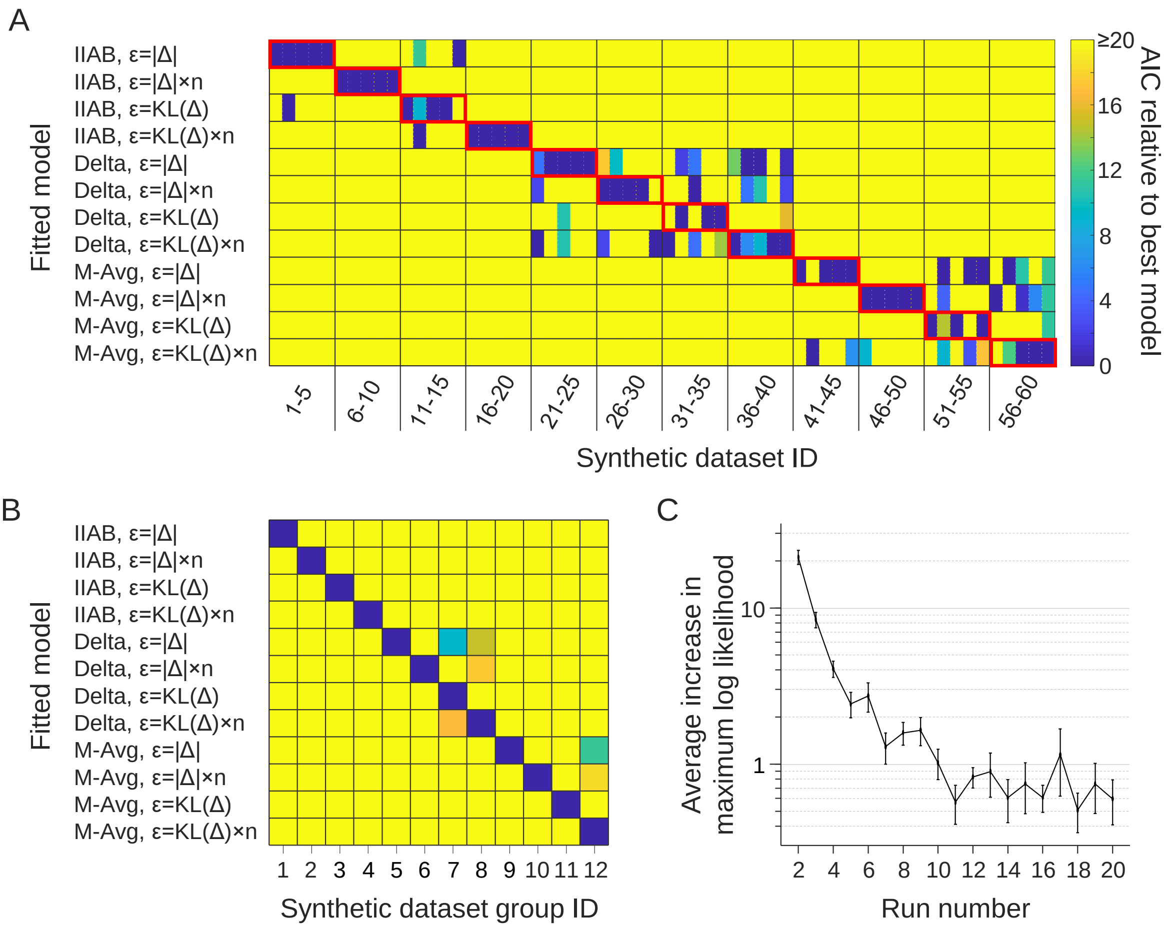

Figure S1 | Model recovery results. (A) AIC-based model comparison at the level of individual datasets. The colours indicate the AIC value of each individual fit relative to the best-fitting model in the respective dataset. Each column has a single best-fitting model, which by definition has a relative AIC value equal to 0. The red boxes indicate for each group of datasets which model generated them. In 46 of the 60 synthetic datasets, the correct model was selected (dark blue cells in the red boxes). In the remaining 14 datasets, an error was made in the inference of the mechanism behind the computation of *E*. No errors were made in the inference of the updating core mechanism (IIAB, Delta, M-Avg), meaning that these mechanism are highly identifiable, even at the level of individual subjects. (B) Relative AIC values averaged within each group of synthetic datasets that share the same generative model. In all 12 groups, the generative model was correctly selected as the model with lowest average AIC. Hence, all 12 models are identifiable at the group level, even when the group contains as few as 5 subjects. (C) The results in panels A and B were obtained by fitting each model 20 times with different initial parameter estimates. To assess how many runs are required for stable model comparison performance, this panel shows the average increase in maximum log likelihood as a function of the number of times each model was fitted. After approximately 10 runs, the average increase in maximum log likelihood rarely exceeds 1. In our analysis of human data, we fit each model 30 times.

1210

1211

1212

APPENDIX D – Cross validation results

In our main analysis, we fitted the models to only the first 750 trials in each dataset. Model comparison based on the log likelihood of the remaining trials (Figure D) are largely consistent with the AIC-based results (Figure 3).

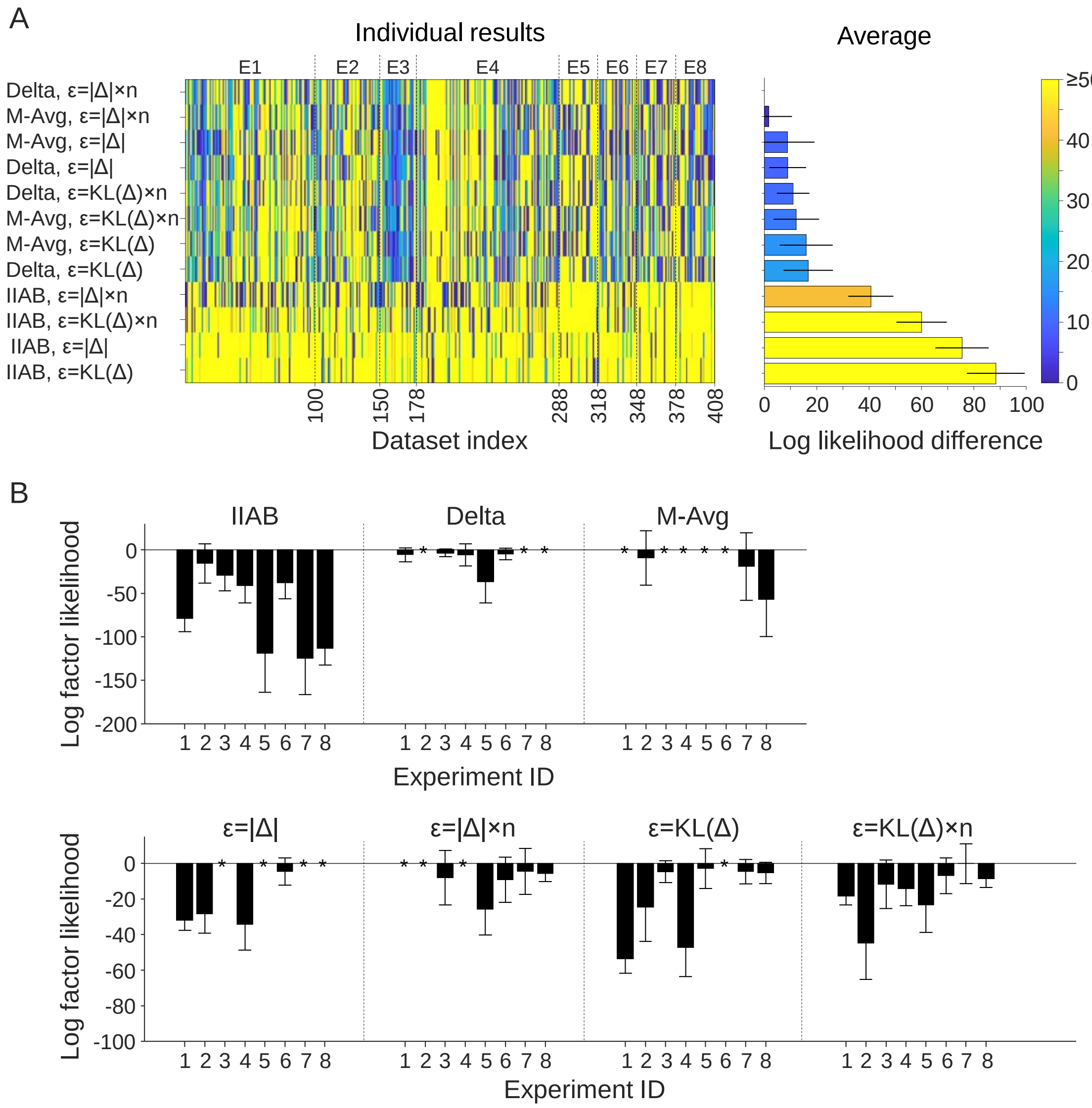

Figure S2 | Model comparison based on cross-validated log likelihoods. (A) Left: Log likelihood values relative to the best-fitting model for individual datasets. Right: Relative log likelihood values averaged across datasets. One may notice that the cross-validated log likelihood differences are smaller than the AIC differences presented in Figure 4. There are two reasons for this. First, AIC is defined as (roughly) twice the log likelihood and, second, the AIC values were based on three times the number of trials (750 vs 250). Hence, to make the cross-validated log likelihoods comparable to the AIC-based results, one should multiply them by a factor of 6. (B) Factor level comparison based on cross-validated log likelihoods. Top: Evidence for each level in the first factor, combined across all models. Bottom: Evidence for each level in the second factor, combined across all models. The most successful levels in each experiment are indicated by asterisks.

APPENDIX E – Maximum-likelihood parameter estimates

Table E1. *Maximum-likelihood Estimates of the Parameters of the IIAB Models.*

Model	Parameter	25% Quartile	Median	75% Quartile
IIAB, ε=|Δ|	μ_{T1}	8.79×10^{-5}	9.15×10^{-3}	4.05×10^{-2}
	σ_{T1}	1.23×10^{-2}	2.08×10^{-2}	3.63×10^{-2}
	T_2	1.17	1.60	7.62
	$\sigma_{unexplained}$	6.04×10^{-2}	8.50×10^{-2}	0.120
IIAB, ε=|Δ|×n	μ_{T1}	2.37×10^{-3}	0.693	1.52
	σ_{T1}	1.01	1.68	3.53
	T_2	0.573	0.927	4.23
	$\sigma_{unexplained}$	4.31×10^{-2}	6.33×10^{-2}	9.80×10^{-2}
IIAB, ε=KL|Δ|	μ_{T1}	2.04×10^{-4}	1.31×10^{-3}	1.12×10^{-2}
	σ_{T1}	2.23×10^{-3}	8.97×10^{-3}	2.29×10^{-2}
	T_2	1.04	1.53	5.66
	$\sigma_{unexplained}$	5.54×10^{-2}	8.15×10^{-2}	0.117
IIAB, ε=KL|Δ|×n	μ_{T1}	2.83×10^{-4}	0.199	0.904
	σ_{T1}	0.274	0.637	1.23
	T_2	0.736	0.984	1.79
	$\sigma_{unexplained}$	4.33×10^{-2}	6.69×10^{-2}	9.70×10^{-2}

Table E2. *Maximum-likelihood Estimates of the Parameter Values of the Delta-rule Models.*

Model	Parameter	25% Quartile	Median	75% Quartile
Delta, ε=|Δ|	μ_{T1}	0.142	0.373	0.807
	σ_{T1}	5.59×10^{-2}	0.162	0.374
	λ	2.94×10^{-2}	9.27×10^{-2}	0.150
	$\sigma_{unexplained}$	2.91×10^{-2}	6.03×10^{-2}	9.66×10^{-2}
Delta, ε=|Δ|×n	μ_{T1}	0.918	5.74	27.2
	σ_{T1}	0.465	3.02	13.0
	λ	1.95×10^{-2}	8.81×10^{-2}	0.147
	$\sigma_{unexplained}$	3.87×10^{-2}	6.61×10^{-2}	0.101
Delta, ε=KL|Δ|	μ_{T1}	0.152	0.689	2.41
	σ_{T1}	7.90×10^{-2}	0.456	2.22
	λ	3.40×10^{-2}	9.67×10^{-2}	0.155
	$\sigma_{unexplained}$	2.92×10^{-2}	6.02×10^{-2}	9.04×10^{-2}
Delta, ε=KL|Δ|×n	μ_{T1}	0.848	5.81	46.5

	σ_{T1}	0.638	3.73	26.1
	λ	2.39×10^{-2}	8.75×10^{-2}	0.140
	$\sigma_{unexplained}$	3.83×10^{-2}	6.44×10^{-2}	0.101

Table E3. *Maximum-likelihood Estimates of the Parameter Values of the Memory-averaging Models.*

Model	Parameter	25% Quartile	Median	75% Quartile
M-Avg, ε=|Δ|	μ_{T1}	0.253	0.470	0.854
	σ_{T1}	5.75×10^{-2}	0.207	0.402
	α	0.854	0.911	0.969
	$\sigma_{unexplained}$	2.98×10^{-2}	5.64×10^{-2}	8.22×10^{-2}
M-Avg, ε=|Δ|×n	μ_{T1}	1.52	6.36	33.0
	σ_{T1}	0.678	3.18	13.4
	α	0.866	0.919	0.982
	$\sigma_{unexplained}$	3.83×10^{-2}	6.56×10^{-2}	9.81×10^{-2}
M-Avg, ε=KL|Δ|	μ_{T1}	0.223	0.761	2.75
	σ_{T1}	7.07×10^{-2}	0.494	2.63
	α	0.843	0.908	0.962
	$\sigma_{unexplained}$	3.24×10^{-2}	5.70×10^{-2}	8.43×10^{-2}
M-Avg, ε=KL|Δ|×n	μ_{T1}	1.42	7.34	41.0
	σ_{T1}	0.848	4.36	21.5
	α	0.858	0.913	0.971
	$\sigma_{unexplained}$	3.81×10^{-2}	6.35×10^{-2}	9.21×10^{-2}