# Asymmetric Belief Updating in Instrumental Tasks

University College London

Thesis for MSc in Cognitive and Decision Sciences

Candidate Name: Mahbod Mehrvarz

Supervisor: Dr Maarten Speekenbrink

Submission Date: 1/9/2021

# Index

# Abstract

Using computational Reinforcement learning models, numerous studies have reported that the magnitude of human learning is biased towards whether the outcome is good or bad. However, there are contradictory reports toward which outcome has a more significant effect on learning, while some suggest the learning bias may be a statistical artifact caused by the fitted model ignoring choice hysteresis. Additionally, previous studies have mainly investigated learning bias in the two-armed bandit paradigm, using only variants of the Q-learning model with fixed initial expectations, and categorizing good and bad outcomes by the sign of the reward prediction error. The current study will generalize the learning bias to new experimental paradigms, use Bayesian Updating models, allow the initial expectations to vary, and employ a reward magnitude-based categorizing of good and bad outcomes. By applying 12 different Reinforcement learning models to three open datasets, we will show that the reward magnitude-based categorizing works as well as the reward prediction error-based categorizing, and allowing the initial expectations to vary, eliminates the learning bias found in previous studies. Furthermore, we investigate the relationship between learning bias and the exploration-exploitation trade-off.

# Introduction

A ubiquitous characteristic of human (and other animals) behaviour is the tendency to seek actions that lead to positive outcomes and avoid ones that lead to negative outcomes, e.g., Thorndike's "Law of Effect" (Thorndike & Bruce, 2017). While this characteristic is crucial for survival, humans must adaptively decide how to act, i.e., by interacting with and learning from their environment (O'Doherty, Cockburn, & Pauli, 2017). For example, when growing up, we learn not to touch the stove through trial and error (interacting with the environment) as it leads to pain and injury (negative outcomes). This desire to seek positive and avoid negative outcomes influences our behaviour to such a degree that it is used at the societal and cultural level to encourage good and discourage bad behaviours. Prosocial actions are rewarded (e.g., tax credits for charity donations), while antisocial actions are punished (e.g., prison time for a burglary). However, it is unclear if there is a systematic bias (difference) in the way in which we learn from positive and negative outcomes. In other words, is there an asymmetry in the magnitude of human learning with respect to rewards and punishments?

While previous studies have provided strong evidence for dissociable learning mechanisms concerning rewards and punishments at the neurobiological level (e.g. Frank, Moustafa, Haughey, Curran, & Hutchison, 2007; Frank, Seeberger, & O'reilly, 2004; Palminteri & Pessiglione, 2017), the phenomenon has been less well captured at the cognitive level. Particularly, there are contradictory reports to the direction in which this asymmetry points, with some suggesting that we learn more from negative outcomes (e.g. Gershman, 2015; Niv, Edlund, Dayan, & O'Doherty, 2012) and others suggesting the opposite (e.g. Chambon et al., 2020; Lefebvre, Lebreton, Meyniel, Bourgeois-Gironde, & Palminteri, 2017; Palminteri, Lefebvre, Kilford, & Blakemore, 2017). Furthermore, recent studies (e.g. Katahira, 2018; Sugawara & Katahira, 2021) indicate that the learning bias found could be a statistical artifact caused by methodological misspecification.

The contradictory findings and the methodological disagreements derive from the complicated nature of investigating human learning at the cognitive level. Human learning is a latent phenomenon; it cannot be observed or measured directly. For this reason, to investigate biases in human learning, previous researchers have relied on computational modelling of adaptive behaviour. In behavioural sciences, computational modelling allows for a better understanding of the data by using precise mathematical models (Wilson & Collins, 2019). It provides a deeper and nuanced enquiry into the underlying algorithms that contribute to the emergence of a particular behaviour (Farrell & Lewandowsky, 2018). In the case of learning asymmetry, by modelling participants' behaviour in an adaptive task, we can disentangle how feedback contributes to learning the value of taking an action for a particular stimulus, and assess whether there is a bias towards the received feedback. Furthermore, modelling allows for precise predictions, which can falsify our theoretical assumptions or hypotheses (e.g., there is a learning asymmetry), and we can compare different proposed underlying mechanisms of a specific behaviour through model comparison (e.g., comparing models with and without learning asymmetry) (Farrell & Lewandowsky, 2018).

For adaptive decision making, computational Reinforcement Learning (RL) provides a theoretical framework to understand how agents (artificial or natural) can learn from their environment to make decisions that maximise the potential of obtaining reward and avoiding punishment (Gureckis & Love, 2015). While computational RL is an interdisciplinary subject, in cognitive science, computational RL models have been primarily employed to understand latent learning-related phenomena at the behavioural level (e.g. Master et al., 2020). RL models are utilised as cognitive models that implement explicit hypotheses about the underlying cognitive processes that lead to the emergence of a particular behaviour. Through the quantification of verbal assumptions, RL models can approximate human behaviour, and if they fit the behavioural data well, it can be inferred that perhaps the participant used the supposed RL models' algorithm cognitively (Eckstein, Wilbrecht, & Collins, 2021). As such, RL models are not simply a descriptive account of adaptive decision-making; instead, they provide an overall insight into the function and objective, referred to as rational analysis (Gureckis & Love, 2015; Marr, 1982). In the next section, we will provide a review of learning asymmetry and its investigation within the computational RL framework.

## Background

One of the most significant findings in human and animal RL has been the role of dopamine in learning (Niv, 2009). While performing a conditioning experiment, Schultz, Tremblay, and Hollerman (1998) found that the firing rate of dopamine neurons increased in the midbrain of monkeys after receiving a **rewarding stimulus (RS)**. However, this firing rate disappeared when a **conditioned stimulus (CS)** preceded the RS. Instead, the dopamine neurons' firing rates were observable when the CS was presented. Furthermore, when the

CS was not followed by the RS, shortly after the omission of the expected RS, the activity of the dopamine neurons fell below the baseline (Sutton & Barto, 1990) Ultimately this pattern was interpreted using temporal-difference based prediction error, which manifests in reinforcement learning as **Reward Prediction Error (RPE)**, the difference between the predicted and received reward (Schultz, 2016). Based on this finding, Computational RL models of human decision-making assume that the values of States, Actions, and Policies are learned and updated using the RPE; similar to a delta-rule in supervised learning (Rescorla, 1972; Sutton & Barto, 2018).

Previous enquiries into human asymmetrical learning used positive and negative RPEs to distinguish between good and bad outcomes to analyze the magnitude in which values are updated with respect to the sign of the estimated RPE. The rationale for this investigation was built on the robust foundation that there are dissociable neural systems for learning from good and bad outcomes (Palminteri, 2021). Frank et al.(2004) demonstrated that patients with Parkinson's disease, who have lower dopamine count in the basal ganglia, display abnormal behaviour in instrumental tasks–which need learning from trial and error. Furthermore, they found that patients who were off medication were better at learning to avoid decisions that led to negative outcomes, while the patients on dopamine medication displayed the opposite bias. Additional genetic analysis research showed that independent dopaminergic mechanisms predicted learning from rewards and punishments in humans (Frank et al., 2007). Notably, a polymorphism in the DARPP-32 gene predicted learning from positive outcomes, while the C957T polymorphism of the DRD2 gene predicted learning to avoid negative outcomes. Subsequent studies further strengthened the finding of dissociable neural systems for learning from positive and negative outcomes (Collins & Frank, 2012; Dabney et al., 2020; Palminteri & Pessiglione, 2017).

Building on this dissociation, several studies began to investigate learning biases in instrumental tasks, using the two-armed bandit paradigm (see Methods section for more details on this paradigm), at the cognitive level. Learning bias was represented in the RL models as differential values for the learning rates parameter (See Equation 4), where the parameter holds different values depending on the sign of the RPE. However, there was no clear direction in which this asymmetry may be greater–i.e., do humans learn more from positive or negative outcomes. Early studies found that the learning rate for negative RPEs to be higher than positive RPEs (Gershman, 2015; Niv et al., 2012). They concluded that this finding could reflect risk aversion: a general bias toward avoiding uncertain options (Rabin & Thaler, 2001). That is, participants learned more from negative outcomes because they had the desire to avoid the more risky options.

Meanwhile, motivated by the various findings that human tend to overestimate the likelihood of desired events, such as the Optimism Bias (Sharot, Korn, & Dolan, 2011) and the Good News–Bad News effect (Eil & Rao, 2011), Palminteri et al. investigated over the course of eight experiments whether the bias towards positive outcome is also apparent in the low-level reinforcement process (Palminteri, 2021). For their investigation, they used the two-armed bandit paradigm (like previous research) with various probability distributions for reward, as well as a factual and a counterfactual paradigm. However, unlike previous studies in learning bias, they found the learning rate parameter's value to be higher for positive RPEs than for negative RPEs (Chambon et al., 2020; Lefebvre et al., 2017; Palminteri et al., 2017). They interpreted their results in terms of positivity bias (for the factual experiments) and confirmation bias (for the counterfactual experiments). Positivity bias is the tendency to give more weight to positive outcomes, and confirmation bias is the tendency to give more weight to outcomes consistent with an already held belief.

A behavioural characteristic of the positivity and confirmation bias is the tendency to repeat previous choices. Therefore, under the computational RL framework for asymmetrical models, the estimated value for the positive learning parameter could be higher when participants choose a favoured option more intensely, despite learning unbiased estimates of action values. Choice repetition is also a behavioural characteristic of choice hysteresis—response repetition biases (Bonaiuto, de Berker, & Bestmann, 2016). The prominent neurophysiological and psychological account for choice hysteresis is a form of habit learning, where agents learn a state-action relation without value computation (Miller, Shenhav, & Ludvig, 2019). The key behavioural manifestation of habit learning is the insensitivity to the outcome of actions. An agent with choice hysteresis (hereafter referred to as autocorrelation) will repeat the same action despite negative outcomes, displaying the same tendency as an agent with positivity bias.

Due to the overlap in behavioural characteristics, it is difficult to disentangle asymmetrical learning from autocorrelation and identify the actual underlying cognitive process. Through simulation analysis, Katahira (2018) demonstrated that fitting an asymmetrical RL model could lead to a pseudo-positivity bias when there is genuine choice autocorrelation, and the absence of asymmetrical learning can lead to a pseudo-choice hysteresis. To remedy this, Katahira proposed a hybrid model that includes both choice-autocorrelation and asymmetrical learning rate (see Equations 9 10). Consequently, Sugawara and Katahira (2021) failed to replicate the positivity bias when fitting the hybrid model and concluded that previous findings of learning asymmetry were due to statistical bias resulting from model misspecification. It is important to note that Palminteri and colleagues did control for choice autocorrelation by including a stickiness parameter. However, the way in which Sugawara and Katahira controlled for autocorrelation in their model is more sophisticated.

The stickiness parameter incorporates choice autocorrelation in a way that it only exerts influence from the most recent trial (e.g. Gillan, Kosinski, Whelan, Phelps, & Daw, 2016). On the other hand, Katahira incorporates choice autocorrelation by adopting a model of gradual choice perseverance from studies in perceptual decision making. This model of choice autocorrelation tracks for each option a choice trace which either increments or decrements at each decision point depending on whether an option is picked. As such, it exerts influence beyond just the previous decision point. The gradual choice perseverance model can be modified to accommodate the stickiness version of choice autocorrelation by fixing the decay rate (the rate at which choice trace increments or decrements) to 1, known as the impulsive choice perseverance model. While Palminteri (2021) criticizes gradual choice perseverance on the basis of parsimony (as the model keeps a separate memory to compute choice trace) and neurophysiological and psychological plausibility, we believe this model provides a better representation of choice hysteresis than the impulsive variant.

First, Katahira (2018) demonstrated that the gradual version provides a better fit of participants' behavioural data. Second, if we assume that choice hysteresis reflects a form of habit learning (as already mentioned), we believe that the gradual version provides a better theoretical account. Consider the following example: if an agent picks an option consecutively for a period of eight trials, and at the ninth trial picks another option, then at the tenth trial, the impulsive choice perseverance model ignores the behaviour of the agent for the first eight trials and exerts maximum influence from the option picked only at the latest trial while the gradual choice perseverance model, depending on the value of the decay rate, keeps track of the preceding trials and exerts influence for both options more reasonably (see Figure 1). We believe it is more likely that the choice history of an option that was picked constantly and consecutively exerts more influence than the choice history of an option that was only picked once. Furthermore, a relatively recent model of habit formation (Miller et al., 2019) resembles the gradual choice perseverance model.

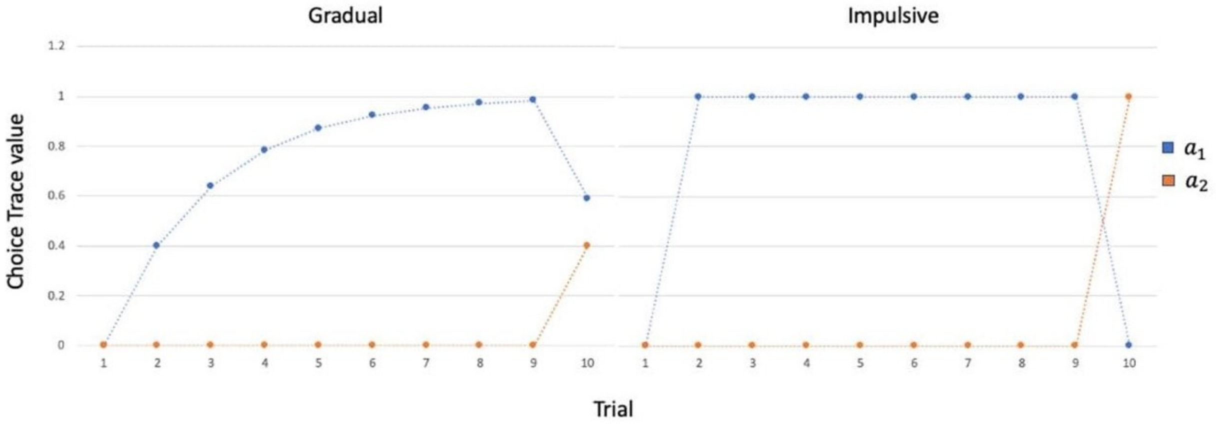
## Study Rationale

In the current study, we aim to address gaps in the emerging research on learning bias. We have identified several such gaps in knowledge, which we will express in this section.

First, previous studies only used model-free Q-learning models. While this class of models capture adaptive decision making adequately, we believe examining the learning bias in more sophisticated models can provide more insight. This, to some extent, was demonstrated by

**Figure 1**

*Difference in choice trace influence between gradual (left) and impulsive (right) choice perseverance models*



Katahira's work mentioned in the previous section. The more sophisticated model incorporating gradual choice perseverance revealed critical insight both at the conceptual and methodological levels. Therefore, in our study, we will extend our model range to include Model-based Bayesian Updating models (see Methods for more details). Our reason for including this class of models is directly linked to the theoretical question proposed by Harada (2020): is learning bias related to the exploration-exploitation trade-off?

Performing well in instrumental tasks requires a fine balance between exploration and exploitation (cf. Sutton & Barto, 2018). The agent must explore different alternatives to learn the values across the action space, and once enough evidence is gathered, the agent can exploit the most rewarding option. As such, agents learn more during the exploration phase than the exploitation phase. The Q-learning models have an unvarying learning parameter which is agnostic towards this trade-off. Therefore, it could very well be that the learning bias in these models results from the model's inability to capture the trade-off. On the other hand, the Bayesian Updating models we employed have varying learning rates that adjust to the magnitude of uncertainty about a given option (Speekenbrink & Konstantinidis, 2015). As such, this model captures the trade-off, and if we find a learning bias in the asymmetrical variants of this model, we can answer the question proposed by Harada.

However, just employing this model is not enough. Learning bias has been mainly investigated in the context of the two-armed bandit with fixed binary reward probabilities. Due to its nature (only having two arms and a fixed probability distribution for rewards), this environment limits the degree of need for exploration and, therefore, learning. A reasonable estimate of action values for the two options can be learned relatively quickly and maintained throughout the experiment. This drastically limits the Bayesian Updating models' ability to capture participants behaviour. Additionally, it would be beneficial to examine learning bias in more dynamic environments with more options. As such, we will generalize our investigation to two other instrumental task paradigms.

Moreover, previous studies have only used an RPE-based dissociation for models with asymmetrical learning. We have not found any evidence that the RPE-based dissociation is the exact manner in which humans differentiate learning from positive and negative outcomes. Therefore, we explore an alternative by including a reward-based dissociation for the asymmetrical learning variants (see Equation 7). Finally, previous studies have fixed the initial expectations of the value of each action to zero. If in fact initial expectations differ from zero, a pseudo learning bias might emerge. We could not find any empirical evidence to justify fixing these values to zero. As such, we will investigate learning bias while controlling for initial expectations and compare models with and without fixed initial values.

## Current Study

The present study aims to investigate learning bias in instrumental tasks by fitting 12 different RL models across three different paradigms. To achieve this aim, our analysis will consist of two steps. We will first examine if there is a systematic difference between the two learning rates for the asymmetrical models across participants. Second, we will assess asymmetrical models' plausibility in terms of parsimony by conducting model comparison. The two steps will provide a standard to which learning bias can be falsified. More specifically, the first step provides evidence of systematic bias at a group level, while the second step speaks to the likelihood of the asymmetrical models representing the underlying cognitive process for participants.

We have split this study into three parts. In the first part, we assess learning bias according to the two steps described above. We will look to replicated previous findings while including additional models; reward-based asymmetrical variants and Updating Bayesian models. Additionally, we will generalize our findings to two new experimental paradigms. Furthermore, following previous studies procedures, we will fix the initial estimates for action values for each option to zero (hereafter referred to as models with fixed initial expectations).

In the second part, we will conduct the same analysis as the first part, but we allow models' initial estimates for action values to vary (hereafter referred to as models with unfixed initial expectations). In the third part, we will directly investigate the effect of initial estimates for action values on model performance by comparing models with fixed and unfixed initial expectations. Finally, we will assess the individual model's features (e.g., autocorrelation) and characteristics (e.g., learning method) in more detail. Particularly looking at their interaction and main effect on model performance. Our expectations for our investigation are as follows:

(a) We expect the asymmetrical models to demonstrate a learning bias across participants and fit better than their symmetrical counterparts in model comparison if there is a true learning bias.

(b) If this learning bias is a statistical artifact of not including autocorrelation, we expect that there will be no learning bias across participants for asymmetrical models with autocorrelation. Additionally, we expect the symmetrical models with autocorrelation to fit better than their asymmetrical variants.

(c) If learning bias is an artifact of not properly addressing exploration and exploitation, we expect the Bayesian Updating models to not demonstrate a learning bias across participants or fit better than their symmetrical counterparts. Furthermore, we expect not to find a learning bias across participants in the second and third datasets, and we expect the symmetrical models to perform better than asymmetrical models. On the other hand, if there is an association between learning bias and exploration and exploitation, we expect the opposite.

(d) We expect the learning bias to be a statistical artifact of not allowing initial expectations to vary if there is no learning bias across participants for the asymmetrical models with unfixed initial expectations. Additionally, we expect the models with unfixed initial expectations to perform better than those with fixed initial expectations.

# Method

## Datasets

**Experimental Paradigm:** We conducted our analysis using open data sets collected by previous studies. These studies employed different variants of the **Multi-Armed Bandit (MAB)** task. In this experimental paradigm, a number of choices (referred to as arms) are presented to the participant. On each trial, the participants must select an arm, which then pays out a randomly drawn reward from a probability distribution specific to each arm. The goal is for the participants to maximise the total accumulated reward (cf. Lee, Zhang, Munro, & Steyvers, 2011). MAB tasks reveal many aspects of human instrumental learning and allow for the analysis of the underlying mechanism that dictates participants' behaviour by computational modelling, mainly using RL models (Gureckis & Love, 2015).

**Dataset 1 (Sugawara & Katahira, 2021):** This data set is comprised of the open data reported by Sugawara and Katahira (2021). They examined asymmetrical learning and autocorrelation in factual and counterfactual learning contexts. Our analysis of this data set was limited only to the factual learning context (N = 143). The particular variant of the MAB task they used is the two-armed bandit task with fixed binary-reward distributions. The participants were exposed to four pairs of arms (for 48 trials each), each associated with different fixed probabilities of obtaining either 10 or -10 points, for 192 trials in total (see Figure 2.A for an example of the reward distribution across trials). One further distinction between their experimental procedure and other data sets is that there was a time constraint of 1500 ms for the participants to make a decision (cf. Sugawara & Katahira, 2021).

**Dataset 2 (Steingroever, Fridberg, et al., 2015):** This data set is comprised of open data collected from 10 independent studies (Fridberg et al., 2010; Horstmann, Villringer, & Neumann, 2012; Kjome et al., 2010; Maia & McClelland, 2004; Premkumar et al., 2008; Steingroever, Šmíra, Lee, & Pachur, 2015; Wetzels, Vandekerckhove, Tuerlinckx, & Wagenmakers, 2010; Worthy, Pang, & Byrne, 2013; Wrosch, Bauer, & Scheier, 2005; **?**). These studies examined the performance of healthy participants in the Iowa Gambling Task (IGT). The IGT, introduced by Bechara, Damasio, Damasio, and Anderson (1994), is a widely-used variant of an MAB, containing four arms (referred to as decks). While across these studies, three distinct reward schemes were used, the underlying nature of the task was the same. Each deck contains a finite number of cards with a numerical reward and punishment associated with it, ranging from 0 to 1100. From the four decks, two of the decks are advantageous, while the other two are disadvantageous (see Figure 2.B for an example of the reward distribution across decks). Participants start with 2000 points and attempt to maximise their rewards (cf. Steingroever, Fridberg, et al., 2015). Our analysis was contained to the experiments that had 100 trials (N = 504).

**Dataset (Speekenbrink & Konstantinidis, 2015):** This data set is comprised of the open data reported by Speekenbrink and Konstantinidis (2015). They examined the role of uncertainty in exploration using a restless variant of the MAB task. In the restless MAB task, reward for each arm is determined according to a Gaussian process:

$$R_t(a_j) = \mu_t(a_j) + \epsilon_t(a_j) \qquad \epsilon_t(a_j) \sim N(0, \sigma_\epsilon) \tag{1}$$

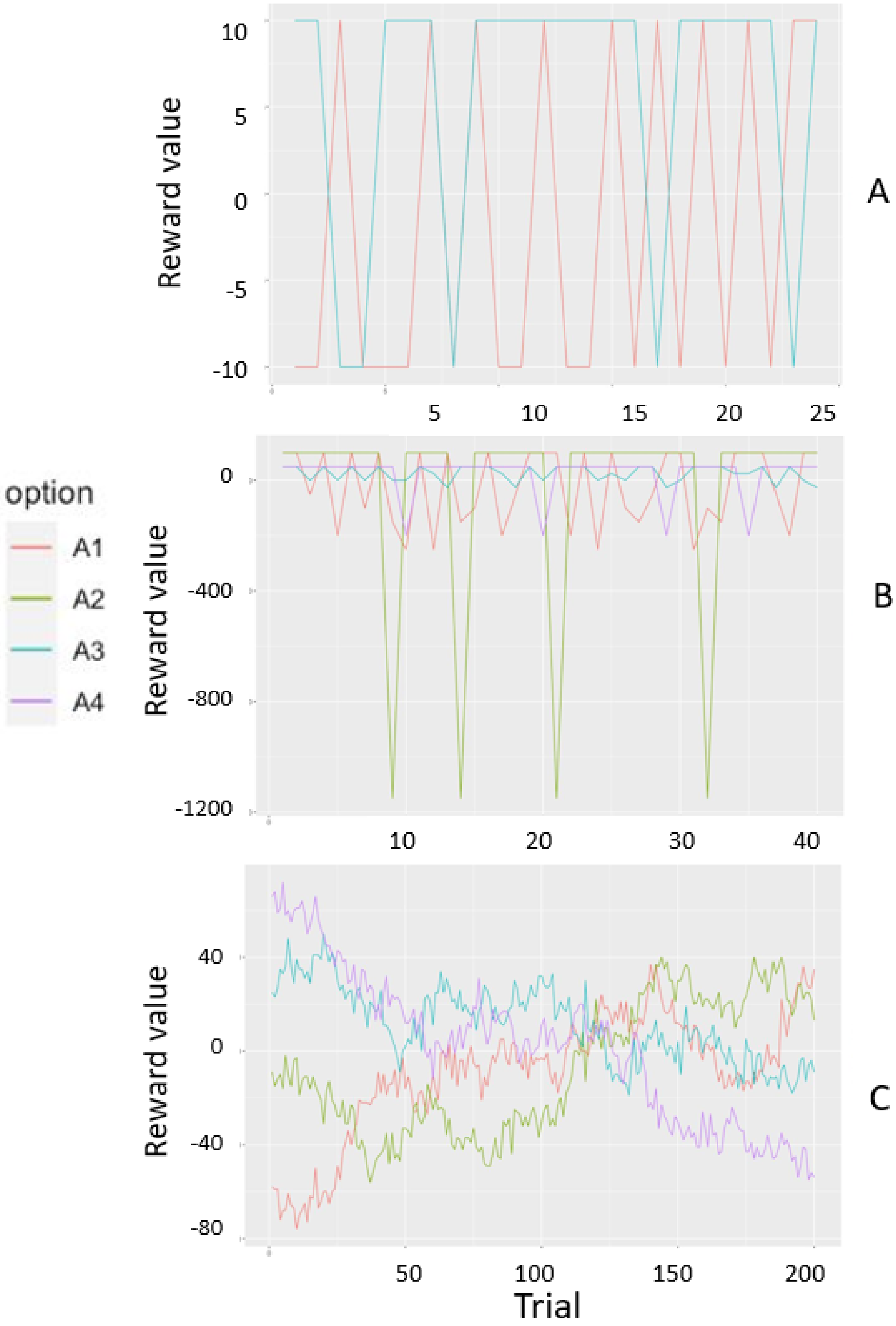
Where the reward, $R_t(\cdot)$, of a particular arm, $a_j$, at trial, $t$ is determined by the sum of mean reward ($\mu_t$) and an error term ($\epsilon_t$). The mean reward varies over trials and is determined according to a random walk:

$$\mu_t(a_j) = \mu_{t-1}(a_j) + \xi_t(a_j) \qquad \xi_t(a_j) \sim N(0, \sigma_\xi) \tag{2}$$

Where the mean reward at trial $t$ is calculated as the sum of the mean reward at the previous trial ($t-1$) and an innovation term ($\xi_t$). While this study has four conditions, our analysis was contained to a single condition, that is the “no-trend stable” condition ($N = 20$). Speekenbrink and Konstantinidis (2015) simulated the experimental environment for this condition by fixing the error and innovation variance to 16 (see Figure 2.c for an example of the reward distribution across trials).

**Figure 2**

*Example rewards for Dataset 1 (A), Dataset 2 (B), and Dataset 3 (C)*



### Models

We fit several models to assess learning asymmetry. The model space includes the Simple Q-learning model (also referred to as the Rescorla-Wagner model), the Bayesian updating Model, and modified versions of both.

The Q-learning model is a model-free RL algorithm that learns the action value $V$ of the chosen arm (option) $a$ at each trial according to the following equation:

$$V_{t+1}(a_j) = V_t(a_j) + \alpha\delta_t(a_j) \tag{3}$$

The action value of arm $a_j$ at trial $t$ is denoted by $V_t(a_j)$ and the learning rate is denoted by $\alpha$, with $\alpha \in [0, 1]$. The RPE is denoted by $\delta$ , which is calculated as:

$$\delta_t(a_j) = R_t(a_j) - V_t(a_j) \tag{4}$$

i.e. the difference between the outcome of the chosen arm $a_j$ at trial $t$, denoted as $R_t(a_j)$, and the estimated value action. This model is referred to as the **Symmetrical Q-learning (SQ)** model. The asymmetrical version of this model allows for differential learning rate:

$$V_{t+1}(a_j) = V_t(a_j) + \alpha^{\pm}\delta_t(a_j) \tag{5}$$

Where $\alpha^{\pm}$ holds different values depending on the sign of the RPE:

$$\begin{aligned} &\alpha^{+} \qquad \text{if } \delta_t(a_j) \geq 0 \\ &\alpha^{-} \qquad \text{if } \delta_t(a_j) < 0 \end{aligned} \tag{6}$$

This model hereafter is referred to as the **Asymmetrical prediction Q-learning (ApQ)** model. We also modified this model to allow for differential learning rates depending on the sign of the obtained reward rather than the RPE:

$$\begin{aligned} &\alpha^{+} \qquad \text{if } R_t(a_j) \geq 0 \\ &\alpha^{-} \qquad \text{if } R_t(a_j) < 0 \end{aligned} \tag{7}$$

This variant is referred to as the **Asymmetrical reward Q-learning (ArQ)** model. Given each action value, the probability of selecting each arm is estimated using the softmax rule:

$$P(C_t(a_j) = 1) = \frac{\exp(\gamma V_t(a_j))}{\sum_{k=1}^{K} \exp(\gamma V_t(a_k))} \tag{8}$$

Where $C_t(a_j)$ is an indicator function which takes the value 1 when option $a_j$ was chosen on trial $t$, and 0 otherwise. $K$ is the number of arms in the option space, and $\gamma$ is the inverse temperature parameter that determines the sensitivity of the choice probability to difference between action values, with $\gamma \in [0, \infty]$.

We also modified the three mentioned models to incorporate autocorrelation. We also modified the three mentioned models to incorporate autocorrelation. The action value was estimated in the same manner as the previous model. However, the softmax choice rule incorporated a choice trace function $CT(\cdot)$, based on the gradual choice perseverance model, which is calculated as:

$$P(C_t(a_j) = 1) = \frac{\exp(\gamma(V_t(a_j) + \varphi CT_t(a_j)))}{\sum_{k=1}^{K} \exp(\gamma(V_t(a_k) + \varphi CT_t(a_k)))} \tag{9}$$

Where $\varphi$ is the perseverance parameter that determines the tendency to repeat or avoid the recently chosen option, with $\varphi \in [-\infty, \infty]$. The choice trace function is calculated in the following way:

$$CT_t(A_j) = CT_{t-1}(a_j) - \tau CT_{t-1}(a_j) + \tau C_t(a_j) \tag{10}$$

The decay rate, denoted by $\tau$, determines to what extent preceding choices influence the current choice, with $\tau \in [0, 1]$. Recall that the indicator function $C_t(a_j)$ takes either the value of 1 if $a_j$ is chosen at trial $t$, and the value of 0 if not. Most RL models incorporate choice autocorrelation in a way that it only exerts influence of the most recent trial (e.g., 31). The gradual choice perseverance model can be modified to accommodate this version of choice autocorrelation by fixing the decay rate to $\tau = 1$, known as the impulsive Choice perseverance model. However, for the reasons detailed earlier, for this study, we adopted the gradual choice perseverance for autocorrelation. We also modified this model, in the same manner as the SQ model, to incorporate asymmetrical learning. We denote the prediction-based iteration as the **Asymmetrical prediction Q-learning with Autocorrelation (ApQwA)** model and the reward-based as the **Asymmetrical reward Q-learning with Autocorrelation (ArQwA)** model.

The remaining six models are Bayesian Updating models, a model-based learning strategy which assumes that each arms' reward is determined according to a Gaussian process (Eq. 1). The action value action for each arm is estimated using the Kalman filter (Kalman, 1960). The Kalman filter in the Bayesian Updating model can be thought of as a variant of the delta-rule with a time-varying learning rate (Speekenbrink, 2019a). This learning rate is adjusted based on the uncertainty of the agent and the perceived variability of the environment. The value action for each arm is calculated by the following equation:

$$V_{t+1}(a_j) = V_t(a_j) + C_t(a_j)K_t(a_j)\delta_t(a_j) \tag{11}$$

Where again $C(\cdot)$ is the indicator function and $\delta_t(a_j)$ is the RPE for arm $a_j$. The Kalman gain term $K_t(a_j)$ functions similarly to the learning rate in Q-learning models but is computed as:

$$K_t(a_j) = \frac{S_t(a_j) + \sigma_\xi^2}{S_t(a_j) + \sigma_\xi^2 + \sigma_\epsilon^2} \tag{12}$$

Where $\sigma_\epsilon^2$ is the error variance and $\sigma_\xi^2$ the innovation variance. The variance of the posterior distribution of each arm, denoted by $S_t(a_j)$, informs the agent about how much uncertainty is associated with each arm (where the magnitude of the variance reflects the magnitude the uncertainty), computed as:

$$S_t(a_j) = [1 - C_t(a_j)K_t(a_j)][S_{t-1}(a_j) + \sigma_\xi^2] \tag{13}$$

Like the Q-learning model, the probability of selecting each arm is estimated using the softmax rule. However, since the softmax rule is less sensitive towards uncertainty, we incorporated an exploration bonus. This way the softmax rule is similar to a popular RL decision rule, **Upper Confidence Bound** rule (Speekenbrink & Konstantinidis, 2015), but with more randomness. The softmax probability of each arm for the Updating Bayesian model is calculated by the following equation:

$$P(C_t(a_j) = 1)\frac{\exp(\gamma_t(V_{t+1}(a_j) + \beta_t(a_j)))}{\sum_{k=1}^{n} \exp(\gamma_t(V_{t+1}(a_k) + \beta_t(a_j)))} \tag{14}$$

Where $\beta_t(a_j)$ is the exploration bonus, which promotes exploration for the arms with high uncertainty, computed as:

$$\beta_t(a_j) = \beta_0\sqrt{S_t(a_j) + \sigma_\xi^2} \tag{15}$$

Where $\beta_0$ reflects the width of the confidence interval for the upper confidence bound, with $\beta_0 \in [0, \infty]$. The described Bayesian updating model is henceforward referred to as the **Symmetrical Kalman Filter (SKF)** model. We modified this model to allow for differential learning. The way in which we achieved this was through letting the error variance hold different values:

$$\begin{aligned} \sigma_\epsilon^2 &= \eta^+ \qquad \text{if } \delta_t(a_j) \geq 0 \\ \sigma_\epsilon^2 &= \eta^- \qquad \text{if } \delta_t(a_j) < 0 \end{aligned} \tag{16}$$

In the SKF model, the amount of change in the agent's belief about the value action of each arm is dependent on the prior belief in comparison to the noise perceived about how the rewards are determined, represented as $\sigma_\epsilon^2$. Therefore, the degree of noise determines the rate at which the received reward influences the posterior estimation of the value action for each arm. This property allows us to differentiate learning rates for the Bayesian Updating model, referred to hereafter as the **Asymmetrical reward Kalman Filter (ArKF) model**. We modified this model to again allow for differential learning rates based on received reward, referred to as the **Asymmetrical reward Kalman Filter (ArKF) model**, where:

$$\begin{aligned} \sigma_\epsilon^2 &= \eta^+ \qquad \text{if } R_t(a_j) \geq 0 \\ \sigma_\epsilon^2 &= \eta^- \qquad \text{if } R_t(a_j) < 0 \end{aligned} \tag{17}$$

Furthermore, we modified the three models to incorporate autocorrelation, using the gradual choice perseverance model, in the same way we did for the Q-learning model. From this point on, we refer to the three models as the **Symmetrical Kalman Filter with Autocorrelation (SKFwA)**, the **Asymmetrical prediction Kalman Filter with Autocorrelation (ABwA)**, and the **Asymmetrical reward Kalman Filter with Autocorrelation (ArKFwA)**, respectively to way which three Bayesian updating models were introduced.

### Parameter Estimation

We estimated the values for the free parameters using **Maximum Likelihood Estimation (MLE)**. The MLE searches for parameters, denoted as , that optimizes the models' probability assignment for each decision point (Speekenbrink, 2019b). In the case of this study, the models assign a probability to participants; choice $C_t$ at specific trial based on previous experience, that is, previous choices $C_{1:(t-1)}$ and the associated rewards $R_{1:(t-1)}$, formally expressed as:

$$P(C_t|C_{1:(t-1)}, R_{1:(t-1)}, \theta) \tag{18}$$

Using the likelihood function $L(\cdot)$ for the set of parameters for each model, we estimate the joint probability of all choices, formally expressed as:

$$L(\theta|C_{1:t}, R_{1:t}) = \prod_{t=1}^{T} P(C_t|C_{0:(t-1)}, R_{0:(t-1)}, \theta) \tag{19}$$

where $C_0$ and $R_0$ are empty "null" values. The maximum likelihood estimate for the set of parameters is the set of parameter values ($\theta$) the maximise the likelihood function, formally expressed as:

$$\hat{\theta} = \arg\max_{\theta} L(\theta|C_{1:t}, R_{1:t}) \tag{20}$$

We computed the likelihood estimate using the "DEoptim" package in R (Katharine M. Mullen, 2011). The "DEoptim" function in this package conducts a stochastic search within specified range to find the global minimum, i.e., $\arg\max_\theta$.

### Model Comparison

The fitted models were evaluated using **Akaike's Information Criterion (AIC)** and the **Bayesian Information Criterion (BIC)**,

$$\begin{aligned} AIC_i &= -2\log L_i + 2g_i \\ BIC_i &= -2\log L_i + g_i \log(n_i) \end{aligned} \tag{21}$$

Where $L_i$ is the likelihood value determined by the likelihood function (Eq. 19), $g$ is the number of estimated parameter and $n$ is the number observation for model $i$. While both criteria compute score based on descriptive accuracy and parsimony, they have a different underlying assumption. The BIC assumes true model is in the set of models being compared, while the AIC assumes the true model is not in the model set (Wagenmakers & Farrell, 2004). Since we cannot be certain of BIC's assumption, we rely on the AIC. However we also rely on the BIC since it penalizes models more harshly, which is suitable for our purposes of assessing asymmetrical learning models with a strict standard of parsimony. Therefore, the AIC will be used as a measure of predicative validity, while the BIC will be used as a measure of true model likelihood. Both scores are also transformed into weighted AIC and weighted BIC:

$$\begin{aligned} \Delta_i(AIC) &= AIC_i - \min AIC \\ w_i(AIC) &= \frac{\exp(-\frac{1}{2}\Delta_i(AIC))}{\sum_{k=1}^{K}(\exp(-\frac{1}{2}\Delta_k(AIC))} \end{aligned} \tag{22}$$

$$\begin{aligned} \Delta_i(BIC) &= BIC_i - \min BIC \\ w_i(BIC) &= \frac{\exp(-\frac{1}{2}\Delta_i(BIC))}{\sum_{k=1}^{K}(\exp(-\frac{1}{2}\Delta_k(BIC))} \end{aligned} \tag{23}$$

The weighted scores provide a more straightforward interpretation. Particularly, they can be thought of as the probability of a given model being the best approximating model (Symonds & Moussalli, 2011). Finally for comparing morels with and without fixed initial estimates for action values we use the Likelihood Ratio Test. This measures the goodness of fit between two a general model and a nested model, where the nested model is special version of the general model with one or more additional parameter (Farrell & Lewandowsky, 2018). This is suitable for our purposes, as we treat the models with fixed initial means as the general model and the models with varying initial expectations as the nested models, computed as:

$$\chi^2 \approx -2 \log L_{\text{specific}} - [-2 \log L_{\text{general}}] \tag{24}$$

The estimated $\chi^2$ is then compared to critical value on the $\chi^2$ distribution with 1 degree of freedom.

### Model Recovery

While model identifiability for the models with and without differential learning rates and autocorrelation was validated by Sugawara and Katahira (2021) for their data set, we conducted model recovery to assess identifiability for the other two data sets using simulated data. We simulated 110 runs (a total of 3960 for Dataset 2 and 1320 for Dataset 3) for each model, parameters of which were corresponding to actual participants. While model identifiability was weak across the whole range of models for both data set (see Appendix A), model identifiability was possible for classes of models.

For Dataset 2 model identifiability between Symmetrical Q-learning model and Asymmetrical Q-learning models was possible 76.81% of simulated the runs. The percentage of the model identifiability for Symmetrical Q-learning and Asymmetrical Q-learning models with autocorrelation was possible for 71.36% of the runs. For the Bayesian-updating models, model identifiability between the Symmetrical Asymmetrical variants was possible 69.09% of the runs, while for the Bayesian-updating models with autocorrelation, the rate of model identifiability was 55.91%. The counterpart models with and without autocorrelations were recoverable only less than 52% for the Q-learning models. However, for the Bayesian updating models, this was higher at 91%. Finally, the model identifiability rate between the Q-learning models and Bayesian updating models was 67.87%.

For Dataset 3 model identifiability between Symmetrical Q-learning model and Asymmetrical Q-learning models was possible 74.09% of simulated the runs. The percentage of the model identifiability for Symmetrical Q-learning and Asymmetrical Q-learning models with autocorrelation was possible for 74.55% of the runs. For the Bayesian-updating models model identifiability 86.36% and for the Bayesian-updating models with autocorrelation, the rate model identifiability was 63.18% between the symmetrical and asymmetrical variants. The counterpart models with and without autocorrelations were only dissociable less than 50% for the Q-learning models. However, for the Bayesian updating models, it was possible for 90% of the simulated runs. Moreover, the model identifiability rate between Kalman-filter models and Q-learning models was 67.87%. Overall, model identifiability was hard to achieve for both environments. However, it was slightly better for the IGT.

### Statistical Tests

We measured the existence of learning bias across participants. Like previous studies, we quantified the bias as the difference between positive and negative values of the learning parameter in asymmetrical models; denoted as $\Delta\alpha^{\pm}$ for the Q-learning models and as $\Delta\eta^{\pm}$ for the Bayesian Updating models. It is important to note that the $\Delta\eta^{-}$ reflects positivity bias, as it implies less uncertainty about the environment, which increases the learning rate (see the model). Using these values, previous studies conducted a two-tailed one-sample t-test to assess whether the mean differs from zero. However, unlike previous studies, we use a one-sample two-tailed Binomial Proportion test. This test examines if the proportion of learning rate differences that are greater (or smaller) than 0 differs from an expected value of $P = .5$.

We believe this test is better suited for our analysis than the t-test. First, to establish learning bias at a group level, we are not concerned with the magnitude of the bias. There is no difference for this analysis if the participants are slightly bias or significantly bias. As such, the Binomial Proportion test captures what we are interested in, i.e., if a significant number of participants demonstrate learning bias. Second, extreme values of $\Delta\alpha^{\pm}$ or $\Delta\eta^{\pm}$ could make the test biased. Lastly, the learning parameters have a Beta distribution for the Q-learning models and log-normal distribution for the Bayesian updating models. The Binomial Proportion test is nonparametric. Thus, we can avoid any bias that may result from tests that require certain assumptions about the family distribution of the values being tested.

We used linear mixed-effects models to investigate models features and characteristics. These models incorporated: (a) fixed effects that reflect the average effects within and difference between models with fixed and unfixed initial expectations, (b) random effects that considers participants' individual variability with respect to model class, parameters, and fixed initial expectations. These models included effect coding for all factors of interest. Specifically, effect coding was used for model learning (Bayesian Updating = 1, Q-learning = -1), initial estimates (fixed = 1, unfixed = -1), autocorrelation (yes = 1, no = -1), and two sets of contrasts for learning bias, as follows:

| Factors | Contrast 1 | *Contrast 2* |
|---|---|---|
| Prediction-based | 1 | 0 |
| Reward-based | 0 | 1 |
| Symetrical learning | -1 | -1 |

Our models included main predictors of the factors and their interaction. We conducted this analysis for both AIC and BIC scores across all three datasets. The analysis was conducted using the afex package (Singmann et al., 2016), which depends on the lme4 (Bates, Maechler, Bolker, Walker, et al., 2014) and emmeans package (Lenth, Singmann, Love, Buerkner, &

Herve, 2019).

# Results

## Part 1: Learning asymmetry and autocorrelation

### *Learning Bias*

We measured learning bias as the difference between the positive and negative learning rate values, denoted as $\Delta\alpha^{\pm}$ in the asymmetrical Q-learning models (see Figure 3). To assess whether participants demonstrated learning bias in Dataset 1, we conducted a one-sample Binomial Proportion test. Like previous studies, we found learning bias for the asymmetrical Q-learning models. The significant two-tailed test indicated that the proportion of positive and negative $\Delta\alpha^{\pm}$ differs from 0.5 for both the ApQ model ($k = 94, N = 143, p < .001$) and the ArQ model ($k = 95, N = 143, p < .001$). The effect size for both the ApQ model (h =.32, 95% CI [0.57, 0.73]) and the ArQ model (h =.34, 95% CI [0.58, 0.74]) was medium with precise confidence intervals. The learning bias was eliminated for the prediction-based asymmetrical Q-learning model once we included autocorrelation; k = 64, N = 143, p =.24. However, the learning bias persisted for the reward-based asymmetrical Q-learning model with autocorrelation ($k = 64, N = 143, p < .01$) with a small effect size but precise confidence intervals; h =.22, 95% CI [0.53, 0.69].

For the asymmetrical Bayesian Updating models learning bias was measured as the difference between positive and negative error terms, denoted as $\Delta\eta^{\pm}$. We replicated the same results for the models without autocorrelations, with the proportion of positive and negative $\Delta\eta\pm$ for both ApKF ($k = 42, N = 143, p < .001$) and ArKF ($k = 41, N = 143, p < .001$) models being different from .5 . The effect size was medium with precise confidence intervals for both ApKF (h =.22, 95% CI [0.22, 0.38]) and ArKF (h =.44, 95% CI [0.21, 0.37]) models. However, unlike the Q-learning models the learning bias was eliminated with inclusion of autocorrolation for both the prediction based ($k = 72, N = 143, p = .99$) and reward based

$(k = 72, N = 143, p = 0.24)$ Bayesian Updating models.

We conducted the same analysis for the other two datasets. For the Dataset 2, the results from the significant two tailed tests indicates that the values of positive and negative $\Delta\alpha^{\pm}$ for the Q-learning models, and positive and negative $\Delta\eta^{\pm}$ for Bayesian Updating models differed from 0.5, with large effect size for all the Q-learning models, medium effect size for the Bayesian updating models without autocorrelations and small effect size for the Bayesian Updating Models (see Table 1 for the results).

Our finding from Dataset 3 suffered from the opposite problem. The two-tailed Binomial test comparing the proportion of positive and negative $\Delta\alpha^{\pm}$ for ArQwA model $(k = 95, N = 143, p < .001)$ to .5 was significant, with a medium effect size and relatively imprecise confidence intervals; h =.44, 95% CI [0.21, 0.37]) . The results from the tests for the seven remaining models were not significant (see Table 1). While the power analysis for the ArQwA model indicated that the ability to detect a significant result was 0.82, the remaining models had low powers, ranging from 0.06 to 0.26. As such, the possibility of Type I error is high for these tests.

**Table 1**

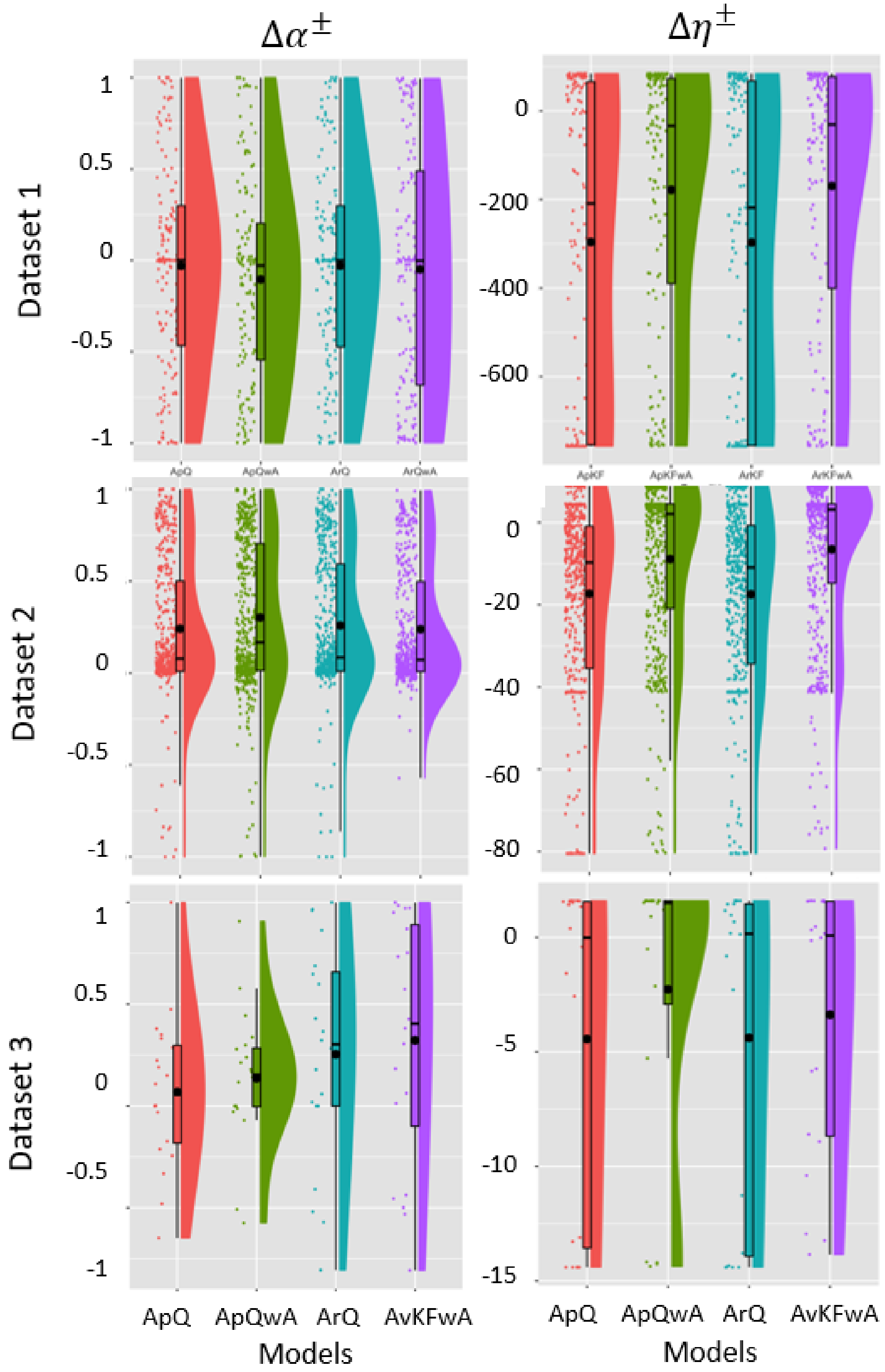
*The results from the proportion binomial test for models with fixed initial expectations.*

| Dataset | Models | $N$ | $k$ | $p$ | Lower CI | Upper CI | $h$ |
|---|---|---|---|---|---|---|---|
| Dataset 1 | ApQ | 143 | 94 | 0.00 | 0.57 | 0.73 | 0.32 |
| | ArQ | 143 | 95 | 0.00 | 0.58 | 0.74 | 0.34 |
| | ApQwA | 143 | 64 | 0.24 | 0.36 | 0.53 | -0.11 |
| | ArQwA | 143 | 87 | 0.01 | 0.52 | 0.69 | 0.22 |
| | ApKF | 143 | 42 | 0.00 | 0.22 | 0.38 | -0.43 |
| | ArKF | 143 | 41 | 0.00 | 0.21 | 0.37 | -0.44 |
| | ApKFwA | 143 | 72 | 1.00 | 0.42 | 0.59 | 0.01 |
| | ApKFwA | 143 | 64 | 0.24 | 0.36 | 0.53 | -0.11 |
| Dataset 2 | ApQ | 504 | 448 | 0.00 | 0.86 | 0.92 | 0.89 |
| | ArQ | 504 | 448 | 0.00 | 0.86 | 0.92 | 0.89 |
| | ApQwA | 504 | 408 | 0.00 | 0.77 | 0.84 | 0.67 |
| | ArQwA | 504 | 417 | 0.00 | 0.79 | 0.86 | 0.71 |
| | ApKF | 504 | 122 | 0.00 | 0.21 | 0.28 | -0.54 |
| | ArKF | 504 | 117 | 0.00 | 0.20 | 0.27 | -0.57 |
| | ApKFwA | 504 | 308 | 0.00 | 0.57 | 0.65 | 0.22 |
| | ApKFwA | 504 | 279 | 0.02 | 0.51 | 0.60 | 0.11 |
| Dataset 23 | ApQ | 20 | 11 | 0.82 | 0.32 | 0.77 | 0.10 |
| | ArQ | 20 | 16 | 0.01 | 0.56 | 0.94 | 0.64 |
| | ApQwA | 20 | 13 | 0.26 | 0.41 | 0.85 | 0.31 |
| | ArQwA | 20 | 13 | 0.26 | 0.41 | 0.85 | 0.31 |
| | ApKF | 20 | 11 | 0.82 | 0.32 | 0.77 | 0.10 |
| | ArKF | 20 | 10 | 1.00 | 0.27 | 0.73 | 0.00 |
| | ApKFwA | 20 | 10 | 1.00 | 0.27 | 0.73 | 0.00 |
| | ApKFwA | 20 | 12 | 0.50 | 0.36 | 0.81 | 0.20 |

Where confidence interval is around the proportion mean and the effect size is estimated according Cohen’$h$

**Figure 3**

*Delta distribution for models with fixed expectations*



### ***Model Comparison***

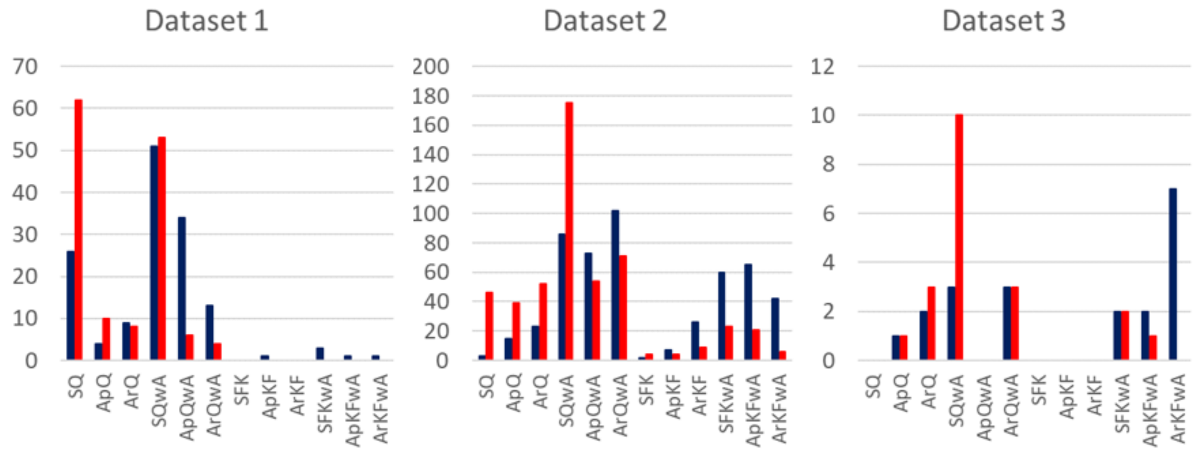
The fitted models were evaluated based on the AIC and BIC. Figure 3 shows the number of participants best fit by each model for all three data sets. Consistent with our hypothesis, in Dataset 1 the Q-learning models were the best-fitting ones almost unanimously for all participants, 95% according to the AIC and 100% according to the BIC. Furthermore, models with symmetrical learning-rate were the best fitting model for most participants according to the BIC (80%). However, there was a near even split between the symmetrical and asymmetrical models according to the AIC; with the symmetrical models having the lowest (best) score for 56% of the participants. Lastly, the models with autocorrelation had the lowest score for 73% of participants, however there was a near even split according to the BIC.

In the case of specific models, the SQwA had the highest predictive validity ($wAIC = .27$) and the second highest likelihood of being the true model ($wBIC = .32$). Meanwhile, the SQ model had the highest likelihood of being the true model ($wBIC = .36$), however it had relatively weak predicative accuracy ($wAIC = .1$). Contrary to what we predicted, the prediction-based models performed slightly better than the reward-based asymmetrical models according to the AIC, particularly the ApQwA had the second highest predicative validity ($wAIC = .25$). There was no notable difference between the prediction-based and reward-based asymmetrical models according to the BIC.

In Dataset 2, models with autocorrelation had fitted more participants according to both criteria; 85% and 70% for the AIC and BIC respectively. There was no noticeable difference between symmetrical and asymmetrical models according to the AIC, except for the SQwA model. The SQwA had the highest likelihood across participants of being the true model ($wBIC = .32$), while the ArQwA had the highest average predictive validity ($wAIC = .26$). Contradictory to what we predicted, Q-learning models performed better than the Bayesian Updating counter parts. Furthermore, unlike Dataset 1, the reward-based asymmetrical

**Figure 4**

*Model comparison according to AIC and BIC for models with fixed initial expectations*



models performed slightly better than the prediction-based asymmetrical models, except for the Bayesian Updating models with autocorrelation.

In Dataset 3, According to the AIC, the ArKFwA outperformed every other model by being the best fitted mode for 35% of the participants. However, this model did not have a strong predictive validity; $wAIC = .28$. According to the BIC the SQwA outperformed every other model, being the best fitted model for 50% of the participants. This model had on average a medium likelihood of being the best fitted model; $wBIC = .42$. The Q-learning models performed better than the Baysian Updating model according to the BIC, capturing 85% of participants. However, there was near even split between the two class of models according the AIC, with Q-learning model capturing only 45% of the total participants. The models with autocorrelation also performed better capturing 85% according to the AIC and 80% according to the BIC of participants. Lastly, the asymmetrical models performed well according to the AIC (75%), but not according to the BIC (40%).

## Part 2: Including initial expectations

### *Learning Bias*

We re-examined learning bias in our models with asymmetrical learning using the Binomial Proportion test, however, this time we allowed for the initial action value estimates to vary. As we predicted, the learning bias was eliminated for the modified Q-learning models. The results from the two-tailed Binomial proportion test indicate a non-significant result for both the ApQ model ($k = 76, N = 143, p = .05$) and ArQ model ($k = 73, N = 143, p = .86$). However, the bias persisted for the asymmetrical Bayesian Updating Models. The results indicate that the proportion of positive and negative $\Delta\eta^{\pm}$ differs from 0.5 for the ApKF model ($k = 56, N = 143, p < .01$) and ArKF model ($k = 55, N = 143, p < .01$). The effect size was weak with precise confidence intervals for both models; h =.22, 95% CI [0.31, 0.48] (ApFK) and h =.23, 95% CI [0.3, 0.47]) (ArFK). Again this bias was captured by the inclusion of model as the results for the two tailed test were not significant for the ApKFwA ($k = 73, N = 143, p = .86$) and ArKFwA ($k = 73, N = 143, p = .86$) models. See table 2 for complete results.

The results for Dataset 2 were the same as when the models had fixed initial expectations. The results from the two tailed Binomial Proportion test were significant for all the models (see Table 2). The results the Dataset 3 indicated that the proportion of positive and negative $\Delta\alpha^{\pm}$ differs from 0.5 for the prediction based asymmetrical model without autocorrelation ($k = 76, N = 143, p = .05$) and with autocorrelation ($k = 76, N = 143, p = .05$). However, there was no learning bias in the reaming models. However, it should be again noted that the two-tailed tests were under-powered for these models; ranging from 0.02 to 43. See table 2 for complete results.

**Table 2**

*The results from the proportion binomial test for models with varying initial expectations.*

| Dataset | Models | N | k | p | Lower CI | Upper CI | h |
|---|---|---|---|---|---|---|---|
| Dataset 1 | ApQ | 143 | 76 | 0.50 | 0.45 | 0.62 | 0.06 |
| | ArQ | 143 | 73 | 0.87 | 0.43 | 0.59 | 0.02 |
| | ApQwA | 143 | 68 | 0.62 | 0.39 | 0.56 | -0.05 |
| | ArQwA | 143 | 67 | 0.50 | 0.38 | 0.55 | -0.06 |
| | ApKF | 143 | 56 | 0.01 | 0.31 | 0.48 | -0.22 |
| | ArKF | 143 | 55 | 0.01 | 0.30 | 0.47 | -0.23 |
| | ApKFwA | 143 | 66 | 0.40 | 0.38 | 0.55 | -0.08 |
| | ApKFwA | 143 | 64 | 0.24 | 0.36 | 0.53 | -0.11 |
| Dataset 2 | ApQ | 504 | 402 | 0.00 | 0.76 | 0.83 | 0.64 |
| | ArQ | 504 | 415 | 0.00 | 0.79 | 0.86 | 0.70 |
| | ApQwA | 504 | 384 | 0.00 | 0.72 | 0.80 | 0.55 |
| | ArQwA | 504 | 409 | 0.00 | 0.77 | 0.84 | 0.67 |
| | ApKF | 504 | 152 | 0.00 | 0.26 | 0.34 | -0.41 |
| | ArKF | 504 | 126 | 0.00 | 0.21 | 0.29 | -0.52 |
| | ApKFwA | 504 | 212 | 0.00 | 0.38 | 0.47 | -0.16 |
| | ApKFwA | 504 | 225 | 0.02 | 0.40 | 0.49 | -0.11 |
| Dataset 3 | ApQ | 20 | 16 | 0.01 | 0.56 | 0.94 | 0.64 |
| | ArQ | 20 | 14 | 0.12 | 0.46 | 0.88 | 0.41 |
| | ApQwA | 20 | 15 | 0.04 | 0.51 | 0.91 | 0.52 |
| | ArQwA | 20 | 14 | 0.12 | 0.46 | 0.88 | 0.41 |
| | ApKF | 20 | 7 | 0.26 | 0.15 | 0.59 | -0.31 |
| | ArKF | 20 | 10 | 1.00 | 0.27 | 0.73 | 0.00 |
| | ApKFwA | 20 | 6 | 0.12 | 0.12 | 0.54 | -0.41 |
| | ApKFwA | 20 | 9 | 0.82 | 0.23 | 0.68 | -0.10 |

Where confidence interval is around the proportion mean and the effect size is estimated according Cohen'$h$

**Figure 5**

*Delta distribution for models with unfixed expectations*

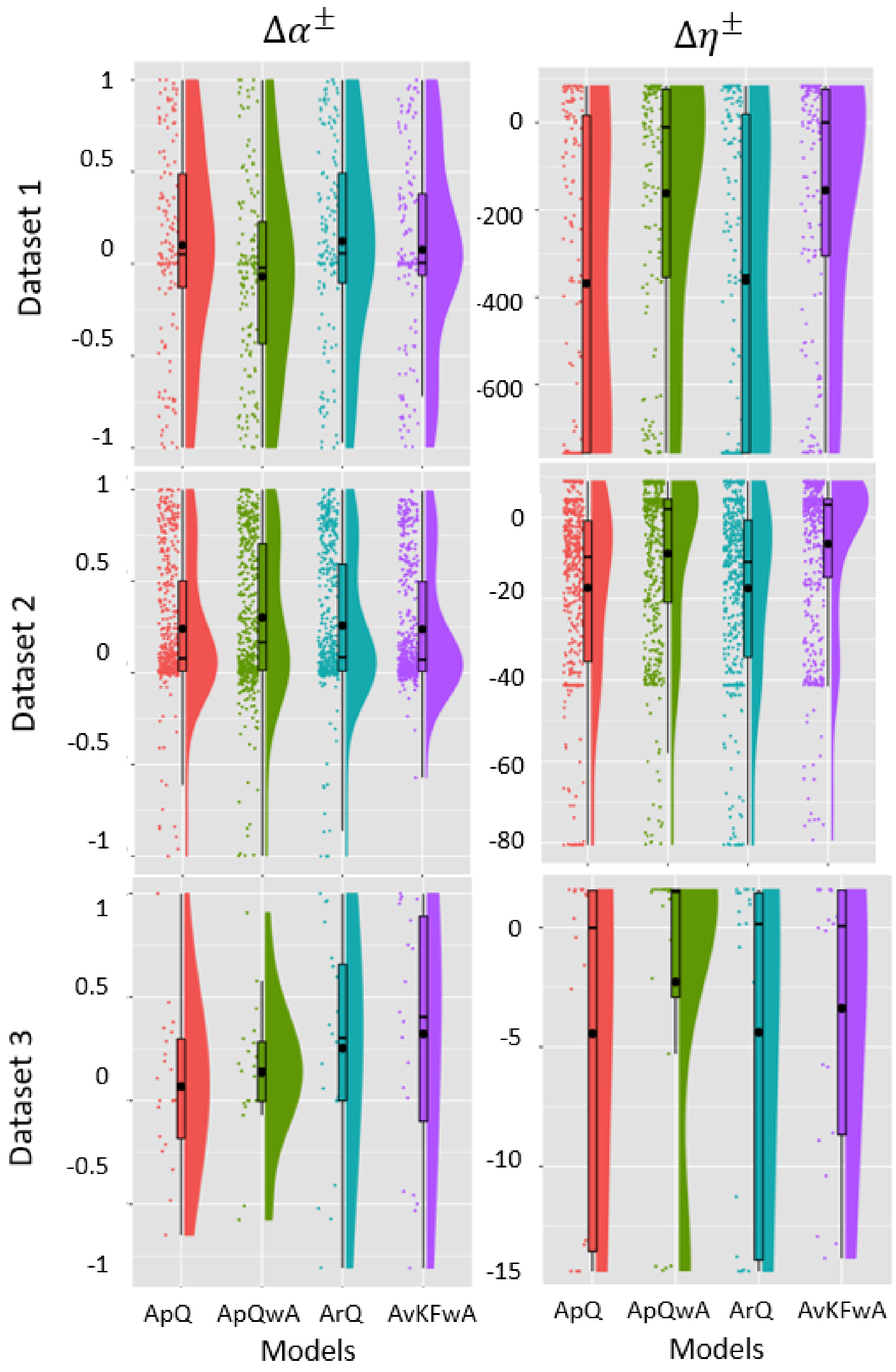
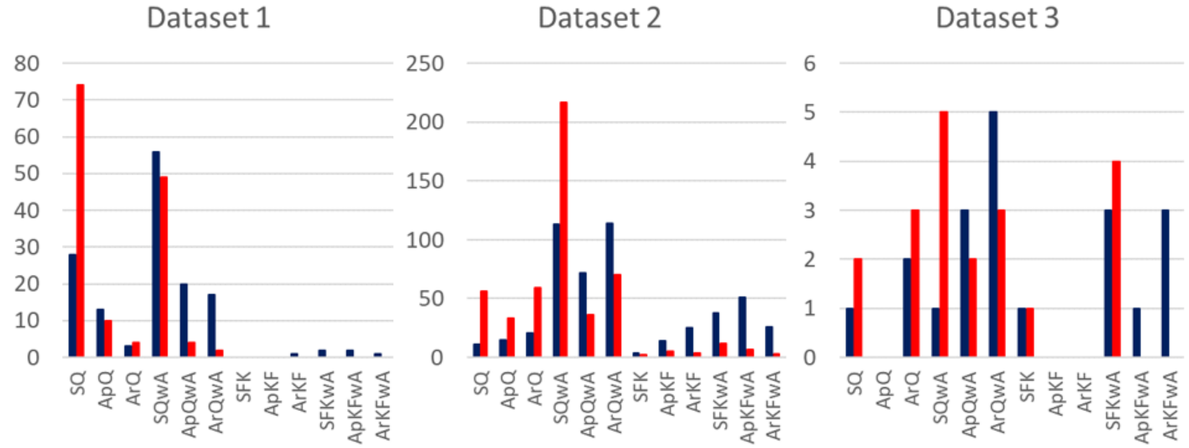


**Figure 6**

*The results from the proportion binomial test for models with varying initial expectations.*



## Model comparison

The results from model comparison for the models with varying initial estimates of action values were very similar to the results from model comparison for models with fixed initial expectations (see Figure 6). However, they are a few difference worth noting. For Dataset 1 there was notable increase in fit according to the BIC for the SQ model, which now had a higher likelihood of being the true model (wBIC =.4). For Dataset 2 the SQwA and ArQwA were the best fitted model for more participants, while the Updating Bayesian models with autocorrelation were the best fitted model for less participants. According to the BIC, the SQwA fitted the most participants.

For the third data set, the number of participants for whom the SQwA captured according to BIC and the ArKFwA captured according to the AIC by 50% and 58% respectively. However, it still remained that the Q-learning models and models with autocorrelation performed better than the Bayesian Updating model and models without autocorrelation. The distinction between models symmetrical and asymmetrical models stayed the same (See Appendix B).

### Part 3: The effect of allowing for varying initial expectations

In this part, we investigated the effect of fixing the initial expectations of action values on model fit. We first looked at this effect across the whole model range.We used the likelihood ratio test to assess if the models with unfixed initial expectations better fit participants' behavioural data than models with fixed initial expectations. We conducted this test for each pair of models across the three paradigms (see Table 3). In Dataset 1, the inclusion of varying initial expectations significantly improved the likelihood values for all the symmetrical and asymmetrical models (both Q-learning and Bayesian Updating) without autocorrelation. However, except for the ArQwA model ($x^2 = 700, df = 143, p < 0.001$), the likelihood values for all the models with autocorrelation, remained relatively unchanged.

For Dataset 2, the results from the likelihood ratio test indicated that the inclusion of varying initial estimates for action values significantly improved the likelihood values across the whole model range. Likewise, for Dataset 3, the inclusion of varying initial estimates significantly improved the likelihood values for all models. Finally, a linear mixed-effect model analysis was conducted for AIC and BIC scores indecently, combining the results from models with and without fixed initial expectations. Table 4 to 6 shows the results for each of the three datasets respectively.

**Table 3**

*The results from the likelihood ratio test*

| Dataset | Models | *L specific (Mean)* | *L specific (SD)* | *L General (Mean)* | *L General (SD)* | $X^2$ | P-value |
|---|---|---|---|---|---|---|---|
| Dataset 1 | SQ | 118.53 | 14.10 | 115.68 | 16.43 | 815.05 | 0.00 |
| | ApQ | 115.06 | 16.14 | 114.14 | 16.86 | 262.83 | 0.00 |
| | ArQ | 115.07 | 16.14 | 114.12 | 16.86 | 271.40 | 0.00 |
| | SQwA | 109.80 | 20.19 | 109.27 | 20.44 | 149.12 | 0.35 |
| | ApQwA | 108.61 | 20.34 | 108.25 | 20.51 | 102.44 | 1.00 |
| | ArQwA | 111.30 | 17.94 | 108.86 | 19.98 | 700.05 | 0.00 |
| | SFK | 117.92 | 14.40 | 115.18 | 17.08 | 783.14 | 0.00 |
| | ApKF | 115.06 | 15.97 | 113.58 | 17.24 | 425.20 | 0.00 |
| | ArKF | 115.06 | 15.98 | 113.56 | 17.27 | 428.69 | 0.00 |
| | SFKwA | 109.12 | 20.21 | 108.80 | 20.39 | 93.04 | 1.00 |
| | ApKFwA | 108.39 | 20.07 | 108.05 | 20.25 | 98.56 | 1.00 |
| | ArKFwA | 108.43 | 20.08 | 108.06 | 20.18 | 108.47 | 0.99 |
| Dataset 2 | SQ | 126.70 | 17.99 | 117.78 | 21.78 | 8991.03 | 0.00 |
| | ApQ | 118.16 | 21.35 | 114.77 | 22.86 | 3417.10 | 0.00 |
| | ArQ | 117.82 | 21.48 | 114.28 | 22.75 | 3572.19 | 0.00 |
| | SQwA | 104.40 | 28.41 | 101.69 | 28.54 | 2738.99 | 0.00 |
| | ApQwA | 102.22 | 28.15 | 100.44 | 28.33 | 1792.19 | 0.00 |
| | ArQwA | 105.07 | 29.88 | 103.23 | 29.61 | 1855.93 | 0.00 |
| | SFK | 123.29 | 19.71 | 117.18 | 23.08 | 6162.06 | 0.00 |
| | ApKF | 114.61 | 22.12 | 111.74 | 23.39 | 2893.51 | 0.00 |
| | ArKF | 114.04 | 22.41 | 111.07 | 23.63 | 2994.17 | 0.00 |
| | SFKwA | 101.44 | 27.97 | 100.31 | 28.33 | 1145.98 | 0.00 |
| | ApKFwA | 99.77 | 27.84 | 98.98 | 27.88 | 792.66 | 0.00 |
| | ArKFwA | 99.91 | 27.87 | 98.95 | 27.95 | 965.63 | 0.00 |
| Dataset 3 | SQ | 129.08 | 48.13 | 124.62 | 51.33 | 178.35 | 0.00 |
| | ApQ | 126.96 | 49.97 | 122.07 | 50.36 | 195.72 | 0.00 |
| | ArQ | 123.38 | 49.50 | 117.85 | 49.41 | 220.95 | 0.00 |
| | SQwA | 107.88 | 43.40 | 104.03 | 43.76 | 154.06 | 0.00 |
| | ApQwA | 107.23 | 43.81 | 102.19 | 43.45 | 201.61 | 0.00 |
| | ArQwA | 111.07 | 45.69 | 105.55 | 43.67 | 220.58 | 0.00 |
| | SFK | 125.93 | 50.25 | 121.83 | 50.99 | 164.05 | 0.00 |
| | ApKF | 124.05 | 50.10 | 121.61 | 50.71 | 97.56 | 0.00 |
| | ArKF | 122.04 | 51.24 | 117.83 | 50.28 | 168.35 | 0.00 |
| | SFKwA | 104.47 | 42.00 | 100.06 | 42.24 | 176.76 | 0.00 |
| | ApKFwA | 103.57 | 41.60 | 99.69 | 41.87 | 155.50 | 0.00 |
| | ArKFwA | 102.45 | 42.02 | 99.13 | 41.23 | 132.81 | 0.00 |

**Table 4**

*Results from the linear mixed-effect model for Dataset 1*

| | AIC | | | | | BIC | | | | |
|---|---|---|---|---|---|---|---|---|---|---|
| | Estimate | Std. Error | df | t value | Pr(>\|t\|) | Estimate | Std. Error | df | t value | Pr(>\|t\|) |
| (Intercept) | 235.49 | 2.96 | 142.10 | 79.50 | 0.00 | 253.89 | 2.96 | 142.08 | 85.70 | 0.00 |
| KF | 2.38 | 0.10 | 78.45 | 23.79 | 0.00 | 7.25 | 0.10 | 98.26 | 70.01 | 0.00 |
| Fixed | 0.23 | 0.15 | 132.57 | 1.54 | 0.13 | -1.39 | 0.15 | 135.80 | -9.19 | 0.00 |
| Autocor. | -4.34 | 0.61 | 141.96 | -7.09 | 0.00 | -1.09 | 0.61 | 141.99 | -1.78 | 0.08 |
| Asym_p | -0.70 | 0.15 | 174.59 | -4.63 | 0.00 | 0.38 | 0.15 | 174.40 | 2.48 | 0.01 |
| Asym_r | 0.12 | 0.14 | 180.18 | 0.87 | 0.39 | 1.20 | 0.14 | 180.36 | 8.53 | 0.00 |
| KF×*Fixed* | -0.11 | 0.07 | 2813.60 | -1.46 | 0.15 | -0.11 | 0.07 | 2829.46 | -1.46 | 0.15 |
| KF×*Autocor.* | -0.25 | 0.07 | 2813.60 | -3.44 | 0.00 | -0.25 | 0.07 | 2829.46 | -3.44 | 0.00 |
| Fixed×*Autocor.* | -0.51 | 0.07 | 2813.60 | -6.94 | 0.00 | -0.51 | 0.07 | 2829.46 | -6.95 | 0.00 |
| KF×*Asym_p* | 0.38 | 0.10 | 2813.60 | 3.66 | 0.00 | 0.38 | 0.10 | 2829.46 | 3.67 | 0.00 |
| KF×*Asym_r* | -0.44 | 0.10 | 2813.60 | -4.26 | 0.00 | -0.44 | 0.10 | 2829.46 | -4.26 | 0.00 |
| Fixed×*Asym_p* | -0.46 | 0.10 | 2813.60 | -4.44 | 0.00 | -0.46 | 0.10 | 2829.46 | -4.45 | 0.00 |
| Fixed×*Asym_r* | 0.08 | 0.10 | 2813.60 | 0.81 | 0.42 | 0.08 | 0.10 | 2829.46 | 0.82 | 0.42 |
| Autocor.×*Asym_p* | 0.20 | 0.10 | 2813.60 | 1.94 | 0.05 | 0.20 | 0.10 | 2829.46 | 1.95 | 0.05 |
| Autocor.×*Asym_r* | 1.04 | 0.10 | 2813.60 | 10.15 | 0.00 | 1.04 | 0.10 | 2829.46 | 10.16 | 0.00 |
| KF×*Fixed* × *Autocor.* | -0.27 | 0.07 | 2813.60 | -3.76 | 0.00 | -0.27 | 0.07 | 2829.46 | -3.76 | 0.00 |
| KF×*Fixed* × *Asym_p* | 0.24 | 0.10 | 2813.60 | 2.37 | 0.02 | 0.24 | 0.10 | 2829.46 | 2.38 | 0.02 |
| KF×*Fixed* × *Asym_r* | -0.27 | 0.10 | 2813.60 | -2.66 | 0.01 | -0.27 | 0.10 | 2829.46 | -2.66 | 0.01 |
| KF×*Autocor* .×*Asym_p* | 0.29 | 0.10 | 2813.60 | 2.79 | 0.01 | 0.29 | 0.10 | 2829.46 | 2.79 | 0.01 |
| KF×*Autocor* .×*Asym_r* | -0.52 | 0.10 | 2813.60 | -5.09 | 0.00 | -0.52 | 0.10 | 2829.46 | -5.09 | 0.00 |
| Fixed×*Autocor.* ×*Asym_p* | 0.08 | 0.10 | 2813.60 | 0.77 | 0.44 | 0.08 | 0.10 | 2829.46 | 0.77 | 0.44 |
| Fixed×*Autocor.* ×*Asym_r* | 0.60 | 0.10 | 2813.60 | 5.83 | 0.00 | 0.60 | 0.10 | 2829.46 | 5.83 | 0.00 |
| KF×*Fixed* × Autocor.×*Asym_p* | 0.13 | 0.10 | 2813.60 | 1.25 | 0.21 | 0.13 | 0.10 | 2829.46 | 1.25 | 0.21 |
| KF×*Fixed* × Autocor.×*Asym_r* | -0.38 | 0.10 | 2813.60 | -3.70 | 0.00 | -0.38 | 0.10 | 2829.46 | -3.70 | 0.00 |

**Table 5**

*Results from the linear mixed-effect model for Dataset 2*

| | AIC | | | | | BIC | | | | |
|---|---|---|---|---|---|---|---|---|---|---|
| | Estimate | Std. Error | df | t value | Pr(>\|t\|) | Estimate | Std. Error | df | t value | Pr(>\|t\|) |
| (Intercept) | 229.49 | 2.09 | 503.05 | 109.70 | 0.00 | 244.25 | 2.09 | 503.10 | 116.76 | 0.00 |
| KF | 0.06 | 0.29 | 503.23 | 0.20 | 0.84 | 3.97 | 0.29 | 503.22 | 13.66 | 0.00 |
| Fixed | 2.09 | 0.15 | 503.80 | 13.54 | 0.00 | 0.78 | 0.15 | 503.82 | 5.08 | 0.00 |
| Autocor. | -13.42 | 0.64 | 503.12 | -20.82 | 0.00 | -10.81 | 0.64 | 502.89 | -16.78 | 0.00 |
| Asym_p | -2.31 | 0.17 | 720.79 | -13.83 | 0.00 | -1.45 | 0.17 | 720.82 | -8.64 | 0.00 |
| Asym_r | -1.39 | 0.23 | 539.64 | -6.00 | 0.00 | -0.53 | 0.23 | 539.60 | -2.26 | 0.02 |
| KF×*Fixed* | -0.61 | 0.09 | 9557.02 | -6.60 | 0.00 | -0.61 | 0.09 | 9556.94 | -6.60 | 0.00 |
| KF×*Autocor.* | -0.01 | 0.09 | 9557.02 | -0.10 | 0.92 | -0.01 | 0.09 | 9556.94 | -0.10 | 0.92 |
| Fixed×*Autocor.* | -1.55 | 0.09 | 9557.02 | -16.68 | 0.00 | -1.55 | 0.09 | 9556.94 | -16.68 | 0.00 |
| KF×*Asym_p* | 0.31 | 0.13 | 9557.02 | 2.39 | 0.02 | 0.31 | 0.13 | 9556.94 | 2.39 | 0.02 |
| KF×*Asym_r* | -1.17 | 0.13 | 9557.02 | -8.88 | 0.00 | -1.17 | 0.13 | 9556.94 | -8.88 | 0.00 |
| Fixed×*Asym_p* | -0.88 | 0.13 | 9557.02 | -6.70 | 0.00 | -0.88 | 0.13 | 9556.94 | -6.70 | 0.00 |
| Fixed×*Asym_r* | -0.76 | 0.13 | 9557.02 | -5.77 | 0.00 | -0.76 | 0.13 | 9556.94 | -5.76 | 0.00 |
| Autocor.×*Asym_p* | 0.95 | 0.13 | 9557.02 | 7.25 | 0.00 | 0.95 | 0.13 | 9556.94 | 7.25 | 0.00 |
| Autocor.×*Asym_r* | 2.91 | 0.13 | 9557.02 | 22.14 | 0.00 | 2.91 | 0.13 | 9556.94 | 22.14 | 0.00 |
| KF×*Fixed* × *Autocor.* | 0.04 | 0.09 | 9557.02 | 0.40 | 0.69 | 0.04 | 0.09 | 9556.94 | 0.40 | 0.69 |
| KF×*Fixed* × *Asym_p* | 0.24 | 0.13 | 9557.02 | 1.79 | 0.07 | 0.24 | 0.13 | 9556.94 | 1.79 | 0.07 |
| KF×*Fixed* × *Asym_r* | 0.25 | 0.13 | 9557.02 | 1.89 | 0.06 | 0.25 | 0.13 | 9556.94 | 1.89 | 0.06 |
| KF×*Autocor* .×*Asym_p* | 0.68 | 0.13 | 9557.02 | 5.17 | 0.00 | 0.68 | 0.13 | 9556.94 | 5.17 | 0.00 |
| KF×*Autocor* .×*Asym_r* | -0.61 | 0.13 | 9557.02 | -4.62 | 0.00 | -0.61 | 0.13 | 9556.94 | -4.62 | 0.00 |
| Fixed×*Autocor.* ×*Asym_p* | 0.63 | 0.13 | 9557.02 | 4.76 | 0.00 | 0.63 | 0.13 | 9556.94 | 4.76 | 0.00 |
| Fixed×*Autocor.* ×*Asym_r* | 0.62 | 0.13 | 9557.02 | 4.72 | 0.00 | 0.62 | 0.13 | 9556.94 | 4.72 | 0.00 |
| KF×*Fixed* × Autocor.×*Asym_p* | -0.16 | 0.13 | 9557.02 | -1.18 | 0.24 | -0.16 | 0.13 | 9556.94 | -1.18 | 0.24 |
| KF ×*Fixed* × Autocor.×*Asym_r* | -0.11 | 0.13 | 9557.02 | -0.87 | 0.38 | -0.11 | 0.13 | 9556.94 | -0.87 | 0.38 |

**Table 6**

*Results from the linear mixed-effect model for Dataset 3*

| | AIC | | | | | BIC | | | | |
|---|---|---|---|---|---|---|---|---|---|---|
| | Estimate | Std. Error | df | t value | Pr(>\|t\|) | Estimate | Std. Error | df | t value | Pr(>\|t\|) |
| (Intercept) | 238.38 | 20.57 | 19.04 | 11.59 | 0.00 | 253.15 | 20.55 | 19.08 | 12.32 | 0.00 |
| KF | -0.27 | 0.73 | 14.72 | -0.37 | 0.72 | 3.64 | 0.73 | 17.37 | 4.98 | 0.00 |
| Fixed | 3.30 | 0.52 | 11.94 | 6.36 | 0.00 | 2.00 | 0.52 | 17.15 | 3.85 | 0.00 |
| Autocor. | -17.16 | 2.94 | 19.00 | -5.83 | 0.00 | -14.55 | 2.94 | 19.02 | -4.94 | 0.00 |
| Asym_p | 0.46 | 0.58 | 47.10 | 0.79 | 0.43 | 1.33 | 0.58 | 47.15 | 2.28 | 0.03 |
| Asym_r | -1.56 | 0.78 | 19.20 | -1.99 | 0.06 | -0.69 | 0.78 | 19.22 | -0.88 | 0.39 |
| KF×*Fixed* | -0.58 | 0.37 | 376.98 | -1.55 | 0.12 | -0.58 | 0.37 | 379.64 | -1.55 | 0.12 |
| KF×*Autocor.* | -1.49 | 0.37 | 376.98 | -4.01 | 0.00 | -1.49 | 0.37 | 379.64 | -4.01 | 0.00 |
| Fixed×*Autocor.* | 0.03 | 0.37 | 376.98 | 0.09 | 0.93 | 0.03 | 0.37 | 379.64 | 0.09 | 0.93 |
| KF×*Asym_p* | 0.89 | 0.53 | 376.98 | 1.68 | 0.09 | 0.89 | 0.53 | 379.64 | 1.68 | 0.09 |
| KF×*Asym_r* | -0.83 | 0.53 | 376.98 | -1.57 | 0.12 | -0.83 | 0.53 | 379.64 | -1.57 | 0.12 |
| Fixed×*Asym_p* | -0.24 | 0.53 | 376.98 | -0.46 | 0.65 | -0.24 | 0.53 | 379.64 | -0.46 | 0.65 |
| Fixed×*Asym_r* | 0.34 | 0.53 | 376.98 | 0.64 | 0.52 | 0.34 | 0.53 | 379.64 | 0.64 | 0.52 |
| Autocor.×*Asym_p* | -1.34 | 0.53 | 376.98 | -2.55 | 0.01 | -1.34 | 0.53 | 379.64 | -2.55 | 0.01 |
| Autocor.×*Asym_r* | 3.44 | 0.53 | 376.98 | 6.53 | 0.00 | 3.44 | 0.53 | 379.64 | 6.53 | 0.00 |
| KF×*Fixed* × *Autocor.* | 0.11 | 0.37 | 376.98 | 0.30 | 0.76 | 0.11 | 0.37 | 379.64 | 0.30 | 0.76 |
| KF×*Fixed* × *Asym_p* | -0.33 | 0.53 | 376.98 | -0.62 | 0.54 | -0.33 | 0.53 | 379.64 | -0.62 | 0.54 |
| KF×*Fixed* × *Asym_r* | -0.30 | 0.53 | 376.98 | -0.57 | 0.57 | -0.30 | 0.53 | 379.64 | -0.57 | 0.57 |
| KF×*Autocor* .×*Asym_p* | 0.79 | 0.53 | 376.98 | 1.51 | 0.13 | 0.79 | 0.53 | 379.64 | 1.51 | 0.13 |
| KF×*Autocor* .×*Asym_r* | -1.93 | 0.53 | 376.98 | -3.66 | 0.00 | -1.93 | 0.53 | 379.64 | -3.66 | 0.00 |
| Fixed×*Autocor.* ×*Asym_p* | 0.36 | 0.53 | 376.98 | 0.69 | 0.49 | 0.36 | 0.53 | 379.64 | 0.69 | 0.49 |
| Fixed×*Autocor.* ×*Asym_r* | -0.26 | 0.53 | 376.98 | -0.49 | 0.62 | -0.26 | 0.53 | 379.64 | -0.49 | 0.62 |
| KF×*Fixed* × Autocor.×*Asym_p* | 0.21 | 0.53 | 376.98 | 0.40 | 0.69 | 0.21 | 0.53 | 379.64 | 0.40 | 0.69 |
| KF×*Fixed* × Autocor.×*Asym_r* | -0.33 | 0.53 | 376.98 | -0.63 | 0.53 | -0.33 | 0.53 | 379.64 | -0.63 | 0.53 |

# Discussion

In the first part of our analyses, we were able to replicate previous findings in Dataset 1, using models with fixed initial expectations. Particularly, there was a learning bias across participants in the prediction-based asymmetric Q-learning (ApQ) model, and this learning bias was eliminated with the inclusion of autocorrelation (ApQwA model). Furthermore, we were able to generalize this finding to the reward-based asymmetrical Q-learning models (ArQ, ArQwA) in this experimental paradigm. Similarly, there was a group-level learning bias in the asymmetrical Bayesian Updating models, both prediction-based (ApKF) and reward-based (ArFK); however, the learning bias was eliminated once models incorporated autocorrelation (ApKFwA, ArKFwA).

Furthermore, we conducted the same analyses for all the models in this paradigm, but this time allowing the models' estimates for initial action values to vary. Contradictory to our results from part 1, there was no learning bias for the ApQ and ArQ models. Naturally, the same finding was observed in their autocorrelation counterparts. However, there was a learning bias for the ApKF and ArKF models, but the learning bias was eliminated by including autocorrelation in the ApKFwA and ArKFwA models. Therefore, based on our findings from group-level analyses, we were unable to confirm a true existence of learning bias in the two-armed bandit task with fixed binary-reward distribution. While it is true that the simple asymmetrical models detect a learning bias, the learning bias is a statistical artifact that the inclusion of autocorrelation can eliminate. Furthermore, we demonstrated that the allowing for individual variation in initial expectations of the options' values also eliminates the learning bias.

The results from model comparison further strengthened our rejection of the learning bias at the group level. We compared models using both the AIC and the BIC. According to both measures, the symmetrical models (both with fixed and unfixed initial expectations) performed considerably better than their asymmetrical variants. Therefore, we cannot confirm the existence of asymmetrical learning in terms of parsimony for Dataset 1.

Subsequently, we extended our analyses to a new paradigm (IGT). However, in Dataset 2, learning bias was observable across all participants for the entire range of models, with both fixed and unfixed initial expectations. The significant findings from this dataset indicate that a more dynamic environment is required to capture learning asymmetry, based on the theoretical plausibility that there is an association between learning asymmetry and the exploration-exploitation trade-off (see study rationale). However, the model comparison results do not provide support for learning bias in light of parsimony. According to both criteria employed, there was no noticeable difference between the symmetrical models and their asymmetrical variants with both unfixed and fixed initial expectations, with the exception of the SQwA model, which did perform better among models with fixed and unfixed initial expectations based on the BIC. While model comparison does not provide strong evidence for learning bias, it is important to note that the model recovery rate was less than ideal for this dataset. As such, we should treat the model comparison results with some caution, and hence can not be convinced that a learning bias is absent for this dataset. However, there is evidence for it at a group level, and from this we conclude that there is thus some plausibility for the association between learning bias and the exploration-exploitation trade-off. But this requires further investigation.

Additionally, we looked at the restless multi-armed bandit paradigm using Dataset 3. With the exception of the ApQ model with unfixed initial expectations and the ArQ model with fixed initial expectations, the remaining models indicated that there is no learning bias across participants. However, power analysis indicated a very low ability to detect a significant result for all the models. As such, we refrain from interpreting the results as they are prone to Type I errors. The model comparison between models with fixed initial expectations indicated two winners according to both criteria. According to the AIC, the ArKFwA had the highest predictive validity, while the SQwA had the highest likelihood of being the true model according to the BIC. However, when the models were modified to allow for varying initial expectations of action values, there was no clear winner for either criterion. As such, we were not able to find clear evidence for or against learning bias from this dataset. However, an interesting observation to note is that the inclusion of varying initial expectations lowered the number of times the models with autocorrelation were the best-fitted models.

In the third part of our analyses, we looked at the effect of allowing initial expectations to vary on model prediction using the likelihood ratio test. The results indicated that allowing for unfixed initial expectations in the models without autocorrelation significantly improved model fit in Dataset 1. However, there was no significant improvement for the models with autocorrelation, except for the ArQwA model. Furthermore, we investigated the effect of initial expectations and autocorrelation on model performance using the mixed-effect analysis. While autocorrelation had a significant main effect on models' predictive validity (i.e., AIC score), it had no effect on the likelihood of the model being the true model (i.e., BIC score). More specifically, the inclusion of autocorrelation decreased (improved) AIC scores on average. Meanwhile, fixing the initial expectations for action values did not overall affect AIC scores. However, it had a significant main effect on the BIC scores. Furthermore, their interaction revealed that allowing initial estimates to vary improves both the AIC and BIC scores more for models without autocorrelation.

Furthermore, we compared the effect of allowing for initial expectations to vary on model performance in Dataset 2 and 3. Every single model with and without autocorrelation was significantly improved. Additionally, the linear-mixed effect model revealed that allowing the initial expectations to vary significantly improved AIC and BIC scores. Furthermore, both AIC and BIC scores were reduced (improved) with the inclusion of varying initial expectations for models with autocorrelation. From these results, we conclude that unfixed initial expectations generally improve model performance. However, this improvement was less for models with autocorrelation. It also appears that there is overlap in the behavioural characteristics that autocorrelation and unfixed initial expectations capture. This is evident by their inverse relationship in how they affect the model performance on different classes of models. Similarly, both inclusion of autocorrelation and initial expectations eliminate the learning bias in the asymmetrical models for Dataset 1.

We also investigated the difference between prediction-based and reward-based asymmetrical learning. As already mentioned, initial expectations and autocorrelation affected these learning variants differently. Furthermore, the main effect of prediction-based and reward-based asymmetrical learning on AIC and BIC scores was sometimes in the opposite direction. However, there was no clear consistency to allow meaningful interpretation of these differences. Ultimately, neither form of learning asymmetry provided better model predictions, except for the Q-learning models with autocorrelation. While there was no noticeable difference in how asymmetrical learning variants influenced model performance, the different manner in which they interacted with autocorrelation and initial estimates suggest that perhaps they have different behavioural characteristics. As mentioned earlier, there is no empirical evidence to privilege the prediction-based method, as previous studies have done. In our study, we failed to provide one either.

Overwhelmingly the evidence suggests that the learning bias found in the two-armed bandit paradigm is a statistical artifact caused by model misspecification. However, it is unclear if the observed pattern is caused by the omission of autocorrelation or by the fixing of initial expectations to zero. Both are theoretically plausible, and both improved model prediction. Furthermore, they are not mutually exclusive. It could very well be that some of the participants were expecting the reward for the initially chosen option to be either greater or less than zero, which the model with fixed expectation overcorrects by estimating a pseudo learning bias. Similarly, it could be a case of choice hysteresis where some participants learn the state-action relation without value computation. However, further research is needed to differentiate between the two.

The learning bias persisted in the IGT paradigm despite the inclusion of autocorrelation or varying initial estimates. However, it is unclear if this finding is, in fact, reflective of an actual learning bias. Firstly, we used two standards to assess learning bias on group-level and parsimony. According to the latter, there was no evidence for or against learning bias. However, since learning bias is the alternative hypothesis, it must provide significant evidence against the null hypothesis, i.e., symmetrical learning. Furthermore, model identifiability results lessen the reliability of our findings. Nevertheless, learning bias should be explored in more dynamic environments.

Furthermore, we believe the way in which we measurement learning bias should become more sophisticated. We do not believe that RPE is an excellent method to categorize positive and negative outcomes. While it is effective for computational modelling, there is no evidence that humans differentiate good and bad outcomes based on whether the RPE is positive or negative. We tested a novel method of categorizing good and bad outcomes (i.e., the reward-based method), but it did not significantly differ from the RPE method. It could also very well be that the approach to categorization based on the reward received for a single item is wrong. To demonstrate our point imagine the following scenario: if the participant had estimated the value action of an option as 5 points, and after picking that option, she receives 10 points, by both the prediction-based and the reward-based method, the value received is categorized as a good outcome. However, if the participant had picked a different option previously at the previous trial and received a value of 100, is the 10 points received at the subsequent trial indeed a positive outcome? We believe there is a need for a more nuanced and sophisticated method of differentiating between good and bad outcomes. Future research should investigate the role of other rewards received on determining whether an outcome is good or bad (as the example demonstrated) and the role of context and individual differences.

# References

Bates, D., Maechler, M., Bolker, B., Walker, S., et al. (2014). *lme4: Linear mixed-effects models using eigen and s4. r package version 1.1-7.*

Bechara, A., Damasio, A. R., Damasio, H., & Anderson, S. W. (1994). Insensitivity to future consequences following damage to human prefrontal cortex. *Cognition*, *50*(1-3), 7–15.

Bonaiuto, J. J., de Berker, A., & Bestmann, S. (2016). Response repetition biases in human perceptual decisions are explained by activity decay in competitive attractor models. *Elife*, *5*, e20047.

Chambon, V., Théro, H., Vidal, M., Vandendriessche, H., Haggard, P., & Palminteri, S. (2020). Information about action outcomes differentially affects learning from self-determined versus imposed choices. *Nature Human Behaviour*, *4*(10), 1067–1079.

Collins, A. G., & Frank, M. J. (2012). How much of reinforcement learning is working memory, not reinforcement learning? a behavioral, computational, and neurogenetic analysis. *European Journal of Neuroscience*, *35*(7), 1024–1035.

Dabney, W., Kurth-Nelson, Z., Uchida, N., Starkweather, C. K., Hassabis, D., Munos, R., & Botvinick, M. (2020). A distributional code for value in dopamine-based reinforcement learning. *Nature*, *577*(7792), 671–675.

Eckstein, M. K., Wilbrecht, L., & Collins, A. G. (2021). What do reinforcement learning models measure? interpreting model parameters in cognition and neuroscience. *Current Opinion in Behavioral Sciences*, *41*, 128–137.

Eil, D., & Rao, J. M. (2011). The good news-bad news effect: asymmetric processing of objective information about yourself. *American Economic Journal: Microeconomics*, *3*(2), 114–38.

Farrell, S., & Lewandowsky, S. (2018). *Computational modeling of cognition and behavior.* Cambridge University Press.

Frank, M. J., Moustafa, A. A., Haughey, H. M., Curran, T., & Hutchison, K. E. (2007). Genetic triple dissociation reveals multiple roles for dopamine in reinforcement learning.

*Proceedings of the National Academy of Sciences*, *104*(41), 16311–16316.

Frank, M. J., Seeberger, L. C., & O'reilly, R. C. (2004). By carrot or by stick: cognitive reinforcement learning in parkinsonism. *Science*, *306*(5703), 1940–1943.

Fridberg, D. J., Queller, S., Ahn, W.-Y., Kim, W., Bishara, A. J., Busemeyer, J. R., … Stout, J. C. (2010). Cognitive mechanisms underlying risky decision-making in chronic cannabis users. *Journal of mathematical psychology*, *54*(1), 28–38.

Gershman, S. J. (2015). Do learning rates adapt to the distribution of rewards? *Psychonomic bulletin & review*, *22*(5), 1320–1327.

Gillan, C. M., Kosinski, M., Whelan, R., Phelps, E. A., & Daw, N. D. (2016). Characterizing a psychiatric symptom dimension related to deficits in goal-directed control. *Elife*, *5*, e11305.

Gureckis, T. M., & Love, B. C. (2015). Computational reinforcement learning. *The Oxford handbook of computational and mathematical psychology*, 99–117.

Harada, T. (2020). Learning from success or failure?–positivity biases revisited. *Frontiers in Psychology*, *11*.

Horstmann, A., Villringer, A., & Neumann, J. (2012). Iowa gambling task: there is more to consider than long-term outcome. using a linear equation model to disentangle the impact of outcome and frequency of gains and losses. *Frontiers in Neuroscience*, *6*, 61.

Kalman, R. E. (1960). A new approach to linear filtering and prediction problems.

Katahira, K. (2018). The statistical structures of reinforcement learning with asymmetric value updates. *Journal of Mathematical Psychology*, *87*, 31–45.

Katharine M. Mullen, D. L. G. D. W. J. C., David Ardia. (2011). *Deoptim: An r package for global optimization by differential evolution.*

Kjome, K. L., Lane, S. D., Schmitz, J. M., Green, C., Ma, L., Prasla, I., … Moeller, F. G. (2010). Relationship between impulsivity and decision making in cocaine dependence. *Psychiatry research*, *178*(2), 299–304.

Lee, M. D., Zhang, S., Munro, M., & Steyvers, M. (2011). Psychological models of human and optimal performance in bandit problems. *Cognitive Systems Research*, *12*(2), 164–174.

Lefebvre, G., Lebreton, M., Meyniel, F., Bourgeois-Gironde, S., & Palminteri, S. (2017). Behavioural and neural characterization of optimistic reinforcement learning. *Nature Human Behaviour*, *1*(4), 1–9.

Lenth, R., Singmann, H., Love, J., Buerkner, P., & Herve, M. (2019). *Emmeans: Estimated marginal means, aka least-squares means. 2018; r package version 1.3. 1.*

Maia, T. V., & McClelland, J. L. (2004). A reexamination of the evidence for the somatic marker hypothesis: what participants really know in the iowa gambling task. *Proceedings of the National Academy of Sciences*, *101*(45), 16075–16080.

Marr, D. (1982). Vision: A computational investigation into the human representation and processing of visual information, henry holt and co. *Inc., New York, NY*, *2*(4.2).

Master, S. L., Eckstein, M. K., Gotlieb, N., Dahl, R., Wilbrecht, L., & Collins, A. G. (2020). Disentangling the systems contributing to changes in learning during adolescence. *Developmental cognitive neuroscience*, *41*, 100732.

Miller, K. J., Shenhav, A., & Ludvig, E. A. (2019). Habits without values. *Psychological review*, *126*(2), 292.

Niv, Y. (2009). Reinforcement learning in the brain. *Journal of Mathematical Psychology*, *53*(3), 139–154.

Niv, Y., Edlund, J. A., Dayan, P., & O'Doherty, J. P. (2012). Neural prediction errors reveal a risk-sensitive reinforcement-learning process in the human brain. *Journal of Neuroscience*, *32*(2), 551–562.

O'Doherty, J. P., Cockburn, J., & Pauli, W. M. (2017). Learning, reward, and decision making. *Annual review of psychology*, *68*, 73–100.

Palminteri, S. (2021). Choice-confirmation bias and gradual perseveration in human reinforcement learning.

Palminteri, S., Lefebvre, G., Kilford, E. J., & Blakemore, S.-J. (2017). Confirmation bias in human reinforcement learning: Evidence from counterfactual feedback processing. *PLoS computational biology*, *13*(8), e1005684.

Palminteri, S., & Pessiglione, M. (2017). Opponent brain systems for reward and punishment learning: causal evidence from drug and lesion studies in humans. In *Decision neuroscience* (pp. 291–303). Elsevier.

Premkumar, P., Fannon, D., Kuipers, E., Simmons, A., Frangou, S., & Kumari, V. (2008). Emotional decision-making and its dissociable components in schizophrenia and schizoaffective disorder: a behavioural and mri investigation. *Neuropsychologia*, *46*(7), 2002–2012.

Rabin, M., & Thaler, R. H. (2001). Anomalies: risk aversion. *Journal of Economic perspectives*, *15*(1), 219–232.

Rescorla, R. A. (1972). A theory of pavlovian conditioning: Variations in the effectiveness of reinforcement and nonreinforcement. *Current research and theory*, 64–99.

Schultz, W. (2016). Dopamine reward prediction-error signalling: a two-component response. *Nature reviews neuroscience*, *17*(3), 183–195.

Schultz, W., Tremblay, L., & Hollerman, J. R. (1998). Reward prediction in primate basal ganglia and frontal cortex. *Neuropharmacology*, *37*(4-5), 421–429.

Sharot, T., Korn, C. W., & Dolan, R. J. (2011). How unrealistic optimism is maintained in the face of reality. *Nature neuroscience*, *14*(11), 1475–1479.

Singmann, H., Bolker, B., Westfall, J., Aust, F., Højsgaard, S., Fox, J., ... Love, J. (2016). afex: analysis of factorial experiments. r package version 0.16-1. *R Package Version 0.16*, *1*.

Speekenbrink, M. (2019a). Modeling reinforcement learning (part i): Defining and simulating rl models. [blog post]. Retrieved from `https://speekenbrink-lab.github.io/modelling/2019/02/28/fit_kf_rl_1.html`

Speekenbrink, M. (2019b). Modeling reinforcement learning (part ii): Maximum likelihood estimation. [blog post]. Retrieved from `https://speekenbrink-lab.github.io/modelling/2019/08/29/fit_kf_rl_2.html`

Speekenbrink, M., & Konstantinidis, E. (2015). Uncertainty and exploration in a restless bandit problem. *Topics in cognitive science*, *7*(2), 351–367.

Steingroever, H., Fridberg, D. J., Horstmann, A., Kjome, K. L., Kumari, V., Lane, S. D., ... others (2015). Data from 617 healthy participants performing the iowa gambling task: A" many labs" collaboration. *Journal of Open Psychology Data*, *3*(1), 340–353.

Steingroever, H., Šmíra, M., Lee, M., & Pachur, T. (2015). Do intuitive and deliberate decision makers perform differently on the iowa gambling task? *Retrieved from Steingroever et al.*

Sugawara, M., & Katahira, K. (2021). Dissociation between asymmetric value updating and perseverance in human reinforcement learning. *Scientific reports*, *11*(1), 1–13.

Sutton, R. S., & Barto, A. G. (1990). Time-derivative models of pavlovian reinforcement.

Sutton, R. S., & Barto, A. G. (2018). *Reinforcement learning: An introduction.* MIT press.

Symonds, M. R., & Moussalli, A. (2011). A brief guide to model selection, multimodel inference and model averaging in behavioural ecology using akaike's information criterion. *Behavioral Ecology and Sociobiology*, *65*(1), 13–21.

Thorndike, L., & Bruce, D. (2017). *Animal intelligence: Experimental studies.* Routledge.

Wagenmakers, E.-J., & Farrell, S. (2004). Aic model selection using akaike weights. *Psychonomic bulletin & review*, *11*(1), 192–196.

Wetzels, R., Vandekerckhove, J., Tuerlinckx, F., & Wagenmakers, E.-J. (2010). Bayesian parameter estimation in the expectancy valence model of the iowa gambling task. *Journal of Mathematical Psychology*, *54*(1), 14–27.

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

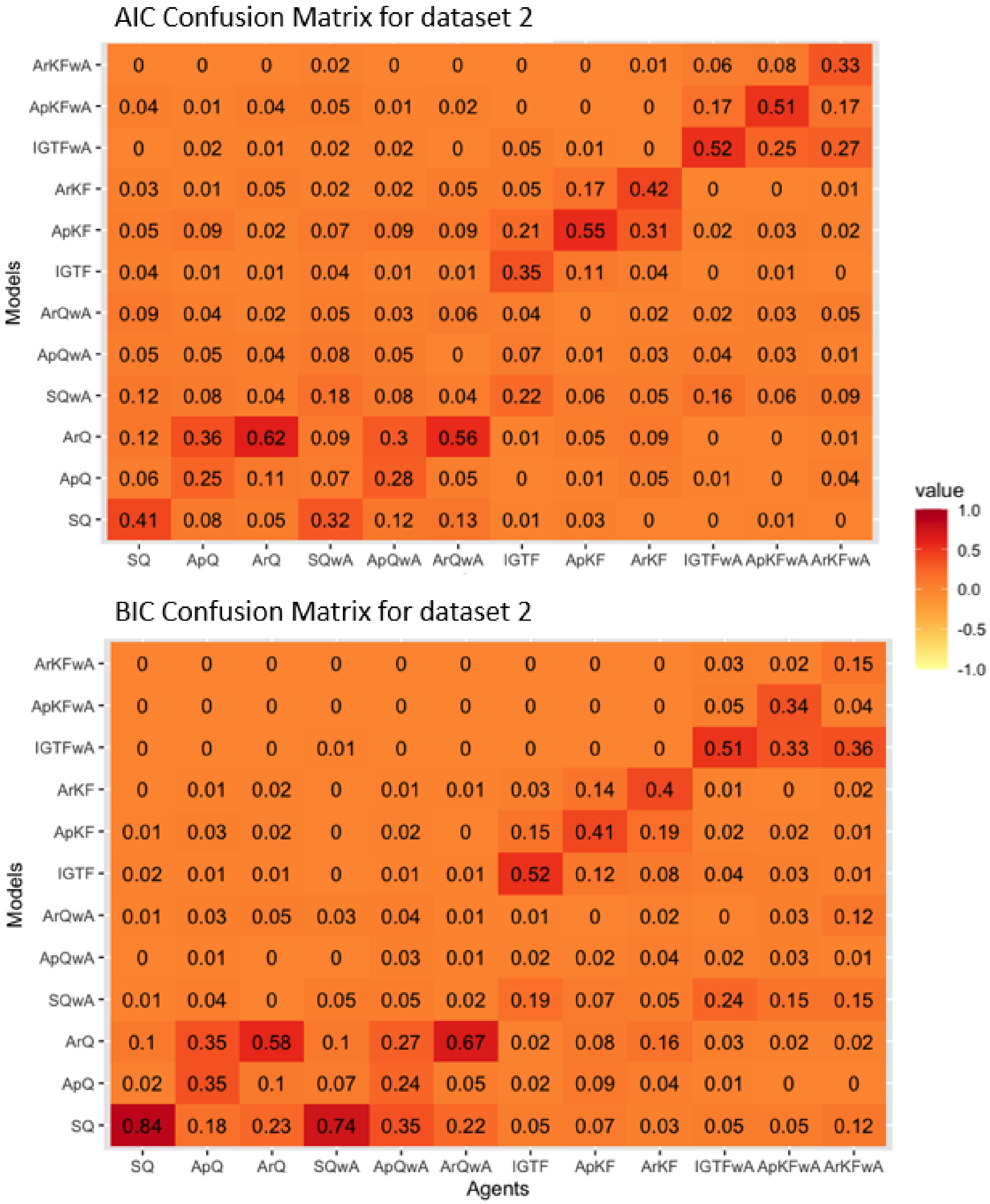
Worthy, D. A., Pang, B., & Byrne, K. A. (2013). Decomposing the roles of perseveration and expected value representation in models of the iowa gambling task. *Frontiers in* *psychology*, *4*, 640.

Wrosch, C., Bauer, I., & Scheier, M. F. (2005). Regret and quality of life across the adult life span: the influence of disengagement and available future goals. *Psychology and aging*, *20*(4), 657.

# Appendix A

**Figure 7**

*Model Identifiability for Dataset 2*



AIC Confusion Matrix for dataset 2

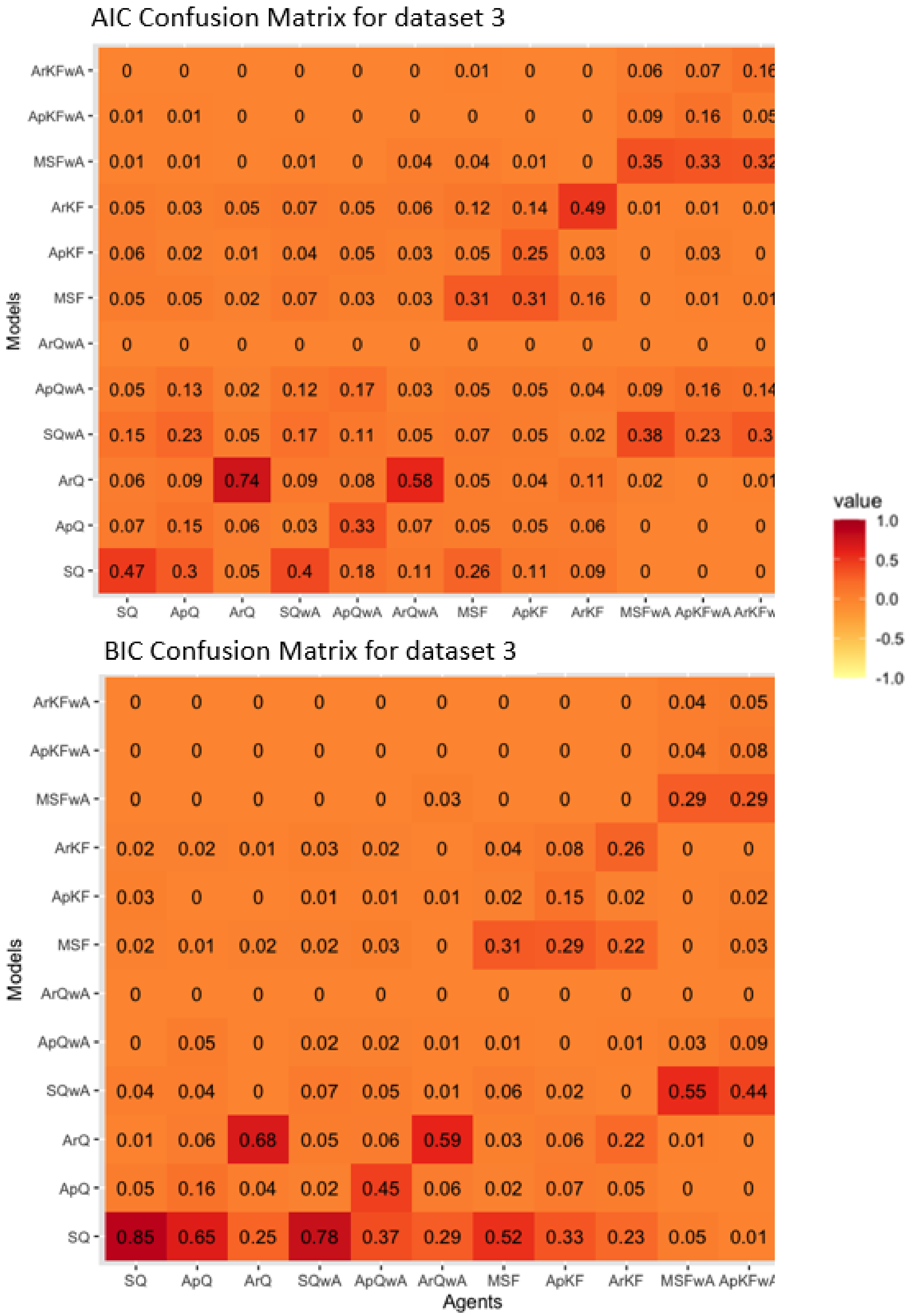
| | SQ | ApQ | ArQ | SQwA | ApQwA | ArQwA | IGTF | ApKF | ArKF | IGTFwA | ApKFwA | ArKFwA |
|---|---|---|---|---|---|---|---|---|---|---|---|---|
| ArKFwA | 0 | 0 | 0 | 0.02 | 0 | 0 | 0 | 0 | 0.01 | 0.06 | 0.08 | 0.33 |
| ApKFwA | 0.04 | 0.01 | 0.04 | 0.05 | 0.01 | 0.02 | 0 | 0 | 0 | 0.17 | 0.51 | 0.17 |
| IGTFwA | 0 | 0.02 | 0.01 | 0.02 | 0.02 | 0 | 0.05 | 0.01 | 0 | 0.52 | 0.25 | 0.27 |
| ArKF | 0.03 | 0.01 | 0.05 | 0.02 | 0.02 | 0.05 | 0.05 | 0.17 | 0.42 | 0 | 0 | 0.01 |
| ApKF | 0.05 | 0.09 | 0.02 | 0.07 | 0.09 | 0.09 | 0.21 | 0.55 | 0.31 | 0.02 | 0.03 | 0.02 |
| IGTF | 0.04 | 0.01 | 0.01 | 0.04 | 0.01 | 0.01 | 0.35 | 0.11 | 0.04 | 0 | 0.01 | 0 |
| ArQwA | 0.09 | 0.04 | 0.02 | 0.05 | 0.03 | 0.06 | 0.04 | 0 | 0.02 | 0.02 | 0.03 | 0.05 |
| ApQwA | 0.05 | 0.05 | 0.04 | 0.08 | 0.05 | 0 | 0.07 | 0.01 | 0.03 | 0.04 | 0.03 | 0.01 |
| SQwA | 0.12 | 0.08 | 0.04 | 0.18 | 0.08 | 0.04 | 0.22 | 0.06 | 0.05 | 0.16 | 0.06 | 0.09 |
| ArQ | 0.12 | 0.36 | 0.62 | 0.09 | 0.3 | 0.56 | 0.01 | 0.05 | 0.09 | 0 | 0 | 0.01 |
| ApQ | 0.06 | 0.25 | 0.11 | 0.07 | 0.28 | 0.05 | 0 | 0.01 | 0.05 | 0.01 | 0 | 0.04 |
| SQ | 0.41 | 0.08 | 0.05 | 0.32 | 0.12 | 0.13 | 0.01 | 0.03 | 0 | 0 | 0.01 | 0 |

BIC Confusion Matrix for dataset 2

| | SQ | ApQ | ArQ | SQwA | ApQwA | ArQwA | IGTF | ApKF | ArKF | IGTFwA | ApKFwA | ArKFwA |
|---|---|---|---|---|---|---|---|---|---|---|---|---|
| ArKFwA | 0 | 0 | 0 | 0 | 0 | 0 | 0 | 0 | 0 | 0.03 | 0.02 | 0.15 |
| ApKFwA | 0 | 0 | 0 | 0 | 0 | 0 | 0 | 0 | 0 | 0.05 | 0.34 | 0.04 |
| IGTFwA | 0 | 0 | 0 | 0.01 | 0 | 0 | 0 | 0 | 0 | 0.51 | 0.33 | 0.36 |
| ArKF | 0 | 0.01 | 0.02 | 0 | 0.01 | 0.01 | 0.03 | 0.14 | 0.4 | 0.01 | 0 | 0.02 |
| ApKF | 0.01 | 0.03 | 0.02 | 0 | 0.02 | 0 | 0.15 | 0.41 | 0.19 | 0.02 | 0.02 | 0.01 |
| IGTF | 0.02 | 0.01 | 0.01 | 0 | 0.01 | 0.01 | 0.52 | 0.12 | 0.08 | 0.04 | 0.03 | 0.01 |
| ArQwA | 0.01 | 0.03 | 0.05 | 0.03 | 0.04 | 0.01 | 0.01 | 0 | 0.02 | 0 | 0.03 | 0.12 |
| ApQwA | 0 | 0.01 | 0 | 0 | 0.03 | 0.01 | 0.02 | 0.02 | 0.04 | 0.02 | 0.03 | 0.01 |
| SQwA | 0.01 | 0.04 | 0 | 0.05 | 0.05 | 0.02 | 0.19 | 0.07 | 0.05 | 0.24 | 0.15 | 0.15 |
| ArQ | 0.1 | 0.35 | 0.58 | 0.1 | 0.27 | 0.67 | 0.02 | 0.08 | 0.16 | 0.03 | 0.02 | 0.02 |
| ApQ | 0.02 | 0.35 | 0.1 | 0.07 | 0.24 | 0.05 | 0.02 | 0.09 | 0.04 | 0.01 | 0 | 0 |
| SQ | 0.84 | 0.18 | 0.23 | 0.74 | 0.35 | 0.22 | 0.05 | 0.07 | 0.03 | 0.05 | 0.05 | 0.12 |

**Figure 8**

*Model Identifiability for Dataset 3*



AIC Confusion Matrix for dataset 3

| | SQ | ApQ | ArQ | SQwA | ApQwA | ArQwA | MSF | ApKF | ArKF | MSFwA | ApKFwA | ArKFw |
|---|---|---|---|---|---|---|---|---|---|---|---|---|
| ArKFwA | 0 | 0 | 0 | 0 | 0 | 0 | 0.01 | 0 | 0 | 0.06 | 0.07 | 0.16 |
| ApKFwA | 0.01 | 0.01 | 0 | 0 | 0 | 0 | 0 | 0 | 0 | 0.09 | 0.16 | 0.05 |
| MSFwA | 0.01 | 0.01 | 0 | 0.01 | 0 | 0.04 | 0.04 | 0.01 | 0 | 0.35 | 0.33 | 0.32 |
| ArKF | 0.05 | 0.03 | 0.05 | 0.07 | 0.05 | 0.06 | 0.12 | 0.14 | 0.49 | 0.01 | 0.01 | 0.01 |
| ApKF | 0.06 | 0.02 | 0.01 | 0.04 | 0.05 | 0.03 | 0.05 | 0.25 | 0.03 | 0 | 0.03 | 0 |
| MSF | 0.05 | 0.05 | 0.02 | 0.07 | 0.03 | 0.03 | 0.31 | 0.31 | 0.16 | 0 | 0.01 | 0.01 |
| ArQwA | 0 | 0 | 0 | 0 | 0 | 0 | 0 | 0 | 0 | 0 | 0 | 0 |
| ApQwA | 0.05 | 0.13 | 0.02 | 0.12 | 0.17 | 0.03 | 0.05 | 0.05 | 0.04 | 0.09 | 0.16 | 0.14 |
| SQwA | 0.15 | 0.23 | 0.05 | 0.17 | 0.11 | 0.05 | 0.07 | 0.05 | 0.02 | 0.38 | 0.23 | 0.3 |
| ArQ | 0.06 | 0.09 | 0.74 | 0.09 | 0.08 | 0.58 | 0.05 | 0.04 | 0.11 | 0.02 | 0 | 0.01 |
| ApQ | 0.07 | 0.15 | 0.06 | 0.03 | 0.33 | 0.07 | 0.05 | 0.05 | 0.06 | 0 | 0 | 0 |
| SQ | 0.47 | 0.3 | 0.05 | 0.4 | 0.18 | 0.11 | 0.26 | 0.11 | 0.09 | 0 | 0 | 0 |

BIC Confusion Matrix for dataset 3

| | SQ | ApQ | ArQ | SQwA | ApQwA | ArQwA | MSF | ApKF | ArKF | MSFwA | ApKFwA |
|---|---|---|---|---|---|---|---|---|---|---|---|
| ArKFwA | 0 | 0 | 0 | 0 | 0 | 0 | 0 | 0 | 0 | 0.04 | 0.05 |
| ApKFwA | 0 | 0 | 0 | 0 | 0 | 0 | 0 | 0 | 0 | 0.04 | 0.08 |
| MSFwA | 0 | 0 | 0 | 0 | 0 | 0.03 | 0 | 0 | 0 | 0.29 | 0.29 |
| ArKF | 0.02 | 0.02 | 0.01 | 0.03 | 0.02 | 0 | 0.04 | 0.08 | 0.26 | 0 | 0 |
| ApKF | 0.03 | 0 | 0 | 0.01 | 0.01 | 0.01 | 0.02 | 0.15 | 0.02 | 0 | 0.02 |
| MSF | 0.02 | 0.01 | 0.02 | 0.02 | 0.03 | 0 | 0.31 | 0.29 | 0.22 | 0 | 0.03 |
| ArQwA | 0 | 0 | 0 | 0 | 0 | 0 | 0 | 0 | 0 | 0 | 0 |
| ApQwA | 0 | 0.05 | 0 | 0.02 | 0.02 | 0.01 | 0.01 | 0 | 0.01 | 0.03 | 0.09 |
| SQwA | 0.04 | 0.04 | 0 | 0.07 | 0.05 | 0.01 | 0.06 | 0.02 | 0 | 0.55 | 0.44 |
| ArQ | 0.01 | 0.06 | 0.68 | 0.05 | 0.06 | 0.59 | 0.03 | 0.06 | 0.22 | 0.01 | 0 |
| ApQ | 0.05 | 0.16 | 0.04 | 0.02 | 0.45 | 0.06 | 0.02 | 0.07 | 0.05 | 0 | 0 |
| SQ | 0.85 | 0.65 | 0.25 | 0.78 | 0.37 | 0.29 | 0.52 | 0.33 | 0.23 | 0.05 | 0.01 |

# Appendix B

## Table 7

*The AIC and BIC results for Dataset 1.*

| | Model | $\Delta AIC$ | $\Delta AIC$ | $w(AIC)$ | $w(AIC)$ | $n(AIC)$ | $\Delta(BIC)$ | $\Delta(BIC)$ | $w(BIC)$ | $w(BIC)$ | $n(BIC)$ |
|---|---|---|---|---|---|---|---|---|---|---|---|
| | | Mean | SD | Mean | SD | | Mean | SD | Mean | SD | |
| Fixed initial expectations | SQ | 10.97 | 15.19 | 0.12 | 0.15 | 28.00 | 6.94 | 13.24 | 0.40 | 0.35 | 74 |
| | ApQ | 9.90 | 13.80 | 0.09 | 0.10 | 13.00 | 9.12 | 12.09 | 0.10 | 0.12 | 10 |
| | ArQ | 9.86 | 13.52 | 0.09 | 0.10 | 3.00 | 9.08 | 11.79 | 0.10 | 0.12 | 4 |
| | SQwA | 2.16 | 2.95 | 0.25 | 0.22 | 56.00 | 4.63 | 4.13 | 0.29 | 0.36 | 49 |
| | ApQwA | 2.12 | 1.87 | 0.20 | 0.16 | 20.00 | 7.83 | 4.38 | 0.06 | 0.11 | 4 |
| | ArQwA | 3.33 | 3.38 | 0.14 | 0.16 | 17.00 | 9.04 | 4.52 | 0.04 | 0.08 | 2 |
| | SFK | 15.99 | 14.81 | 0.01 | 0.01 | 0.00 | 21.69 | 12.90 | 0.00 | 0.00 | 0 |
| | ApKF | 14.77 | 12.91 | 0.01 | 0.03 | 0.00 | 23.73 | 11.17 | 0.00 | 0.00 | 0 |
| | ArKF | 14.74 | 12.95 | 0.01 | 0.03 | 1.00 | 23.69 | 11.21 | 0.00 | 0.00 | 0 |
| | SFKwA | 7.21 | 3.56 | 0.03 | 0.06 | 2.00 | 19.41 | 4.61 | 0.00 | 0.01 | 0 |
| | ApKFwA | 7.72 | 3.52 | 0.02 | 0.04 | 2.00 | 23.16 | 4.55 | 0.00 | 0.00 | 0 |
| | ArKFwA | 7.73 | 3.32 | 0.02 | 0.04 | 1.00 | 23.17 | 4.38 | 0.00 | 0.00 | 0 |
| Unfixed initial expectations | SQ | 15.83 | 22.91 | 0.10 | 0.15 | 26.00 | 11.28 | 21.33 | 0.36 | 0.35 | 62 |
| | ApQ | 10.90 | 14.41 | 0.08 | 0.10 | 4.00 | 9.60 | 12.79 | 0.11 | 0.13 | 10 |
| | ArQ | 10.92 | 14.44 | 0.09 | 0.10 | 9.00 | 9.62 | 12.82 | 0.11 | 0.13 | 8 |
| | SQwA | 2.37 | 3.21 | 0.27 | 0.23 | 51.00 | 4.31 | 4.14 | 0.32 | 0.36 | 53 |
| | ApQwA | 2.00 | 1.88 | 0.25 | 0.20 | 34.00 | 7.19 | 4.45 | 0.08 | 0.13 | 6 |
| | ArQwA | 7.38 | 8.96 | 0.10 | 0.17 | 13.00 | 12.57 | 8.36 | 0.03 | 0.08 | 4 |
| | SFK | 20.62 | 23.20 | 0.01 | 0.01 | 0.00 | 25.81 | 21.68 | 0.00 | 0.00 | 0 |
| | ApKF | 16.90 | 15.00 | 0.01 | 0.03 | 1.00 | 25.34 | 13.35 | 0.00 | 0.00 | 0 |
| | ArKF | 16.89 | 14.69 | 0.01 | 0.03 | 0.00 | 25.33 | 13.03 | 0.00 | 0.00 | 0 |
| | SFKwA | 7.02 | 3.54 | 0.04 | 0.07 | 3.00 | 18.70 | 4.64 | 0.00 | 0.01 | 0 |
| | ApKFwA | 7.56 | 3.43 | 0.03 | 0.04 | 1.00 | 22.49 | 4.53 | 0.00 | 0.00 | 0 |
| | ArKFwA | 7.64 | 3.44 | 0.03 | 0.04 | 1.00 | 22.57 | 4.55 | 0.00 | 0.00 | 0 |

Values of n() are the total number of participants best fit for the model

**Table 8**

*The AIC and BIC results for Dataset 2.*

| | Model | $\Delta AIC$ | $\Delta AIC$ | $w(AIC)$ | $w(AIC)$ | $n(AIC)$ | $\Delta(BIC)$ | $\Delta(BIC)$ | $w(BIC)$ | $w(BIC)$ | $n(BIC)$ |
|---|---|---|---|---|---|---|---|---|---|---|---|
| | | Mean | SD | Mean | SD | | Mean | SD | Mean | SD | |
| Fixed initial expectations | SQ | 46.37 | 31.33 | 0.03 | 0.11 | 1.00 | 38.68 | 29.63 | 0.09 | 0.24 | 2.00 |
| | ApQ | 43.27 | 28.55 | 0.02 | 0.05 | 0.00 | 38.19 | 26.90 | 0.03 | 0.06 | 0.00 |
| | ArQ | 34.83 | 27.11 | 0.07 | 0.18 | 2.00 | 29.75 | 24.79 | 0.13 | 0.27 | 3.00 |
| | SQwA | 9.19 | 8.20 | 0.08 | 0.14 | 1.00 | 6.71 | 8.31 | 0.25 | 0.33 | 5.00 |
| | ApQwA | 7.51 | 7.39 | 0.15 | 0.27 | 3.00 | 7.64 | 7.49 | 0.12 | 0.22 | 2.00 |
| | ArQwA | 14.23 | 14.86 | 0.20 | 0.31 | 5.00 | 14.36 | 13.34 | 0.15 | 0.32 | 3.00 |
| | SFK | 46.79 | 30.65 | 0.02 | 0.08 | 1.00 | 46.91 | 29.58 | 0.03 | 0.15 | 1.00 |
| | ApKF | 48.35 | 33.26 | 0.02 | 0.06 | 0.00 | 51.08 | 31.96 | 0.00 | 0.02 | 0.00 |
| | ArKF | 40.79 | 26.85 | 0.01 | 0.03 | 0.00 | 43.52 | 25.70 | 0.00 | 0.02 | 0.00 |
| | SFKwA | 7.24 | 6.15 | 0.16 | 0.22 | 3.00 | 12.57 | 9.28 | 0.13 | 0.26 | 4.00 |
| | ApKFwA | 8.50 | 5.79 | 0.08 | 0.12 | 1.00 | 16.44 | 9.04 | 0.02 | 0.03 | 0.00 |
| | ArKFwA | 7.40 | 5.96 | 0.16 | 0.26 | 3.00 | 15.34 | 9.26 | 0.03 | 0.08 | 0.00 |
| Unfixed initial expectations | SQ | 45.53 | 26.87 | 0.00 | 0.00 | 0.00 | 37.20 | 25.88 | 0.01 | 0.03 | 0.00 |
| | ApQ | 43.29 | 28.19 | 0.03 | 0.11 | 1.00 | 37.57 | 26.87 | 0.05 | 0.18 | 1.00 |
| | ArQ | 36.12 | 26.27 | 0.05 | 0.15 | 2.00 | 30.40 | 24.78 | 0.11 | 0.25 | 3.00 |
| | SQwA | 7.13 | 6.68 | 0.13 | 0.20 | 3.00 | 4.02 | 5.04 | 0.42 | 0.39 | 10.00 |
| | ApQwA | 7.83 | 6.60 | 0.07 | 0.10 | 0.00 | 7.32 | 4.72 | 0.06 | 0.06 | 0.00 |
| | ArQwA | 15.50 | 15.09 | 0.16 | 0.28 | 3.00 | 14.99 | 13.86 | 0.14 | 0.30 | 3.00 |
| | SFK | 45.23 | 29.08 | 0.01 | 0.03 | 0.00 | 44.72 | 28.15 | 0.00 | 0.02 | 0.00 |
| | ApKF | 43.46 | 26.13 | 0.00 | 0.01 | 0.00 | 45.56 | 25.05 | 0.00 | 0.00 | 0.00 |
| | ArKF | 39.44 | 25.67 | 0.00 | 0.01 | 0.00 | 41.54 | 24.57 | 0.00 | 0.00 | 0.00 |
| | SFKwA | 6.31 | 4.86 | 0.11 | 0.15 | 2.00 | 11.02 | 6.64 | 0.09 | 0.23 | 2.00 |
| | ApKFwA | 6.51 | 5.40 | 0.14 | 0.25 | 2.00 | 13.82 | 7.75 | 0.06 | 0.21 | 1.00 |
| | ArKFwA | 4.27 | 5.03 | 0.28 | 0.30 | 7.00 | 11.58 | 7.58 | 0.04 | 0.05 | 0.00 |

Values of n() are the total number of participants best fit for the model

**Table 9**

*The AIC and BIC results for Dataset 3.*

| | Model | $\Delta AIC$ | $\Delta AIC$ | $w(AIC)$ | $w(AIC)$ | $n(AIC)$ | $\Delta(BIC)$ | $\Delta(BIC)$ | $w(BIC)$ | $w(BIC)$ | $n(BIC)$ |
|---|---|---|---|---|---|---|---|---|---|---|---|
| | | Mean | SD | Mean | SD | | Mean | SD | Mean | SD | |
| Fixed initial expectations | SQ | 46.37 | 31.33 | 0.03 | 0.11 | 1.00 | 38.68 | 29.63 | 0.09 | 0.24 | 2.00 |
| | ApQ | 43.27 | 28.55 | 0.02 | 0.05 | 0.00 | 38.19 | 26.90 | 0.03 | 0.06 | 0.00 |
| | ArQ | 34.83 | 27.11 | 0.07 | 0.18 | 2.00 | 29.75 | 24.79 | 0.13 | 0.27 | 3.00 |
| | SQwA | 9.19 | 8.20 | 0.08 | 0.14 | 1.00 | 6.71 | 8.31 | 0.25 | 0.33 | 5.00 |
| | ApQwA | 7.51 | 7.39 | 0.15 | 0.27 | 3.00 | 7.64 | 7.49 | 0.12 | 0.22 | 2.00 |
| | ArQwA | 14.23 | 14.86 | 0.20 | 0.31 | 5.00 | 14.36 | 13.34 | 0.15 | 0.32 | 3.00 |
| | SFK | 46.79 | 30.65 | 0.02 | 0.08 | 1.00 | 46.91 | 29.58 | 0.03 | 0.15 | 1.00 |
| | ApKF | 48.35 | 33.26 | 0.02 | 0.06 | 0.00 | 51.08 | 31.96 | 0.00 | 0.02 | 0.00 |
| | ArKF | 40.79 | 26.85 | 0.01 | 0.03 | 0.00 | 43.52 | 25.70 | 0.00 | 0.02 | 0.00 |
| | SFKwA | 7.24 | 6.15 | 0.16 | 0.22 | 3.00 | 12.57 | 9.28 | 0.13 | 0.26 | 4.00 |
| | ApKFwA | 8.50 | 5.79 | 0.08 | 0.12 | 1.00 | 16.44 | 9.04 | 0.02 | 0.03 | 0.00 |
| | ArKFwA | 7.40 | 5.96 | 0.16 | 0.26 | 3.00 | 15.34 | 9.26 | 0.03 | 0.08 | 0.00 |
| Unfixed initial expectations | SQ | 45.53 | 26.87 | 0.00 | 0.00 | 0.00 | 37.20 | 25.88 | 0.01 | 0.03 | 0.00 |
| | ApQ | 43.29 | 28.19 | 0.03 | 0.11 | 1.00 | 37.57 | 26.87 | 0.05 | 0.18 | 1.00 |
| | ArQ | 36.12 | 26.27 | 0.05 | 0.15 | 2.00 | 30.40 | 24.78 | 0.11 | 0.25 | 3.00 |
| | SQwA | 7.13 | 6.68 | 0.13 | 0.20 | 3.00 | 4.02 | 5.04 | 0.42 | 0.39 | 10.00 |
| | ApQwA | 7.83 | 6.60 | 0.07 | 0.10 | 0.00 | 7.32 | 4.72 | 0.06 | 0.06 | 0.00 |
| | ArQwA | 15.50 | 15.09 | 0.16 | 0.28 | 3.00 | 14.99 | 13.86 | 0.14 | 0.30 | 3.00 |
| | SFK | 45.23 | 29.08 | 0.01 | 0.03 | 0.00 | 44.72 | 28.15 | 0.00 | 0.02 | 0.00 |
| | ApKF | 43.46 | 26.13 | 0.00 | 0.01 | 0.00 | 45.56 | 25.05 | 0.00 | 0.00 | 0.00 |
| | ArKF | 39.44 | 25.67 | 0.00 | 0.01 | 0.00 | 41.54 | 24.57 | 0.00 | 0.00 | 0.00 |
| | SFKwA | 6.31 | 4.86 | 0.11 | 0.15 | 2.00 | 11.02 | 6.64 | 0.09 | 0.23 | 2.00 |
| | ApKFwA | 6.51 | 5.40 | 0.14 | 0.25 | 2.00 | 13.82 | 7.75 | 0.06 | 0.21 | 1.00 |
| | ArKFwA | 4.27 | 5.03 | 0.28 | 0.30 | 7.00 | 11.58 | 7.58 | 0.04 | 0.05 | 0.00 |

Values of n() are the total number of participants best fit for the model