Exploring the Link between Perception and Expectations in Associative Learning: A Bayesian Analysis of Hierarchical Models

Jesper Fischer Ehmsen (201910213) (JFE)

Jiaqi Zhao (202202798) (JZ)

School of Communication and Culture, Aarhus University, Data Science exam, 27/05/2023

Abstract (JZ, JFE)

Studies to date have investigated the relationship between expectation and perception of humans through perceptual associative learning tasks. It has shown that changes in expectations relate to changes in precepts, but few studies have investigated the link between how percepts might also influence expectations in a bidirectional manner evolving in time. The current study we apply four hierarchical computational models to explore the relationship between perception and expectation. Three of the four models implemented assume a bidirectional link between perception and expectation whereas the last served as control, assuming a unidirectional link. The models were run through a thorough Bayesian workflow and then fitted to experimental data from an associative thermal learning task that was repeated twice for test-retest reliability check. We show that the unidirectional model was the model best describing the data using Pareto smoothed importance sampling leave one out cross validation, outperforming the other models by many standard errors. Lastly, we show that only some of the subject specific parameter estimates were stable across time, most notably the learning rate for the unidirectional model showed a strong correlation between sessions with a highest posterior density interval of r = [0.68; 0.79].

Jesper Fischer Ehmsen (201910213)
Jiaqi Zhao (202202798)

Introduction (JFE)

Understanding how and why humans make decisions has been investigated for centuries and has been one of the cornerstones in psychology and cognitive science research. Many concepts and phenomena studied in psychology and cognitive science are heavily involved with human decision making, including cognitive biases, perception, persuasion, belief updating and perhaps mental disorders. One of the common ways of studying decision making dates back to Ivan Pavlov, who studied, what has now been termed classical condition, in dogs (Rehman et al., 2023). Pavlov observed that when his dogs were presented with food they would begin to salivate and discovered that when the food was paired with another stimulus (i.e., a cue) then he could make the dogs salivate, after several repetitions, from just presenting the cue. His discovery had a considerate impact on cognitive science, and a huge impact on the behaviourist paradigm in the early 20th century. Later work was done by Robert Rescorla and Allan Wagner to formulate, what Pavlov had discovered, into a mathematical model which has dominated the decision-making literature in cognitive science (Rescorla & Wagner, 1972). More sophisticated models of decision making have evolved in the last decade, many of which leverage that the human brain in many aspects appears to behave as a Bayesian agent, that is both in how we perceive the world, but also in how we make decisions (Mathys et al., 2011, Nosofsky, 2011, Piray & Daw, 2020).

In this paper we will argue that when conducting associative learning experiments where perception is involved, the link between the perceptual experience and the expectations generated in the experiment is crucial as this can gain additional insights to perception and expectations. We will present four hierarchical models previously described in the literature, that we will first modify to the current associative learning experiment and then run them through a thorough Bayesian workflow (Wilson & Collins, 2019). Pareto smoothed importance sampling leave one out cross validation will be used to compare these models on the current data, which will help determine how participants weigh the sources of information in the study (Vehtari et al., 2017). Lastly, we will compare subject-level parameter estimation between two sessions with the same participants to calculate the correlation between the parameter values from the two sessions, effectively doing a test-retest reliability of the parameters of the fitted models.

Perception (JZ)

Throughout history and to an extent still in the public it is believed that perception is a one-to-one representation of the outside world, meaning that an agent is observing directly what "is out there" and that contextual factors play no role in this process. Research has however repeatedly shown that this cannot be the case as our representation of the world changes with the context and in times of very little sensory stimulation we infer what is happening. One of the most striking examples of this phenomena is the Rubber Hand Illusions (RHI). In the RHI experiment, participants are viewing a rubber hand being touched, while the participant's own hand is hidden and touched synchronously. Repeated stimulation of both the rubber hand and the real hand makes the participants associate the rubber hand with their own, and at some point stimulation of only the rubber hand elicits a percept in some participants (Petkova & Ehrsson, 2008, Zopf et al., 2013). The RHI paradigm suggests that there are other factors involved in the perceptual process than solely sensory stimuli. Another well-known example is the McGurk Effect. In the experiment, a video showed a man vocalising a sound repetitively with the mouth shape when saying "bo", while in another video, the mouth shape is as if he is trying to pronounce "fo". People reported hearing different sounds when watching the two videos regardless of the fact that the audios for the two videos are identical (Mcgurk & Macdonald, 1976, Magnotti & Beauchamp, 2017). These phenomena serve as evidence that human perception does not only take the sensory stimulus into account but also other contextual factors.

To investigate the underlying factors of perception, it is necessary to look at the perceptual process, in other words, how perception gets formed and shaped cognitively. This brings up the concept of the "top-down" and "bottom-up" processes that are ubiquitously used in neuroscience, cognitive science and psychology. Due to their broad application over decades, the two terms have evolved into numerous definitions under different contexts. Here we take inspiration from related definitions and summarise the following concepts from a cognitive point of view: Top-down process is a descending/feedforward influence that is delivered from a higher-level representation (such as expectations/beliefs) to a lower-level sensation (e.g. sensory representation). Whereas the "Bottom-up" process refers to the information transition from the opposite direction, i.e. the lower-level features that serve as an input to modify the higher-level interpretation through a feedback/ascending path. The concepts of top-down and bottom-up imply that there is a hierarchy regarding information processing, where the sensory stimuli are seen as a lower level, whereas more integrated information is presented as a higher

level. Although the concepts of the two terms haven't changed much over the last three decades, the discussion of their relationship has evolved (Rauss & Pourtois, 2013). Decades ago, top-down and bottom-up processes were seen as dichotomic pairs, so research usually tried to define a certain process (e.g., perceptual process) as either top-down or bottom-up. However, recent studies argued that the dichotomy view should be avoided, as using this simple binary classification cannot fully capture the complex process that involves intention, expectation and numerous other components (Theeuwes, 2010, Rauschenberger, 2010). Moreover, there seems to be a large overlap between the two processes, taking the visual system for example, some (Barlow, 1997, Rauss & Pourtois, 2013) interpret the top-down process as a way to react to the present sensory stimulus with the prior knowledge, while the bottom-up process is the interaction between the intrinsic structure i.e. both sensory organs and neural connects and the environment, but the structure itself also encodes the memory i.e. the previous knowledge, which overlaps with top-down processes. Meanwhile, studies have shown that the interactive information processing between hierarchical neural systems plays an important role in perception (Di Lollo et al., 2000), which adds more evidence that, instead of differentiating a perceptual process as either top-down or bottom-up, one should investigate the interactive dynamics between them.

Predictive coding (JFE)

The emergence of the predictive coding theory provides a framework to understand perceptual processing, which gives room to dynamic interactions between bottom-up and top-down processes. Firstly used to describe visual processes in the retina, predictive coding theory now has been expanded into a widely-used framework that tries to explain how hierarchical neural units interact by passing neuronal messages (e.g. signals), using a predictive model to represent signals (Rao & Ballard, 1999, Friston, 2010). It's assumed that the system always tries to make the prediction that matches the environment i.e., the sensory stimulus in order to minimise prediction error. When it comes to perception, predictive coding also assumes that the top-down prediction transfers from higher to lower levels whereas the prediction error, i.e. the difference between the current stimulus and prediction, is passed in the opposite direction as a bottom-up process (Friston, 2010, Rauss & Pourtois, 2013). Based on the predictive coding theory, Rauss & Pourtois (2013) goes on to propose that the dynamics between a higher-level system that provides predictions and a lower-level region that receives the stimulus can be seen

as a closed feedback loop, where information exchange between them is bidirectional in order to integrate the internal prediction with sensory information from the environment, which facilitates optimising prediction errors. Therefore, they argue that the predictive-coding theory can be utilised as a conceptual framework for top-down and bottom-up beyond the binary classification.

Bayesian brain theories (JFE)

To further understand the perceptual inference process, it is important to address how the brain works regarding processing information and interacting with the environment. It's self-evident that the world is full of uncertainty, so it is reasonable to assume the brain has equipped or developed the function that learns to adapt to the uncertain world. Bar (2009) demonstrates the brain as a proactive engine that constantly makes predictions of the world under the "analogy → association → prediction" process, which aligns with the predictive coding theory. Moreover, not only evaluating new information, but the brain also extracts previously restored information i.e., memory that associates with the new information to minimise the prediction error. This view can rest on the Bayesian brain hypothesis, a concept viewing the human brain in the Bayesian probability inference paradigm. The concept states that the brain utilises generative models to make the optimal interaction with the world that is, minimising prediction errors in the future, by initiating actions, making decisions, etc., Then the probabilistic inference towards a sensory input i.e. perception, is generated by the posterior probability using Bayes theorem, which can be inferred by the likelihood (i.e. the probability of the input given the current condition) together with prior (i.e. the probability of the current environmental condition) (Knill & Pouget, 2004, Friston, 2010). The Bayesian inference framework has been widely applied in many domains including neuroscience, cognitive science, and the clinic, where it has shown promising results and applications. (Geisler & Diehl, 2003, Körding & Wolpert, 2004, Rao, 2004, Kulkarni et al., 2023). More examples will be presented in the following session.

The link between expectation and perception (JZ)

Together with the predictive coding theory, Bayesian inference theory intuitively leads to the concept of prior belief / expectation that one learns from previous experience. Through decades

of investigation, it's now a consensus in most areas of research that expectation has a strong impact on perception, taking the classical placebo and nocebo effects as examples (Price et al., 2008, Colloca et al., 2008). Therefore, the underlying interactive mechanism between expectation and perception can be formed into the Bayesian inference framework, where an agent's perceptual knowledge is a result of previous exposure to the world (i.e., prior knowledge) and the sensory input (i.e. likelihood) from the world. Along with the predictive coding theory, the updated perception then gets passed ascendingly to generate a new expectation/belief based on the prediction error. Here we see again the bidirectional influences between expectation and perception. Indeed many widely applied cognitive models manage to capture these underlying mechanisms through mathematical equations. For example, the feedback path of updating expectations on prediction error is used in the classical reinforcement learning model - Rescorla Wagner (RW) model.

This view of the brain as a Bayesian observer that constantly updates its belief based upon prior knowledge and sensory evidence also has shown promise in describing mental disorders. For example, Jardri & Denève (2013) and Jardri et al. (2017) investigated Schizophrenia (SCZ) patients under a Bayesian inference framework. Psychotic symptoms can be interpreted as an imbalance of the weight put on either the prior belief or the sensory evidence. For example, when a person aberrantly sticks to their prior beliefs of the world, they perceive what they're expecting and not what is actually being presented. While putting a rather high weight on every new information even when it's not reliable, would cause a person to constantly switch beliefs and be in constant internal conflict. Based on this mechanism, Jardri & Denève (2013) and Jardri et al. (2017) extend the concept as "circular inference", the misinterpretation during the bidirectional information exchanges and extreme mutual effects will end up making the subject "see what we expect" and / or "expect what we see" (Jardri et al. 2017). In the study, the circular inference model best captured the participants' responses for both SCZ and healthy controls. The results also showed that the SCZ group did put higher weight on their priors compared to the healthy controls. Similarly, studies (Jepma et al., 2018, Onysk et al., 2023) also show a positive dynamic feedforward and feedback interplay between expected pain and pain perception, namely, a higher expected pain leads to higher perceived pain from participants' self-reported pain scale but also from their neural responses to pain. In the studies in question, only the computational models that present the positive feedback loop between pain perception and expectation succeed in capturing the main pattern of the experimental data, which contributes more evidence that there's a dynamic positive feedback loop between

expectation and perception. Results from the parameter estimation of the Rescorla Wagner (RW) model also revealed that participants with chronic back pain weighted their prior expectation more compared to healthy controls. Revealing a new way to understand the underlying cognitive mechanisms of pain, these findings have significant implications for pain regulation.

Indeed, the Bayesian brain theory provides a significantly supportive interpretation of the relationship between expectation and perception. Many studies to date have investigated this relationship through perceptual associative learning tasks. The core idea of these experiments is that by introducing associations between a stimulus and a cue one can investigate the proactive prediction making process in the brain, as the cue-stimulus association will be encoded as prior knowledge, which is used to make future prediction (Bar, 2009). Therefore, participant's expectation can be artificially manipulated by different cue-stimuli associations. Using this paradigm, studies have shown that changes in expectations relate to changes in precepts through computational models, but most of the studies only investigate the impacts of expectation on percept and not the other way around. Based on perceptual mechanism under the Bayesian inference framework, and the results from the previous associative learning studies (Jepma et al., 2018, Onysk et al., 2023), we argue that a computational model should represent the bidirectional interactions, which might serve as a better interpretation of the process, especially given the fact that participants have to learn the cue-stimulus association through their percepts, meaning that the expectation of the stimulus will influence the percept, but the percept of the stimulus will determine how the belief of the cue-stimulus association will change.

Bayesian workflow in computational modelling (JFE)

In recent decades computational models have seen a surge in popularity both due to the flexibility that this modelling approach offers, but also due to the accessibility of more powerful computers (Roberts & Hutcherson, 2019). It has been hypothesised by several authors that to move towards more theory-based and perhaps more rigorous understandings of the human mind and cognition, moving beyond linear and generalised linear models is necessary (Press et al., 2022, Guest & Martin, 2021). In the domain of decision making, computational modelling has seen great promise in providing additional nuance to simple associative learning

experiments, but also in more complex social decision-making tasks (Simonsen et al., 2021, Iglesias et al., 2013). These computational models are supposed to offer a gauge of latent variables which govern the decision process of the participants, which cannot be obtained by the usual outcome-focused analysis using generalised linear models. These latent variables have been hypothesised to be related to psychiatric disorders like schizophrenia or health conditions like depression or chronic pain (Karvelis et al., 2023). However, caution must be taken before interpreting the parameters of such complex computational models and especially inferring that a difference is due to differences in underlying disorders without properly investigating their implications and validity (Karvelis et al., 2023).

It has been argued that to ensure that these complex mathematical models behave the way we expect them to and provide meaningful information about the human mind and behaviour, we should test and validate them in a fail-safe environment where the latent parameters are known, utilising simulated data, here called the Bayesian workflow (Wilson & Collins, 2019, Gelman et al., 2020). From a practical point of view this entails stimulating behaviours for many different parameter values of the computational models and then verifying that the observed behaviours are firstly plausible, but also something to be expected from participants performing the experimental task. After having verified that the models provide reasonable responses it is recommended that the simulated responses are fit to the generative model that, given the data, gives the most likely parameter values. This step of first running the model forward, simulating responses from known parameter values, and then running the data back into the generative model and getting parameter values is known as parameter recovery (Wilson & Collins, 2019). This step is a crucial step in computational modelling as researchers interpret directly on these parameters, or on latent variables calculated from the parameters. If models do not recover any or some of the parameters, rethinking the experimental paradigm, the number of subjects, the number of trials or the model itself is advised. As the interpretation of these parameters when fit to real data can become next to meaningless if the model cannot obtain similar values to those that it was simulated from. After having achieved successful parameter recovery for the models tested, the next step is to ensure that the tested models in the experimental paradigm are well distinguished. This step ensures that if an agent follows the rule from one of the models implemented, then model comparison would be able to find this distinction. This step therefore involves simulating data from all models used for testing and then fitting all models to all the simulated data producing a square matrix of the size of the model space (i.e. the number of models tested) (Wilson & Collins, 2019, Gelman et al., 2020).

The beauty and the curse about the flexibility of these models is that all these steps can be performed by simply simulating data and should therefore ideally be done before data collection. This is because in the end the experimental paradigm will determine whether models are recoverable, distinguishable and useful. However, this convenience that everything boils down to the experimental paradigm is also the curse, as it's often the case that researchers take existing experimental paradigms and modify them to answer their specific hypothesis of interest. This in an idealised world would therefore entail rerunning all the initial steps to ensure internal consistency in the models themselves (parameter recovery) but also discriminability between the models tested (model recovery).

Another concern of not only computational models, but statistical models in general, is whether parameter values are stable over time. This has recently gotten attention in the literature showing that for some experimental paradigms and some models, parameters are stable across time (Schaaf et al., 2023, Waltmann et al., 2022), but not all (Mkrtchian et al., 2023). It is to be noted that many of these studies report that fitting hierarchical model and utilising the pooling effects that these models exert is crucial to either get a good correlation between the parameter estimates between sessions or a high intraclass correlation, which is normally used in test-retest reliability studies.

Here we implement 4 computational models that have previously been used in other experimental paradigms. Specifically, we implement a modified version of the original Rescorla Wagner model (MRW) that was together with a modified Kalman filter (KF) both introduced by Jepma et al. (2018). The modification to these models entails a connected loop between expectation and perception. We also implement what has been called a weighted Bayes (WB) agent (inspired from the model introduced by Jardri & Denève (2013). As a base line for model comparisons, we additionally implement a Simple RW model (SRW) where only the perception gets influenced by expectation not the other way around. Lastly, we also test the test-retest reliability of our models as the experimental paradigm was performed twice with a week between visits.

Methods (JZ, JFE)

Experimental procedure (JZ)

The experiment was a thermosensory associative learning task containing 2 parts. In the study, a thermo stimulator was placed on the participant's volar forearm providing heat and warm stimulation. Before the main experiment (the associative learning task) was conducted participants went through 40 trials of receiving stimuli of varying intensity to determine their sensitivity to when a temperature was felt as burning. The 40 intensities were determined by an adaptive Bayesian algorithm that computes the temperature (intensity value) that will minimise entropy in the parameter estimates (slope and threshold) of the psychometric function the most. After having estimated the slope and threshold for each participant the temperatures used in the main experiment were set to 3 standard deviations above and below the mean (threshold) for the heat pain and warm stimulus respectively. It should be noted that in the study, the term "burning" is not referring to the feeling when the skin got burned or any sensations resulting from actual skin damage, instead, it's about an unpleasant or noxious sensation. While "not burning" refers to the comfortable warmth. All participants acknowledged this definition before giving responses. This procedure was done for each participant twice, with a week between visits, like the rest of the study.

Training phase (JFE)

To familiarise the participants with the task, every subject went through 10 practice trials of the main experiment. During these trials, participants were told in advance that a stimulus would be given after an arbitrary cue and there were two types of cues, one of which was associated with burning stimuli while another was associated with the warm stimulus. Participants were explicitly told that this association would be deterministic for the training phase, but probabilistic for the main experiment. Lastly, to have the participants experience a reversal of cue-stimulus association the first 6 trials featured that cue 1 was associated with the hot stimulus and in the last 4 it was cue 2 that was associated with the hot stimulus. For a full list of cues used in the experiment see appendix A. On each trial, after participants received the cue, they were asked to predict whether the following stimulus would be burning or not. Then they received the stimulus and provided feedback on whether they felt the stimulus as burning or not.

Main experiment (JZ)

Two cues, novel from those in the training phase, were introduced to avoid bias learned from the previous phase. Like the training phase, participants were asked to make predictions based on one of the cues presented. Then the stimulus would be presented, and participants were to indicate what kind of stimulus they thought they received. Besides giving a binary response of the stimulus i.e., burning vs warm, they were also asked to give a rating on how intense they felt it. This subjective rating was provided on a visual analog scale ranging from "did not detect any sensation" to "the most intense I have ever experienced". Another difference to the training phase, which was explicitly told to every participant was that the association between cues and stimuli was probabilistic and therefore not fixed, in other words, even during the periods where cue 1 is associated with the burning stimuli, there would still a probability that a warm stimulus would follow. To help participants not get discouraged about the difficulty of the task, it was explicitly stated that the task would not be easy and getting many predictions wrong was normal, and that they just had to do their best. Participants then went through 160 trials of cue-stimulus association, which were divided into 8 sub-sessions i.e. breaks after every 20 trials to account for potential habituation and sensitisation effects of the skin. To account for the fact that participants performed the experiment twice the underlying cue-stimulus association was never the same for both visits for one participant and the cues used would also change between visits. This change in cue-stimulus association and cues were counterbalanced.

Participants (JFE)

The participants in the current study are healthy subjects without reported physical or mental diseases, e.g., cancer, psychiatric diseases, neurological diseases, chronic pain, etc. nor a history of those diseases. All the participants were between 18 to 30 years old. Participants gave informed consent and answered a questionnaire relating to their health status. The study recruited 50 participants including 15 females. These 50 participants received a mean temperature of 48 $\pm$1.2 (mean$\pm$sd) degrees for the high intensity stimulus and 44$\pm$2.3 degrees celsius for the low intensity stimulus.

Computational Models (JZ)

We develop four different computational models with different assumptions about how participants learn and update their beliefs. All models include a perceptual and a response

model, the former describing how agents come to their perceptual experiences and the latter being how these are translated into the responses that are recorded in the experiment. In the following section we represent our four models and their update equations together with how these models relate and differ in assumptions made about how agents learn and perceive. To get a quick overview of these models readers are referred to appendix B where a plate notation for each model is presented together with prior values for each parameter.

The Rescorla Wagner model is a mathematical model of how classical conditions are thought to occur. The basic idea is that an agent updates his belief based on a weighted prediction error which is the difference between the last cue-stimulus association and the last belief. The weighing of the prediction error is generally called the learning rate as this parameter determines the degree to which an agent utilizes the new information to update his belief. This learning rate α is for the Rescorla Wagner agent assumed to be constant.

$$E_t = E_{t-1} + \alpha * \underbrace{(A_{t-1} - E_{t-1})}_{\text{Prediction Error (PE)}} \quad (1)$$

Where E_t is the expected value of trial t and A_{t-1} being the outcome association observed by the participant. For our first computational model, which we call the simple Rescorla Wagner (SRW) we assume that participants learn about the cue stimulus association using (1). This expected value is then transformed into a belief of either warm or hot stimulus, using the cue, which is used to generate responses in a generalized linear framework, to predict the continuous and binary responses of the participant. See appendix B for a full description of this model.

For our next model we modify equation (1) to encapsulate the interaction between precept and expectations, we simply call this model as the Modified Rescorla Wagner (MRW). We do this by modifying how the prediction error is calculated as now the prediction error is going to be calculated from the precept ψ_t that the expectation and the stimulus elicits. We assume for the MRW agent that the weighting of stimulus and expectation is governed by a single parameter γ, like the learning rate α it determines the weight that is put on the stimulus vs expectation in determining what the participant felt on the current trial ψ_t

$$\psi_t = \gamma \cdot S_t + (1 - \gamma) \cdot E_t \quad (2)$$

This percept is then used to generate a prediction error that updates the belief of the agent for the next trial.

$$PE_t = \psi_t - E_{t-1} \quad (3)$$

$$E_t = E_{t-1} + \alpha * \mathrm{PE_t} \quad (4)$$

The weighted Bayes model (WB) is inspired by Bayes theorem where two sources of information a prior and a likelihood are combined to a posterior belief. The difference between this and the Rescorla Wagner model is that the percept which is used to update the belief on each trial is calculated using a simplified Bayes theorem, where the logit of the posterior is being set equal to the sum of the weighted logits of the sources of information. This entails that the two sources of information are weighted independently, resembling a normal linear regression.

$$\psi_t = S\big(w_1 * S^{-1}(S_t) + w_2 * S^{-1}(E_1)\big) \quad (5)$$

where $S(x) = \frac{1}{1+e^{-x}}$. The rest of the update equations remain similar to the Resorla Wagner, essentially also utilizing the RW model to update the expectation on each trial with the prediction error being (3)

Kalman filter (JFE)

The Kalman filter can be thought of as a generalization of the MRW model, where the assumption of the learning rate being constant is not met. From a normative perspective it makes sense that agents should update the rate at which they learn depending on the context of the experiment. In the Kalman filter model it is assumed that the learning rate is trial specific and dependent on the uncertainty of the prediction made, where higher uncertainty means a higher learning rate and lower uncertainty a lower learning rate. It is assumed in the Kalman filter that the agent keeps track of both belief as in the MRW model, but also the uncertainty of that belief in order to update his belief based on both.

To arrive at the update equations for the Kalman filter we start off by assuming that each agent observes the Stimulus with Gaussian noise:

$$S_t \sim \mathcal{N}(\psi_t, \sigma_s^2) \tag{6}$$

Next we assume that the mean of this normal distribution is sampled from another normal distribution with the expectation at trial t as its mean

$$\psi_t \sim \mathcal{N}\left(E_t, \sigma_\psi^2\right) \tag{7}$$

Where this mean is updated based on the stimulus given the cue on previous trials

$$E_t \sim \mathcal{N}\left(E_{t-1}, \sigma_\eta^2\right) \tag{8}$$

Lastly this prior expectation, i.e., the expectation of the last trial is given by all the previous inputs which the agent keeps track of.

$$E_t|S_{t-1} \sim \mathcal{N}\left(\mu_{E_{t-1}}, \sigma_{E_t}^2\right) \tag{9}$$

Now combining (9) and (7) we get a prior for the precept at a given trial, which is before getting the stimulus but after getting the cue

$$\psi_t|S_{t-1} \sim \mathcal{N}\left(\mu_{E_{t-1}}, \sigma_{E_t}^2 + \sigma_\psi^2\right) \tag{10}$$

After the stimulus is observed (the likelihood) (5) we can combine this with the prior (9) using bayes rule to get a posterior for percept.

$$\psi_t|S_t \sim \mathcal{N}\left(\frac{\sigma_s^2 * \mu_{E_{t-1}} + \left(\sigma_\psi^2 + \sigma_{E_t}^2\right) * S_t}{\sigma_s^2 + \sigma_\psi^2 + \sigma_{E_t}^2}, \frac{\sigma_s^2 + \sigma_\psi^2 + \sigma_{E_t}^2}{\sigma_s^2 * \left(\sigma_\psi^2 + \sigma_{E_t}^2\right)}\right) \tag{11}$$

The mean rating of the agent on trial t is therefore given by the mean, which can also be written as:

$$\mu_{\psi_t} = \frac{\sigma_s^2}{\sigma_s^2 + \sigma_\psi^2 + \sigma_{E_t}^2} * \mu_{E_{t-1}} + \frac{\sigma_\psi^2 + \sigma_{E_t}^2}{\sigma_s^2 + \sigma_\psi^2 + \sigma_{E_t}^2} * S_t \tag{12}$$

where it becomes evident that the mean rating at trial t is a weighting between the expectation and the stimulus on the current trial.

After the posterior for the percept is calculated the belief is then updated which can then be used as priors for the next trial.

$$E_t|S_t \sim \mathcal{N}\left(\frac{\left(\sigma_s^2+\sigma_\psi^2\right)*\mu_{E_{t-1}}+\sigma_{E_t}^2*S_t}{\sigma_s^2+\sigma_\psi^2+\sigma_{E_t}^2}, \frac{\left(\sigma_s^2+\sigma_\psi^2\right)*\sigma_{E_t}^2}{\sigma_s^2+\sigma_\psi^2+\sigma_{E_t}^2}+\sigma_\eta^2\right) \quad (13)$$

It can be shown that these update equations for the expectation and the mean percept are a generalization of the above mentioned MRW model if we define two trial by trial parameters γ_t and α_t

$$\gamma_t = \frac{\sigma_s^2}{\sigma_s^2+\sigma_\psi^2+\sigma_{E_t}^2} \quad (14)$$

$$\alpha_t = \frac{\sigma_{E_t}^2}{\sigma_\psi^2+\sigma_{E_t}^2} \quad (15)$$

one can arrive at the following update equations for the percept and the expectation.

$$\psi_t = (1-\gamma_t)\cdot S_t + \gamma_t \cdot E_t \quad (16)$$

$$E_t = E_{t-1} + \alpha_t * (\psi_t - E_{t-1}) \quad (17)$$

For the derivation of equations (16) and (17) readers are referred to appendix E. As can be seen equations (16) and (17) closely resemble the structure of equation 2 and 4 however with a trial instead of subject dependent learning rate as well as a trial dependent weighting of stimulus vs expectations.

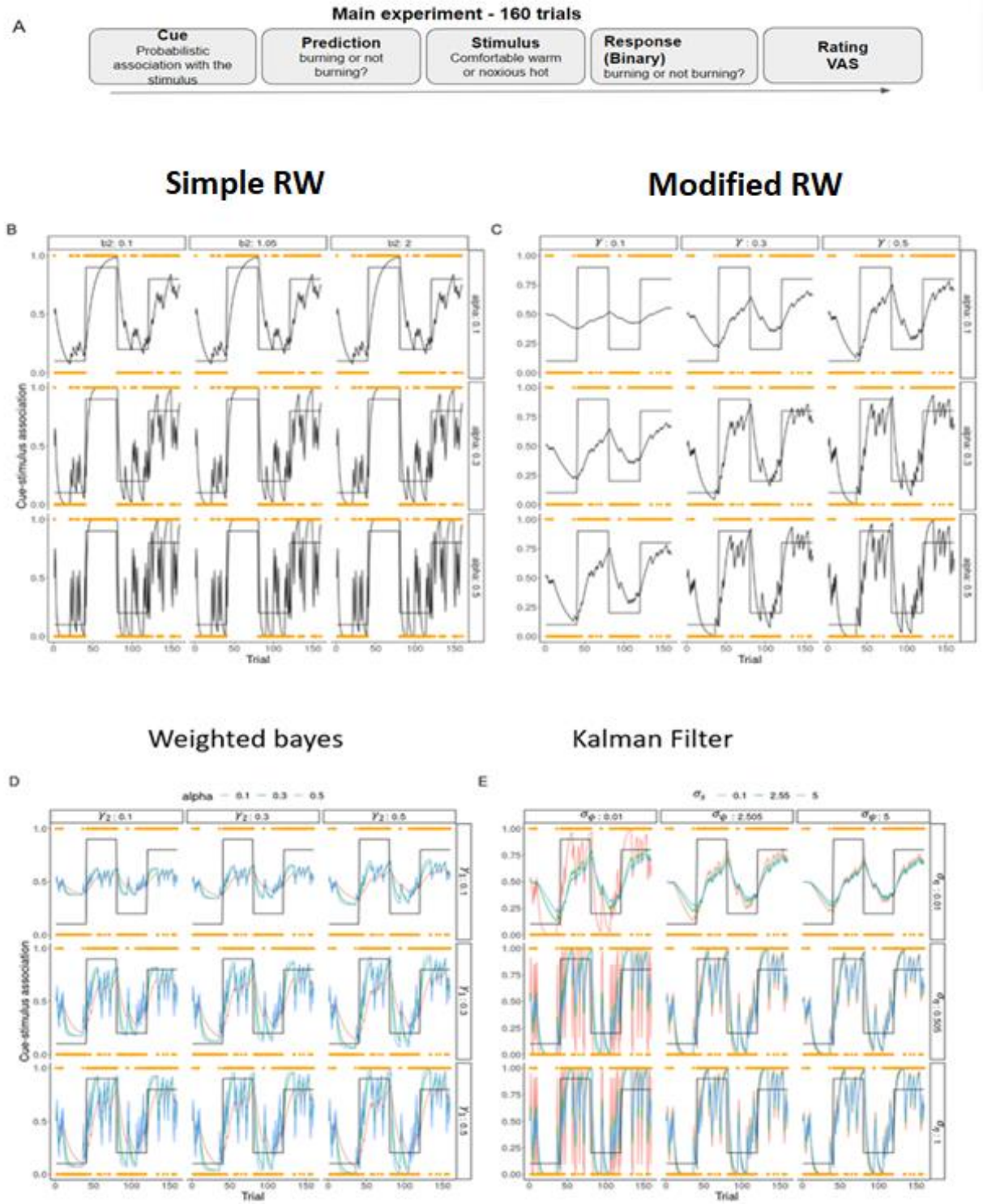


Figure 1. *The experimental paradigm and cue-stimulus association stimulations.* ***A*** *illustrates the trail structure of the main experiment. The two experiment sessions for test-retest reliability examination share the same trail structure.* ***B, C, D, E*** *show the real association between each cue and stimulus (solid black line), together with the simulated agent's inferred association along trails, with different parameter values, of the Simple RW, the Modified Rescorla Wagner, the Weighted Bayes and the Kalman filter model.*

Sampling (JZ)

To invert our models, we used Hamiltonian Markov chain Monte Carlo (HMCMC) sampling implemented in cmdStanR 0.5.2 (Carpenter et al., 2017). All models were run with 4 chains, 1000 warm up iterations and 1000 samples. Model quality and convergence was checked by plotting the chains of parameters as well as investigating the Rhat value where values below 1.1 are typically used as a guidance for good convergence (Brooks & Gelman, 1998). Lastly divergent transitions were investigated which can be present if the posterior parameter space is hard or perhaps impossible to explore (Betancourt, 2016).

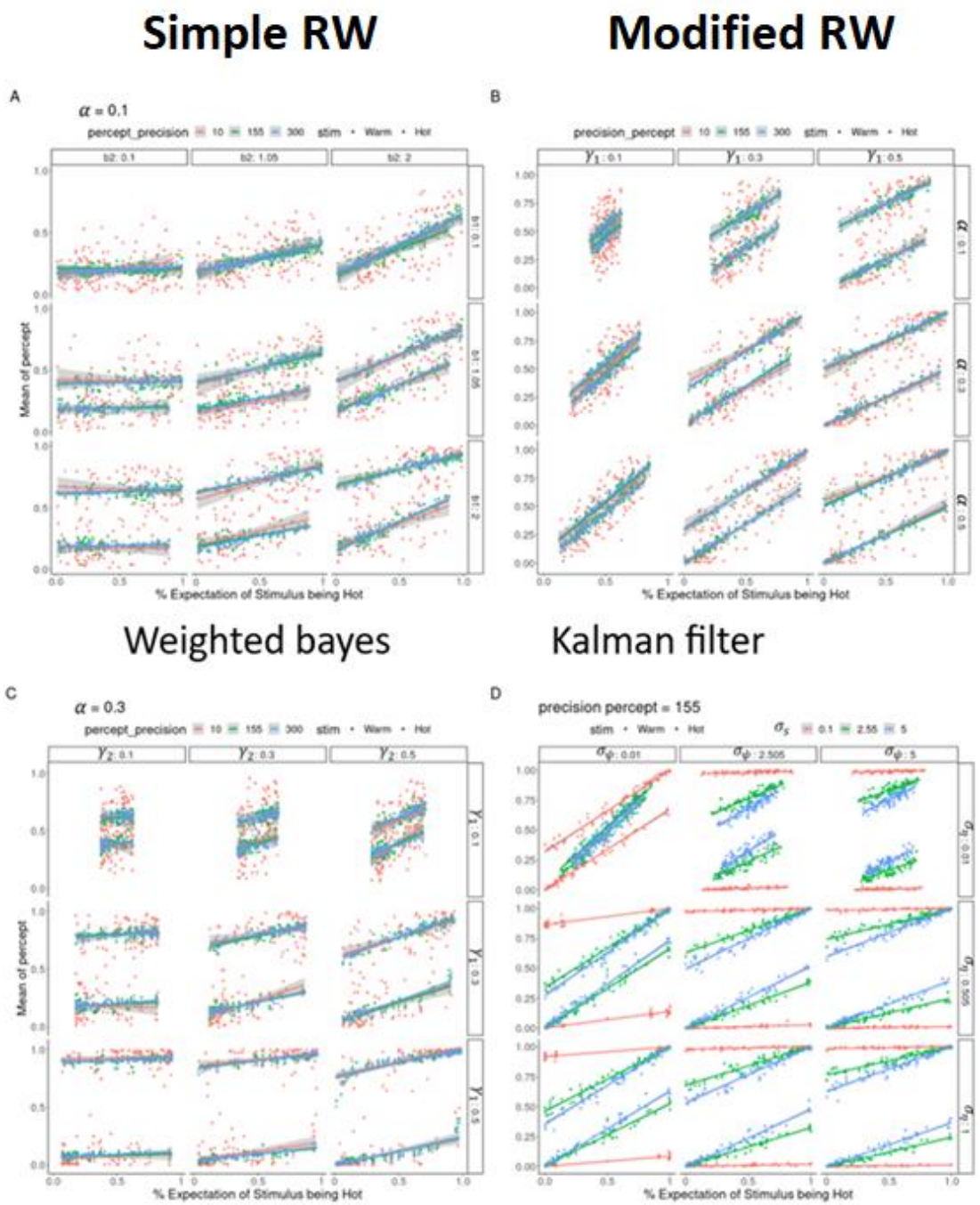


Figure 2*. Expectation (X-axis) and perception (Y-axis) stimulations of the four computational models with different parameter values. A: for Simple Rescorla Wagner, B: for Modified Rescorla Wagner, C: for the Weighted Bayes, D: for the Kalman filter*

Results (JFE)

To show internal validity of our implemented models we first simulate agents using the generative scheme outlined by the models with different parameter values and simulate their behaviour. Figure 1 displays how the four different models learn the cue-stimulus association and Figure 2 shows how expectation and perception are interacting. Next, we performed parameter recovery for each of the models, for a full list of parameters simulated see appendix C. Parameter recovery for the four models shows that 3 of the 4 models have internal validity both in the hierarchical level (see Figure 3) and on the subject level see appendix D. Given the fact that the Kalman filter with some combination of parameter values took several hours to run and always included many divergent transitions in the sampling process we decided to discard this model completely. Next, we checked the discriminability between models by performing model recovery of which the results can be seen in Table 1. After having performed validity checks of our models we fit the models to our experimental data. This step was performed in two steps, first we fit our models separately to each of the two sessions to determine which of the models best explained the data from the two sessions independently. Table 2 shows the winning model of session 1 and session 2 using Pareto smoothed importance sampling leave one out cross validation. Lastly, we fit our models to both sessions and in this calculated the correlation between the subject specific parameters from session 1 and session 2. Figure 4 shows how the most interesting subject specific parameters correlate between the two sessions for the three models tested. The highest posterior density interval was calculated and is displayed in Table 3.

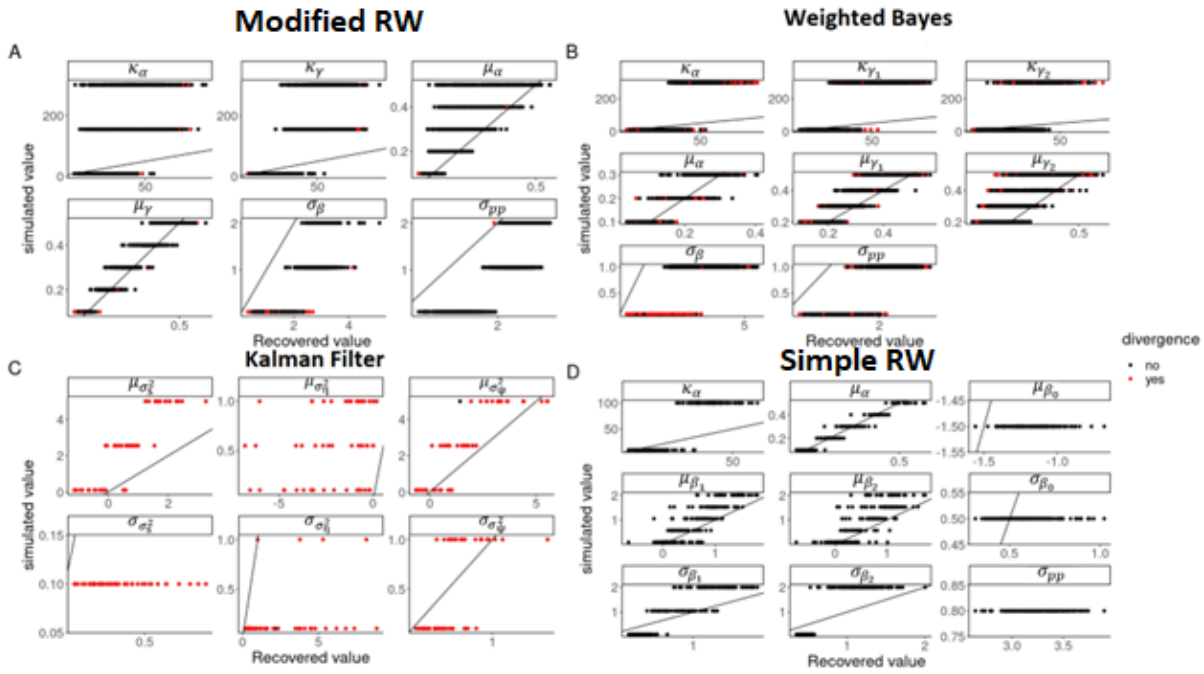


Figure 3*. Parameter recovery of the computational models. Note the excessive number of divergences in the Kalman filter model even with only 50 simulated parameter values.*

Fitted models - - - - - - - - - - - - - - Fitted data	Modified Resorla Wagner	Weighted Bayes	Simple Resorla Wagner
Modified Rescorla Wagner	0	100	0
Weighted Bayes	0	100	0
Simple Rescorla Wagner	0	0	100

Table 1 *Model recovery for the Modified Rescorla Wagner, Weighted Bayes and Simple Rescorla Wagner models*

model	Session 1			Session 2		
	ΔELPD	SE	Weight	ΔELPD	SE	Weight
Modified Rescorla Wagner	-5441	110	0.018	-5691	116	0.018
Weighted Bayes	-2152	92	0.018	-2394	100	0.017
Simple Rescorla Wagner	0	NA	0.964	0	NA	0.965

Table 2 *Model comparison for the three models on the two sessions. ELPD = the expected log predictive density of the model, calculated using Pareto smoothed importance sampling leave one out cross validation (PSIS-LOO). SE = standard error in the estimate of the difference. Weight indicates the stacking weights based on PSIS-LOO, i.e. the preference for a given model under the data*

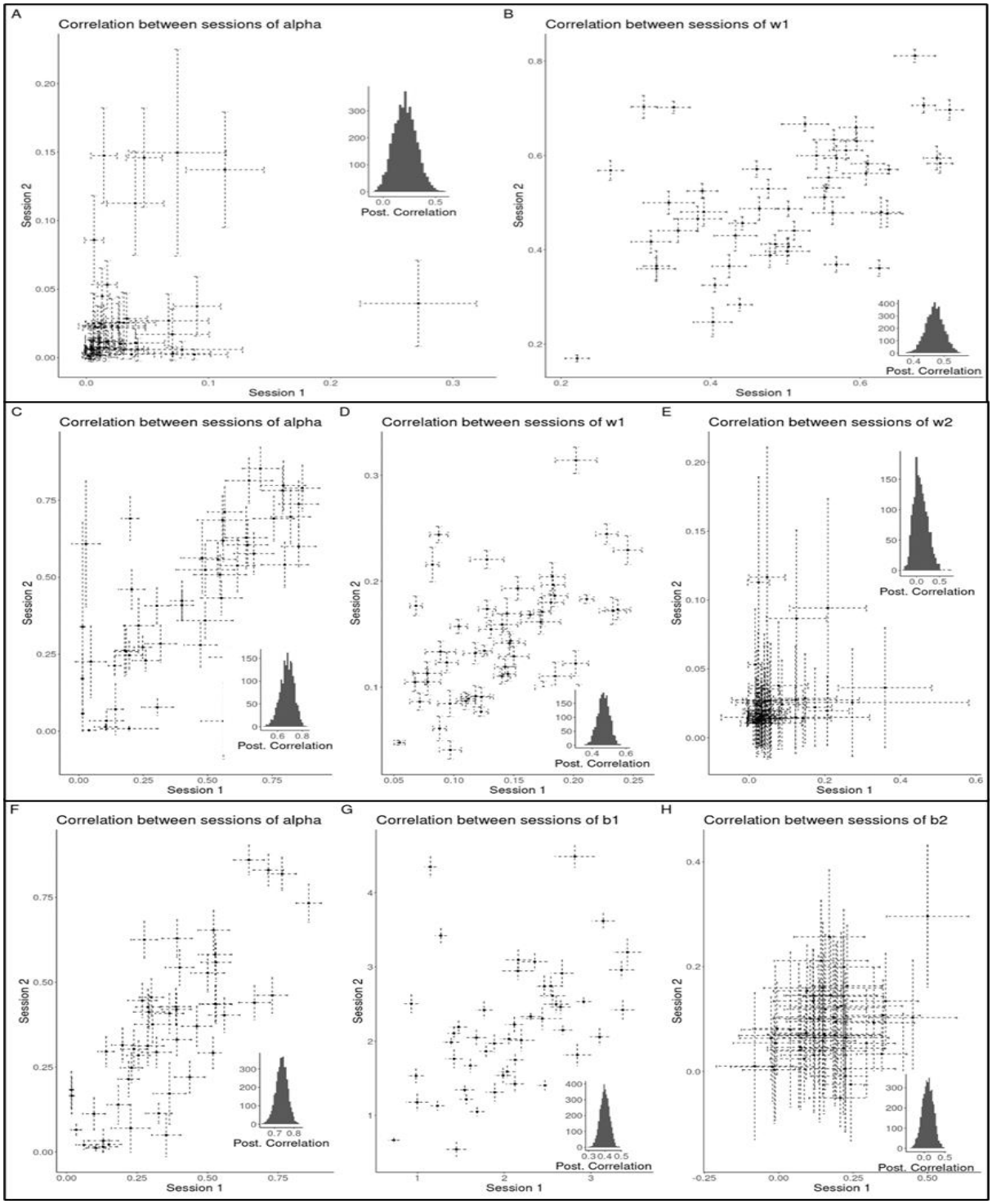


Figure 4. *Parameter correlation for the two sessions, for three computational models investigated. All figures display one point for each subject and the uncertainty, i.e. one standard deviation, for the given session. Lastly all figures display 4000 posterior draws from the correlation coefficient as a histogram. The first row* ***(A, B)*** displays the correlation for the Modified Rescorla Wagner model, the second row ***(C, D, E)*** shows the correlation for the weighted Bayes model and lastly the third row ***(F, G, H)*** displays the correlation for the simple Rescorla Wagner model.

model	Modified Rescorla Wagner		Weighted Bayes			Simple Rescorla Wagner		
parameter	α	γ	α	γ_1	γ_2	α	β_1	β_2
highest posterior density interval	[0; 41]	[0.41; 0.53]	[0.55; 073]	[0.40; 0.54]	[-0.18; 0.36]	[0.68; 0.79]	[0.34. 0.46]	[-0.19; 0.33]

Table 3 *Highest posterior density intervals of the correlation for some of the subject specific parameters of the three models tested.*

Discussion (JZ, JFE)

Unidirectional link between expectation and perception (JZ)

Computational models allow for a flexible theory-driven approach to how data from experimental paradigms are analysed. In the current study we utilise an associative learning paradigm using thermal stimuli to investigate 2 main questions. Do participants' expectation and perception have a mutual impact on one another or is it a unidirectional influence from expectation to perception? Secondly are the parameters that determine these relationships stable over time. To answer these questions, we built 4 computational models with differing assumptions about the underlying mechanisms of participants learning cue-stimulus associations, using thermal stimuli. Our first model (i.e., Simple Rescorla Wagner (SRW)) was a baseline model that assumed that only expectations influence perception which is commonly assumed in these experimental paradigms. Our last three models assumed that perception and expectation about the upcoming stimulus have a mutual impact on each other, meaning there is a bidirectional effect between expectation and perception. After performing parameter recovery and model recovery of these novel computational models, we show that the Kalman filter could not recover its parameters and that the remaining models assuming a mutual impact between expectations and perception could not be distinguished when performing model recovery. Even though the Modified Rescorla Wagner (MRW) model and the weighted Bayes (WB) model couldn't be distinguished using model recovery we still used both models in model comparison. Using the expected log predictive density, we evaluated the three models on each of the two sessions and found that for both sessions the SRW outperformed the weighted Bayes by over 23 standard errors and the MRW model by over 49 standard errors. This was also reflected in the model weight which is the strength of preference for a given model, where over 95% of the weight was given to the SRW model for both sessions. Lastly, to investigate how the parameters of these models varied across time, we fitted the three models to both experimental sessions and calculated the correlation between the subject specific parameters, see Table 3.

Test re-test reliability (JFE)

The MRW model only had a reliable correlation between the two sessions for one parameter, the weighting of stimulus and expectations on perception. The learning rate for the model showed a positive correlation but with a high posterior density interval, meaning that it was not consistent between sessions, which is also evident from Figure 4. The WB model has a reliable correlation between sessions for two parameters namely the learning rate and the

weight for how the stimulus affects perception, but not for the weight for how expectation influences perception. Lastly the SRW, the winning model in model comparison, showed the same pattern as the WB. These analyses show that in the current experiment the link between perception and expectation seems to be very weak as the winning model assuming a unidirectional link between perception and expectation outperformed the remaining models that assume a mutual influence. A reason for this pattern could be that in the experimental paradigm, the stimulus used was tailored to each individual participant and the intensity was chosen such that there were 6 standard deviations between them, meaning that the difference between the two stimuli was very big. This could serve as a ceiling effect for how much expectations matter in the perception of the participants, as participants might simply rate the two stimuli as two discrete entities as the difference between them is so big, leaving very little for the expectation of the participant to explain. This interpretation would also explain why the previously mentioned studies found that models with a bidirectional link were superior given that in their experimental setup used 47 and 48 degrees. Whereas the temperatures used in the current experiment was for the high temperature 48 $\pm$1.2 and 44$\pm$2.3 degrees celsius for the low temperatures.

Limitations and Future Directions (JZ, JFE)

Model limitations (JZ)

The current study implemented 3 computational models and 1 baseline model for comparison, trying to investigate the optimal model to describe the relationship between expectation and perceptions. Among them, 3 models worked properly and were able to get run through the Bayesian workflow, namely: parameter recovery, model recovery, model comparison, parameter estimation for the experimental data and the test-retest reliability test. The Kalman Filter (KF) model could not produce meaningful parameter recovery, due to a large amount of divergent transitions (figure 3C). Previous studies have applied the KF with similar associative tasks, with a similar modification as applied in the current study, however these studies do not report on parameter recovery or on model convergence besides the Rhat value (Jepma et al., 2018, Onysk et al., 2023). Future work should look into the KF model and perhaps try and fix some of the free parameters to make the model space easier to explore, which was done in one of the two previously cited papers.

With the models that ran through all phases (i.e., MRW, WB and the SRW), due to the considerate time consumption, model recovery was done with only one set of parameters for 100 simulations and both parameter recovery and model recovery was done with only 10 subjects instead of the 50 participants the experiment was conducted with. The main limitation of running the model recovery with only one set of parameter values is that the given set of parameter values might not distinguish two models as was the case for the MRW and the WB, but that another set of parameter values would. To completely investigate the difference and similarities of these models it would therefore be preferable to run with many different sets of parameter values as was done for parameter recovery, see appendix C.

Experimental limitations (JFE)

Besides the simulation, it is also worth attention that there are some potential limitations in the experiment design. As described, there are 160 trials in the main experiment and 50 more trials before that (i.e., the threshold measuring and the training phase). Given the repetitive nature of the experimental design, fatigue, boredom sensitization and habituations are crucial variables to account for. To help combat fatigue and boredom, future designs may consider giving participants some incentives to concentrate and keep them motivated, for instance by gamifying the task (Robison et al., 2021).

Another experimental limitation lies on that the stimulator was placed at the same place of the participant's forearm during the whole experiment, which may cause sensitization or habituation of the skin. According to previous studies on pain, when participants received repetitive noxious heat stimuli, their perception of the pain will either increase, steady, or decrease, varying from each individual (Naert et al., 2008, Hollins et al., 2011). The experimental setup tried to account for these effects by having breaks every 20 trials, however the previously cited literature suggests that these effects occur even after 3-5 trials. We would advise against including more breaks in the experiment as participants have to remember the cue-stimulus association and extended breaks might make them forget the association they ended with. Therefore, to combat sensitization and habituation effects in the experiment one would ideally have moved the thermal stimulator after 3-5 trials such that a different area of the forearm was stimulated. Another approach to solve the issue of sensitization and habituation effects would be to include them in the computational model.

Conclusion (JZ, JFE)

The current study investigates the link between expectation and perception through an associative learning task. 3 computational models (another one is discarded due to the divergence issue) using the Bayesian inference framework that assume either unidirectional or bidirectional influence are examined by parameter recovery, model recovery and model comparison. Results from parameter recovery imply that all 3 models' parameters can be properly recovered indicating internal validity of the models. The winning model turned out to be the Simple Rescorla Wagner model that presents a unidirectional link between expectation and perception, which counters the results from previous studies. The clear distinction between the hot and warm stimuli could be a potential reason for the current results as when the differences between the sensory stimuli are salient, the expectation may not be taken into much account. Therefore, we suggest future studies should investigate the relationship between expectation and perception with painful and non-painful stimuli that are not as distinct. Using these computational models, we also examined the test-retest reliability of the parameters of these models, of which the results showed that some, but not all, parameters were highly correlated. Several limitations regarding the simulation and the experimental design are also presented for future directions.

Code and data availability (JZ, JFE)

All data and code to rerun the whole analysis can be found in the following GitHub:
https://github.com/JesperFischer/Data-Science-Prediction-and-Forecasting.git

References

A simple model for learning in volatile environments | PLOS Computational Biology. (n.d.). Retrieved May 24, 2023, from https://journals.plos.org/ploscompbiol/article?id=10.1371/journal.pcbi.1007963

Bar, M. (2009). The proactive brain: Memory for predictions. *Philosophical Transactions of the Royal Society B: Biological Sciences*, *364*(1521), 1235–1243. https://doi.org/10.1098/rstb.2008.0310

Barlow, H. B. (1997). The knowledge used in vision and where it comes from. *Philosophical Transactions of the Royal Society of London. Series B: Biological Sciences*, *352*(1358), 1141–1147. https://doi.org/10.1098/rstb.1997.0097

Betancourt, M. (2016). *Diagnosing Suboptimal Cotangent Disintegrations in Hamiltonian Monte Carlo* (arXiv:1604.00695). arXiv. https://doi.org/10.48550/arXiv.1604.00695

Brooks, S., & Gelman, A. (1998). General Methods for Monitoring Convergence of Iterative Simulations. *J. Comput. Graphi. Stat.*, *7*, 434–455. https://doi.org/10.1080/10618600.1998.10474787

Carpenter, B., Gelman, A., Hoffman, M. D., Lee, D., Goodrich, B., Betancourt, M., Brubaker, M., Guo, J., Li, P., & Riddell, A. (2017). Stan: A Probabilistic Programming Language. *Journal of Statistical Software*, *76*, 1–32. https://doi.org/10.18637/jss.v076.i01

Colloca, L., Sigaudo, M., & Benedetti, F. (2008). The role of learning in nocebo and placebo effects. *Pain*, *136*(1), 211–218. https://doi.org/10.1016/j.pain.2008.02.006

Di Lollo, V., Enns, J. T., & Rensink, R. A. (2000). Competition for consciousness among visual events: The psychophysics of reentrant visual processes. *Journal of Experimental Psychology: General*, *129*(4), 481–507. https://doi.org/10.1037/0096-3445.129.4.481

Friston, K. (2010). The free-energy principle: A unified brain theory? *Nature Reviews Neuroscience*, *11*(2), 127–138. https://doi.org/10.1038/nrn2787

Geisler, W. S., & Diehl, R. L. (2003). A Bayesian approach to the evolution of perceptual and cognitive systems. *Cognitive Science*, *27*(3), 379–402. https://doi.org/10.1207/s15516709cog2703_3

Gelman, A., Vehtari, A., Simpson, D., Margossian, C. C., Carpenter, B., Yao, Y., Kennedy, L., Gabry, J., Bürkner, P.-C., & Modrák, M. (2020). *Bayesian Workflow* (arXiv:2011.01808). arXiv. https://doi.org/10.48550/arXiv.2011.01808

Guest, O., & Martin, A. E. (2021). How Computational Modeling Can Force Theory Building in Psychological Science. *Perspectives on Psychological Science*, *16*(4), 789–802. https://doi.org/10.1177/1745691620970585

Hollins, M., Harper, D., & Maixner, W. (2011). Changes in pain from a repetitive thermal stimulus: The roles of adaptation and sensitization. *Pain*, *152*(7), 1583–1590. https://doi.org/10.1016/j.pain.2011.02.049

Hutmacher, F. (2019). Why Is There So Much More Research on Vision Than on Any Other Sensory Modality? *Frontiers in Psychology*, *10*, 2246. https://doi.org/10.3389/fpsyg.2019.02246

Iglesias, S., Mathys, C., Brodersen, K. H., Kasper, L., Piccirelli, M., den Ouden, H. E. M., & Stephan, K. E. (2013). Hierarchical Prediction Errors in Midbrain and Basal Forebrain during Sensory Learning. *Neuron*, *80*(2), 519–530. https://doi.org/10.1016/j.neuron.2013.09.009

Jardri, R., & Denève, S. (2013). Circular inferences in schizophrenia. *Brain*, *136*(11), 3227–3241. https://doi.org/10.1093/brain/awt257

Jardri, R., Duverne, S., Litvinova, A. S., & Denève, S. (2017). Experimental evidence for circular inference in schizophrenia. *Nature Communications*, *8*(1), 14218. https://doi.org/10.1038/ncomms14218

Jepma, M., Koban, L., Van Doorn, J., Jones, M., & Wager, T. D. (2018). Behavioural and neural evidence for self-reinforcing expectancy effects on pain. *Nature Human Behaviour*, *2*(11), 838–855. https://doi.org/10.1038/s41562-018-0455-8

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

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

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

Kulkarni, K. R., O'Brien, M., & Gu, X. (2023). Longing to act: Bayesian inference as a framework for craving in behavioral addiction. *Addictive Behaviors*, *144*, 107752. https://doi.org/10.1016/j.addbeh.2023.107752

Magnotti, J. F., & Beauchamp, M. S. (2017). A Causal Inference Model Explains Perception of the McGurk Effect and Other Incongruent Audiovisual Speech. *PLOS Computational Biology*, *13*(2), e1005229. https://doi.org/10.1371/journal.pcbi.1005229

Mathys, C., Daunizeau, J., Friston, K., & Stephan, K. (2011). A Bayesian Foundation for Individual Learning Under Uncertainty. *Frontiers in Human Neuroscience*, *5*. https://www.frontiersin.org/articles/10.3389/fnhum.2011.00039

Mcgurk, H., & Macdonald, J. (1976). Hearing lips and seeing voices. *Nature*, *264*(5588), 746–748. https://doi.org/10.1038/264746a0

Mkrtchian, A., Valton, V., & Roiser, J. P. (2023). *Reliability of Decision-Making and Reinforcement Learning Computational Parameters* (No. 1). *7*(1), Article 1. https://doi.org/10.5334/cpsy.86

Naert, A. L. G., Kehlet, H., & Kupers, R. (2008). Characterization of a novel model of tonic heat pain stimulation in healthy volunteers. *Pain*, *138*(1), 163–171. https://doi.org/10.1016/j.pain.2007.11.018

Nosofsky, R. M. (2011). The generalized context model: An exemplar model of classification. In E. M. Pothos & A. J. Wills (Eds.), *Formal Approaches in Categorization* (1st ed., pp. 18–39). Cambridge University Press. https://doi.org/10.1017/CBO9780511921322.002

Onysk, J., Whitefield, M., Gregory, N., Jain, M., Turner, G., Seymour, B., & Mancini, F. (2023). *Statistical learning in acute and chronic pain* [Preprint]. Pain Medicine. https://doi.org/10.1101/2023.03.23.23287656

Palmer, S. E. (1999). *Vision science: Photons to phenomenology*. MIT Press.

Petkova, V. I., & Ehrsson, H. H. (2008). If I Were You: Perceptual Illusion of Body Swapping. *PLoS ONE*, *3*(12), e3832. https://doi.org/10.1371/journal.pone.0003832

Piray, P., & Daw, N. D. (2020). A simple model for learning in volatile environments. *PLOS Computational Biology*, *16*(7), e1007963. https://doi.org/10.1371/journal.pcbi.1007963

Press, C., Yon, D., & Heyes, C. (2022). Building better theories. *Current Biology*, *32*(1), R13–R17. https://doi.org/10.1016/j.cub.2021.11.027

Price, D. D., Finniss, D. G., & Benedetti, F. (2008). A Comprehensive Review of the Placebo Effect: Recent Advances and Current Thought. *Annual Review of Psychology*, *59*(1), 565–590. https://doi.org/10.1146/annurev.psych.59.113006.095941

Rao, R. P. N. (2004). Bayesian Computation in Recurrent Neural Circuits. *Neural Computation*, *16*(1), 1–38. https://doi.org/10.1162/08997660460733976

Rao, R. P. N., & Ballard, D. H. (1999). Predictive coding in the visual cortex: A functional interpretation of some extra-classical receptive-field effects. *Nature Neuroscience*, *2*(1), 79–87. https://doi.org/10.1038/4580

Rauschenberger, R. (2010). Reentrant processing in attentional guidance—Time to abandon old dichotomies. *Acta Psychologica*, *135*(2), 109–111. https://doi.org/10.1016/j.actpsy.2010.04.014

Rauss, K., & Pourtois, G. (2013). What is Bottom-Up and What is Top-Down in Predictive Coding? *Frontiers in Psychology*, *4*. https://doi.org/10.3389/fpsyg.2013.00276

Rehman, I., Mahabadi, N., Sanvictores, T., & Rehman, C. I. (2023). Classical Conditioning. In *StatPearls*. StatPearls Publishing. http://www.ncbi.nlm.nih.gov/books/NBK470326/

Rescorla, R., & Wagner, A. (1972). A theory of Pavlovian conditioning: The effectiveness of reinforcement and non-reinforcement. *Classical Conditioning: Current Research and Theory*.

Roberts, I. D., & Hutcherson, C. A. (2019). Affect and Decision Making: Insights and Predictions from Computational Models. *Trends in Cognitive Sciences*, *23*(7), 602–614. https://doi.org/10.1016/j.tics.2019.04.005

Robison, M. K., Unsworth, N., & Brewer, G. A. (2021). Examining the effects of goal-setting, feedback, and incentives on sustained attention. *Journal of Experimental Psychology. Human Perception and Performance*, *47*(6), 869–891. https://doi.org/10.1037/xhp0000926

Schaaf, J., Weidinger, L., Molleman, L., & van den Bos, W. (2023). *Test-retest reliability of reinforcement learning parameter*. https://doi.org/10.31234/osf.io/chq5a

Simonsen, A., Fusaroli, R., Petersen, M. L., Vermillet, A.-Q., Bliksted, V., Mors, O., Roepstorff, A., & Campbell-Meiklejohn, D. (2021). Taking others into account: Combining directly experienced and indirect information in schizophrenia. *Brain*, *144*(5), 1603–1614. https://doi.org/10.1093/brain/awab065

Taking others into account: Combining directly experienced and indirect information in schizophrenia | Brain | Oxford Academic. (n.d.). Retrieved May 24, 2023, from https://academic.oup.com/brain/article/144/5/1603/6214913?login=false

Theeuwes, J. (2010). Top–down and bottom–up control of visual selection. *Acta Psychologica*, *135*(2), 77–99. https://doi.org/10.1016/j.actpsy.2010.02.006

Vehtari, A., Gelman, A., & Gabry, J. (2017). Practical Bayesian model evaluation using leave-one-out cross-validation and WAIC. *Statistics and Computing*, *27*(5), 1413–1432. https://doi.org/10.1007/s11222-016-9696-4

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

Wilson, R. C., Bonawitz, E., Costa, V. D., & Ebitz, R. B. (2021). Balancing exploration and exploitation with information and randomization. *Current Opinion in Behavioral Sciences*, *38*, 49–56. https://doi.org/10.1016/j.cobeha.2020.10.001

Wilson, R. C., & Collins, A. G. (2019a). Ten simple rules for the computational modeling of behavioral data. *ELife*, *8*, e49547. https://doi.org/10.7554/eLife.49547

Wilson, R. C., & Collins, A. G. (2019b). Ten simple rules for the computational modeling of behavioral data. *ELife*, *8*, e49547. https://doi.org/10.7554/eLife.49547

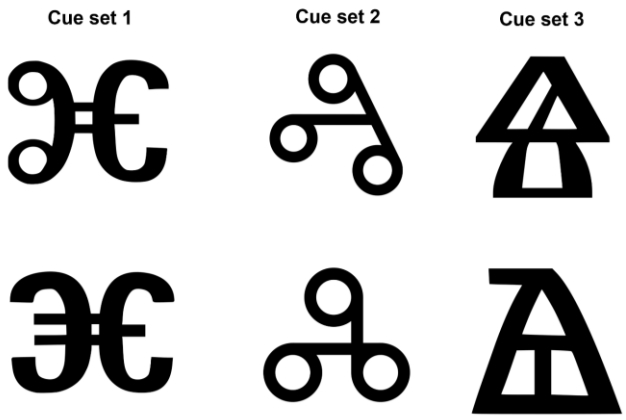
Zopf, R., Savage, G., & Williams, M. A. (2013). The Crossmodal Congruency Task as a Means to Obtain an Objective Behavioral Measure in the Rubber Hand Illusion Paradigm. *Journal of Visualized Experiments*, *77*, 50530. https://doi.org/10.3791/50530

Jesper Fischer Ehmsen (201910213)
Jiaqi Zhao (202202798)

Appendix

A

Cues used in the experimental task.



B

Plate notation and prior values for the four computational models

Here we present the plate notation for the four computational models introduced. Note that hierarchical level parameters are depicted as means, standard deviations, or precisions of the given parameter. Furthermore, priors can be seen in the left most column, these were chosen based on knowledge about the parameters i.e. learning rate is bounded between 0 and 1 as well as investigating what parameter values make reasonable behaviour (see Figure 2 in main text)

Jesper Fischer Ehmsen (201910213)
Jiaqi Zhao (202202798)

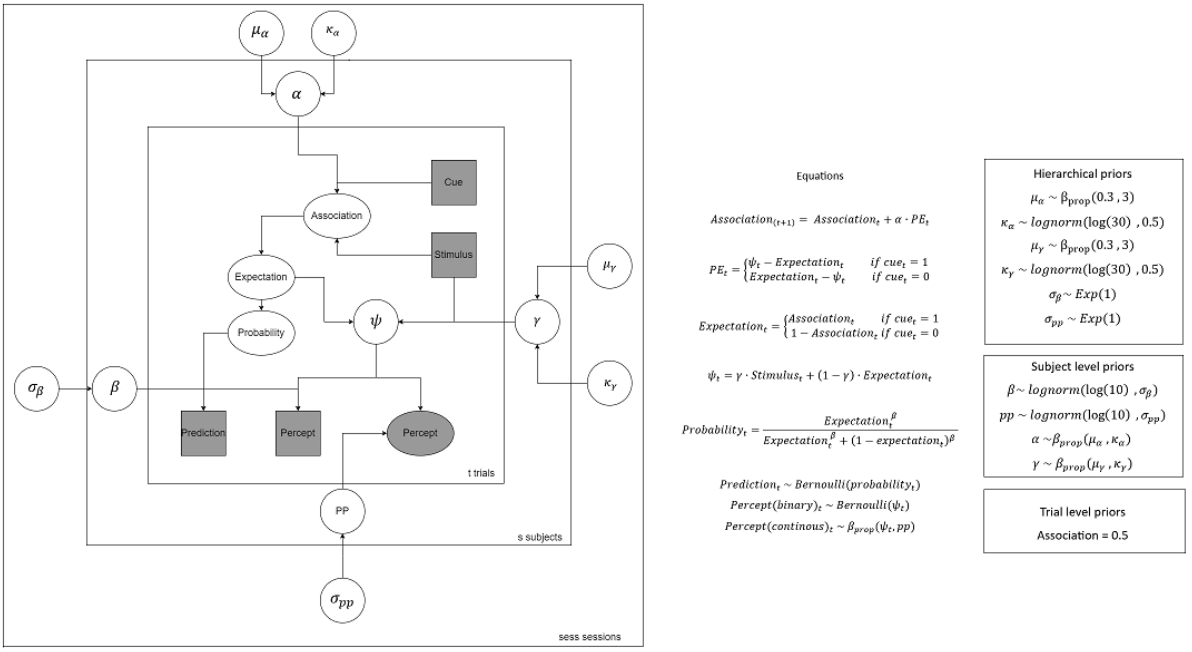


Figure 1: Notation for the hierarchical MRW model

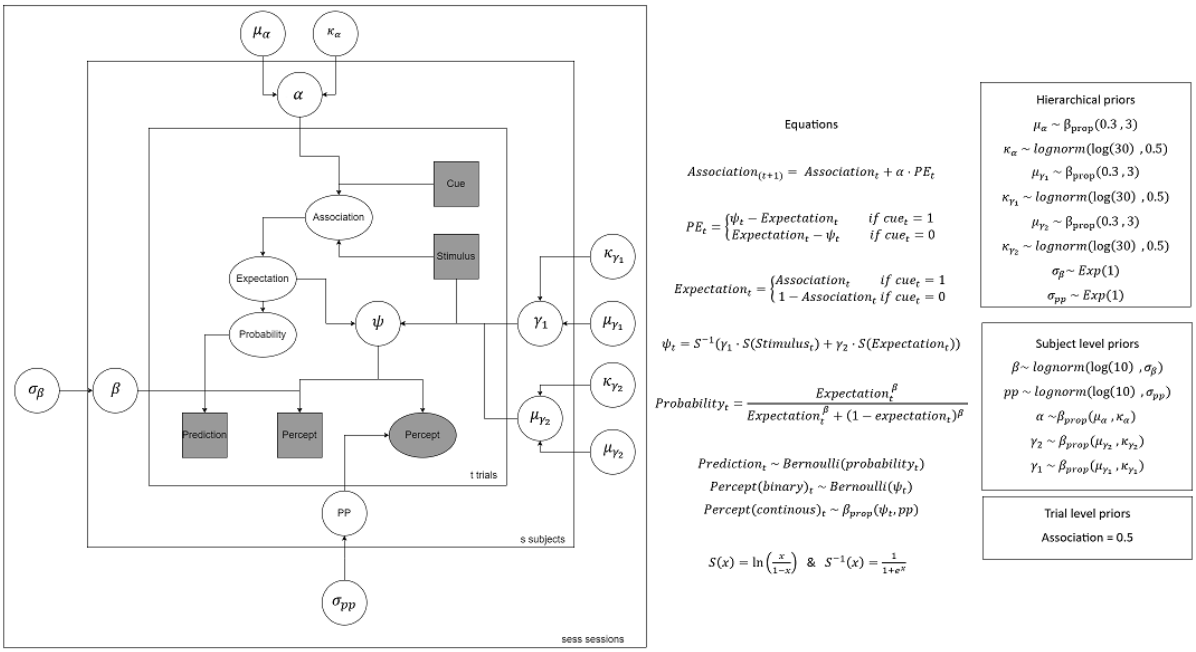


Figure 2: Notation for the hierarchical WB model

Jesper Fischer Ehmsen (201910213)
Jiaqi Zhao (202202798)

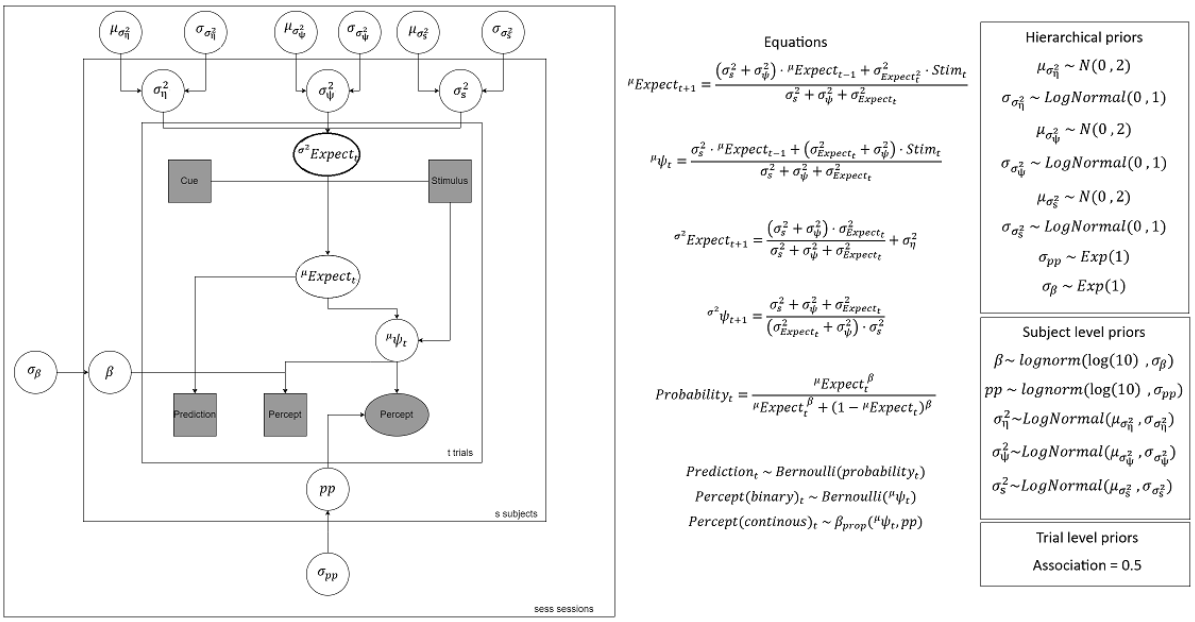


Figure 3: Notation for the hierarchical Kalman filter model

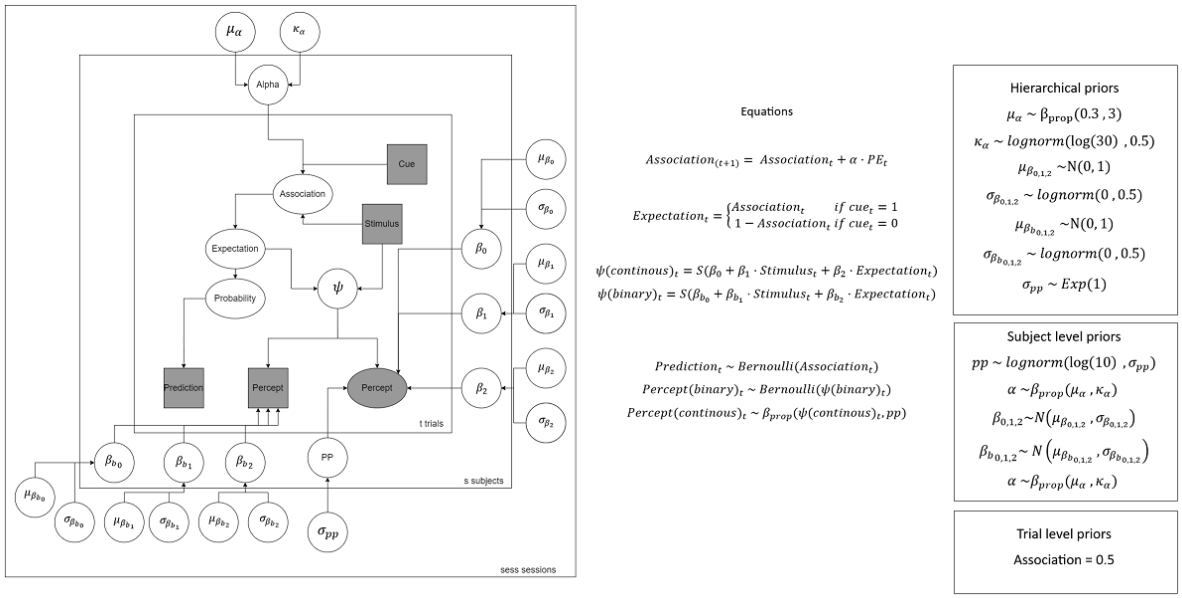


Figure 4: Notation for the hierarchical SRW model

Jesper Fischer Ehmsen (201910213)
Jiaqi Zhao (202202798)

C

List of parameter values for parameter recovery

Modified Rescorla Wagner

$$\mu_\alpha = \{0.1,0.2,0.3,0.4,0.5\}$$

$$\kappa_\alpha = \{10,155,300\}$$

$$\mu_\gamma = \{0.1,0.2,0.3,0.4,0.5\}$$

$$\kappa_\gamma = \{10,155,300\}$$

$$\sigma_{pp} = \{0.1,1.05,2\}$$

$$\sigma_\beta = \{0.1,1.05,2\}$$

Weighted bayes

$$\mu_\alpha = \{0.1,0.2,0.3,0.4,0.5\}$$

$$\kappa_\alpha = \{10,155,300\}$$

$$\mu_{\gamma_1} = \{0.1,0.2,0.3,0.4,0.5\}$$

$$\kappa_{\gamma_1} = \{10,155,300\}$$

$$\mu_{\gamma_2} = \{0.1,0.2,0.3,0.4,0.5\}$$

$$\kappa_{\gamma_2} = \{10,155,300\}$$

$$\sigma_{pp} = \{0.1,1.05,2\}$$

$$\sigma_\beta = \{0.1,1.05,2\}$$

Kalman filter

$$\mu\sigma_\eta = \{0.1\ , 0.6\ , 1\}$$

$$\sigma\sigma_\eta = \{0.1\ , 1\}$$

$$\mu\sigma_\psi = \{0.1\ , 2.6\ , 5\}$$

$$\sigma\sigma_\psi = \{0.1\ , 1\}$$

$$\mu\sigma_s = \{0.1\ , 2.6\ , 5\}$$

$$\sigma\sigma_s = \{0.1\ , 1\}$$

$$\sigma_{pp} = \{0.1\ , 1\}$$

$$\sigma_\beta = \{0.1\ , 1\}$$

Simple Rescorla Wagner

$$\mu_\alpha = \{0.1,0.2,0.3,0.4,0.5\}$$

$$\kappa_\alpha = \{10,100\}$$

$$\mu_{\beta_0} = -1.5\ ;\ \sigma_{\beta_0} = 0.5$$

$$\mu_{\beta_1} = \{0.1,0.57, 1.05, 1.53,2\}$$

$$\sigma_{\beta_1} = \{0.1,1.1,2\}$$

$$\mu_{\beta_2} = \{0.1,0.57, 1.05, 1.53,2\}$$

$$\sigma_{\beta_2} = \{0.1,1.1,2\}$$

$$\mu_{\beta_{b_0}} = -4\ ;\ \sigma_{\beta_{b_0}} = 0.5\ ;\ \mu_{\beta_{b_1}} = 7$$

$$\sigma_{\beta_{b_1}} = 0.5\ ;\ \mu_{\beta_{b_2}} = 0.7\ ;\ \sigma_{\beta_{b_2}} = 0.5\ ;\ \sigma_{pp} = 0.8$$

List of parameter values for model recovery

Modified Rescorla Wagner

$$\mu_\alpha = 0.3\ ;\ \kappa_\alpha = 50\ ;\ \mu_\gamma = 0.3$$

$$\kappa_\gamma = 50;\ \sigma_{pp} = 0.5$$

$$\sigma_\beta = 0.5$$

Weighted Bayes

$$\mu_\alpha = 0.3;\ \kappa_\alpha = 50$$

$$\mu_{\gamma_1} = 0.3;\ \kappa_{\gamma_1} = 50$$

$$\mu_{\gamma_2} = 0.3;\ \kappa_{\gamma_2} = 50$$

$$\sigma_{pp} = 0.5;\ \sigma_\beta = 0.5$$

Simple Rescorla Wagner

$$\mu_\alpha = 0.3 \; ; \kappa_\alpha = 50 \; ; \mu_{\beta_0} = -1.5 \; ; \sigma_{\beta_0} = 0.5$$

$$\mu_{\beta_1} = 2 \; ; \sigma_{\beta_1} = 1; \mu_{\beta_2} = 1; \sigma_{\beta_2} = 1$$

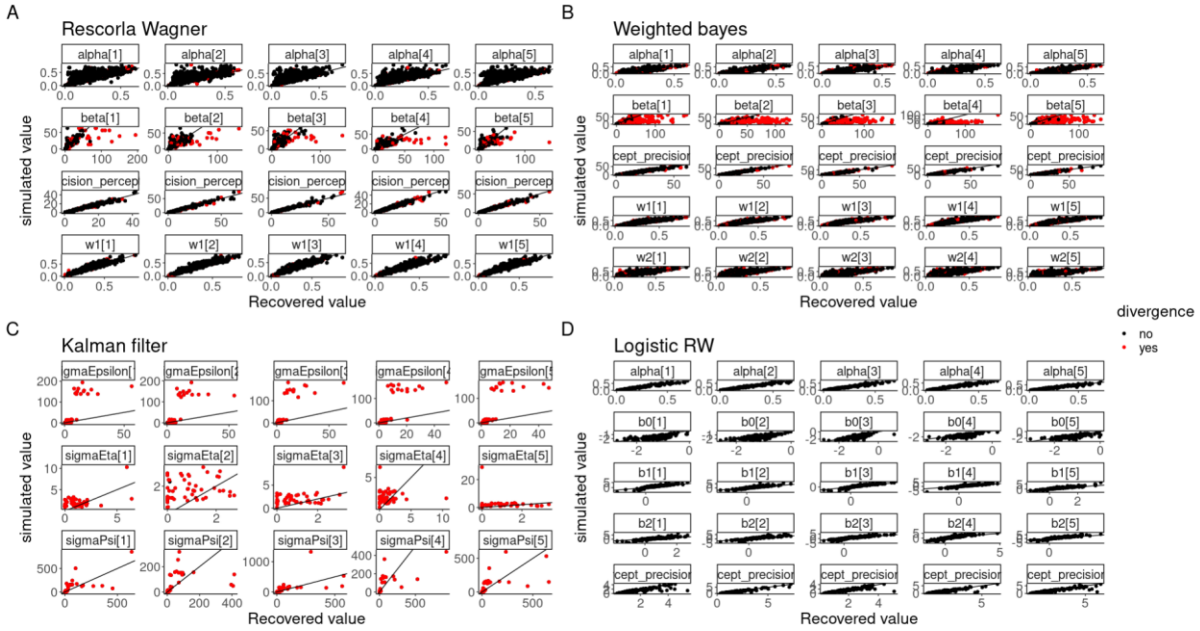
$$\mu_{\beta_{b_0}} = -3; \sigma_{\beta_{b_0}} = 0.5 \; ; \mu_{\beta_{b_1}} = 5; \sigma_{\beta_{b_1}} = 0.5$$

$$\mu_{\beta_{b_2}} = 0.7 \; ; \sigma_{\beta_{b_2}} = 0.5 \; ; \sigma_{pp} = 0.5$$

Jesper Fischer Ehmsen (201910213)
Jiaqi Zhao (202202798)

D

Subject level parameter recovery



E

Showing equation (16) and (17)

Starting of with the mean of the precept

$$\mu_{\psi_t} = \frac{\sigma_s^2}{\sigma_s^2 + \sigma_\psi^2 + \sigma_{E_t}^2} * \mu_{E_{t-1}} + \frac{\sigma_\psi^2 + \sigma_{E_t}^2}{\sigma_s^2 + \sigma_\psi^2 + \sigma_{E_t}^2} * S_t$$

using (14) we need to show the following.

$$(1-\gamma) = \frac{\sigma_\psi^2 + \sigma_{E_t}^2}{\sigma_s^2 + \sigma_\psi^2 + \sigma_{E_t}^2}$$

Now use the fact that if $x = \frac{a}{b}$ then $1 - x = \frac{(b-a)}{b}$

$$(1-\gamma) = \frac{\sigma_s^2 + \sigma_\psi^2 + \sigma_{E_t}^2 - \sigma_s^2}{\sigma_s^2 + \sigma_\psi^2 + \sigma_{E_t}^2}$$

which gives exactly:

$$(1-\gamma) = \frac{\sigma_\psi^2 + \sigma_{E_t}^2}{\sigma_s^2 + \sigma_\psi^2 + \sigma_{E_t}^2}$$

The same logic goes for equation (17)