EBERHARD KARLS
UNIVERSITÄT
TÜBINGEN

MPI for Biological Cybernetics, Computational Principles of Intelligence
Graduate School of Neural Information Processing
University of Tübingen

Uncertainty Preferences Change From Exploration to Exploitation

Alica Leonie Guzmán

Supervision: Kristin Witte, Dr. Mirko Thalmann and Dr. Eric Schulz
Second Reader: Prof. Dr. Kou Murayama

Abstract.
When making decisions in everyday life we constantly face the explore-exploit dilemma: should we try out something new or take something that we already know and like? This dilemma has been studied extensively using cognitive tasks such as multi-armed bandits. While research in cognitive science suggests that people seek uncertainty during exploration, research in economics mostly finds uncertainty avoidance. We suggest that these findings are not mutually exclusive, but that uncertainty preferences change over time as individuals move from exploration to exploitation. Recent literature leads us to expect that uncertainty seeking turns into uncertainty aversion during learning for gains. In the loss domain, we expect a continued preference for uncertain options, given evidence from the economic literature of increased uncertainty seeking in the loss domain. To investigate these hypotheses, we introduce a new experimental design in which we control for differences in variances and means in a two-armed bandit task that includes both gains and losses. We manipulate people's uncertainty about the arms by inducing irreducible variances and thereby irreducible variance differences between arms. We find that people use uncertainty more in exploration than in exploitation, switching on average from uncertainty seeking to uncertainty aversion, with high individual variability. Surprisingly, this trend is mainly driven by behaviour in the loss domain. However, in line with our expectations, the modelling results support the notion of increased uncertainty seeking for losses relative to gains.

Submission date: 11-03-2024

Contents

1 Introduction

In our everyday lives, we are constantly faced with the explore-exploit dilemma (cf. Cohen et al., 2007; Daw et al., 2006; Hills et al., 2015; Sutton & Barto, 2018): when deciding what kind of ice cream to get at our favourite ice-cream parlour, when looking for a holiday apartment, or simply when buying groceries in the shop, should we try something new or take something we already know (and like)?

This dilemma has been studied extensively from a variety of approaches using different types of cognitive tasks and models of reinforcement learning. Probably the most commonly used task to investigate explore-exploit decisions is the multi-armed bandit, which is a controllable variation of a casino slot machine. In a classical multi-armed bandit task, participants have to maximise their cumulative rewards over a certain number of trials. On each trial, they choose one of several options, often referred to as arms of the multi-armed bandit. When a choice is made, a reward is drawn from the underlying reward distribution, the properties of which can be specified by the experimenter. While static reward distributions (cf. Cogliati Dezza et al., 2019; Gershman, 2018; Wilson, Geana, et al., 2014), which do not change over time, provide a simple and stable framework for analysing learning and decision making, drifting distributions (cf. Daw et al., 2006; Speekenbrink & Konstantinidis, 2015), which change over time, are studied to capture how people adapt to changing environments. Since optimal policies are only known for very simple versions of the tasks (Cohen et al., 2007) with known reward distributions and few arms, one needs a strategy to maximise cumulative rewards by balancing exploration and exploitation.

It has been shown that people take uncertainty into account when deciding when to explore (Knox et al., 2012). In addition, cognitive experiments have shown that people actually use uncertainty strategically to guide exploration. Specifically, individuals tend to bias their choices in favour of the more informative option in order to reduce their uncertainty (Hogeveen et al., 2022; Speekenbrink, 2022) about that option. In other words, people may have a preference for an option that they have chosen less often in order to gain information about that option. By making such choices, people actively seek out uncertainty as part of the learning process. This strategy is known as directed exploration (Gershman, 2018; Schulz & Gershman, 2019; Wilson, Bonawitz, et al., 2021; Wilson, Geana, et al., 2014). The literature distinguishes directed exploration from so-called random exploration, which in contrast is independent of relative uncertainty and can be viewed as decision noise added to each decision, describing a more general form of behavioural variability.

In contrast to the evidence that people seek out uncertainty when using directed exploration is the literature on decision making in economics. In economics, the uncertainty of an outcome is also referred to as risk (Aven & Renn, 2009). Economists find that rather than seeking out uncertainty, people tend to avoid uncertain or risky options when it comes to gains, and only seek out uncertainty when it comes to losses (Hertwig & Erev, 2009; Kahneman & Tversky, 1979; Platt & Huettel, 2008). Uncertainty aversion is found not only in decisions presented by description, but also in decisions under risk arising from experience (description-experience gap cf. Hertwig, Barron, et al., 2004; Hertwig & Erev, 2009) by sampling rewards over successive trials (Niv et al., 2012). A decision from description refers to a choice between several options, where each option is given by probabilities that certain outcomes will occur, whereas in decisions from experience these probabilities have to be learned. In this vein, in the risk literature, risk can be further distinguished from ambiguity, where risk refers to known outcome probabilities, while ambiguity refers to unknown ones. This means that in the context of our work, we are more specifically talking about ambiguity aversion (Platt & Huettel, 2008).

A major difference between the risk literature and the exploration literature is that sampling paradigms of decisions under risk, typically feature binary rewards and an uncertainty free safe option, where the risk of the uncertain alternative is quickly identifiable. This makes the reducible uncertainties very small. This could reduce the need for exploration to estimate outcome uncertainties compared to more complex exploration tasks and thus induce a faster entry into exploitation. Accordingly, divergent findings on uncertainty preferences could be related to the trade-off between exploration and exploitation. This idea is also consistent with findings from the horizon task (Wilson, Geana, et al., 2014; Zaller et al., 2021), a cognitive task that disentangles random and directed exploration. After four forced-choice trials, which are used to manipulate the information participants have about each arm, they have one or six more free choices (the number of trials is called horizon), to manipulate the future prospect. Assuming the means of the two arms observed over the forced choice trials are the same, participants tend to choose the option with higher reducible uncertainty more often than chance when they have the future prospect of five more trials (exploration), while they tend to choose the more certain option more often than chance when they have no future prospect (exploitation).
We argue that directed exploration and uncertainty avoidance are not mutually exclusive. Exploration can still be directed towards uncertain options, while exploitation is uncertainty avoidant. Thus, we hypothesise that people seek uncertainty in the beginning of exploration and switch uncertainty preferences when it comes to exploitation, so that uncertainty preferences change during learning. Therefore, we would expect in an explore-exploit setting that uncertainty preferences decrease with successive trials. This trend could also depend on the reward domain, as will be explained below.
In economics, the domain of loss is well studied. Prospect theory postulates that people seek uncertainty for losses, while they are uncertainty averse to gains, which is called the reflection effect and will be explained in the following. Assuming a convex value function for losses and a concave value function for gains, as proposed by prospect theory (Kahneman & Tversky, 1979), higher variances shift the value of both gains and losses towards zero. This favours uncertainty seeking for losses and uncertainty aversion for gains, as we want to minimise losses and maximise gains.
Similarly, research on learning and decision making in the cognitive sciences finds complementary results when studying uncertainty preferences. Charpentier et al. (2017) find a greater tolerance for uncertainty in losses compared to gains. In line with this, Krueger et al. (2017) find increased uncertainty seeking for losses in comparison to gains in the horizon task. The authors argue that this uncertainty bias leaves directed exploration unaffected and should be seen as baseline uncertainty seeking arising from Bayesian shrinkage that is driven by a prior that is optimistic for losses and pessimistic for gains. This Bayesian shrinkage hypothesis provides a different explanation than prospect theory, but both lead to the same predictions. According to both two theoretical approaches, we expect that people will continue to seek uncertainty as they move from exploration to exploitation for losses, while they will turn from uncertainty seeking to aversion for gains.

To shed light on how uncertainty preferences are affected by moving from exploration to exploitation for both gains and losses, we introduce an experimental paradigm consisting of a two-armed bandit task using stationary but unknown reward distributions, studying gains and losses in separate blocks. We manipulated estimates of perceived uncertainty using a 2x2 factorial design with equal and unequal means and variances. In contrast to previous two-armed bandit tasks (Gershman, 2018; Wilson, Geana, et al., 2014), we studied a comparatively large number of trials (20) to increase the chances that people converge to exploitation.

In our analysis, we first tested whether people biased their decision towards the uncertain arm by comparing a number of standard models. We then fitted each participant with a Kalman Filter (Kalman, 1960) and different variants of the Upper Confidence Bound (Auer et al., 2002) choice rule to test how uncertainty preferences change over the course of learning. Our results suggest, that people seek uncertainty during exploration but avoid it during exploitation. Participants preferred uncertainty especially in early trials (exploration) and less in later trials (exploitation), with high individual variability. We also found increased choice determinism over the course of learning, consistent with increased random exploration in exploratory trials compared to exploitative trials. Comparing gains and losses revealed that people preferred the arm with higher uncertainty for gains but avoided it for losses. However, comparing model estimates for early and late trials suggests that, contrary to these initial observations, people exhibit increased uncertainty seeking for losses compared to gains, aligning with our predictions from the literature.

2 Methods

2.1 Participants

We collected data from 102 participants online at prolific (www.prolific.com). Inclusion criteria were that participants' primary language was English and that they were between 18 and 45 years old. In our analyses, we included all participants who passed at least three of the four attention checks administered in the study and who reported not cheating on the bandit task in final questionnaires. One participant did not meet the inclusion criteria and was therefore excluded from the analysis. This left 101 participants, 49 female and 52 male. The mean age was (26.90 ± 0.55) years. The mean time for the whole experiment was $47\,\mathrm{min}$ and $23\,\mathrm{min}$ for the bandit task. Participants were paid $6\,£$ for their participation. A bonus of a maximum of $5\,£$ was paid in addition to the base payment, depending on performance on the working memory span task and the bandit task. The average bonus payment was $(3.46 \pm 0.05)\,£$.

2.2 Task

Participants performed three tasks in the following order: operation span, two armed bandit, and questionnaires. Due to the time constraints of this thesis, we focus on the two-armed bandit task, which was presented as the second task for all participants. The other two tasks (operation span and psychiatric questionnaires) are described in more detail in appendix A.2 and contributed to the experimental design.

The two armed bandit tasks consisted of two blocks of one practice and 16 test games of 20 trials each. On every trial, participants had to choose between two different slot machines represented by green squares. The chosen machine gave feedback on a number of points drawn from its underlying Gaussian reward distribution, which remained constant over the course of a game. Participants were told that the average reward for each arm was stationary within each game, and that the rewards would be drawn with different amounts of noise in each game. They were shown which trial they were on in each round, how many games were left, and a counter that tracked points within the same block. We paid a bonus based on their overall performance. The understanding of this knowledge was ensured with a comprehension check.

The two different blocks represented different reward conditions (gain and loss) and their order was counterbalanced across participants to control for fatigue and practice effects. Out of the 101 included participants, 50 started with the gain block and 51 with the loss block.

Figure 1: Screenshot of the two-armed bandit task. Each machine is represented by one green square. Left: Game 1, Trial 11, before a choice is made. Right: Game 1, Trial 11, after selecting the right machine.

Within each block, the 16 test games spanned a 2x2 factorial design of mean and variance differences, resulting in the 4 conditions: equal variances, equal means; equal variances, unequal means; unequal variances, equal means and unequal variances, unequal means, each consisting of 4 games.
A screenshot of the task is shown in fig. 1.

2.3 Procedure and Design

Procedure. To select one of the two machines shown in fig. 1, participants clicked either the left or right arrow key on the keyboard. After selecting a machine, they received feedback for 600 ms showing the number of points scored on the selected machine before the next trial began. To motivate participants and provide cumulative feedback, we displayed a point counter that was reset for each block. While all points in the gain block were positive and added to a point counter, all points in the loss block were negative and subtracted from the start counter. The start counter in the loss block was set to the maximum number of points that could be lost, and to zero in the gain block.

Design. To examine individual differences in exploration and exploitation behaviour and how these are mediated by working memory span, anxiety and depression (see appendix A.2), we decided for a fully within-subjects design, when designing the experiment. Therefore, we aimed to introduce as little variance as possible between participants and to reduce order and sequence effects (Dale & Arnell, 2013; Goodhew & Edwards, 2019; Hedge et al., 2018). We therefore fixed the game order and the reward sequences for all participants so that if two participants clicked the left and right arm the same number of times during a game (regardless of order), they would receive the same rewards. Therefore, the rewards for each game were drawn prior to running the experiment and all participants encountered exactly the same stationary Gaussian reward distributions for each game.
The Gaussian distributions for sampling were constructed as follows. For the equal mean conditions, the generative means were identical. For the unequal mean conditions, we observed during piloting that a mean difference of 8 well separated performance in early and late trials. Thus, for the unequal mean conditions, the difference was set to 8. We controlled that the empirical deviation from the generative mean was maximal 1 after 6, 10, and 20 trials to avoid means that were very different from the intended means. One mean was randomly drawn from a uniform distribution between 40 plus the mean difference (either 0 or 8) and 60 and randomly

assigned to one of the arms. The mean of the other arm was determined by subtracting the mean difference of that game from the first mean, so that all means were between 40 and 60.
Generative variances were set to 8 and 40 for unequal variances and either both 8 or both 40 for equal variances, both counterbalanced. The variances were based on previous piloting. We controlled that the empirical deviation from the generative variances was no more than 10 after 6, and 2 after 10 and 20 trials.
For better comparability, we kept the generative means and variances for both gain and loss blocks identical for all games, with the exception of the sign. However, to avoid learning and memory effects, we shuffled the order of the games and drew different reward sequences for both blocks. The conditions within a block were interleaved, but all participants encountered the games within each block in exactly the same order to minimise sequence and order effects between participants.
Means, variances and rewards were drawn in `JavaScript`. The task was coded using `HTML 5`, `CSS` and `JavaScript`.

2.4 Modelling Learning and Decision Making

We modelled human behaviour in the two-armed bandit task to investigate how people change their uncertainty preferences as they learn.
To model human behaviour during learning, one needs to choose a learning rule that captures updates to expectation estimates, and a decision rule that makes a decision based on those estimates. We modelled learning using learning rules commonly used to fit human data in multi-armed bandits (Wilson & Collins, 2019): the Kalman Filter (Kalman, 1960) and the Delta Rule (Rescorla & Wagner, 1972). For decision making, we used different variants of the softmax choice rule.
As a prerequisite for investigating the use of uncertainty over the course of learning, we first tested whether people actually use uncertainty to bias their decision during learning by comparing models that incorporate uncertainty with models that do not. We then compared models that incorporate uncertainty during learning to test different assumptions about how people use uncertainty during learning.
The fitted learning and decision rules will be described in the following paragraphs. All models were implemented in `python 3.9.4`.

Kalman Filter. The Kalman Filter (Kalman, 1960) or also called Bayesian learner or ideal observer, is a model-based reinforcement learning algorithm that is designed to estimate the dynamics of systems under the assumption that these dynamics can be modelled linearly. It further assumes that rewards are drawn from Gaussian distributions and that the means of these Gaussian distributions change over time according to a stochastic process with Gaussian process noise (Speekenbrink & Konstantinidis, 2015). It tracks the expectations of both arms, as well as uncertainty estimates (Danwitz et al., 2022), and updates them by computing the posteriors trial by trial by Bayesian optimal integration of the observation on that trial and its priors. The learning rate changes with these estimates and is updated on each trial.
For simplicity, we fixed exact values for both the process noise and the variance of the Gaussian sampling distribution in the fitting procedure. In our case, we had stationary arms, which was shared knowledge with the participants, so we assumed zero process noise. The trial-wise learning rate is called *Kalman Gain* with $k_{\mathrm{j}}(t) \in [0, 1]$ for arm $j = 0, 1$ on trial $t = 1, \ldots, 20$.

Under our assumptions its updates are given by

$$k_{\mathrm{j}}(t) = \begin{cases} \frac{v_{\mathrm{j}}(t)}{v_{\mathrm{j}}(t)+\sigma^2_{\mathrm{j,\,obs}}} & \text{if } c(t) = j \\ 0 & \text{otherwise.} \end{cases} \tag{1}$$

$c(t)$ denotes the choice (binary in our case) on trial t, $v_{\mathrm{j}}(t)$ denotes the variance estimate of arm j on trial t and $\sigma^2_{\mathrm{j,\,obs}}$ the observation variance which we set to the irreducible variance of arm j. The mean $m_{\mathrm{j}}(t)$ and variance $v_{\mathrm{j}}(t)$ estimation updates on trial t are given by

$$m_{\mathrm{j}}(t) = m_{\mathrm{j}}(t-1) + k_{\mathrm{j}}(t) \cdot (R(t) - m_{\mathrm{j}}(t-1)) \tag{2}$$

and

$$v_{\mathrm{j}}(t) = (1 - k_{\mathrm{j}}(t)) \cdot v_{\mathrm{j}}(t-1). \tag{3}$$

$R(t)$ is the reward drawn on trial t.
The initial values of the mean and variance estimates were set to $m_{\text{gain,j}}(0) = 50$, $m_{\text{loss,j}}(0) = -50$ and $v_{\text{gain,j}}(0) = v_{\text{loss,j}}(0) = 100$, respectively. $m_j(0)$ is the average reward observed in each trial across all games and arms within each block, as stated in the task introductions, and $v_j(0) = 100$ was used previously (see Speekenbrink and Konstantinidis (2015)).
Using fixed values for the process noise, $\sigma^2_{\mathrm{j,\,obs}}$, $m_j(0)$ and $v_j(0)$ allowed us to reduce free parameters and obtain all mean and variance estimates of a participant without parameter fitting. Previous simulations also suggested limited parameter identifiability when using free parameters in the Kalman Filter part of the modelling process.
Obtaining variance estimates for each arm allowed us to use a choice rule that incorporates uncertainty preferences. While the softmax choice policy from eq. (4) only captures the randomness of a choice and thus random exploration, we can also introduce a term accounting for uncertainty preferences with the free parameter β, called the exploration bonus. We then obtain eq. (5), which is a probabilistic version of the Upper Confidence Bound (UCB) algorithm (Auer et al., 2002). These two policies estimate the probabilities of choosing each arm $j = 0, 1$ on trial t using the value estimates $m_{\mathrm{j}}(t)$ and $v_{\mathrm{j}}(t)$ of the two arms on that trial. The only free parameter for the softmax choice rule is the inverse softmax temperature γ, which ranges between 0 and ∞ and encodes the stochasticity of the choices, with higher values indicating more deterministic choices. For the UCB we additionally fitted the exploration bonus β. While positive values of β indicate uncertainty seeking, negative values indicate uncertainty avoidance.

$$P_{\text{softmax,j}}(t) = \frac{\exp(\gamma \cdot m_{\mathrm{j}}(t))}{\sum_{k=0}^{1} \exp(\gamma \cdot m_{\mathrm{k}}(t))} \tag{4}$$

$$P_{\text{UCB,j}}(t) = \frac{\exp(\gamma \cdot (m_{\mathrm{j}}(t) + \beta \cdot \sqrt{v_{\mathrm{j}}(t)}))}{\sum_{k=0}^{1} \exp(\gamma \cdot (m_{\mathrm{k}}(t) + \beta \cdot \sqrt{v_{\mathrm{k}}(t)}))} \tag{5}$$

In this thesis, we compare different models that account for a change in uncertainty preferences during learning. We have therefore taken the classical probabilistic UCB decision rule from eq. (5), which captures uncertainty preferences, and adapted it to allow for changing uncertainty preferences over trials.
The most manual way is to separate early and late trials. To do this, we fitted a model with 4 instead of 2 free parameters, one pair each of inverse softmax temperature γ and exploration bonus β for early trials and one pair for late trials. To achieve this, we fitted the first 7 trials with γ_{early} and β_{early} and took the current mean and variance estimates $m_j(7), v_j(7)$ as the initial values for trial 8, from which we continued with the parameters γ_{late} and β_{late}. We call this variant

the *4-parameter step* variant. The threshold here was chosen post-hoc, based on observations in our behavioural analysis. Similarly, we could allow only the exploration bonus to vary between early and late trials, resulting in 3 free parameters (*3-parameter step*).
Another way to capture changing uncertainty preferences during learning is to allow the exploration bonus β to change continuously over trials. We have therefore implemented two different methods to obtain an effective exploration bonus $\beta_{\text{eff.}}$ to feed into eq. (5). These are given by eq. (6) and eq. (7) and will be referred to as *linear* and *reciprocal* variants, respectively. Whether or not we allowed an intercept term c, we ended up with models with three or two free parameters, respectively.

$$\beta_{\text{eff.}} = a_1 \cdot t + c_1 \tag{6}$$

$$\beta_{\text{eff.}} = \frac{a_2}{t} + c_2 \tag{7}$$

To test whether people use uncertainty for learning, we compared the Kalman Filter with probabilistic UCB with models that do not take uncertainty into account: the Kalman Filter with softmax choice and the Delta Rule model with softmax choice, which will be presented below.

Delta Rule. The Delta Rule is a model-free learning rule based on the Rescorla-Wagner model from Rescorla and Wagner (1972) (Danwitz et al., 2022). It is a simple reinforcement learning rule based on updating the expected value Q_j for arm j. It has one free parameter: the learning rate α, which ranges from 0 to 1 and indicates the updating strength from low to high. The update is given by

$$Q_j(t+1) = Q_j(t) + \begin{cases} \alpha \cdot \delta(t) & \text{if } j = c(t) \\ 0 & \text{otherwise.} \end{cases} \tag{8}$$

$\delta(t) = R(t) - Q_c(t)$ denotes the reward prediction error and is given by the difference between the reward $R(t)$ received on trial t and the value estimate $Q_c(t)$ on trial t and the chosen arm $c(t)$ on that trial. Note that only the value of the chosen arm is updated on a given trial. The initial values were set to $Q_{\text{gain,j}}(0) = 50$ in the gain block and $Q_{\text{loss,j}}(0) = -50$ in the loss block. Participants were informed that these values could be expected on average within each block.
We modelled the decision based on the mean estimates $m_{\text{j}}(t) = Q_{\text{j}}(t)$ using the softmax policy from eq. (4).

2.5 Model Estimation and Comparison

To fit the models to the data, we used Maximum Likelihood Estimation (MLE). MLE is a statistical method for estimating the parameters of a model by maximising the likelihood function. The likelihood function measures how likely the given data is under different parameter values, looking for parameter values that make the observed data most likely. If we denote the set of parameters as θ, then $L(\theta)$ is the likelihood function of that set of parameters, and MLE solves $\hat{\theta} = \arg\max_\theta L(\theta) \hat{=} \arg\min_\theta(-\log(L(\theta)))$. We implemented MLE in `python 3.9.4` using the `minimize` function from `scipy.optimize 1.8.1` (Virtanen et al., 2020) with default settings. We constrained the model fitting with the following bounds: $\gamma, \alpha \in [0.01, 1]$, $\beta, a_1, c_1, c_2 \in [-10, 10]$ and $a_2 \in [-1, 1]$.
To compare the models, we used the Akaike Information Criterion (AIC) and the Bayesian Information Criterion (BIC), which strike a balance between goodness of fit and model complexity. AIC and BIC are given by AIC $= -2\log(L) + 2k$ and BIC $= -2\log(L) + k\log(n)$, where k is the number of free parameters and n is the sample size. While only BIC scales with the size

of the dataset and favours simpler models to avoid overfitting, AIC is more prone to complex models, especially for larger datasets.

2.6 Further Data Analysis

Data analyses were performed in Python, version `3.9.4`.

Regression Models. Statistical analyses involved the use of linear mixed effects regression models implemented using the `mixedlm` function from `statsmodels.formula.api 0.13.2` (Seabold & Perktold, 2010). For further analysis of the regression results we also performed t-tests for single sample analyses using the `ttest_1samp` function and for repeated samples analyses using the function `ttest_rel` within the same `statsmodels` library.
Logistic mixed effects regression models were used to analyse binary outcomes, using `Lmer` from the `pymer.models 0.8.1` (Jolly, 2018), specifying 'family' as 'binomial'. This function interfaces with the `lme4 1.1.35.1` (Bates et al., 2015) from `R 4.0.5` for calculation. In all specified models, the participant's ID served as a random grouping factor to account for individual differences.
For both linear and logistic mixed effects models we started with the maximal effect structure and then removed single predictors to reach convergence. We entered all non-categorical predictors as standardised floats, and categorical predictors, mainly design factors, as centred floats, coding the unequal means condition, early trials, gain block and block order starting with the loss block as -0.5, and the equal means, late trials, loss block and block order starting with the gain block as 0.5.
For logistic mixed effects models predicting participants' better arm choices and high variance arm choices, we included the same fixed and random effects for all models, except that the mean and variance differences were changed to the respective differences of the better minus worse and high minus low variance arms. When predicting high variance arm choices for all conditions, we could not achieve convergence with random intercept when block was entered as a centred float, so in this exception we entered block as an object. The model without random intercept lost an AIC comparison with the model with categorical block predictor. However, not centering the predictors requires extra care in subsequent analyses as it leads to restrictions in the interaction effects.

Confidence Intervals. To display confidence intervals in the within-subject design, we excluded within-subject variability from the across-participant averages. To do this, we processed our data X_{pg}, compromising arm choices for all participants, games and trials, through a transformation given by eq. (9) (Cousineau, 2005).

$$Y = X_{\text{pg}} - \bar{X}_{\text{p}} + \bar{X} \tag{9}$$

where $\bar{X}_{\text{p}}$ is the participant mean and $\bar{X}$ is the overall group mean across participants and games.

3 Results

Learning, Exploration and Exploitation. To show that participants were engaged in our experimental paradigm, we first examined two behavioural measures: performance and choice stickiness. The average cumulative proportion of better arm choices (bac) across participants, which we chose as the performance measure, is represented by the thick lines in fig. 2 (left)

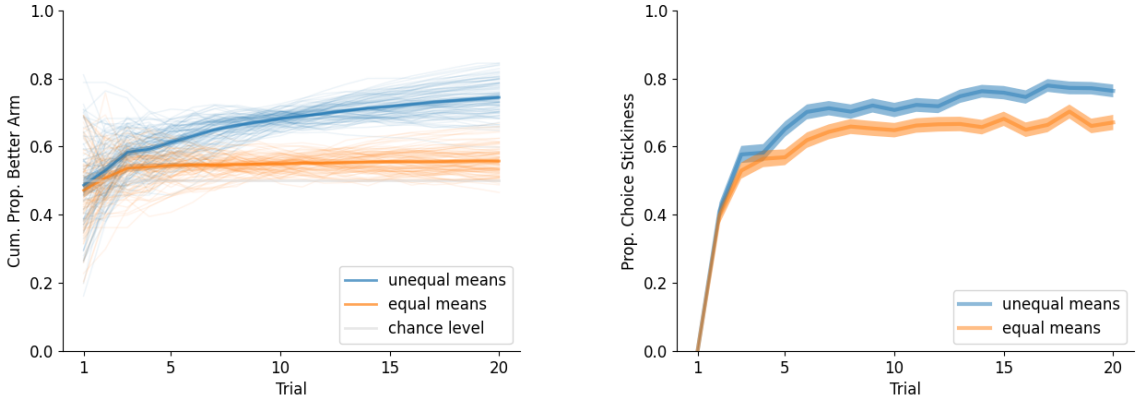

Figure 2: Left: Average cumulative proportion of better arm choices across trials for equal and unequal mean conditions. The thick line describes the average across participants and games, reported with $95\,\%$ confidence intervals in within-subject design. Thin lines show subjects' averages across all games within conditions. Right: Average choice stickiness across participants across trials for equal and unequal mean conditions with $95\,\%$ confidence intervals in within-subject design. Choice stickiness is defined as boolean value of sticking with the choice on the last trial and is averaged within participants for all games for each condition and then across subjects.

across trials. In this thesis, cumulative arm choices per trial refer to the cumulative number of choices for a given arm up to each trial divided by the number of trials.
The thin lines in fig. 2 (left) represent the average cumulative proportions of better arm choices for each individual subject, showing variability in performance and a clear separation between equal and unequal mean conditions for most participants. The cumulative proportion of better arm choices increases with the trial number, suggesting that participants learned within games. As expected, overall average performance is increased for unequal means compared to the more difficult equal mean conditions. We observe above chance performance even for equal generative means, which is enabled by a sampled mean difference below 1.
To examine whether participants used both exploration and exploitation, we looked at the average choice stickiness across participants, shown in fig. 2 (right). Choice stickiness for a trial in a game was defined as a binary number, where 1 indicates that the same arm was chosen in the previous trial and 0 indicates that the other arm was chosen in the previous trial. We obtained the average choice stickiness for each participant on each trial by averaging over all games within that condition and then averaging over all participants. We observe increasing choice stickiness across trials ($\beta = 0.112$, $p < 0.001$), indicating the switch from exploration to exploitation, approaching a value of around 0.7. On average, choice stickiness was slightly lower for equal means than for unequal means ($\beta = -0.064$, $p < 0.001$), in line with increasing task difficulty. Furthermore, we found a smaller increase in choice stickiness with trial number for equal than for unequal means ($\beta = -0.019$, $p < 0.001$), indicating that participants stuck to their choice faster for the easier unequal means condition than for the more difficult equal means condition. When comparing equal to unequal variances, we also observe increased choice stickiness for unequal variances compared to equal variances ($\beta = -0.017$, $p < 0.001$), suggesting increased exploration for equal variances (cf. appendix A.1 fig. 10). We did not find an interaction between trial number and variance condition ($\beta = -0.006$, $p = 0.121$).
Having tested that participants engaged in our task by learning, exploring and exploiting within

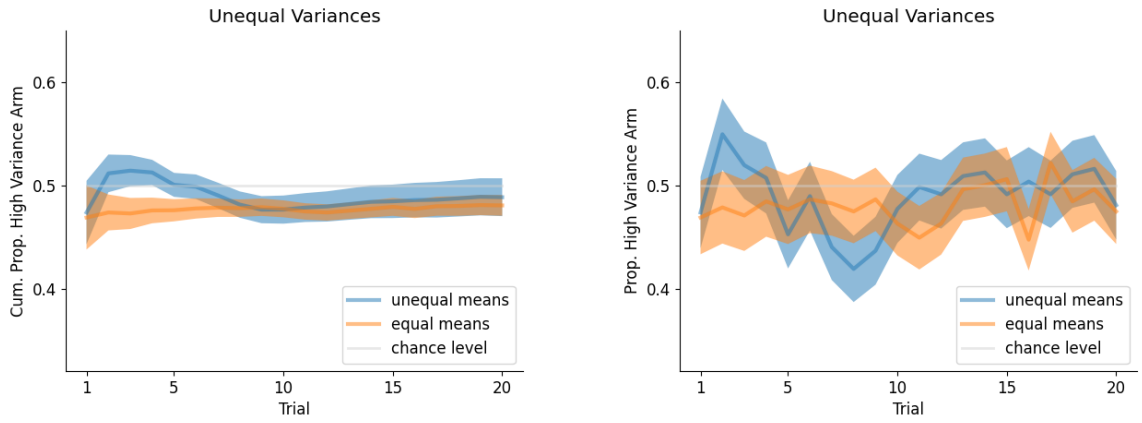

Figure 3: Left: Average cumulative proportion of high variance arm choices across trials for unequal variance conditions. The thick line describes the average across participants and games, reported with $95\,\%$ confidence intervals in within-subject design. Right: Corresponding plot for the non-cumulative measure.

games, we can now examine uncertainty preferences during learning and how these evolve across successive trials.

3.1 (How) Do Uncertainty Preferences Change When Moving From Exploration to Exploitation?

Behavioural analysis. We start with a model agnostic analysis of the data. To examine uncertainty preferences, we inspected the unequal variance conditions and adapted the mean proportion of high variance arm choices across trials as a metric, with higher values indicating a greater preference for the arm with higher uncertainty. More specifically, with high variance arm choices, we refer to the percentage of choices made for the arm with the higher empirical variance within each game.

We computed the average proportion of high variance arm choices for each trial by taking into account the results of all games for each participant and then across all participants. The average cumulative and absolute proportions of high variance arm choices for both unequal variance conditions are shown in fig. 3 left and right panels, respectively. In line with our expectations, we observe an increase in high variance arm choices in early trials and a subsequent decrease in late trials, especially for unequal means, indicating changing uncertainty preferences during learning. Towards the end of the games, participants seem to have become slightly more uncertainty averse. Note that so far we have been looking at averages across all games, including both gains and losses.

The behavioural plots suggest that uncertainty preferences do indeed change over the course of learning. To further investigate if and how uncertainty preferences would change over the course of learning, we ran simulations of model variants implementing different assumptions about how uncertainty preferences change over the course of learning (see section 2) to qualitatively compare model behaviour with behavioural data.

Simulations. To simulate behaviour in the unequal variance conditions, we first fitted different variants of the probabilistic UCB to our behavioural data to obtain parameter estimates for our simulations. To capture individual behaviour, we then simulated data for each participant

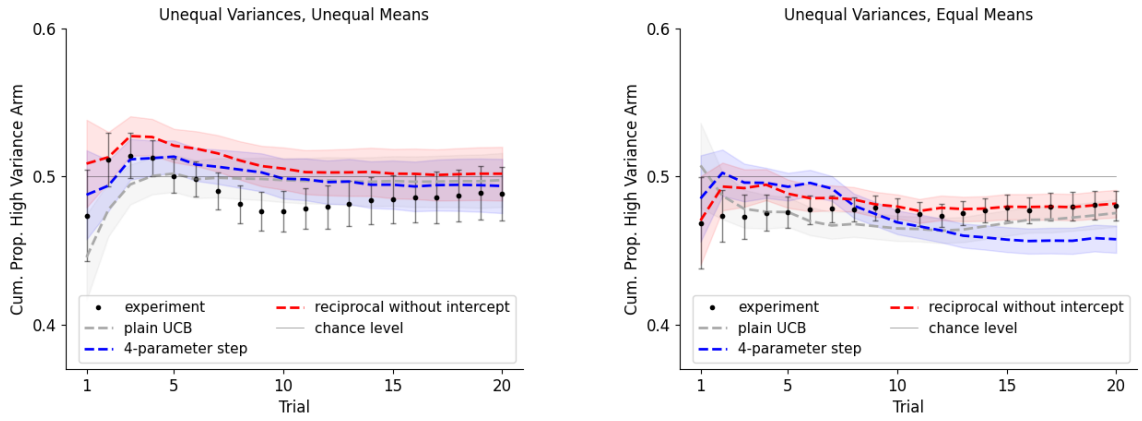

Figure 4: Left: Average cumulative proportion of experimental and simulated high variance arm choices across trials for unequal variances and unequal means. The thick line describes the average across participants and games, reported with $95\,\%$ confidence intervals in within-subject design. Right: Corresponding simulations for the unequal variances and equal means condition.

separately and then computed the overall average. A selection of simulations is shown in fig. 4 for the unequal means (left) and for the equal means (right) conditions. Individual exploration bonus parameter estimates of different model variants are depicted in fig. 11 (see appendix A.1) and show considerable variability between participants, indicating that not all participants have a positive exploration bonus in early trials.
While the plain UCB algorithm (solid grey line) in fig. 4 with constant exploration bonus β qualitatively only accounts for an increase in the average cumulative proportion of high variance arm choices, all other models capture the decrease for late trials compared to early trials of the unequal variances, unequal means condition from fig. 4 (left). More specifically, we obtain decreasing uncertainty preferences for late trials relative to early trials for all models that account for a change in uncertainty preference during learning. This can be seen for the exemplary selected models in fig. 4 for both unequal and equal means, resulting in a higher qualitative similarity to the experimental data in the unequal mean conditions. When comparing different uncertainty accounting model variants, only the linear models with $\beta = a_1 \cdot t + c_1$ with and without intercept, the reciprocal model with intercept $\beta = a_2/t + c_2$ and the 3- and 4- parameter step models, both containing two different exploration bonuses for early and late trials, allow for a switch from positive to negative β estimates (cf. fig. 11 in appendix A.1). This switch results in negative values for later trials, reversing the uncertainty preference from uncertainty seeking to uncertainty aversion, leading to higher qualitative similarity for unequal means but lower for equal means.

To further test whether the variances of the two arms were different enough to manipulate individuals' variance estimates, and whether people engage in uncertainty biased exploration and exploitation in our task, and how their biases change over the course of learning, we used computational modelling.

Model-based analysis. To test whether the differences in the generative variances were large enough to ensure different perceived variance estimates for the high and low variance arms, we modelled learning with a Kalman Filter learning rule without free parameters, which gave us a rough estimate of participants' perceived variances on every trial. Therefore, the observation

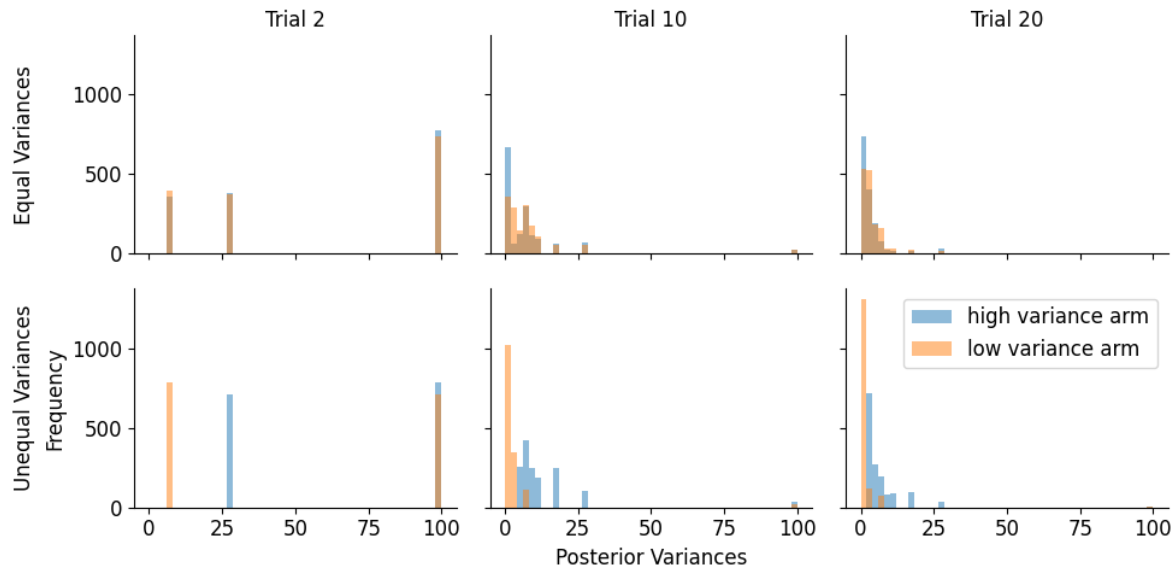

Figure 5: Histogram of participants' posterior variance estimates for equal and unequal variance conditions across trials. Posterior variances were estimated using a Kalman Filter with no free parameters.

variances were set to the true generative variances, the process variance was set to zero, and the initial values to previously used values (cf. section 2). Histograms of the posterior variance estimates for equal variance and unequal variance conditions and high and low variance arms are depicted in fig. 5 over different trial numbers. For the equal variances conditions, higher variance arms indicate that these arms had slightly higher empirical variances, although the generative variances were identical. As expected, for equal variances both high and low variance arms show similar variance estimates across trials. Also, there remains a difference between high and low variance arms for unequal variances, indicating that participants retained a higher uncertainty estimate for the arm with higher variance than for the arm with lower variance.

To investigate whether participants used directed exploration to cope with unequal uncertainties, we compared a different set of standard models described in section 2: a Delta Rule model with a softmax choice policy, a Kalman Filter with a softmax choice policy and a Kalman Filter with probabilistic UCB. The latter is the only model that explicitly allows for uncertainty preferences during learning. If people are guided by their uncertainty preferences, we expected that a reasonable number of participants would be best described by the Kalman Filter with a probabilistic UCB choice rule. The free parameters of all three models were well recoverable on our experimental task within our parameter space (cf. section 2.5), and all models, as well as their parameters, could be recovered based on the set of these three comparative models. The recoveries are shown in appendix A.1.1.
Fitting data from both blocks and all trials, the proportion of participants best described by each model is shown in table 1. Most participants are best described by the Delta Rule model with a softmax choice policy. To investigate whether this might be due to low average uncertainty preferences across trials, we compared early and late trials separately by allowing different parameter estimates for the first 7 and the last 13 trials. Comparing models within trial segments, the Kalman Filter with UCB choice, which takes uncertainty preferences into account, does indeed best describe the highest proportion of participants for early trials (cf. table 1). In

	All Trials		First 7 Trials		Last 13 Trials		
Model	BIC	AIC	BIC	AIC	BIC	AIC	k
Delta Rule + softmax	0.72	0.74	0.34	0.45	0.63	0.67	2
Kalman Filter + UCB	0.16	0.19	0.41	0.48	0.21	0.25	2
Kalman Filter + softmax	0.12	0.07	0.26	0.08	0.16	0.08	1

Table 1: Proportion of participants best described by different models with k parameters according to BIC and AIC of model fits for different trial segments.

UCB Variant	Free Model Parameters	BIC	AIC
Reciprocal without intercept	γ, $\beta = a_2/t$	0.29	0.09
Linear without intercept	γ, $\beta = a_1 \cdot t$	0.25	0.10
4-parameter step	$\gamma_{\text{early}}, \gamma_{\text{late}}, \beta_{\text{early}}, \beta_{\text{late}}$	0.18	0.36
Plain	γ, β	0.13	0.09
Reciprocal	γ, $\beta = c_2 + a_2/t$	0.11	0.20
Linear	γ, $\beta = c_1 + a_1 \cdot t$	0.03	0.08
3-parameter step	γ, $\beta_{\text{early}}, \beta_{\text{late}}$	0.02	0.09

Table 2: Proportion of participants best described by BIC and AIC of model fits with the Kalman Filter learning rule and different variants of the UCB choice rule. γ denotes the inverse softmax temperature, β the exploration bonus and t the trial number. The step variants change the parameter after 7 trials.

contrast, for late trials, the Delta Rule model with softmax choice still best captures the highest proportion of participants, suggesting that uncertainty preferences were lower in later trials, as also observed for unequal means and unequal variances in fig. 3. If we fit models that account for different preferences for uncertainty across learning, we would expect this to be reflected in the estimates of the exploration bonus parameter β, which would be closer to zero for late trials.

Although the UCB model is not the winning model for all trial segments, we used it to investigate uncertainty preferences throughout learning, as it is the only model that accounts for uncertainty and describes a significant number of participants reasonably well.
To test how uncertainty preferences change during learning we performed a model comparison between the model variants of probabilistic UCB choice as described in section 3.1 that account for changing uncertainty preferences during learning.
The comparison of these variants is shown in table 2. The results show considerable individual variability, with different models providing the best fit for different subsets of participants. This suggests that uncertainty preferences and their rate of change during learning vary between participants. Only a few participants are best fit by a plain UCB model, that includes a constant bonus for the uncertain option, further supporting our hypothesis, that uncertainty preferences change as learning progresses.
Finally, we compared our winning model with the plain UCB and found that $59\,\%$ of our participants are best described by the reciprocal exploration bonus without intercept and $41\,\%$ by the plain UCB indicating that indeed most of our participants started with high absolute values for the exploration bonus that decreased as they learned.
As shown in table 2, the winning model according to AIC is the 4-parameter step model, which allows gamma and beta to vary freely, separately for the first 7 and last 13 trials, but has twice as

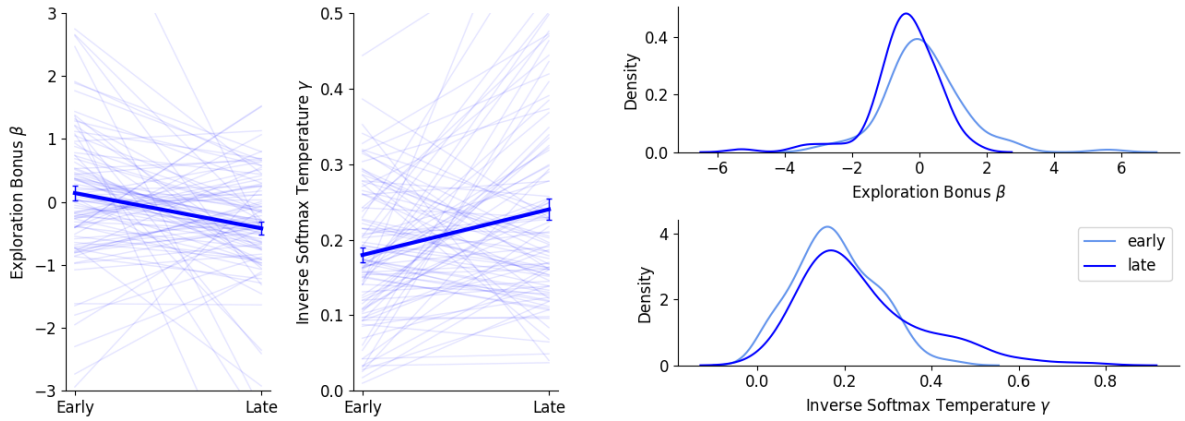

Figure 6: Left: Comparison of parameter estimates for early and late trials from the Kalman Filter with UCB choice rule containing 4 parameters, each an inverse softmax temperature γ and exploration bonus β for early trials and a pair for late trials. Thick lines represent the average change across participants, reported with $95\,\%$ confidence intervals. Thin lines represent individual parameter estimates. Right: Corresponding parameter estimate densities for all participants for early and for late trials.

many parameters fitted. As the reciprocal UCB without intercept does not allow for exploration bonuses for late trials to deviate from zero, it does not allow us to examine uncertainty preferences for late trials, even if they exist. These might be small compared to early exploration preferences. We therefore examined the model parameter estimates of the 4-parameter step model, to have a closer look at how the model parameters change for late trials compared to early trials.

After removing outliers (>3 standard deviations from the mean estimates) from the parameter estimates, we compared the inverse softmax temperature γ and exploration bonus β for early trials with the estimates for late trials. The average and individual slopes, and densities of both parameter spaces are shown in fig. 6.
Consistent with the previous results, we found that uncertainty preferences decreased on average from early to late trials ($\beta = -0.560$, $p < 0.001$). However, we observe considerable individual variability in the slopes in fig. 6 left panel, suggesting that not all participants followed the decreasing trend. Furthermore, when looking at the inverse softmax temperature γ in fig. 6 right panel, we observe increasing estimates for late trials compared to early trials ($\beta = 0.059$, $p < 0.001$), implying an average increase in choice determinism for later trials compared to early trials.
To answer the question of whether we observe uncertainty aversion in late trials, we conducted a one-sample t-test to test whether the exploration bonus in late trials is less than zero. We obtained a significant negative estimate ($t(95) = -4.042$, $p < 0.001$), indicating uncertainty aversion in late trials. The average exploration bonus in late trials is given by $\overline{\beta}_{\text{late}} = (-0.420 \pm 0.104)$.
When examining uncertainty preferences in early trials, fig. 6 suggests that the early exploration bonus is very small for this model. Also, a one sample t-test suggested that the average parameter estimates are not significantly different from zero ($t(95) = 1.146$, $p < 0.127$), which will be discussed in section 4.

So far, we have seen that uncertainty preferences change during learning. We have tested our

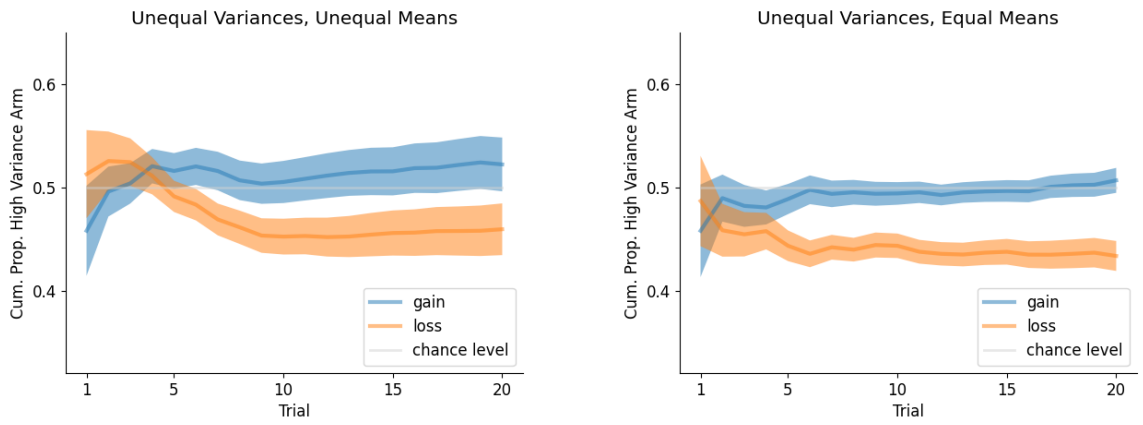

Figure 7: Left: Average cumulative proportion of high variance arm choices across trials for unequal variances and unequal means. The thick line describes the average across participants and games, reported with $95\,\%$ confidence intervals in within-subject design. Right: Corresponding results for the unequal variances and equal means condition.

hypothesis by ignoring the potentially different effects of gains and losses. Now we want to investigate whether the valence of rewards influences uncertainty preferences and their rate of change during learning.

3.2 (How) Do Preferences for Uncertainty Differ Between Gains and Losses?

Behavioural Analysis. The average proportions of cumulative high variance arm choices, for the loss and gain blocks in unequal variance conditions, are depicted in fig. 7 (for absolute values see fig. 12). The results show greater overall aversion for the high variance arm in losses relative to gains, contrary to our predictions from prospect theory (Kahneman & Tversky, 1979) and Krueger et al. (2017), which suggest greater overall uncertainty seeking for losses relative to gains. For unequal means, we observe a greater increase of high variance arm choices in early trials in the loss block relative to gains. Otherwise, the qualitative patterns of the loss block and the gain block are very similar.

To facilitate comparison between the loss and gain blocks, we plotted the mean proportion of better arm choices and high variance arm choices for each participant in the gain block against their mean proportion in the loss block. All conditions of means and variances are depicted separately in fig. 8. The better arm choices are mostly distributed around the identity line, which is plotted in black. This shows that the average performance was similar for gains and losses. For unequal means, the proportions of better arm choices are higher, in line with lower task difficulty.

We can see, that the proportion of selecting the high variance arm is generally lower than the proportion of selecting the better arm. The high values of high variance arm choices for the equal variances and unequal means condition may be due to sampling, as the higher empirical variance happened to fall on the better arm in three out of four games.

In line with the previously observed higher average aversion to the high variance arm for losses compared to gains, we observe that the orange distributions (high variance arm choices) are on average below the identity line for unequal variances.

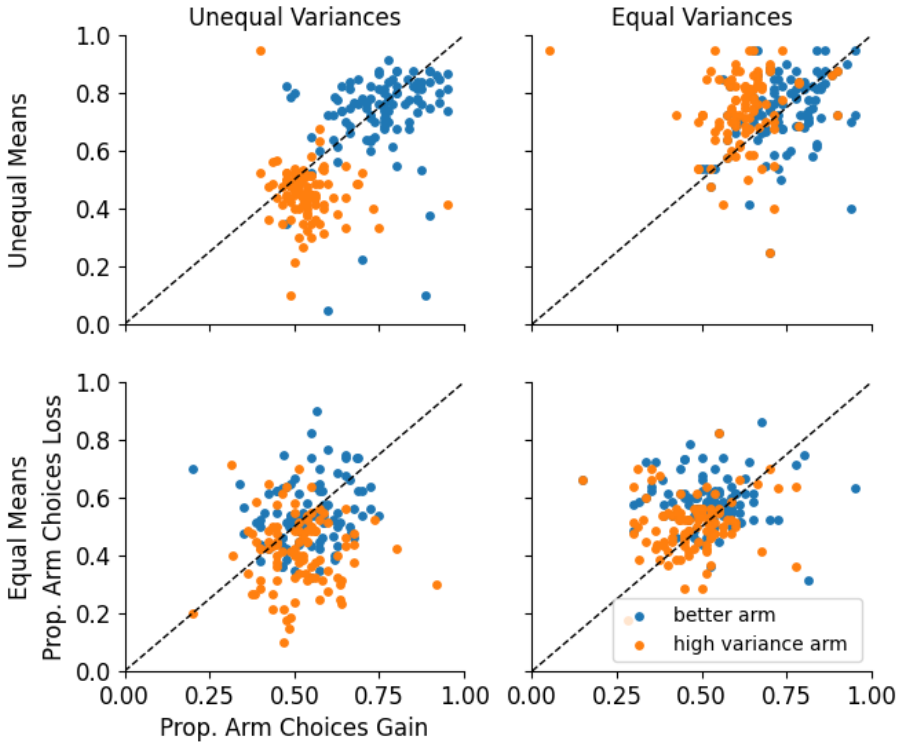

Figure 8: Proportion of high variance arm choices and better arm choices for the gain block compared to the loss block for each pair of mean, variance conditions. Each point represents the average proportion of arm choices across games for one participant. The dashed black line indicates an equal number of arm choices in the gain and loss block.

Model-based analysis. To further test our hypotheses, we ran logistic mixed effects models to predict high variance arm choices and better arm choices by our experimental manipulations. Mean and variance differences were set to the difference between the high variance and low variance arm and the better arm and worse arm, respectively. Detailed results are given in table 3 and table 4 for high variance arm choices, and in table 5 for better arm choices.
As expected, we found an increase in high variance arm choices for larger positive mean differences between high and low variance arms (β=0.827, p <0.001), indicating that the higher the mean difference, the more often people chose the better arm. Furthermore, we found an increase in the selection of high variance arms for increasing variance difference between high and low variance arms (β=0.056, p =0.001). In line with previous observations from fig. 8, we found a decrease in high variance arm choices in the loss block compared to the gain block (β=−0.139, p <0.001). Contrary to our results from section 3.1, we found an overall increase in high variance arm choices across trials (β=0.088, p <0.001). This could be due to the non-linear relationship observed in section 3.1 and fig. 7. We found no interaction between trial and block (β=0.013, p =0.467). So far, we have examined all conditions. However, unequal mean conditions could confound the results as empirically high variance arms were not counterbalanced for unequal means. We therefore ran another logistic mixed effects model to predict high variance arm choices only for unequal variances. The results are shown in table 4. Consistent with the results for all conditions, we obtained an increase in high variance arm choices with increasing mean difference between high and low variance arms (β=0.942, p <0.001). We also

	Fixed Effects					
	Estimate	Std. Err.	z-value	$p > \|z\|$	95% CI	
Intercept	0.240	0.021	11.390	<0.001	0.199	0.282
Mean diff.	0.837	0.034	24.633	<0.001	0.771	0.904
Trial	0.088	0.013	6.939	<0.001	0.063	0.113
Block[Loss]	-0.139	0.029	-4.710	<0.001	-0.196	-0.081
Var. diff.	0.056	0.017	3.329	0.001	0.023	0.089
Game	-0.095	0.015	-6.417	<0.001	-0.123	-0.066
Block order	-0.012	0.031	-0.396	0.692	-0.072	0.048
Mean diff.:Trial	0.298	0.010	28.942	<0.001	0.278	0.319
Mean diff.:Block[Loss]	-0.054	0.020	-2.635	0.008	-0.094	-0.014
Block[Loss]:Var. diff.	-0.205	0.018	-11.081	<0.001	-0.241	-0.169
Trial:Var. diff.	-0.013	0.009	-1.438	0.150	-0.031	0.005
Trial:Block[Loss]	0.013	0.018	0.727	0.467	-0.022	0.048

Table 3: Results of logistic mixed effects regression predicting high variance arm choices. Mean and variance differences are given by the value for the high variance arm minus the value for the low variance arm.

obtained a decrease in high variance arm choices from gain to loss (β=-0.372, p <0.001), more pronounced than for all conditions. In contrast to the effects for all conditions, we found no effect for high variance arm choices with trial (β=0.002, p <0.858), which again could result from a non-linear relationship. We also found no interaction of high variance arm choices with variance difference (β=-0.003, p <0.893), which could be due to the low variability of variance differences for unequal variances, as we controlled for empirical variances having only small deviations from the generative variances. In contrast to the results for all conditions, we obtained a negative interaction of trial by block (β=-0.138, p <0.001). This is consistent with the observed stronger decrease in high variance arm choices with trial number for the loss block compared to the gain block, which we observed in fig. 8 left panel.
In our analysis of the prediction of better arm choices, we found, as expected, an increase in better arm choices with increasing mean difference (β=0.481, p <0.001) as task difficulty decreases. We also found a decrease in better arm choices with increasing variance difference between better and worse arm (β=-0.051, p <0.001), as task difficulty increases with increasing variance difference. As further expected from fig. 8, we find no significant block effect (β=0.037, p =0.352), indicating that people performed similarly well in the gain and the loss block. We also found an increase in better arm choices with increasing trial number (β=0.238, p <0.001), showing that participants learned within games.

To test whether the observed differences between the gain and loss blocks are related to differences in uncertainty preferences and how they change during learning, we performed a model comparison. We compared the reciprocal model without intercept fitting both gain and loss blocks together, which is the winning model of section 3.1 that accounts for changes in uncertainty preferences during learning, with the same model fitting both blocks separately. The results are shown in table 6. Based on both AIC and BIC scores, we found that some participants' behaviour is better captured by modelling the two blocks separately, while others show more consistent behaviour across both blocks.

	Fixed Effects					
	Estimate	Std. Err.	z-value	$p > \|z\|$	95% CI	
Intercept	-0.104	0.024	-4.271	<0.001	-0.152	-0.056
Mean diff.	0.942	0.042	22.415	<0.001	0.859	1.024
Trial	0.002	0.013	0.178	0.858	-0.023	0.028
Block	-0.372	0.047	-7.942	<0.001	-0.464	-0.280
Var. diff.	-0.003	0.023	-0.134	0.893	-0.048	0.042
Game	-0.127	0.026	-4.919	<0.001	-0.178	-0.077
Block order	-0.019	0.042	-0.446	0.656	-0.102	0.064
Mean diff.:Trial	0.350	0.015	23.506	<0.001	0.321	0.379
Mean diff.:Block	-0.016	0.029	-0.538	0.590	-0.074	0.042
Block:Var. diff.	-0.099	0.039	-2.570	0.010	-0.175	-0.024
Trial:Var. diff.	-0.036	0.013	-2.715	0.007	-0.062	-0.010
Trial:Block	-0.138	0.026	-5.319	<0.001	-0.189	-0.087

Table 4: Results of logistic mixed effects regression predicting high variance arm choices for unequal variance conditions. Mean and variance differences are given by the value for the high variance arm minus the value for the low variance arm.

To examine the differences in uncertainty preferences between the gain and loss blocks over trials, we followed an analysis similar to that in fig. 6. This involved comparing the model estimates and parameter densities of a step model with 8 free parameters (4 for each of the gain and loss blocks), as shown in fig. 9, fitted to the data from all conditions.

Looking first at the parameter estimates of the loss block, the exploration bonus decreases from early to late trials (β=-0.588, p <0.001), while the inverse softmax temperature increases from early to late trials (β=0.110, p <0.001). In contrast, for gains there is no significant change in exploration bonus (β=-0.169, p =0.451) or inverse softmax temperature (β=0.004, p =0.825) over trials.

When looking at both block, as expected, we observe an overall decrease in the exploration bonus β from early to late trials (β=-0.382, p =0.007), indicating decreasing uncertainty seeking across trials. In contrast to our behavioural observations fig. 8 for unequal variances, which show a greater selection of high variance arm choices for the gain block compared to the loss block, we found significantly higher exploration bonuses for the loss block compared to the gain block (β=0.471, p =0.001), indicating that uncertainty preferences are higher for losses compared to gains. To test how much exploration bonus estimates differ from gain to loss, we performed two-sample t-tests, which also showed significantly lower exploration bonuses for the gain block compared to the loss block in early trials ($t(91)$=-3.252, p =0.002) and late trials ($t(91)$=-2.128, p =0.036). We found no significant interaction between trial (early/late) and block (gain/loss) (β=-0.422, p =0.137).

For the softmax temperature in fig. 9 we observe an overall increase from early to late trials (β=0.057, $p < 0.001$), indicating increased choice determinism for later trials, consistent with our findings from section 3.1. We also observe higher choice determinism for the gain block compared to the loss block (β=0.065, $p < 0.001$) and a stronger increase from early to late trials for the loss block compared to the gain block, consistent with a positive interaction between trial and block (β=0.105, $p < 0.001$). The shift towards more deterministic choices from early to late trials in the loss block may explain the observed preference for the high variance arm in the gain block, as compared to the loss block, as shown in fig. 8.

	Fixed Effects					
	Estimate	Std. Err.	z-value	$p > \|z\|$	95% CI	
Intercept	0.694	0.030	22.936	<0.001	0.635	0.753
Mean diff.	0.481	0.022	21.623	<0.001	0.438	0.525
Block	0.037	0.040	0.930	0.352	-0.041	0.115
Trial	0.238	0.009	25.910	<0.001	0.220	0.256
Var. diff.	-0.051	0.014	-3.550	<0.001	-0.079	-0.023
Game	-0.035	0.014	-2.530	0.011	-0.063	-0.008
Block order	-0.008	0.047	-0.158	0.874	-0.100	0.085
Mean diff.:Block	-0.120	0.019	-6.421	<0.001	-0.157	-0.084
Mean diff.:Trial	0.204	0.009	21.795	<0.001	0.186	0.222
Block:Var. diff.	-0.127	0.019	-6.815	<0.001	-0.163	-0.090
Trial:Var. diff.	-0.001	0.009	-0.111	0.912	-0.019	0.017
Block:Trial	0.087	0.018	4.832	<0.001	0.052	0.123

Table 5: Results of logistic mixed effects regressions predicting better arm choices. Mean and variance differences are given by the value for the better arm minus the value for the worse arm.

Model	BIC	AIC	k
Gain and loss fitted together	0.61	0.3	2
Gain and loss fitted separately	0.39	0.7	4

Table 6: Proportion of participants best described by BIC and AIC values for model comparison of the reciprocal variant without intercept ($\gamma, \beta = a_2/t$), gain and loss fitted together and separately (k parameters).

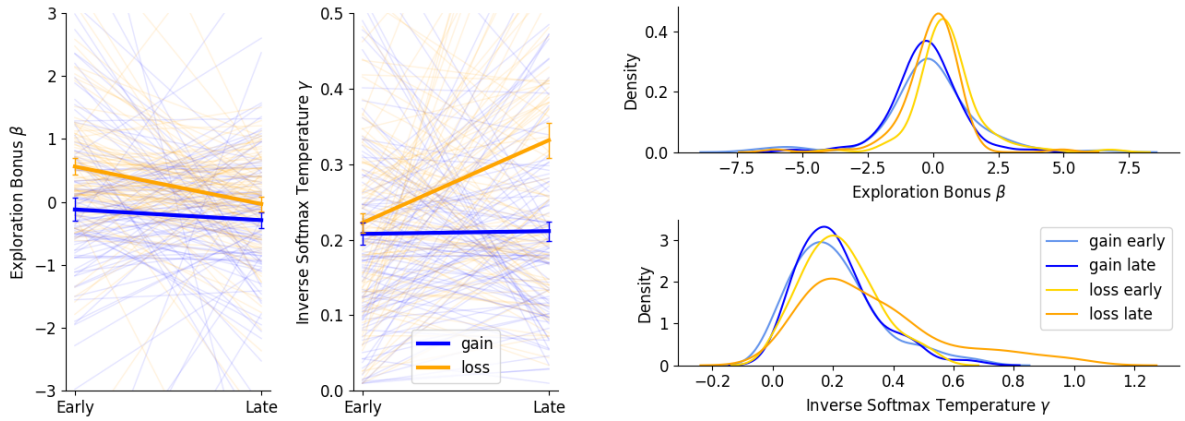

Figure 9: Left: Comparison of parameter estimates after removing the outliers (>3 standard deviations greater than the mean) for early and late trials and gain and loss blocks from the Kalman Filter with UCB choice rule with step, for each gain and loss block each one inverse softmax temperature γ and exploration bonus β for early and one pair for late trials. Thick lines represent the average change across participants, reported with $95\,\%$ confidence intervals. Thin lines represent individual parameter estimates. Right: Corresponding parameter estimate densities for all participants for early and for late trials separately.

4 Discussion

In this thesis, we introduced a novel experimental design, that builds on the existing literature exploring how uncertainty affects learning and decision making processes (Gershman, 2018; Krueger et al., 2017; Wilson, Geana, et al., 2014), to study how uncertainty preferences change during learning, both for gains and for losses. We set up a two-armed bandit task with stationary but unknown Gaussian reward distributions, spanning a 2x2x2 factorial design of reward domain (gain vs. loss), generative mean differences (0 vs. 8) and variance differences (0 vs. 40). We manipulated perceived uncertainty through the variances of the Gaussian reward distributions, which were specified for each arm within each game. Our results indicate that people change their uncertainty preferences during learning, which has been largely neglected in the current literature.

Our experimental design includes the exploration of a longer horizon than in previously studied tasks (e.g., Gershman (2018)), aiming to capture the dynamics of uncertainty preferences until learning is complete. In addition, our factorial design aims to disentangle the effect of mean and variance differences and allows for the exploration of behaviour under conditions of unequal means, highlighting how differences in means can influence uncertainty preferences. While previous work has often neglected the effect of reward valence, recent work suggests that the valence of reward has an influence on learning and decision making (Krueger et al., 2017). In our design, we adapted this approach and examined gains and losses in two separate blocks to picture their effects on decision making. Furthermore, our effort differs from paradigms where risk must be learned from sampling (e.g., Niv et al., 2012), which typically feature binary rewards and an uncertainty free, safe option, where the risk of the uncertain alternative is quickly identifiable. Unlike related studies, our full within-subject design allowed us to examine individual variability in uncertainty preferences that is not confounded with order and sequence effects, as we fixed game and reward order and sequences for all participants.

Within our factorial framework, we found that participants were actively engaged in learning, exploring, and exploiting. Participants showed above-chance performance that increased with decreasing task difficulty of our design conditions, and increased with trial number. They also stuck to their choices more for increasing trial number, and with decreasing task difficulty.
By manipulating the generative variances, we were able to induce different asymptotic uncertainty estimates in the two arms under conditions of unequal variances. We used this variance manipulation to shed light on how perceived uncertainties influence decision making under different conditions.
To investigate how uncertainty preferences change during learning, we fitted a Kalman Filter with different variants of the probabilistic UCB, which are capable of capturing time-varying uncertainty preferences, to our behavioural data. Most participants were indeed best described by allowing the exploration bonus parameter to vary over time, when compared to the plain UCB model with a constant exploration bonus. A qualitative comparison of the predictions of the constant and time-varying exploration bonus models with the behavioural data (see fig. 4) further supports the notion that uncertainty preferences change as learning progresses. On average, uncertainty preferences decreased during learning, so that the average participant was uncertainty seeking in early trials and uncertainty averse in late trials. We also found increased choice determinism from early to late trials, consistent with findings from the horizon task (Wilson, Geana, et al., 2014), suggesting increased random exploration for exploratory (early) trials compared to exploitative (late) trials.

When comparing the effect of reward valence, we found that the participants in our study preferred the high variance arm more for gains than for losses. However, computational modelling suggests that participants were more uncertainty seeking for losses than for gains. This is consistent with predictions from both prospect theory (Kahneman & Tversky, 1979) and a baseline uncertainty bias arising from Bayesian shrinkage (Krueger et al., 2017). The question of where this shift in uncertainty search for losses versus gains comes from needs to be addressed in future research.
The changing uncertainty preferences found in the first part of our analysis (section 3.1) mainly arose from changing uncertainty preferences in the loss block, as we found no significant effect for gains. For losses compared to gains, we also found higher parameters for random exploration. Together with increased directed exploration for losses compared to gains, this supports the notion of increased exploration for losses compared to gains, which is consistent with previous work (Lejarraga & Hertwig, 2016).
In almost all our results, we observed high individual variability, showing individual differences in how people account uncertainty throughout the learning process. In future, we want to link the additionally collected depression, anxiety and operation span scores to individual differences in performance and uncertainty preferences, and how they change during learning (see appendix A.2).

In the following, we discuss open questions from our analysis. First, we observed different patterns of exploration bonuses for early versus late trials when comparing different model variants that account for uncertainty. Different results for the step models could be related to the threshold that determines the point at which we allow parameters to change between early and late trials, which we chose post-hoc based on our behavioural observations. The step variant showed higher bonuses for later trials, whereas the reciprocal model showed higher bonuses for early trials. This difference, which would be related to allowing the parameters to change in the step model after seven trials, is consistent with the behavioural data showing an early peak that returns to baseline in the unequal variances and means condition (fig. 2 left). The reciprocal model accounts for a decreasing beta in the early trials, in contrast to the average estimation approach of the step model, potentially leading to underestimated early parameters. However, in late trials (8 to 20) the step model predictions are more accurate as uncertainty preferences stabilise, as seen in fig. 2. This behaviour in late trials is consistent with expectations from a standard UCB model (fig. 4). The chosen threshold for distinguishing early from late trials is crucial for analysing changes in uncertainty preferences. This suggests that future research should explore how different thresholds affect the parameter estimates and explore other models that examine the point at which uncertainty preferences shift from uncertainty seeking to aversion.
Relatedly, our task lacks a clear separation between exploration and exploitation, which limits our ability to investigate if and how strategies under uncertainty are related to these processes. The observe-or-bet task (Tversky & Edwards, 1966) provides a more definitive separation of exploration and exploitation. This approach provides a promising direction for future studies aimed at disentangling how uncertainty preferences evolve across the exploration-exploitation continuum.
Second, why did the behavioural analysis show a greater selection of the arm with higher uncertainty for gains than for losses, if the computational model suggests a higher uncertainty preference for losses than for gains? First of all, we have to be cautious about comparing the parameters of early and late trials, as we only tested one threshold, which was chosen post-hoc to our behavioural analysis. However, we observed an increased choice determinism from early

to late trials for losses and no increase for gains. Although we aimed to disentangle mean and variance differences by reducing the correlation between uncertainty and reward (a major motivation for the horizon task by Wilson, Geana, et al. (2014)) through our factorial design, there is a small, but present correlation between rewards and uncertainty, that we could not fully diminish with our experimental design. This leads us to suspect that the observations in high variance arm choices could somehow be caused by the effect of random exploration. Because we decided to look at an objective measure (high variance arm choices) that does not include model assumptions for behavioural observations, we neglected peoples perceived variances, although it might be crucial to include them. To justify our approach, we tested that the empirical variances were able to induce these differences in perceived variances, especially for late trials (see fig. 5). Nevertheless, in future investigations, we want to extend our analysis by looking at high variance arm choices of perceived uncertainty, which we can estimate by a Kalman Filter. We want to do this, because sampling the better arm more often could lead to an underestimated variance for the better arm and thus to an even higher correlation between perceived high variance arm choices and better arm choices.

While our research makes an important contribution to understanding how uncertainty preferences change over the course of learning, the generalisability of our findings is subject to certain limitations. In particular, our research design was limited to the same four games per condition for each participant. By adopting a fully within-subjects design, we predetermined the reward sequences and fixed both the reward and game sequences for all participants. This approach was taken to minimise confounding between-subject variability from order and sequence (Dale & Arnell, 2013; Goodhew & Edwards, 2019; Hedge et al., 2018) and to investigate how individual differences in anxiety and depression scores and working memory span affect performance, choice stochasticity and uncertainty preferences (see appendix A.2).
On the one hand, this methodology helped us to better analyse and understand the remarkable individual differences in the development of model parameters during the learning process. It highlights the critical role of accounting for individual differences when interpreting the effectiveness of different theoretical models. On the other hand, this uniformity in experimental design implies that our results may not generalise to different games, game and reward sequences, or stimulus sets beyond those we specifically chose. Thus, while our findings contribute valuable perspectives to the study of uncertainty in learning, they come with a caveat regarding their applicability to broader or different contexts.
In the future, we aim to address this concern by expanding each condition with additional games and new reward sequences.

To expand the scope of our discussion, our study prompts a reevaluation of how uncertainty is considered in learning and decision making.
First, while we used solely irreducible uncertainty to manipulate the participants' behaviour in our learning task, research on directed exploration often manipulates solely reducible uncertainty (Krueger et al., 2017; Wilson, Geana, et al., 2014; Zaller et al., 2021), neglecting possible effects of irreducible uncertainties. As people may consider irreducible and reducible uncertainties differently (Speekenbrink, 2022), we may need to find an experimental task that includes both types of uncertainty but disentangles them. Therefore, one would need to manipulate both types of uncertainty simultaneously but independently. One promising approach might be to modify the horizon task to include more forced-choice trials. Assigning different generative variances (irreducible uncertainties) to each arm and demonstrating them by showing multiple samples, and additionally manipulating reducible uncertainties by showing a different number of samples

for each arm, would allow to study the effects of both types of uncertainty. When considering different types of uncertainty, it may also be critical to use a computational model that provides separate estimates for each type of uncertainty, such as a count based estimate for reducible uncertainty and a value based estimate for irreducible uncertainty.
Second, we have only examined uncertainty preferences in the decision process. However, it is possible that uncertainty preferences are not only due to decision strategies but also to differences in learning. Similarly, recently investigated optimism and pessimism biases (Lefebvre et al., 2017; Niv et al., 2012; Palminteri & Lebreton, 2022) suggest that people learn with different learning rates for positive and negative prediction errors. When dealing with biased learning rules, higher uncertainties could lead to higher (optimistic) or lower (pessimistic) mean updates for higher than expected rewards, such that uncertainty seeking and uncertainty aversion would result not from a bias in decision making, but from biased learning. Crucially, our task is a potential candidate for comparing such a learning bias with a decision bias. In contrast, the horizon task may not be able to capture learning biases, as they appear to be specific to free choices (Chambon et al., 2020). This suggests a broader framework for understanding how people integrate uncertainty during learning in sampling paradigms that integrates both decision rules and learning processes. As a first step, in our task, we could compare a dual learning rate model with softmax choice to the Kalman Filter with reciprocal UCB variant. To further distinguish biased learning from biased decision making, one would need to ask participants for mean estimates from the underlying reward distributions without forcing them to make a choice. While biased learning would lead to overestimated (optimistic) or underestimated (pessimistic) means, biased decisions would have no effect on mean estimates.

Overall, while our results support the idea that individuals change their uncertainty preferences as they learn, they also highlight several avenues for further research. These include exploring alternative models and experimental designs to capture the complexity of learning and decision making under uncertainty, as well as the potential distinctions between reducible and irreducible uncertainty in shaping decision strategies.

5 Conclusion

In summary, our study provides new insights into the dynamics of uncertainty preferences during learning, revealing high individual variability. We find that people change their uncertainty preferences during learning, on average following a decreasing trend across trials. This trend is mainly driven by behaviour in response to losses, with participants becoming more deterministic in their choices as they progress in learning.

6 Acknowledgements

I would like to thank everyone who contributed to my thesis. In particular, I would like to thank my supervisors Eric Schulz, Mirko Thalmann and Kristin Witte, who guided me in finding a topic, designing my experiment and collecting and analysing the data. And I would also like to thank my family and friends for their emotional support, especially Mini, who helped me whenever I needed to talk about a scientific puzzle or struggled with a decision.

References

Aberg, K. C., Toren, I., & Paz, R. (2022). A neural and behavioral trade-off between value and uncertainty underlies exploratory decisions in normative anxiety. *Molecular psychiatry*, *27*(3), 1573–1587. https://doi.org/10.1038/s41380-021-01363-z

Auer, P., Cesa-Bianchi, N., & Fischer, P. (2002). Finite-time analysis of the multiarmed bandit problem [Cited by: 4253; All Open Access, Bronze Open Access]. *Machine Learning*, *47*(2-3), 235–256. https://doi.org/10.1023/A:1013689704352

Aven, T., & Renn, O. (2009). On risk defined as an event where the outcome is uncertain. *Journal of Risk Research*, *12*, 1–11. https://doi.org/10.1080/13669870802488883

Barrouillet, P., Bernardin, S., & Camos, V. (2004). Time constraints and resource sharing in adults' working memory spans. *Journal of experimental psychology: General*, *133*(1), 83. https://doi.org/10.1136/jnnp.50.5.654-a

Bates, D., Mächler, M., Bolker, B., & Walker, S. (2015). Fitting linear mixed-effects models using lme4. *Journal of Statistical Software*, *67*(1), 1–48. https://doi.org/10.18637/jss.v067.i01

Bennett, D., Sutcliffe, K., Tan, N. P.-J., Smillie, L. D., & Bode, S. (2021). Anxious and obsessive-compulsive traits are independently associated with valuation of noninstrumental information. *Journal of Experimental Psychology: General*, *150*(4), 739. https://doi.org/10.1037/xge0000966

Blanco, N. J., Otto, A. R., Maddox, W. T., Beevers, C. G., & Love, B. C. (2013). The influence of depression symptoms on exploratory decision-making. *Cognition*, *129*(3), 563–568. https://doi.org/10.1016/j.cognition.2013.08.018

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. https://doi.org/10.1038/s41562-020-0919-5

Charpentier, C. J., Aylward, J., Roiser, J. P., & Robinson, O. J. (2017). Enhanced risk aversion, but not loss aversion, in unmedicated pathological anxiety. *Biological psychiatry*, *81*(12), 1014–1022. https://doi.org/10.1016/j.biopsych.2016.12.010

Christopher, G., & MacDonald, J. (2005). The impact of clinical depression on working memory. *Cognitive Neuropsychiatry*, *10*, 379–399. https://doi.org/10.1080/13546800444000128

Cogliati Dezza, I., Cleeremans, A., & Alexander, W. (2019). Should we control? the interplay between cognitive control and information integration in the resolution of the exploration-exploitation dilemma. *Journal of Experimental Psychology: General*, *148*(6), 977. https://doi.org/10.1037/xge0000546

Cohen, J. D., McClure, S. M., & Yu, A. J. (2007). Should i stay or should i go? how the human brain manages the trade-off between exploitation and exploration. *Philosophical Transactions of the Royal Society B: Biological Sciences*, *362*(1481), 933–942. https://doi.org/10.1098/rstb.2007.2098

Collins, A., & Frank, M. (2012). How much of reinforcement learning is working memory, not reinforcement learning? a behavioral, computational, and neurogenetic analysis. *The European journal of neuroscience*, *35*, 1024–35. https://doi.org/10.1111/j.1460-9568.2011.07980.x

Cousineau, D. (2005). Confidence intervals in within-subject designs: A simpler solution to loftus and masson's method. *Tutorials in Quantitative Methods for Psychology*, *1*(1), 42–45. https://doi.org/10.20982/tqmp.01.1.p042

Dale, G., & Arnell, K. M. (2013). Investigating the stability of and relationships among global/local processing measures. *Attention, Perception, & Psychophysics*, *75*, 394–406. https://doi.org/10.3758/s13414-012-0416-7

Danwitz, L., Mathar, D., Smith, E., Tuzsus, D., & Peters, J. (2022). Parameter and model recovery of reinforcement learning models for restless bandit problems. *Computational Brain & Behavior*, *5*(4), 547–563. https://doi.org/10.1101/2021.10.27.466089

Daw, N. D., O'doherty, J. P., Dayan, P., Seymour, B., & Dolan, R. J. (2006). Cortical substrates for exploratory decisions in humans. *Nature*, *441*(7095), 876–879. https://doi.org/10.1038/nature04766

Fan, H., Gershman, S. J., & Phelps, E. A. (2023). Trait somatic anxiety is associated with reduced directed exploration and underestimation of uncertainty. *Nature Human Behaviour*, *7*(1), 102–113. https://doi.org/10.1038/s41562-022-01455-y

Gershman, S. J. (2018). Deconstructing the human algorithms for exploration. *Cognition*, *173*, 34–42. https://doi.org/10.1016/j.cognition.2017.12.014

Goodhew, S. C., & Edwards, M. (2019). Translating experimental paradigms into individual-differences research: Contributions, challenges, and practical recommendations. *Consciousness and Cognition*, *69*, 14–25. https://doi.org/10.1016/j.concog.2019.01.008

Hedge, C., Powell, G., & Sumner, P. (2018). The reliability paradox: Why robust cognitive tasks do not produce reliable individual differences. *Behavior research methods*, *50*, 1166–1186. https://doi.org/10.3758/s13428-017-0935-1

Hertwig, R., Barron, G., Weber, E. U., & Erev, I. (2004). Decisions from experience and the effect of rare events in risky choice. *Psychological science*, *15*(8), 534–539. https://doi.org/10.1111/j.0956-7976.2004.00715.x.

Hertwig, R., & Erev, I. (2009). The description–experience gap in risky choice. *Trends in cognitive sciences*, *13*(12), 517–523. https://doi.org/10.1016/j.tics.2009.09.004

Hills, T. T., Todd, P. M., Lazer, D., Redish, A. D., & Couzin, I. D. (2015). Exploration versus exploitation in space, mind, and society. *Trends in Cognitive Sciences*, *19*(1), 46–54. https://doi.org/10.1016/j.tics.2014.10.004

Hogeveen, J., Mullins, T. S., Romero, J. D., Eversole, E., Rogge-Obando, K., Mayer, A. R., & Costa, V. D. (2022). The neurocomputational bases of explore-exploit decision-making. *Neuron*, *110*(11), 1869–1879. https://doi.org/10.1016/j.neuron.2022.03.014

Huys, Q. J. M., Eshel, N., O'Nions, E., Sheridan, L., Dayan, P., & Roiser, J. P. (2012). Bonsai trees in your head: How the pavlovian system sculpts goal-directed choices by pruning decision trees. *PLOS Computational Biology*, *8*(3), 1–13. https://doi.org/10.1371/journal.pcbi.1002410

Jolly, E. (2018). Pymer4: Connecting r and python for linear mixed modeling. *Journal of Open Source Software*, *3*(31), 862. https://doi.org/10.21105/joss.00862

Kahneman, D., & Tversky, A. (1979). Prospect theory: An analysis of decision under risk. *Econometrica*, *47*(2), 263–291. https://doi.org/https://doi.org/10.2307/1914185

Kalman, R. E. (1960). A new approach to linear filtering and prediction problems. https://doi.org/10.1115/1.3662552

Knox, W. B., Otto, A. R., Stone, P., & Love, B. C. (2012). The nature of belief-directed exploratory choice in human decision-making. *Frontiers in psychology*, *2*, 398. https://doi.org/10.3389/fpsyg.2011.00398

Kroenke, K., Spitzer, R. L., & Williams, J. B. (1999). Patient health questionnaire-9. *Cultural Diversity and Ethnic Minority Psychology*. https://doi.org/10.1037/t06165-000

Kroenke, K., Spitzer, R. L., & Williams, J. B. (2001). The phq-9: Validity of a brief depression severity measure. *Journal of general internal medicine*, *16*(9), 606–613. https://doi.org/10.1046/j.1525-1497.2001.016009606.x

Krueger, P. M., Wilson, R. C., & Cohen, J. D. (2017). Strategies for exploration in the domain of losses. *Judgment and Decision Making*. https://doi.org/10.1017/S1930297500005659

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), 0067. https://doi.org/10.1038/s41562-017-0067

Lejarraga, T., & Hertwig, R. (2016). How the threat of losses makes people explore more than the promise of gains. *Psychonomic Bulletin Review*, *24*. https://doi.org/10.3758/s13423-016-1158-7

Moran, T. P. (2016). Anxiety and working memory capacity: A meta-analysis and narrative review. *Psychological bulletin*, *142 8*, 831–864.

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. https://doi.org/10.1523/JNEUROSCI.5498-10.2012

Palminteri, S., & Lebreton, M. (2022). The computational roots of positivity and confirmation biases in reinforcement learning. *Trends in Cognitive Sciences*. https://doi.org/10.1016/j.tics.2022.04.005

Platt, M. L., & Huettel, S. A. (2008). Risky business: The neuroeconomics of decision making under uncertainty. *Nature neuroscience*, *11*(4), 398–403. https://doi.org/10.1038/nn2062

Ree, M. J., French, D., MacLeod, C., & Locke, V. (2008). Distinguishing cognitive and somatic dimensions of state and trait anxiety: Development and validation of the state-trait inventory for cognitive and somatic anxiety (sticsa). *Behavioural and Cognitive Psychotherapy*, *36*(3), 313–332. https://doi.org/10.1111/j.2044-8260.1983.tb00601.x

Rescorla, R., & Wagner, A. (1972). A theory of pavlovian conditioning: Variations in the effectiveness of reinforcement and nonreinforcement.

Rogers, M. A., Kasai, K., Koji, M., Fukuda, R., Iwanami, A., Nakagome, K., Fukuda, M., & Kato, N. (2004). Executive and prefrontal dysfunction in unipolar depression: A review of neuropsychological and imaging evidence. *Neuroscience Research*, *50*(1), 1–11. https://doi.org/10.1016/j.neures.2004.05.003

Schmiedek, F., Hildebrandt, A., Lövdén, M., Wilhelm, O., & Lindenberger, U. (2009). Complex span versus updating tasks of working memory: The gap is not that deep. *Journal of Experimental Psychology: Learning, Memory, and Cognition*, *35*(4), 1089. https://doi.org/10.1037/a0015730

Schulz, E., & Gershman, S. J. (2019). The algorithmic architecture of exploration in the human brain [Machine Learning, Big Data, and Neuroscience]. *Current Opinion in Neurobiology*, *55*, 7–14. https://doi.org/10.1016/j.conb.2018.11.003

Seabold, S., & Perktold, J. (2010). Statsmodels: Econometric and statistical modeling with python. *9th Python in Science Conference*.

Smith, R., Taylor, S., Wilson, R. C., Chuning, A. E., Persich, M. R., Wang, S., & Killgore, W. D. (2022). Lower levels of directed exploration and reflective thinking are associated with greater anxiety and depression. *Frontiers in Psychiatry*, *12*, 782136. https://doi.org/10.3389/fpsyt.2021.782136

Speekenbrink, M. (2022). Chasing unknown bandits: Uncertainty guidance in learning and decision making. *Current Directions in Psychological Science*, *31*(5), 419–427. https://doi.org/10.1177/096372142211050

Speekenbrink, M., & Konstantinidis, E. (2015). Uncertainty and exploration in a restless bandit problem. *Topics in cognitive science*, *7*(2), 351–367. https://doi.org/10.1111/tops.12145

Sutton, R. S., & Barto, A. G. (2018). *Reinforcement learning: An introduction*. MIT press. https://doi.org/10.5555/3312046

Turner, M. L., & Engle, R. W. (1989). Is working memory capacity task dependent? *Journal of memory and language*, *28*(2), 127–154. https://doi.org/10.1016/0749-596X(89)90040-5

Tversky, A., & Edwards, W. (1966). Information versus reward in binary choices. *Journal of Experimental Psychology*, *71*(5), 680. https://doi.org/https://doi.org/10.1037/h0023123

Virtanen, P., Gommers, R., Oliphant, T. E., Haberland, M., Reddy, T., Cournapeau, D., Burovski, E., Peterson, P., Weckesser, W., Bright, J., van der Walt, S. J., Brett, M., Wilson, J., Millman, K. J., Mayorov, N., Nelson, A. R. J., Jones, E., Kern, R., Larson, E., ... SciPy 1.0 Contributors. (2020). SciPy 1.0: Fundamental Algorithms for Scientific Computing in Python. *Nature Methods*, *17*, 261–272. https://doi.org/10.1038/s41592-019-0686-2

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

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

Wilson, R. C., Geana, A., White, J. M., Ludvig, E. A., & Cohen, J. D. (2014). Humans use directed and random exploration to solve the explore–exploit dilemma. *Journal of Experimental Psychology: General*, *143*(6), 2074. https://doi.org/10.1037/a0038199

Witte, K., Schulz, E., Wise, T., & Huys, Q. (2023). People who worry more explore more. *Conference on Cognitive Computational Neuroscience (CCN 2023)*. https://doi.org/10.32470/CCN.2023.1036-0

Zaller, I., Zorowitz, S., & Niv, Y. (2021). Information seeking on the horizons task does not predict anxious symptomatology. *Biological Psychiatry*, *89*(9), S217–S218. https://doi.org/10.1016/j.biopsych.2021.02.550

A Appendix

A.1 Supplementary Material

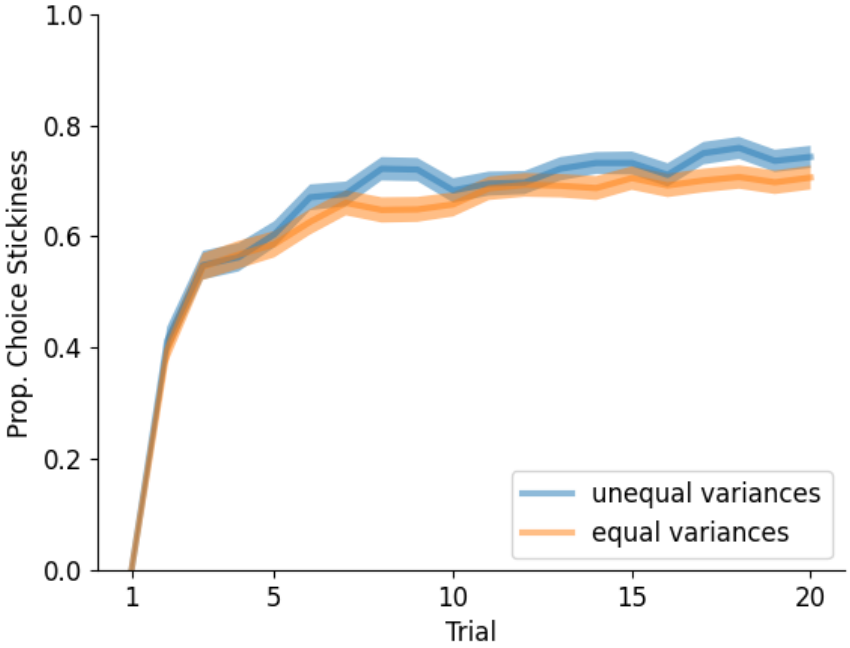

Figure 10: Average choice stickiness across trials and participants with $95\,\%$ confidence intervals in within-subject design for equal and unequal variance conditions. Choice stickiness is defined as boolean value of sticking with the choice on the last trial and is averaged within participants for all games for each condition and then across subjects.

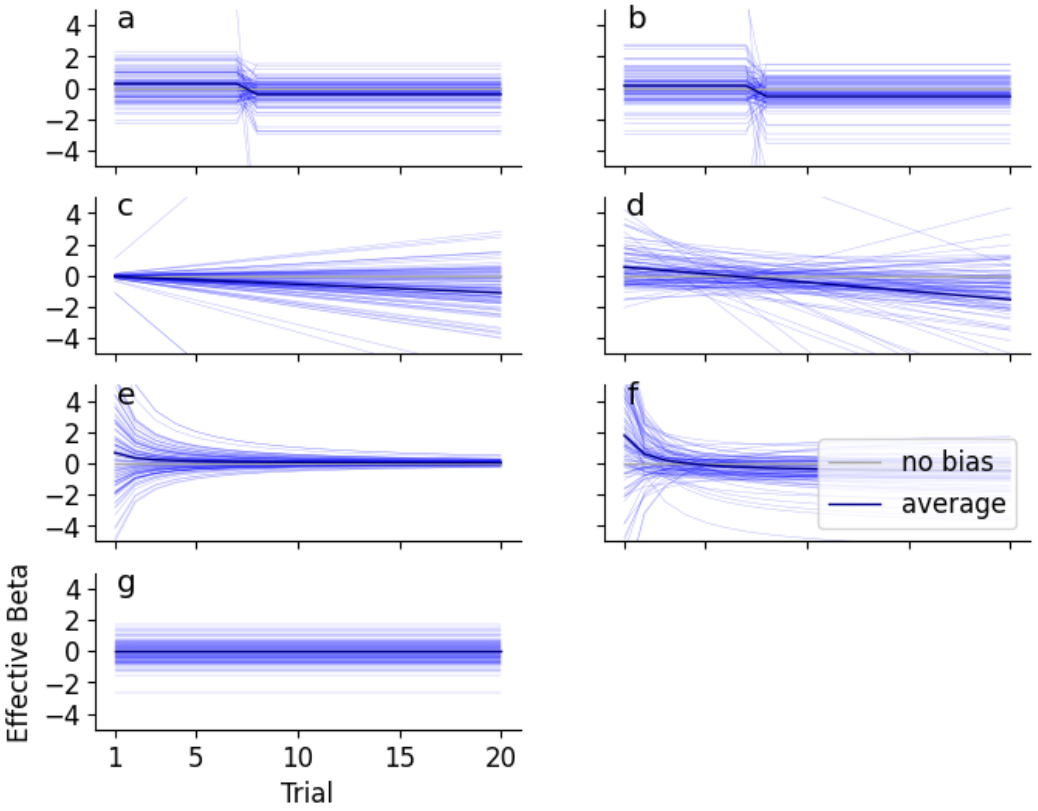

Figure 11: Effective parameter estimates for the exploration bonus β based on fitted parameter estimates for different variants of the probabilistic UCB model for each participant. Enlarged to the range of non-outliers. (a) 3-parameter step (trial $t = 7$) with $\gamma, \beta_{\text{early}}, \beta_{\text{late}}$. (b) 4-parameter step (trial $t = 7$) with $\gamma_{\text{early}}, \gamma_{\text{late}}, \beta_{\text{early}}, \beta_{\text{late}}$ (c) Linear without intercept with $\gamma, \beta = a_1 \cdot t$. (d) Linear with $\gamma, \beta = c_1 + a_1 \cdot t$. (e) Reciprocal without intercept with $\gamma, \beta = a_2/t$. (f) Reciprocal with $\gamma, \beta = c_2 + a_2/t$. (g) Plain UCB with constant γ, β.

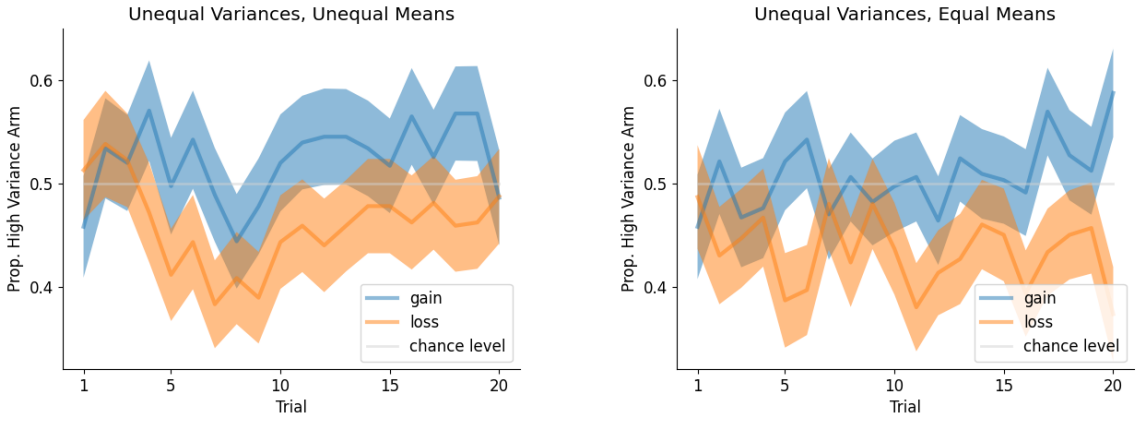

Figure 12: Left: Average proportion of high variance arm choices across trials for unequal variances and unequal means after removing between-subject differences for gains and losses. The thick line describes the average across participants and games, reported with 95 % confidence intervals. Right: Corresponding results for the unequal variances and equal means condition.

A.1.1 Parameter and Model Recoveries

For parameter recovery, we simulated data from 100 participants running through a grid of our parameter space, as described in section 2. We then fitted the simulated participants with the generative model and correlated simulation parameters with the fitted parameters.
For model recovery, we used the simulations obtained for parameter recovery and then fitted the simulations with each of the models. The recovery describes the proportion of simulated participants best described by each model, with 1 indicating that all were best described and 0 indicating that none of the simulated participants were best described by the respective fitted model.

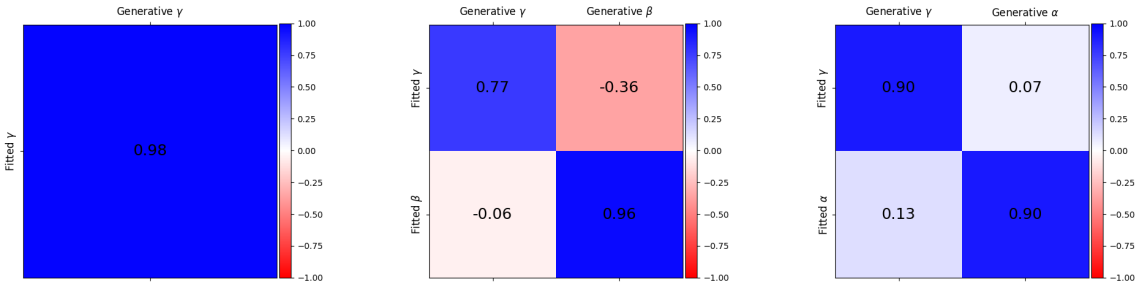

Figure 13: Left: Parameter recovery for Kalman Filter with softmax choice with inverse softmax temperature γ. Middle: Parameter recovery for Kalman Filter and probabilistic Upper Confidence Bound choice with exploration bonus β. Right: Parameter recovery for Delta Rule and softmax choice with learning rate α.

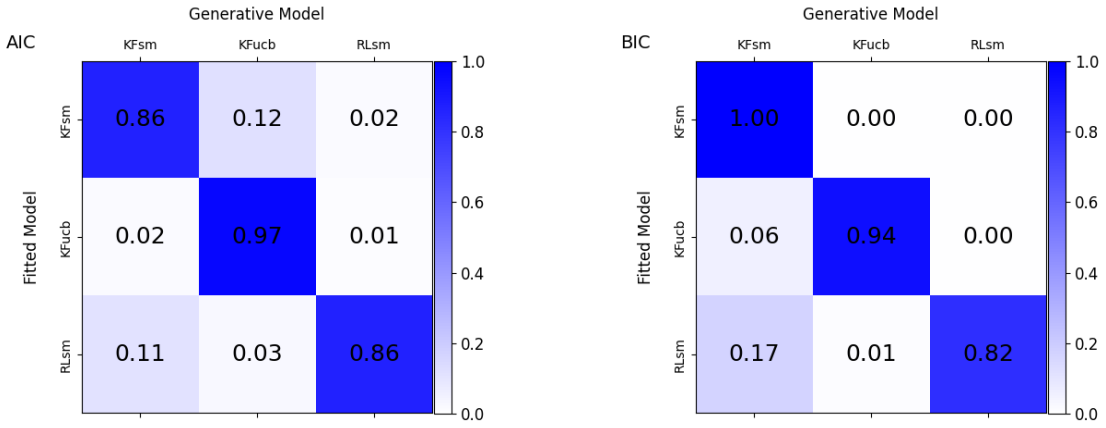

Figure 14: Left: Model recovery for AIC scores: Delta Rule and softmax choice (RLsm), Kalman Filter and softmax choice (KFsm), and Kalman Filter and probabilistic Upper Confidence Bound choice (KFucb) Right: Corresponding recovery for BIC values.

A.2 Individual Differences

Our experimental setup was originally designed to investigate individual differences in exploration strategies during learning. Due to the time constraints of this thesis, we were not able to investigate individual differences in our experimental data. However, we did administer a working memory span test prior to the bandit task to provide an estimate of working memory capacity, and collected depression and anxiety questionnaires following the bandit task. Using simple scores of these additional data, we aimed to investigate how individual differences in exploration strategies are mediated by anxiety and depression. The hypotheses and tasks are described below.

A.2.1 Introduction

To date, empirical evidence on exploration strategies in anxious and depressed patients has been mixed. Some findings suggest that anxiety and depression make people more exploratory (Blanco et al., 2013, Aberg et al., 2022, Witte et al., 2023) and increase uncertainty reduction (Bennett et al., 2021), while other studies find that depression and anxiety lead to less exploration (Huys et al., 2012), and uncertainty avoidance (Charpentier et al., 2017). More recent studies, which more specifically distinguish between random and directed exploration, find a decrease in directed exploration (Smith et al., 2022 Fan et al., 2023). We argue that a decrease in directed exploration with an even steeper increase in random exploration can still lead to an overall net increase in exploration. We want to test this using a computational model that includes parameter estimates for both random and directed exploration, such as the probabilistic UCB model used for the main part of this thesis.
An increase in directed exploration could still be consistent with uncertainty avoidance in exploitation. We can test this by looking at early and late exploration behaviour separately.
In addition, anxious and depressed individuals are known to have working memory deficits (Rogers et al., 2004, Moran, 2016, Christopher & MacDonald, 2005). And working memory load affects performance and exploration behaviour in learning tasks (Collins & Frank, 2012, Cogliati Dezza et al., 2019). We hypothesise that working memory load varies as a function of working memory capacity and aim to investigate whether individual differences in exploration behaviour and performance on our two-armed bandit task in anxious and depressed participants are mediated by working memory capacity. We hypothesize that more anxious and depressed individuals, as well as those with lower working memory capacity scores, will show more random and less directed exploration and a decline in performance. If anxious and depressive behaviour were mediated by working memory capacity, we would expect working memory capacity scores to correlate with anxiety and depression scores.

A.2.2 Operation Span Task

In order to assess individual differences in working memory span, we performed an operation span task first introduced by Turner and Engle (1989). This task has been shown to be a good indicator of individual differences in working memory capacity (Schmiedek et al., 2009).
For all participants, the operation span task was the first task in the experimental sequence. The task consisted of two subtasks: memorising letters in the correct order and performing arithmetic operations between them. Previous work suggests that both processing and maintenance require attention, and thus cognitive load depends on both the number of recalls and the time ratio (Barrouillet et al., 2004). We therefore varied the number of operations and recalls per trial and set a maximum time interval for processing.

Task. The tasks consisted of a practice phase and a test phase. In the practice phase, participants first encountered two trials of pure letter memorisation, followed by two trials of pure solving of arithmetic operations. After completing each arithmetic task, they received immediate feedback on their accuracy. In two further practice trials, participants encountered the full mixed task, which was also the task they encountered in the test phase. Here, a letter was presented for a short time and then an arithmetic task had to be solved within a time interval. There was no immediate feedback in between. This procedure of letter presentation and solving an arithmetic operation was repeated within one trial of the task. After all letters had been presented and all arithmetic operations had been solved, participants had to recall the letters in the correct order. When the recall was completed, feedback was given on the number of letters correctly recalled and the number of arithmetic operations correctly solved. Then, the next trial then began.
After the practice phase, participants encountered a comprehension check and were guided to the start of the test phase if they answered all the questions correctly. Otherwise, they were returned to the beginning of the instructions.
The test phase consisted of 15 trials, 3 for each different set size of 4,5,6,7,8 letters and operations. In order to asses the individual differences, the trial order and test sets remained constant across participants.
Working memory span or recall accuracy were determined as the average proportion of correct recalls across all test trials. Subjects were excluded if they scored on average less than 75% accuracy on all arithmetic operations or if they reported cheating in a final questionnaire at the end of all tasks. Participants received a bonus based on their performance in both processing and recall.

Procedure. Letters were presented for the duration of $1000\,\mathrm{ms}$ and a response on an arithmetic task had to be made within $6000\,\mathrm{ms}$. For the two isolated trials of solving arithmetic operations in the practice phase, participants received immediate feedback on their accuracy for $1000\,\mathrm{ms}$. For the full task, no feedback was given in between. Participants solved arithmetic operations by using the up-arrow key to indicate that the response is 'correct' and the down-arrow key to indicate that the response is 'false'. For the full task, once all the letters had been presented and all the arithmetic operations had been solved, participants had to recall the letters in the correct sequential order by clicking the letters in the correct order on an on-screen keyboard. When the recall was completed, feedback was given on the number of letters correctly recalled and the number of arithmetic operations correctly solved for $1000\,\mathrm{ms}$.

A.2.3 Psychiatric Questionnaires

Psychiatric questionnaires were carried out at the end of the experiment for all participants.
We assessed anxiety scores using the trait version of the State-Trait Inventory for Cognitive and Somatic Anxiety (STICSA) by Ree et al. (2008). STICSA contains of 21 questions, 10 targeting cognitive anxiety and 11 somatic anxiety. It shows re-test reliabilities within subgroup higher than 0.60 and internal consistencies higher than 0.75.
To assess depression scores we used the Patient Health Questionnaire (PHQ9) by Kroenke et al. (1999). PHQ9 contains of 9 items scoring each of the 9 DSM-IV critera on a scale between '0' and '3'. It shows internal consistency higher than 0.86 and re-test reliabilities around 0.84 (Kroenke et al., 2001).
Scoring was carried out by taking the average across all relevant questions.

A.2.4 Data

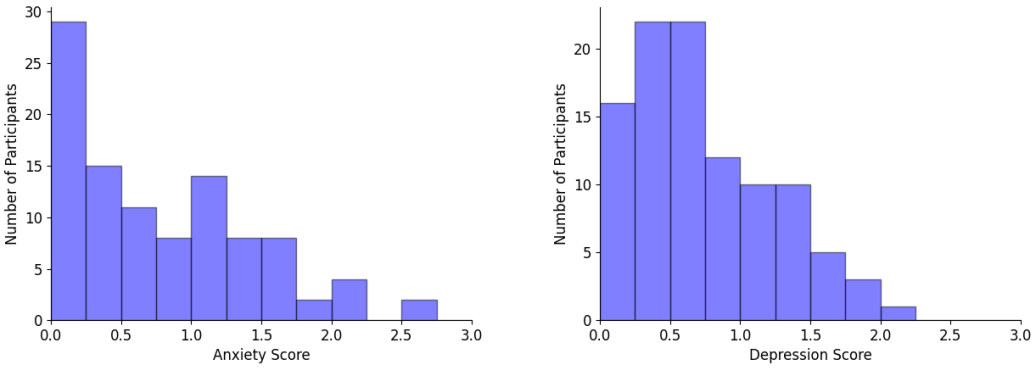

Figure 15: Left: Histogram of anxiety and scores of 101 participants on a scale from 0 to 3. Right: Corresponding histogram for depression scores.

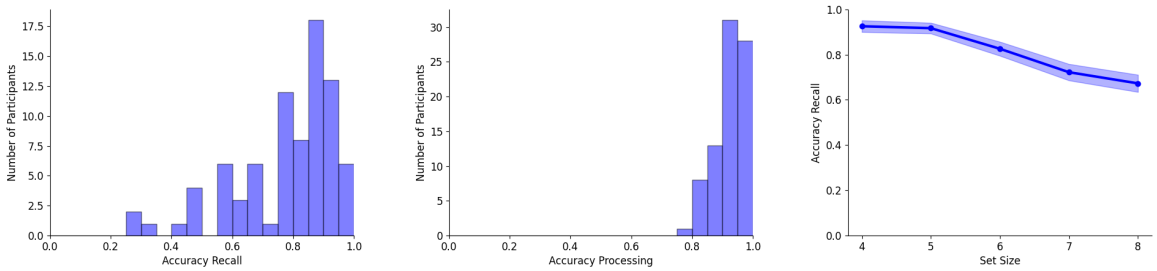

Figure 16: 81 participants were included. Left: Histogram of recall accuracy in the operation span task. Middle: Histogram of processing accuracy in the operation span task. Right: Average recall accuracy per set size, reported with 95 % confidence interval.

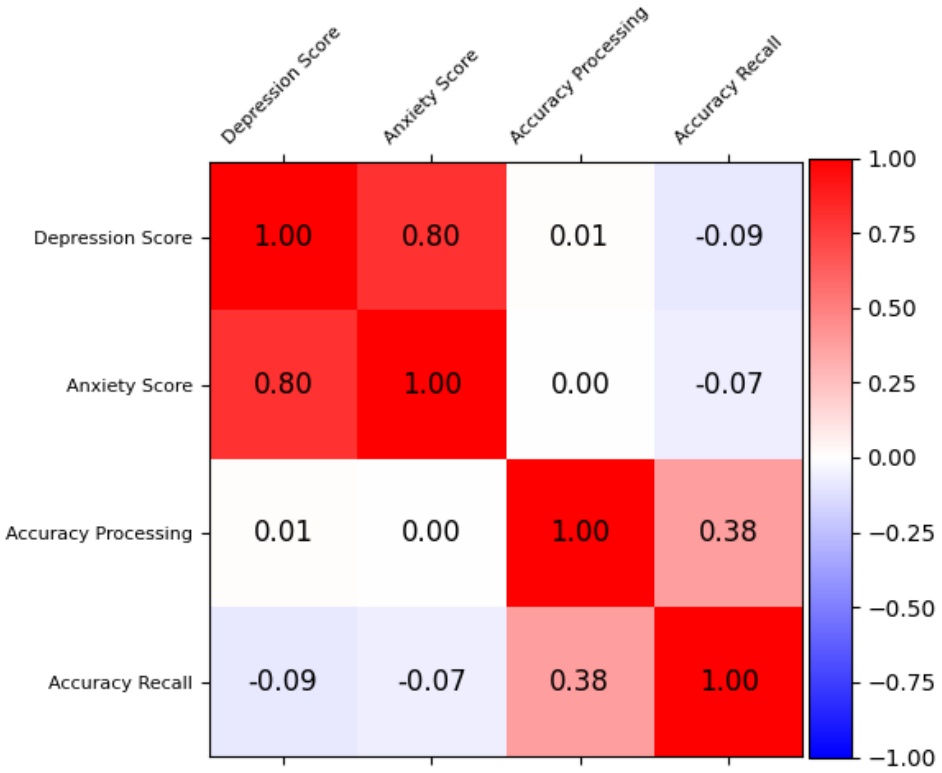

Figure 17: Correlation of depression scores, anxiety scores and recall and processing accuracy of 81 included participants.