This is a translated version of my work's main body, aided by GPT-4. After a rough manual check, there are still numerous remaining grammar errors and incoherence, but I hope the readers can get the general idea. The file name is misleading, but for those proficient in Chinese, I strongly recommend referring to the original version, bandit problems_Chinese version2.pdf, which includes more details like a comprehensive literature review and further discussion of the exploration-exploitation tradeoff.

While not clearly stated, all the reinforcement learning models discussed in this paper belong to temporal difference reinforcement learning. The partial feedback paradigm corresponds to two-armed bandit problems.

Table of Contents

Abstract

In recent years, numerous studies have demonstrated a systematic discrepancy between decisions from description and decisions from experience, especially in the weighting of small probabilities. Although this description-experience gap implies different cognitive processes for these two types of decisions, there were nearly no studies conducting a comprehensive model comparison for decisions from experience in binary choice problems.

Based on the Technion prediction tournament (Erev et al., 2010), this study compared 12 reinforcement learning (RL) models consisting of 2 utility functions, 3 choice functions, and 2 choice sensitivities, as well as 3 heuristics models derived from the win-stay-loss-shift (WSLS) strategy. By comparing the performances of post hoc fit and generalization, the study revealed: (1) For the RL models, the logarithmic utility function outperformed the power utility function. The softmax choice function outperformed the probit choice function, while the potential of the ratio choice function remained to be explored. The trial-independent choice sensitivity usually outperformed the trial-dependent choice sensitivity, but the trial-dependence remained to be explored. (2) For the heuristics models, the WSLS-2 model was always the best, supporting its two assumptions. The first was that the initial stay probabilities after win and loss trials were different. The second was that a single trial would adjust both conditional stay probabilities. (3) The heuristics models dominated in both overall performances and best performances.

The study provides a reference for future behavioral modeling studies, and shows the contribution of mathematical modeling in building new theories or improving existing theories.

Keywords: reinforcement learning, win-stay-loss-shift, mathematical modeling, model comparison

1 Introduction

In real life, people make a variety of risky decisions every day, ranging from whether to cook dinner or eat out, to whether to get married. In some situations, the decision-maker (DM) knows all the outcomes and their corresponding probabilities. For example, a gambler is preparing to buy a lottery ticket. There are two lottery tickets to buy: Lottery A costs 5 yuan, with a probability of 0.05 to win a prize of 100 yuan; lottery B costs 10 yuan, with a probability of 0.01 to win a prize of 1,000 yuan. This gambler knows the information in advance, so the decision problem can be reduced to a choice between option (-5, 0.95; 95, 0.05) and option (-10, 0.99; 990, 0.01) (here (x, p; y, q) represents an option producing outcome x with probability p, producing outcome y with probability q, and producing nothing with probability 1 - p - q). In such problems, the DMs need to understand the structure of options by consulting descriptive information, so they are called decisions from description (DFD).

However, in many situations, the DMs cannot obtain complete information in advance. For example, a worker feels very tired after a long day of work and decides to eat at a restaurant. There are two restaurants near his home: Restaurant A often serves medium-quality dishes, while Restaurant B sometimes serves high-quality dishes but sometimes low-quality dishes. Clearly, this worker has many strategies available. For example, he can score the dishes of different qualities (for example, give high, medium, and low-quality dishes 10, 5, and 0 points respectively), and then estimate the probability of dishes of different qualities appearing in different restaurants. Then it can be similar to the previous “lottery problem”. Of course, he can also directly compare his most recent dining experiences at two restaurants and choose the one that offered better food. But except for a few extremely simple strategies (for example, tossing a coin to decide which restaurant to go to), general strategies require the worker to visit the two restaurants at least once, that is, to make decisions based on the dining experiences in each restaurant. In such problems, the DMs need to infer the option structure based on their own experiences, so they are called decisions from experience (DFE).

In the two situations, if the DMs adopt the same strategy, then when the descriptive and empirical information are equivalent (for example, being told that the probability of getting an

outcome from a specified option is 0.8, or getting the outcome 8 out of 10 choices), they should make similar decisions. However, numerous studies in recent years have revealed a systematic difference in these two types of decisions (Barron & Erev, 2003; Hertwig, Barron, Weber, & Erev, 2004; Wulff, Mergenthaler-Canseco, & Hertwig, 2018). This difference is called the description-experience gap (DEG).

1.1 Description-Experience Gap

In labs, risky decisions are often simplified to binary economic choices. In other words, the DMs will face two options, each corresponding to an outcome distribution, and the two distributions are independent and static. In DFD studies, information about outcome distributions is often presented visually (e.g., pie charts and frequency diagrams) or numerically, and DMs are required to make a single choice, often without feedback (Hertwig et al., 2004; Kahneman & Tversky, 1979). In DFE studies, the DMs need to obtain information about outcomes and corresponding probabilities through repeated choices and feedback, which is used to guide further choices (Hertwig et al., 2004).

Faced with the two types of problems, people show different behavioral patterns. Specifically, in DFD, small probabilities are overweighted. That is, when the DMs evaluate a certain option, the weight of a low-probability outcome seems to be higher than its objective probability. Therefore, phenomena such as the certainty effect and reflection effect can be observed (Tversky & Kahneman, 1981). However, in DFE, small probabilities are underweighted, and the reversal of the above effects can be observed (Barron & Erev, 2003; Wulff et al., 2018).

Due to the DEG, researchers have tried to build different models for the two types of decisions. For DFD, a widely accepted theory in the psychological community is the prospect theory (PT) proposed by Kahneman and Tversky (1979). For DFE, because the history of relevant studies is relatively short and the experimental paradigms are heterogeneous, there is a lack of a unified theory to explain the cognitive processes under them.

1.2 Partial Feedback Paradigm

In DEG studies, DFE are mainly achieved through three experimental paradigms: sampling, partial-feedback, and full-feedback paradigms. Camilleri and Newell (2011) observed differences in DFE under the three paradigms when fixing the number of trials, indicating that the psychological processes triggered may not be the same.

This study focuses on DFE under the partial feedback paradigm. In this paradigm, the DMs initially do not have any prior knowledge about the options, and they need to make a specified number of choices between the two options to obtain information about the outcomes and corresponding probabilities. After each choice, only the chosen option will give feedback. This feedback will be added to the total payoff. The DMs' task is to maximize the total payoff. Figure 1.1 briefly shows the procedure. The psychological processes triggered by the partial feedback paradigm are relatively complex. First, the DMs need to evaluate the options based on experiences, i.e., feedback after each choice. Second, the DMs make repeated decisions, that is, they need to make a series of choices (Hertwig et al., 2004). Third, since only the chosen option gives feedback, the DMs face an exploration-exploitation tradeoff, a tradeoff between maximizing individual payoff based on current information and getting as much information about the options as possible to avoid shortsighted choices (Audibert, Munos, & Szepesvári, 2009).

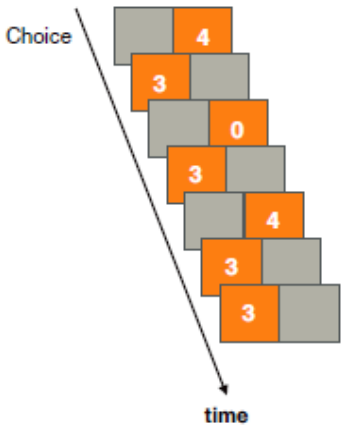

Figure 1.1 Partial feedback paradigm

Note: Quoted from Wulff et al. (2018).

For the partial feedback paradigm, two popular models are the reinforcement learning (RL) models and the heuristics models. The RL models are normative models that can capture the dynamic learning process in DFE, but the functional form of each component and the performances in different tasks are debatable (Yechiam & Busemeyer, 2005). The heuristics models can better predict DFE in some environments, but such static strategies assume that the DMs will not learn the option structure, so they meet difficulties capturing the learning process (Erev & Barron, 2005; Yoon, Vo, & Venkatraman, 2017).

1.3 Reinforcement Learning Models

The RL models generally include three components: utility function, updating function, and choice function. The utility function converts the objective value of an outcome into utility, the updating function updates the expectation of each option based on experiences, and the choice function calculates the probability of choosing each option (Ahn, Busemeyer, Wagenmakers, & Stout, 2008). The utility and choice functions are also considered in most models of DFD.

1.3.1 Utility Function

No matter for DFD or DFE, people's evaluation of the outcomes is not based on objective values, but on utilities. In the PT, Kahneman and Tversky (1982) first used the power function as a mathematical form of the utility function. In DFE, its form is

$$u(t) = \begin{cases} x(t)^{\alpha} & x(t) \geq 0 \\ -\lambda|x(t)|^{\alpha} & x(t) < 0 \end{cases} \tag{1-1}$$

Where $u(t)$ is the utility of the outcome on the t^{th} trial, and $x(t)$ is the outcome on the t^{th} trial. The parameter α controls the shape of the utility function ($0 < \alpha < 1$; for gains $\alpha \to 0$, $u(t) \to 1$; $\alpha \to 1$, $u(t) \to x(t)$). The parameter λ controls the degree of loss aversion ($\lambda > 1$; a larger λ means that the DM is more sensitive to losses).

Another widely used utility function form is the logarithmic function. There are many variations of the logarithmic utility function. A competitive one is the normalized logarithmic

utility function proposed by Scholten and Read (2010). Its form in DFE is

$$u(t) = \begin{cases} \frac{1}{\alpha} \ln (1 + \alpha x(t)) & x(t) \geq 0 \\ -\frac{\lambda}{\alpha} \ln (1 + \alpha |x(t)|) & x(t) < 0 \end{cases} \tag{1-2}$$

The parameter α controls the shape of the utility function ($\alpha > 0$; for gains, $\alpha \rightarrow 0$, $u(t) \rightarrow x(t)$; $\alpha \rightarrow \infty$, $u(t) \rightarrow 0$). Other parameters are the same as those in Formula 1-1.

No studies have compared these two utility functions for DFE. But in DFD, there is a phenomenon called the magnitude effect. Specifically, imagine two problems, the first is between options (m, p) and option (n, q), the second is between options (am, p) and (an, q), where $0 < m < n$, $0 < q < p \leq 1$, $a > 1$ and $pm = qn$. Most DMs have a higher probability of choosing the safe option (i.e., the option with smaller variance, which are (m, p) and (am, p)) in the second than the first problem. The magnitude of this tendency increases as a increases (Markowitz, 1952; Prelec & Loewenstein, 1991; Weber & Chapman, 2005). The logarithmic utility function can explain this phenomenon better than the power utility function (See Appendix for the proof). Scholten and Read (2014), and Bouchouicha and Vieider (2017) compared the two utility functions in DFD within a PT framework, both finding that the logarithmic utility function performed better.

Furthermore, the assumption that the utility is a nonlinear transformation of the objective value is supported in DFE studies. Ahn et al. (2008) conducted a model comparison study using the Iowa gambling task (IGT) and the Soochow gambling task (SGT), two DFE tasks with four options. The results showed that the power utility function performed better than the utility functions assuming linear transformations. It indicates that even in DFE, people's feelings about payoff still exhibit diminishing sensitivity. Similarly, Lejarraga and Hertwig (2017) also found that the power utility function better explained the DFE data than a identity function to map the objective value. Another finding was that the estimate of the loss aversion parameters was smaller than 1, indicating that the loss aversion phenomenon may not exist in DFE.

1.3.2 Updating Function

After each feedback, the DM needs to combine it with past experiences to update the option expectations. The updating function reflects the process of learning. A commonly used updating function is the delta learning rule (Rescorla & Wagner, 1972; Sutton & Barto, 2018), whose form is

$$E_j(t) = E_j(t-1) + A\delta_j(u(t) - E_j(t-1)) \tag{1-3}$$

where $E_j(t)$ represents the expectation of option j (j = 1, 2) on the t^{th} trial. Parameter A controls the degree to which the expectation is updated ($0 < A < 1$; a larger A indicates a stronger recency effect, that is, the greater the impact of the latest outcome). δ_j is a dummy variable that takes 1 when option j is chosen and 0 otherwise.

1.3.3 Choice Function

Stott (2006) reviewed 8 DFD studies that required DMs to make repeated choices. It showed that the choices were not constant for the same problems, indicating the inherent noise in choice behavior. But the frequency to choose the better option increased when the difference in utility increased.

There are two mainstream accounts for the choice randomness. The first is the Luce's choice axiom (LCA; Luce, 1959), which defines the probability of choosing option x in a set of option sets S as

$$P_S(x) = \frac{v(x)}{\sum_{y \in S} v(y)} \tag{1-4}$$

where $v(x)$ is the response strength of option x. In DFE, a intuitive idea is replacing it with option expectation.

$$P\left(D(\mathrm{t}+1)=j\right)=\begin{cases}\dfrac{E_j(t)^{\theta(t)}}{E_j(t)^{\theta(t)}+E_{\sim j}(t)^{\theta(t)}}=\dfrac{1}{1+\left(\dfrac{E_{\sim j}(t)}{E_j(t)}\right)^{\theta(t)}} & gain\ domain\\ \dfrac{|E_{\sim j}(t)|^{\theta(t)}}{|E_j(t)|^{\theta(t)}+|E_{\sim j}(t)|^{\theta(t)}}=\dfrac{1}{1+\left(\dfrac{E_j(t)}{E_{\sim j}(t)}\right)^{\theta(t)}} & loss\ domain\end{cases}$$

(1-5)

Where $P(D(t+1)=j)$ represents the probability of choosing option j (j = 1, 2) on the $(t+1)^{\text{th}}$ trial, $E_{\sim j}(t)$ represents the expectation of the mutually exclusive option on the t^{th} trial (if j = 1, then $\sim j$ = 2, and vice versa). $\theta(t)$ represents the choice sensitivity on the t^{th} trial. A larger $\theta(t)$ indicates a stronger choice sensitivity, that is, the DM is more inclined to choose the currently better option. A gain domain problem refers to one in which all possible outcomes are non-negative. A loss domain problem refers to one in which all possible outcomes are non-positive.

The function is based on the ratio between option expectations, so I call it the ratio choice function. However, it does not apply to the situation when the option expectations have different signs.

A more commonly used LCA function is the softmax choice function, which is

$$P(D(t+1)=j)=\frac{e^{\theta(t)E_j(t)}}{e^{\theta(t)E_j(t)}+e^{\theta(t)E_{\sim j}(t)}}=\frac{1}{1+e^{\theta(t)\left(E_{\sim j}(t)-E_j(t)\right)}}$$
$$=\mathrm{logistic}\left(\theta(t)\left(E_j(t)-E_{\sim j}(t)\right)\right)$$

(1-6)

All parameters are the same as those in Formula 1-5. The softmax function reduces to the logistic function in binary choice tasks. The function is based on the difference between expectations, so it does not require the expectations to have the same sign.

The second account is the random utility theory (RUT). RUT can be traced back to Thurstone's (1927) Case V theory for pairwise comparisons, and has been widely used in choice modeling since the 1960s (Luce & Suppes, 1965; Yellott, 1977). This theory assumes that for each option, the DM's evaluation contains a fixed component and an independent and identically distributed random component, that is,

$$U(x) = u(x) + \varepsilon_x$$

(1-7)

where $u(x)$ is its utility. ε_x follows a zero-centered probability distribution. The DM will choose the option with the largest evaluation value deterministically. When the random component follows a Gaussian distribution, the function is called the probit choice function. That is,

$$P(D(t+1) = j) = P\big(E_j(t) + \varepsilon_j(t) > E_{\sim j}(t) + \varepsilon_{\sim j}(t)\big) = \Phi\left(\frac{E_j(t) - E_{\sim j}(t)}{\theta(t)}\right)$$

(1-8)

where $\Phi(\cdot)$ is the cumulative Gaussian distribution function, $\varepsilon_j(t)$ and $\varepsilon_{\sim j}(t)$ are the random components for option j and its mutually exclusive options (j = 1, 2) respectively on the t^{th} trial. Other parameters are the same as those in Formula 1-5. However, $\theta(t)$ here can also be understood as the variance of the Gaussian distribution on the t^{th} trial. A smaller $\theta(t)$ indicates a smaller variance. Similar to the softmax choice function, the probit choice function is also difference-based.

In DFE studies, the softmax choice function is most widely used as it is computationally simple and can be easily extended to the problems with more than two options (Schulz & Gershman, 2019). In addition, its formulation of the exploration-exploitation trade-off has some neurophysiological plausibility (Daw, O'doherty, Dayan, Seymour, & Dolan, 2006; Collins & Frank, 2014), and indeed showed advantages in some studies (Daw et al., 2006; Yechiam & Busemeyer, 2005).

However, the ratio and probit choice functions may still be interesting. For the former, Worthy, Maddox, and Markman (2008) designed a DFE task in which the distribution of option outcomes dynamically changed. There were three types of problems. The first were the baselines. For the second, a fixed amount was added to each possible outcome (the differences between outcomes were the same). For the third, each possible outcome was multiplied by a fixed amount (the ratios between outcomes were the same). The results showed that the choices in the first and third types of problems were similar but significantly different from those in the second type of problems. It somewhat supported the ratio choice function, with a flaw that the objective values were not equivalent to utilities. For the latter, it is a representative member in RUT. The RUT's assumption that people's evaluation of options is random has been used in

some RL models (Gershman, 2018; Schulz & Gershman, 2019).

1.3.4 Choice Sensitivity

In each choice function, there is a choice sensitivity parameter capturing the other factors influencing the choices. In DFD studies, the choice sensitivity is usually trial-independent.

$$\theta(t) = 3^c - 1 \tag{1-9}$$

The parameter c controls the strength of choice sensitivity ($c > 0$; a larger c indicates a stronger choice sensitivity across trials).

However, the DMs need to make repeated choices in DFE problems, the practice or fatigue effect may be strong, which can be represented by another choice sensitivity

$$\theta(t) = (\frac{t}{10})^c \tag{1-10}$$

The parameter c controls how large the choice sensitivity changes across trials. When $c > 0$,the DM's choice sensitivity increases with more trials. It reflects that the DM is more and more confident about the estimates of expectations, which can be regarded as a practice effect. When $c < 0$, the DM's choice sensitivity weakens with more trials. It reflects that fatigue or boredom leads to random choices (the above inferences are limited to the ratio and softmax choice functions, which is on the contrary for the probit choice function). The denominator of 10 means that in the first 10 trials, the DM is mainly exploring (i.e., understanding option structures). While in subsequent trials, the DM is mainly exploiting (i.e., maximizing payoffs).

In the study of Ahn et al.'s (2008), the best model used trial-independent sensitivity, but the trial-dependent choice sensitivity performed better overall. Beitz, Salthouse and Davis (2014) fitted the RL models on choice data of different age groups in IGT tasks. They found that the trial-independent sensitivity exhibited an advantage in explaining the behavior of adults and elderly people.

In summary, this study will compare 2 (utility function) × 3 (choice function) × 2 (choice sensitivity) = 12 RL models.

1.4 Heuristics Models

Although the RL model is a normative cognitive model, people may not use such a complex strategy when making repeated choices. They are more likely to make decisions based on only a few attributes, like the relative magnitude of possible outcomes or the probability of getting the largest payoff. In other words, they may prefer simple heuristics. For DFE, a representative heuristics is the win-stay-loss-shift (WSLS). This model assumes that people tend to hold their choices if receiving rewards and change their choices if not. That is

$$P(D(t+1)=j)=\begin{cases} p(stay|win) & D(t)=j\ \&\ x(t)\geq x(t-1) \\ 1-p(stay|win) & D(t)=\sim j\ \&\ x(t)\geq x(t-1) \\ p(shift|loss) & D(t)=\sim j\ \&\ x(t)<x(t-1) \\ 1-p(shift|loss) & D(t)=j\ \&\ x(t)<x(t-1) \end{cases} \tag{1-11}$$

When the outcome on the t^{th} trial is not smaller than that on the $(t - 1)^{th}$ trial, this trial is a win trial. On the $(t + 1)^{th}$ trial, the DM will hold the choice (i.e., stay) with a conditional probability of p(stay|win), or change the choice (i.e., shift) with a conditional probability of 1 - p(stay|win) (i.e., p(shift|win)). When the outcome on the t^{th} trial is smaller than that on the $(t - 1)^{th}$ trial, this trial is a loss trial. On the $(t + 1)^{th}$ trial, the DM will shift with a conditional probability of p(shift|loss), or stay with a conditional probability of 1 - p(shift|loss) (i.e., p(stay|loss)). Therefore, the model contains only two parameters. The larger p(stay|win) is, the more inclined the DM is to stay after a win trial. The larger p(shift|loss) is, the more inclined the DM is to shift after a loss trial. Other parameters are the same as those in Formulas 1-1 and 1-5.

Although in some studies, the WSLS model showed similar or even better performance than the RL model (Worthy, Hawthorne, & Otto, 2013; Worthy, Otto, & Maddox, 2012), like other heuristics models, the WSLS model cannot capture the dynamic learning process. To solve this, inspired by the associative learning model of Estes (1950), Worthy and Maddox (2014) proposed a modified model. I call the WSLS-1 model. It assumes that if the t^{th} trial is a win trial, it will make

$$p(stay|win)_{t+1}=p(stay|win)_t+\theta_{p(stay|win)}(1-p(stay|win)_t)$$

$$p(shift|loss)_{t+1}=p(shift|loss)_t$$

(1-12)

$p(\text{stay}|\text{win})_t$ refers to the probability of staying on the t^{th} trial if the $(t - 1)^{\text{th}}$ trial is a win trial. $p(\text{shift}|\text{loss})_t$ refers to the probability of shifting on the t^{th} trial if the $(t - 1)^{\text{th}}$ trial is a loss trial. The parameter $\theta_{p(\text{stay}|\text{win})}$ represents the adjustment of a win trial on the subsequent stay probability ($0 < \theta_{p(\text{stay}|\text{win})} < 1$; a larger $\theta_{p(\text{stay}|\text{win})}$ indicates a larger degree of adjustment).

If the t^{th} trial is a loss trial, it will make

$$p(shift|loss)_{t+1} = p(shift|loss)_t + \theta_{p(shift|loss)}(1 - p(shift|loss)_t)$$

$$p(stay|win)_{t+1} = p(stay|win)_t$$

(1-13)

The parameter $\theta_{p(\text{shift}|\text{loss})}$ represents the adjustment of a loss trial on the subsequent shift probability ($0 < \theta_{p(\text{shift}|\text{loss})} < 1$; a larger $\theta_{p(\text{shift}|\text{loss})}$ indicates a larger degree of adjustment). Other parameters are the same as those in Formula 1-12. The model also allows the DM to have different initial conditional probabilities. In other words, there are two other parameters $p(\text{stay}|\text{win})_1$ and $p(\text{shift}|\text{loss})_1$.

Worthy and Maddox (2014) combined this model with a RL model to form a weighted one and found their weights were similar in a DFE task. They believed it reflected that people would use mixed strategies.

A key assumption in the WSLS-1 model is that a win or loss trial only adjusts its relevant conditional probability. However, Lejarraga and Hertwig (2017) modified it. They believed that both conditional probabilities were adjusted, which is called the WSLS-2 model here. It assumes that if the t^{th} trial is a win trial, it will make

$$p(stay|win)_{t+1} = p(stay|win)_t + \theta_{p(stay|win)}(1 - p(stay|win)_t)$$

$$p(shift|loss)_{t+1} = (1 - \theta_{p(shift|loss)})p(shift|loss)_t$$

(1-14)

If the t^{th} trial is a loss trial, it will make

$$p(shift|loss)_{t+1} = p(shift|loss)_t + \theta_{p(shift|loss)}(1 - p(shift|loss)_t)$$

$$p(stay|win)_{t+1} = (1 - \theta_{p(stay|win)})p(stay|win)_t$$

(1-15)

All parameters are the same as those in Formulas 1-12 and 1-13.

Lejarraga and Hertwig (2017) compared the WSLS-2 model and several RL models in a DFE task, which showed an obvious advantage.

Both the WSLS-1 and WSLS-2 models assume two conditional probabilities. However, there seems to be no studies testing this assumption. In addition, in Estes's (1950) associative learning model, there is only one conditional probability. Therefore, this study proposes a new heuristics model to test the necessity of two conditional probabilities. I call it the simple stay (SS) model. The SS model assumes that the DM only has a simple tendency to stay, but it is adjusted by experiences. If the t^{th} trial is a win trial, it will make

$$p(stay)_{t+1} = p(stay)_t + \theta(1 - p(stay)_t) \tag{1-16}$$

where $p(\text{stay})_t$ refers to the probability of staying on the t^{th} trial, and the parameter θ represents the adjustment of a single trial on the subsequent stay probability ($0 < \theta < 1$; a larger θ indicates a larger degree of adjustment).

If the t^{th} trial is a loss trial, it will make

$$p(stay)_{t+1} = (1 - \theta)p(stay)_t \tag{1-17}$$

All parameters are the same as those in Formula 1-16. The model also contains a parameter $p(\text{stay})_1$ representing the initial stay probability. In fact, the SS model is equivalent to a restricted WSLS-2 model, which requires that $p(\text{stay}|\text{win})_1$ and 1 - $p(\text{shift}|\text{loss})_1$ are equal, and $\theta_{p(\text{stay}|\text{win})}$ and $\theta_{p(\text{shift}|\text{loss})}$ are equal.

In summary, this study will compare 3 heuristics models: WSLS-1, WSLS-2, and SS.

1.5 Research Significance

Most previous DFE studies did not use binary choice tasks, but focused on other tasks such as IGT and probability learning tasks. IGT contains 4 options, and each option will produce a certain gain and a possible loss. In probability learning tasks, the possible outcomes are only 0 and 1. In each trial, there must be an option producing 0 and the other producing 1. So their findings may not apply to the binary choice DFE tasks.

In addition, few studies have systematically compared the combination of different utility functions, choice functions, and choice sensitivities for RL models. A review of DFD models (Stott, 2006) has shown that there may be interactions among different components of models. So the model consisting of the best components may not be the best model. Meanwhile, in the DFE study of Ahn et al. (2008), there was an interaction between the choice sensitivity and the updating function on model performance. Therefore, finding the best RL model still requires testing all combinations of different components.

As for the heuristics model, although Lejarraga and Hertwig (2017) have found the advantages of the WSLS-2 model over some RL models, its assumptions of two conditional probabilities, and that each trial will adjust both conditional probabilities have not been tested. So it is necessary to compare it with the WSLS-1 and SS models to find the most appropriate one.

To fill the gap in the above fields, this study will conduct a systematic model comparison using a DFE dataset based on the partial feedback paradigm. Specifically, this study will use multiple metrics to compare the fitting and generalization performance of 12 RL models and 3 heuristics models. It will analyze the behavioral patterns in DFE from the perspectives of RL and heuristics strategies respectively, and compare the best RL and heuristics models, to figure out the main strategy used by DMs.

In a theoretical sense, mathematical modeling can help researchers infer the underlying cognitive mechanisms of behavior and separate its processes. Model comparison can help to propose new theories or improve existing theories. For example, among the functional form of each RL model, the combination of the power utility function and the ratio choice function is the only one that cannot explain the magnitude effect in DFD (see the Appendix for proof). So at least for DFD, it should not be used to explain the data. If this combination also performed worse in this study, it can be speculated that the magnitude effect also exists in DFE. Erev et al. (2010) pointed out that the establishment of many important theories came from the observation and summary of mathematical laws. For example, the Weber-Fechner law originated from the pattern found by Weber when studying the minimum noticeable difference in weight.

In a practical sense, despite the development of information technology enabling people to obtain descriptive information to make decisions, some personal choices, such as whether to

start a business or whether to go shopping, relies on experiential information. The existence of DEG shows that the two types of decisions are not equivalent. Therefore, this study can help us better understand how we make decisions when learning in a dynamic environment, and can also help relevant institutions to design more appropriate nudges.

2 Model Fitting

2.1 Dataset

The dataset came from the Technion prediction tournament (TPT) organized by Erev et al. (2010), which was the largest public dataset for DFE currently. TPT provided repeated choice data for 120 DFE problems under the partial feedback paradigm. Each problem contained a certain option and a risky option. Each risky option had two possible outcomes. There were 40 problems in the gain domain, 40 problems in the loss domain, and 40 problems in the mixed domain. The mixed domain referred to the situation where the risky option had one positive and one negative possible outcomes. For the certain option in the mixed domain problems, half of them gave gains and the rest gave losses. For all risky options, 40 of them contained a high positive outcome with a small probability (i.e. $p < 0.1$), and 40 of them contained a high negative outcome with a small probability.

The 120 problems were divided into 10 groups, each containing 4 problems in the gain domain, 4 problems in the loss domain, and 4 problems in the mixed domain. Each problem consisted of 100 trials. Each group was completed by 20 participants, and each participant completed all 12 problems. At the end of the experiment, one of the trials completed by the participant was randomly selected, whose outcome was paid to the participant in Sheqel, an Israeli currency (there was still an exploration-exploitation trade-off, because the more times an option was chosen, the more likely the participant could get an outcome from this option).

Among the 200 participants, 7 of them chose an option no more than 2 times in more than half of the problems. This extreme behavior reflected that they might have not completed the experiment carefully, so their data were excluded from subsequent analyses.

2.2 Parameter Estimation

This study used the maximum likelihood method to estimate the parameters. Since each participant completed multiple problems, this study used the choices in the gain domain and loss domain problems as the training set, and the choices in the mixed domain problems as the test set. The models were fitted at the participant level. Additionally, for the following two reasons, only the last 90 choices in the training set were used to calculate the likelihood function: First, in the early stage, the participants' choices were relatively random. For example, they might consistently choose an option in the first few trials. This extreme behavior could lead to a large difference in expectations. If the initialization was also extreme (for example, a large parameter c in the choice sensitivity), it might lead to numerical underflow. Second, for the three heuristics models, win or loss could only be defined after the second trial. Since the likelihood function is also influenced by the number of data points, I fixed the size of training set to make the metrics comparable across models.

To search for the best combination of parameter values, I calculated the log-likelihood function for each model and each participant

$$LL_{model} = \ln L(data|model)$$
$$= \sum_{q=1}^{8} \sum_{t=11}^{100} \sum_{j=1}^{2} \ln\left(P\left(D(t,q) = j \middle| X(t-1,q), Y(t-1,q)\right)\right) \delta_j(t,q) \quad (2\text{-}1)$$

where q referred to the problem number. $Y(t - 1, q)$ referred to the participant's first $(t - 1)$ choices in the q^{th} problem. $X(t - 1, q)$ referred to the first $(t - 1)$ outcomes in the q^{th} problem. $D(t, q) = j$ meant that option j was chosen on the t^{th} trial of the q^{th} problem. $\delta_j(t, q)$ was a dummy variable that took the value of 1 when option j was chosen on the t^{th} trial in the q^{th} problem, and 0 otherwise.

To find parameter values that maximize the log-likelihood function, this study used grid simplex search (Nelder & Mead, 1965). The grids were combinations of quartering points of each parameter. The SS model had 9 grids, and other models had 81 grids. Each grid served as an initialization position. The searched log-likelihood and its corresponding parameters were recorded. The final estimates were the parameters that produced the maximum searched value.

To prevent numerical underflow, it was necessary to limit the parameter ranges: for the power utility function, $0 < \alpha < 1$; for the logarithmic utility function, $0 < \alpha < 5$; and for both utility functions, $0 < \lambda < 5$. For the trial-independent choice sensitivity, $0 < c < 3$ when using the ratio and softmax choice functions, and $0.25 < c < 3$ when using the probit choice function. For the trial-dependent choice sensitivity, $-1.5 < c < 1.5$ when using the ratio and softmax choice functions, and $-0.5 < c < 1.5$ when using the probit choice function (this is because the parameter served a contrary role in probit choice function compared to the first two choice functions). For the three heuristics models, all parameters took values between 0 and 1. Similarly, for the models with the ratio choice function, the initial expectations of the two options in the gain domain problem were 0.0001, and the initial expectations of the two options in the loss domain problem were -0.0001. For other models, the initial expectations were 0.

Parameter estimation and subsequent data analysis were completed using R 4.0.0. The mean parameter estimates of each model were shown in Table 2.1 and Table 2.2.

Table 2.1 Mean parameter estimates with standard deviations of RL models

Model			Mean Estimates			
Utility Function	Choice function	Choice sensitivity	α	λ	A	c
power utility	ratio	trial-independent	0.61(0.43)	0.29(0.95)	0.57(0.30)	1.73(0.79)
		trial-dependent	0.65(0.43)	0.18(0.70)	0.52(0.32)	0.69(0.50)
	softmax	trial-independent	0.47(0.41)	1.30(1.88)	0.47(0.35)	1.47(0.92)
		trial-dependent	0.51(0.31)	1.02(1.64)	0.42(0.32)	0.69(0.60)
	probit	trial-independent	0.51(0.38)	1.13(1.72)	0.43(0.34)	0.67(0.60)
		trial-dependent	0.47(0.27)	1.16(1.67)	0.44(0.33)	-0.25(0.37)
logarithmic utility	ratio	trial-independent	3.39(2.28)	0.39(1.21)	0.44(0.32)	1.74(0.74)
		trial-dependent	3.13(2.39)	0.29(1.01)	0.41(0.33)	0.68(0.47)
	softmax	trial-independent	2.29(2.32)	1.15(1.76)	0.44(0.34)	1.49(0.94)
		trial-dependent	1.05(1.35)	0.98(1.60)	0.45(0.32)	0.64(0.64)
	probit	trial-independent	1.41(1.68)	1.11(1.67)	0.41(0.34)	0.71(0.65)
		trial-dependent	0.98(0.97)	1.10(1.66)	0.37(0.35)	-0.22(0.40)

Table 2.2 Mean parameter estimates with standard deviations of heuristics models

Model	Mean estimates					
	$p(stay\|win)_1$	$p(\text{shift}\|\text{loss})_1$	$\theta p_{(\text{stay}\|\text{win})}$	$\theta p_{(\text{shift}\|\text{loss})}$	$p(\text{stay})_1$	θ
WSLS-1	0.85(0.20)	0.40(0.41)	0.05(0.08)	0.29(0.29)	—	—
WSLS-2	0.82(0.13)	0.35(0.24)	0.02(0.02)	0.02(0.08)	—	—
SS	—	—	—	—	0.75(0.28)	0.07(0.06)

Since this study mainly focused on model comparison rather than parameter differences, no analysis of the parameter estimates was conducted.

3 Fitting Performance Comparison

3.1 Evaluation Metrics

To evaluate the fitting performance, this study used the Bayesian information criterion (BIC; Schwartz, 1978) and the mean square deviation (MSD; Yechiam, & Busemeyer, 2005) as evaluation metrics.

To calculate the *BIC* of a model, the first step was calculating its G^2

$$G^2 = 2(LL_{model} - LL_{baseline}) \tag{3-1}$$

LL_{model} was the final maximum log-likelihood value of the target model, and LL_{baseline} was that of the baseline model. The baseline model was the Bernoulli model, which assumed the risky option was always chosen with probability p.The parameter estimate was equal to the proportion of risky choices in the 720 choices.

BIC gave a penalty for the number of parameters, i.e.

$$BIC = G^2 - \Delta k \ln(N) \tag{3-2}$$

where Δk was the difference in the number of parameters between the target model and baseline models. For the SS model, $\Delta k = 1$, and for other models, $\Delta k = 3$. N was the size of training set, which was 720. When the target model explained the data better than the baseline model, the *BIC* was positive, and it increased as the advantage of the target model increased.

BIC measured the model's ability to predict one step ahead. That is, the predicted choice conditioned on the actual choices and outcomes before. This was a short-range prediction. *MSD* measured the long-term prediction ability of the target model without actual data. Specifically, I used the target model to generate a simulated choices and compared it with the actual choices. Because each participant's initial choices were highly random but could significantly influence

the subsequent choices, the *MSD* was ultimately calculated at the group level for each problem

$$MSD = \frac{1}{200}\sum_{t=1}^{100}\sum_{j=1}^{2}\left(\overline{D}_{exp,j}(t) - \overline{D}_{sim,j}(t)\right)^2$$
$$= \frac{1}{100}\sum_{t=1}^{100}\left(\overline{D}_{exp,risk}(t) - \overline{D}_{sim,risk}(t)\right)^2 \quad (3\text{-}3)$$

where $\overline{D}exp,j(t)$ was the proportion of all participants who chose option *j* on the t^{th} trial in the specified problem, and $\overline{D}sim,j(t)$ was the proportion of simulated choices of option *j* on the t^{th} trial in this problem. Since there were only two options in the task, the calculation could be reduced to the final form of Formula 3-3, where $\overline{D}exp,risk(t)$ was the proportion of all participants who chose the risky option on the t^{th} trial in this problem, and $\overline{D}sim, risk(t)$ was the proportion of simulated risky choices on the t^{th} trial in this problem. A smaller *MSD* indicated a better model performance.

Since *MSD* was generally a metric for continuous variables, the actual and simulated choices needed to be smoothed before calculation. For the actual data, each participant's 4^{th} to 97^{th} choices were replaced with the moving mean of 7 nearby points. For the simulated data, the final choices was the mean of 100 simulations. The study used percentages instead of decimals to demonstrate *MSD*.

Because the empirical distributions of the two metrics might not be Gaussian, I also calculated the mean grade for each model at the participant/problem level like Stott (2006). That was, for each participant or problem, I ranked the models based on the corresponding metric (the best was recorded as 1, the worst was recorded as 15). Then I calculated the mean of this rank for them.

3.2 *BIC* Comparison

The distributions of *BIC* statistics for different models were shown in Table 3.1. It also listed the results of the Shapiro-Wilk normality test (Shapiro & Wilk, 1965) of the *BIC* statistics. Since some RL and heuristics models failed the normality test ($p < .05$), this study used aligned rank transform (ART) repeated measures analysis of variance to compare models. ART was a

non-parametric method that could handle interactions in multi-factor designs and overcame the type I error inflation caused by the direct rank transformation of data in traditional methods (Wobbrock, Findlater, Gergle, & Higgins, 2011).

Table 3.1 *BIC* distributions of different models

Model			*BIC*							
Utility Function	Choice function	Choice sensitivity	10% quantile	Median	90% quantile	Mean	Standard deviation	Mean grade	SW test Significance	BIC > 0 percentage
power utility	ratio	trial-independent	-68.69	72.88	292.17	93.62	147.93	12.27	0.066	76.17
		trial-dependent	-67.38	82.76	298.41	100.67	152.66	11.99	0.179	76.68
	softmax	trial-independent	-31.55	154.03	336.52	144.95	147.01	7.92	0.505	84.97
		trial-dependent	-13.71	165.19	353.98	157.79	144.31	8.06	0.887	87.05
	probit	trial-independent	-24.68	142.47	342.17	140.99	147.76	9.62	0.195	85.49
		trial-dependent	-27.08	138.78	338.31	137.58	146.38	9.80	0.424	84.97
logarithmic utility	ratio	trial-independent	-46.43	147.78	502.98	182.23	210.09	7.74	< .001	82.38
		trial-dependent	-32.84	143.58	431.48	167.25	186.72	8.63	0.006	84.97
	softmax	trial-independent	-31.84	147.80	353.09	147.80	144.39	8.28	0.822	84.97
		trial-dependent	-14.21	158.48	353.44	156.41	143.73	8.37	0.892	86.53
	probit	trial-independent	-32.66	137.77	345.71	136.69	146.45	10.38	0.300	85.49
		trial-dependent	-29.59	130.87	339.49	136.48	145.90	10.16	0.558	84.97
WSLS-1			181.45	509.54	763.27	474.76	230.67	2.41	0.002	98.45
WSLS-2			219.72	525.82	756.87	493.63	214.02	1.89	0.003	98.45
SS			178.94	504.50	760.12	476.97	228.16	2.47	0.001	97.93

Note: For all models, the percentage of *BIC* > 0, was significantly greater than 50% ($\alpha = .05$, one-sided test).

With utility function (power, logarithmic), choice function (ratio, softmax, probit), and choice sensitivity (trial-independent, trial-dependent) as the independent variables, and *BIC* as

the dependent variable, an ART repeated measures analysis of variance was conducted. The results showed that the main effect of the utility function was significant, $F(1, 2112) = 92.35$, $p < .001$. The main effect of the choice function was significant, $F(2, 2112) = 48.76$, $p < .001$. The interaction between the utility function and the choice function was significant, $F(2, 2112) = 103.6$, $p < .001$. The interaction between the choice function and the choice sensitivity was significant, $F(2, 2112) = 3.25$, $p = .039$.

When fixing the choice sensitivity to trial-independent, a subsequent ART repeated measures analysis of variance was conducted. The results showed that the main effect of the utility function was significant, $F(1,960) = 60.88$, $p < .001$. The main effect of the choice function was significant, $F(2, 960) = 17.17$, $p < .001$. The interaction between the utility function and the choice function was significant, $F(2, 960) = 65.23$, $p < .001$.

Because the ART method transformed the data into ranks, it could not directly test simple effects, but it could test the differences of simple effect. I used the difference in mean aligned-rank-transformed *BIC* between the power and logarithmic utility models as the simple effect. The results showed that there was no significant difference in the simple effects between the probit and softmax choice models, $t(960) = -0.924$, $p = .356$. When the choice function was the ratio, the simple effect was significantly lower than that when the choice function was softmax, $t(960) = -9.40$, $p < .001$, and also significantly lower than that when the choice function was probit, $t(960) = -10.32$, $p < .001$.

Figure 3.1 was a violin plot of the *BIC* distributions of RL models when using the trial-independent choice sensitivity. The empirical distribution was estimated using the kernel method. Combining with the above tests and the mean grade, when the choice sensitivity was trial-independent, the best RL model was the logarithmic + ratio + trial-independent model, and the worst RL model was the power + ratio + trial-independent model.

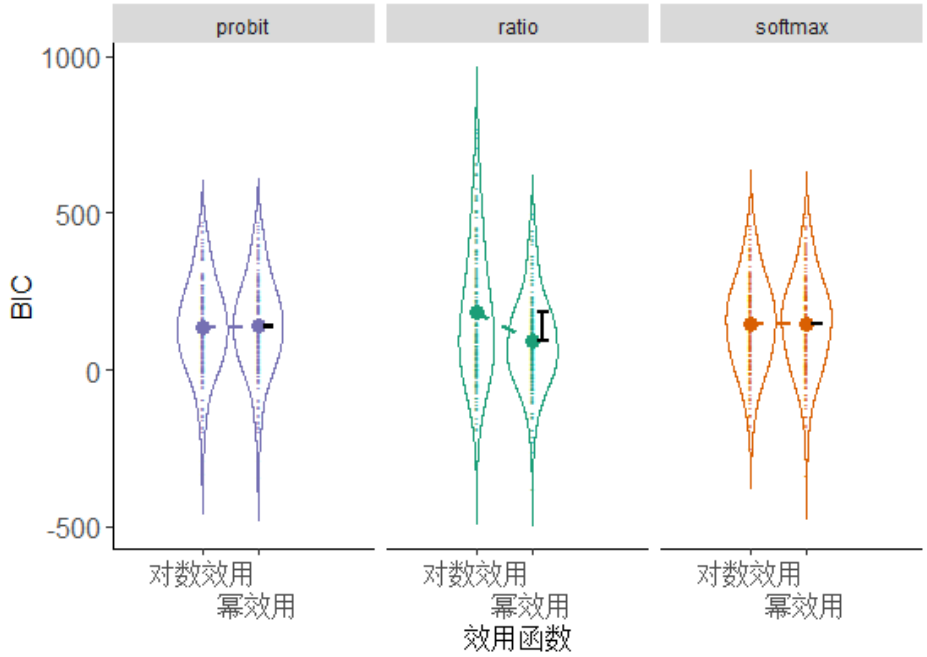

Figure 3.1 *BIC* distributions of RL models (trial-independent choice sensitivity)

Note: The dots represented the *BIC* mean, and the black line indicated the simple effect of the utility function (for original, not aligned-rank-transformed BIC). 对数效用=logarithmic utility, 幂效用=power utility.

When fixing the choice sensitivity to trial-dependent, a subsequent ART repeated measures analysis of variance was conducted. The results showed that the main effect of the utility function was significant, $F(1,960) = 31.88$, $p < .001$. The main effect of the choice function was significant, $F(2, 960) = 32.65$, $p < .001$. The interaction between the utility function and the choice function was significant, $F(2, 960) = 37.61$, $p < .001$.

A further test on the simple effect showed that there was no significant difference in the simple effect between the softmax and probit choice models, $t(960) = -0.090$, $p = .929$. When the choice function was ratio, the simple effect was significantly lower than that when the choice function was softmax, $t(960) = -7.47$, $p < .001$, and also significantly lower than that when the choice function was probit, $t(960) = -7.56$, $p < .001$.

Figure 3.2 was a violin plot of the *BIC* distributions of the RL models when using trial-dependent choice sensitivity. Combining with the above tests and the mean grade, when the choice sensitivity was trial-dependent, the best RL model was the logarithmic + ratio + trial-dependent model, and the worst RL model was the power + ratio + trial-dependent model. .

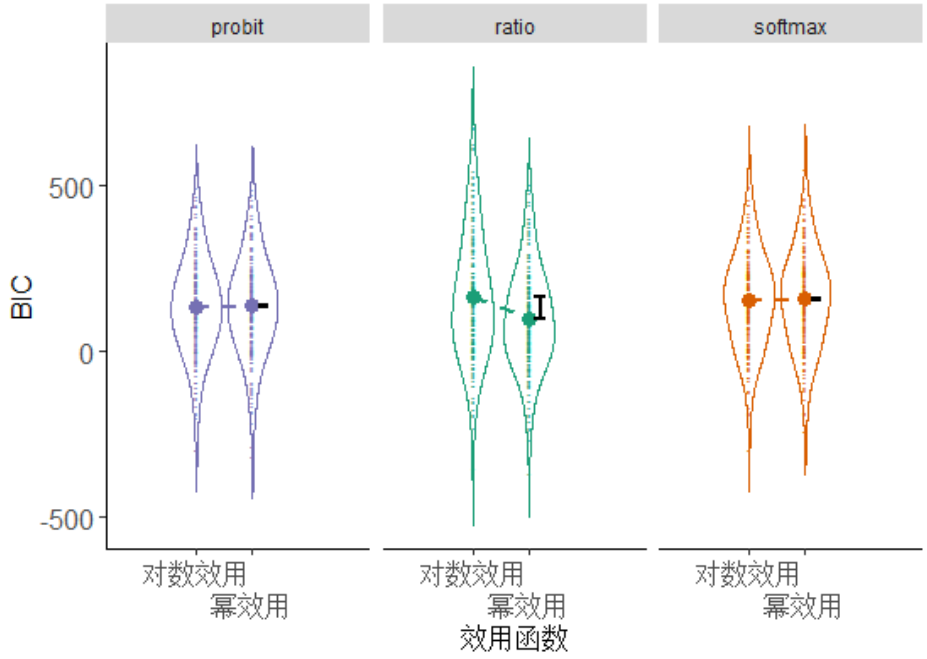

Figure 3.2 *BIC* distributions of RL models (trial-dependent choice sensitivity)

Note: The bold dots represented the mean *BIC*, and the black line indicated the simple effect of the utility function (for original, not aligned-rank-transformed BIC). 对数效用=logarithmic utility, 幂效用=power utility.

Finally, whether considering the mean or the mean grade of the original BIC, the logarithmic + ratio + trial-independent model was obviously better than the logarithmic + ratio + trial-dependent model. So the best RL model based on *BIC* was the logarithmic + ratio + trial-independent model.

With model type (WSLS-1, WSLS-2, SS) as the independent variable, and *BIC* as the dependent variable, an ART repeated measures analysis of variance was conducted. The results showed that the effect of model type was significant, $F(2, 384) = 29.42$, $p < .001$. The post hoc comparison showed no significant difference in mean *BIC* between the WSLS-1 and SS models, $t(384) = -0.71$, $p = .758$. The mean *BIC* of the WSLS-2 model was significantly higher than that of the WSLS-1 model, $t(384) = 6.97$, $p < .001$, and that of the SS model, $t(384) = 6.26$, $p < .001$.

Figure 3.3 was a violin plot of the *BIC* distributions of different heuristics models. Combined with the above tests and the mean grade, WSLS-2 was the best heuristics model.

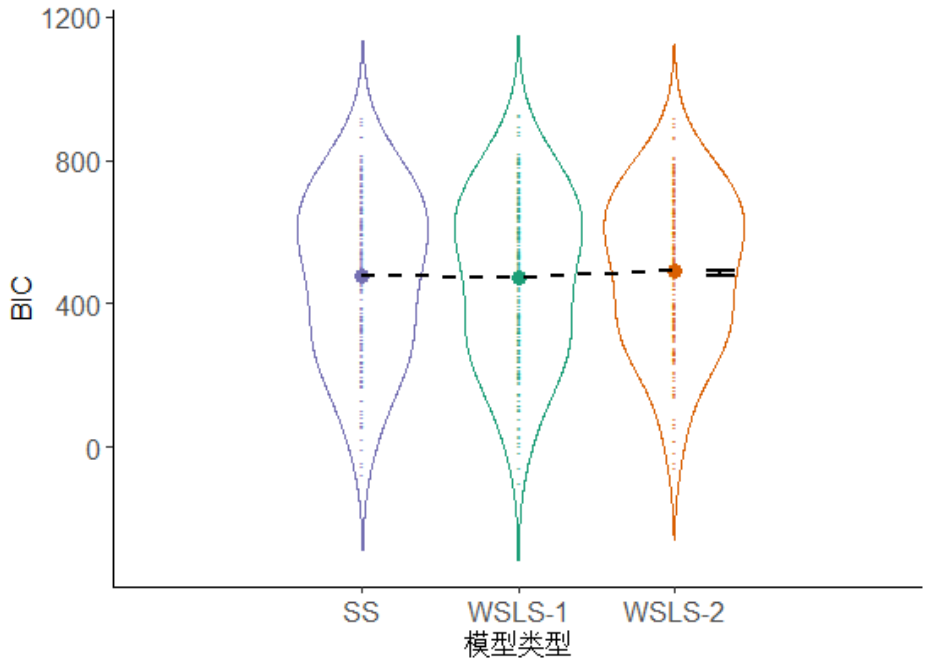

Figure 3.3 *BIC* distributions of heuristics models

Note: The dots represented the *BIC* mean, and the black line indicated the difference between the WSLS-2 model mean *BIC* and the merged mean *BIC* of WSLS-1 model and SS model (for original, not aligned-rank-transformed *BIC*).

A paired sample Wilcoxon signed-rank test was conducted on the *BIC* between the WSLS-2 and the logarithmic + ratio + trial-independent model. The results showed that the mean *BIC* of the WSLS-2 model was significantly higher than that of the logarithmic + ratio + trial-independent model, $V = 14$, $p < .001$.

Therefore, the best heuristics model performed better when using *BIC* as the evaluation metric. In addition, the heuristics models overall also outperformed the RL models in terms of the mean grade.

3.3 Fitted *MSD* Comparison

The distributions of fitted *MSD* statistics of different models were shown in Table 3.2.

Table 3.2 Fitted *MSD* distributions of different models

Model			Fitted *MSD*						
Utility Function	Choice function	Choice sensitivity	10% quantile	Median	90% quantile	Mean	Standard deviation	Mean grade	SW test Significance
power utility	ratio	trial-independent	58.35	238.34	916.04	382.79	405.94	10.54	< .001
		trial-dependent	52.23	207.37	1002.05	351.93	340.10	9.76	< .001

	softmax	trial-independent	43.39	155.25	784.70	307.62	323.65	8.23	< .001
		trial-dependent	37.80	164.98	577.07	250.88	245.39	6.54	< .001
	probit	trial-independent	48.59	177.43	853.44	312.97	327.41	8.69	< .001
		trial-dependent	46.95	165.67	679.26	279.96	288.82	7.38	< .001
	ratio	trial-independent	68.80	260.56	1047.25	402.91	404.41	10.98	< .001
		trial-dependent	57.76	219.66	1014.53	364.88	351.51	9.95	< .001
logarithmic utility	softmax	trial-independent	44.62	167.67	771.19	289.65	306.25	7.69	< .001
		trial-dependent	38.99	162.01	566.97	245.89	241.65	6.80	< .001
	probit	trial-independent	50.46	179.45	692.30	291.09	300.37	8.41	< .001
		trial-dependent	47.51	174.38	640.76	278.36	274.96	8.20	< .001
WSLS-1			31.76	108.79	456.97	202.47	266.25	5.43	< .001
WSLS-2			37.96	105.34	471.01	189.67	257.87	5.48	< .001
SS			38.27	112.91	513.82	224.74	272.35	5.95	< .001

With utility function (power, logarithmic), choice function (ratio, softmax, probit), and choice sensitivity (trial-independent, trial-dependent) as the independent variables, and fitted *MSD* as the dependent variable, an ART repeated measures analysis of variance was conducted. The results showed that the main effect of the choice function was significant, $F(2, 869) = 51.36$, $p < .001$. The main effect of the choice sensitivity was significant, $F(1, 869) = 7.23$, $p = .007$.

A post hoc comparison on the choice function showed there was no significance in the mean fitted *MSD* between the softmax and probit choice models, $t(869) = -1.76$, $p = .182$. The mean fitted *MSD* of the ratio choice models was significantly higher than that of the softmax choice models, $t(869) = -9.53$, $p < .001$, and that of the probit choice models, $t(869) = -7.76$, $p < .001$.

A post hoc comparison on the choice sensitivity showed that the mean fitted *MSD* of trial-independent choice sensitivity models was significantly higher than that of the trial-dependent choice sensitivity models, $t(869) = 2.69$, $p = .007$.

Figure 3.4 was a violin plot of the fitted *MSD* distributions of the RL models. Combining with the above tests and the mean grade, the best RL models were the power + softmax + trial-dependent model and the logarithmic + softmax + trial-dependent model.

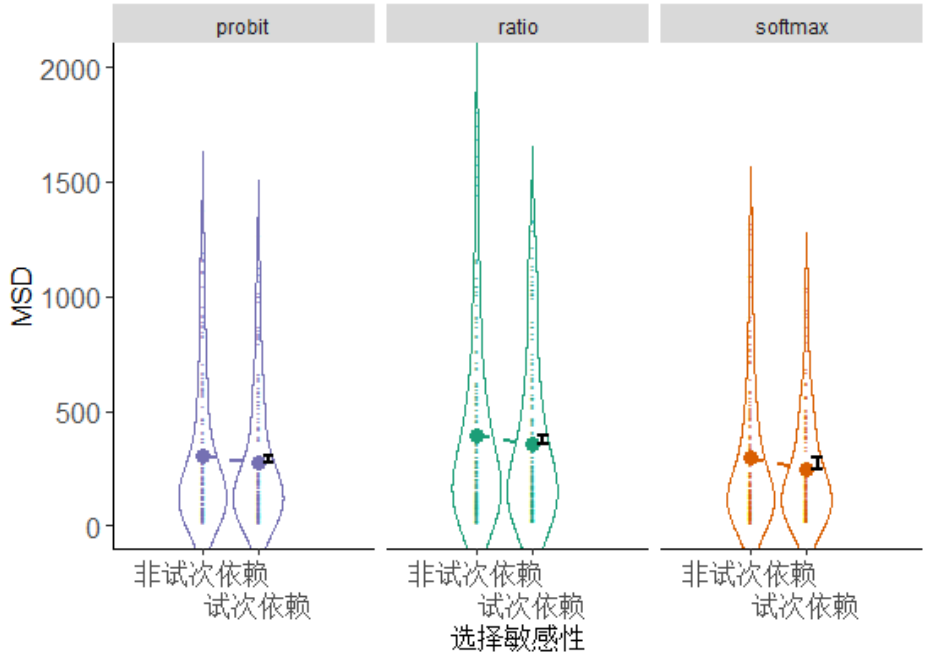

Figure 3.4 Fitted *MSD* distributions of RL models

Note: The effect of utility function was marginalized. The bold dots represented the mean fitted *MSD*, and the black line indicated the simple effect of choice sensitivity (for original, not aligned-rank-transformed *MSD*). 试次依赖=trial-dependent，非试次依赖=trial-independent.

With model type (WSLS-1, WSLS-2, SS) as the independent variable, and fitted MSD as the dependent variable, an ART repeated measures analysis of variance was conducted. The results showed that the effect of model type was not significant, $F(2, 158) = 1.47$, $p = .232$.

Figure 3.5 was a violin plot of the fitted *MSD* distributions of heuristics models. Combining with the above tests and mean grade, the best heuristics models were the WSLS-1 and WSLS-2 models.

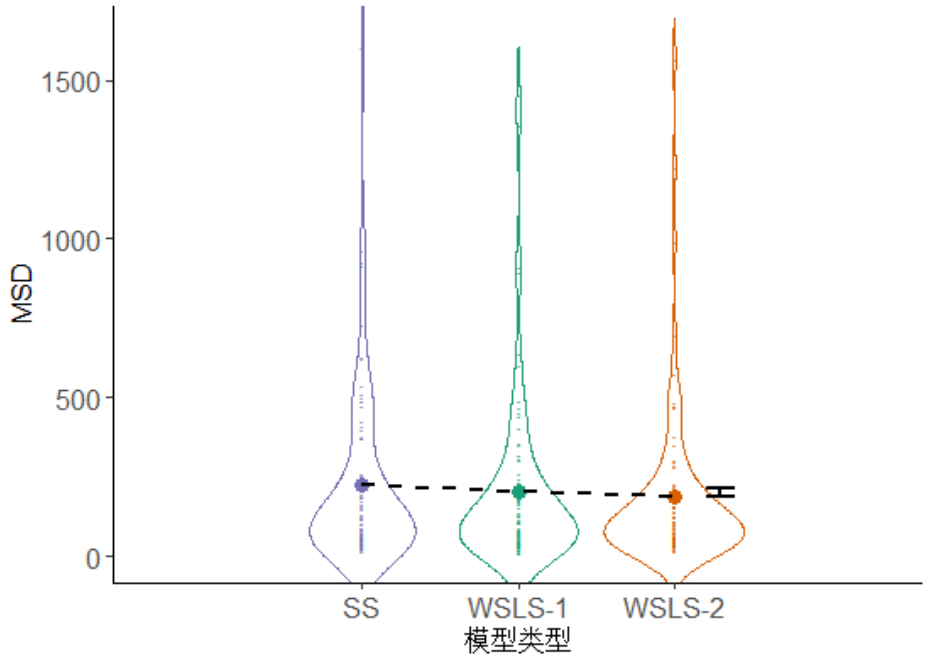

Figure 3.5 Fitted *MSD* distributions of different heuristics models

Note: The dots represented the fitted *MSD* mean, and the black line indicated the difference between the WSLS-2 model mean fitted *MSD* and the merged mean fitted *MSD* of WSLS-1 model and SS model (for original, not aligned-rank-transformed fitted *MSD*).

A paired sample Wilcoxon signed-rank test was conducted on the fitted *MSD* between the best RL and heuristics models. The results showed the mean fitted *MSD* of the WSLS-1 model was significantly lower than those of the two best RL models, V_1=2250, p_1 = .003, V_2=2231, p_2 = .003. The mean fitted *MSD* of the WSLS-2 model was significantly lower than those of the two best RL models, V_1=2144, p_1 = .012, V_2=2162, p_2 = .009.

Therefore, the best heuristics models performed better when using fitted *MSD* as the evaluation metric. In addition, the heuristics model overall also outperformed the RL models in terms of the mean grade, though the advantage was not as obvious as that when using *BIC* as the metric.

3.4 Discussion

Combining the results of the two metrics, the logarithmic utility function was better than the power utility function. The softmax choice function was slightly better than the probit choice function (mainly reflected by the mean grade), while the potential of the ratio choice

function remained to be tested. Whether the choice sensitivity was trial-dependent needed further investigation.

Interestingly, there was a significant interaction between the utility function and the choice function when using *BIC* as the metric. The logarithmic + ratio and the power + ratio models were the best and worst models respectively. In fact, among the 6 combinations of the utility and choice functions, the logarithmic + ratio was the only one that could strictly explain the magnitude effect, while the power + ratio was the only one that could never explain the magnitude effect. It supported the existence of magnitude effect in DFE from the perspective of mathematical modeling.

For the heuristics models, WSLS-2 was always the best model, while WSLS-1 was slightly better than the SS model, which supported the assumption of two conditional probabilities. From the parameter estimates (see Table 2.2), I also found that the sum of $p(\text{stay}|\text{win})_1$ and $p(\text{shift}|\text{loss})_1$ is not close to 1, which was consistent with previous research (Worthy & Maddox, 2014; Worthy et al., 2012). So there was a difference in the initial stay probabilities of the DMs.

As for the DMs' strategies, the heuristics models outperformed the RL models at both the best and overall levels. It suggested that DMs might prefer to rely on the relative rather than absolute magnitudes of outcomes, or that this simple strategy had a larger weight throughout the decision process.

4 Generalization Performance Comparison

A complex model can explain the training data better, either because it captures the patterns of the behavior better, or because it captures the noise. Besides using *BIC* to penalize the parameter numbers, another method is to compare the models on the unseen data. This is the generalization performance. The fitting performance focuses on the posterior explanatory ability, while the generalization performance focuses on the prior explanatory ability (Busemeyer & Wang, 2000). The generalization performance may be more important because the goal of most mathematical models is to accurately predict behavior in different situations, which requires the model to be stable across situations.

4.1 Evaluation metrics

This study used two statistics, G^2 and MSD, as the evaluation metrics of generalization performance. The metrics were calculated on the participant/problem level, with the last 90 trials of each problem in the mixed domain as the test set.

The calculation of G^2 was shown in Formulas 2-1 and 3-1, but two parts needed to be modified. First, the log-likelihood function was calculated using parameters estimated from the training set and data from the test set. Second, the baseline model was replaced with a random selection model, that was, a model assuming that participants were equally likely to choose either option. This model had no free parameters. Since the test set was unseen, there was need to penalize the parameter numbers. A larger G^2 indicated a better generalization performance.

The calculation of *MSD* was shown in Formula 3-3. A smaller *MSD* indicated a better generalization performance.

Since the ratio choice function was not suitable for the mixed domain problems, the related models were not be tested for generalization. So the rank of the worst model was 11.

4.2 G^2 Comparison

The distributions of G^2 statistics of different models were shown in Table 4.1.

Model			G^2							
Utility Function	Choice function	Choice sensitivity	10% quantile	Median	90% quantile	Mean	Standard deviation	Mean grade	SW test Significance	$G^2 > 0$ percentage
power utility	softmax	trial-independent	-1077.36	42.82	241.54	-225.15	968.89	6.80	< .001	59.59
		trial-dependent	-614.54	28.68	241.42	-207.93	1141.27	7.29	< .001	58.03
	probit	trial-independent	-810.54	32.82	223.36	-273.87	1273.00	7.33	< .001	57.51
		trial-dependent	-1070.44	5.50	232.04	-342.06	1689.51	7.99	< .001	51.81
Logarithmic utility	softmax	trial-independent	-240.82	45.02	241.94	-39.72	469.39	6.55	< .001	61.66
		trial-dependent	-501.50	25.80	233.00	-169.45	968.06	7.45	< .001	57.51

	probit	trial-independent	-384.24	47.90	217.04	-81.20	563.39	7.34	< .001	61.66
		trial-dependent	-806.72	12.04	222.84	-349.43	1692.49	8.17	< .001	53.37
WSLS-1			122.50	322.72	444.19	280.19	172.34	2.55	< .001	97.41
WSLS-2			142.20	324.08	441.78	267.06	445.21	1.99	< .001	96.89
SS			111.64	322.98	434.20	288.16	148.01	2.54	< .001	97.92

Table 4.1 G^2 distributions of different models

Note: If the percentage of $G^2 > 0$ was significantly greater than 50% ($\alpha = .05$, one-sided test), the corresponding cell was shaded.

With utility function (power, logarithmic), choice function (softmax, probit), and choice sensitivity (trial-independent, trial-dependent) as the independent variables, and G^2 as the dependent variable, an ART repeated measures analysis of variance was conducted. The results showed that the main effect of the choice function was significant, $F(1, 1344) = 12.22$, $p < .001$, and the main effect of the choice sensitivity was significant, $F(1, 1344) = 10.89$, $p = .001$.

A post hoc comparison on the choice function showed that the mean G^2 of the softmax choice models was significantly higher than that of the probit choice models, $t(1344) = 3.50$, $p = .001$.

A post hoc comparison on the choice sensitivity showed that the mean G^2 of the trial-independent choice sensitivity models was significantly higher than that of the trial-dependent choice sensitivity models, $t(1344) = 3.30$, $p = .001$.

Figure 4.1 was a violin plot of the G^2 distributions of the RL models. Combining the above tests and the mean grade, the best RL models were the power + softmax + trial-independent model and the logarithmic + softmax + trial-independent model.

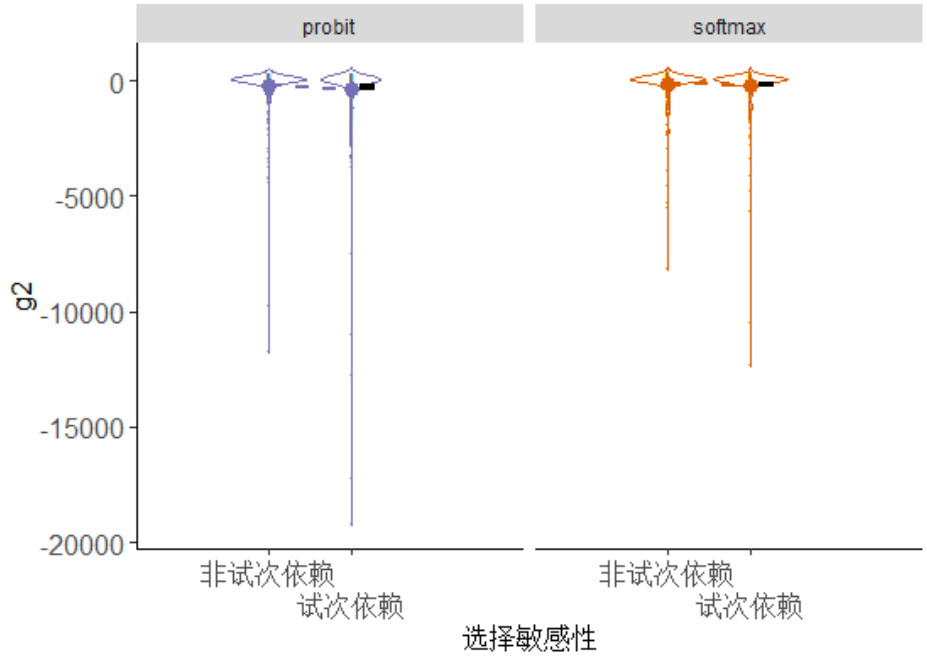

Figure 4.1 G^2 distributions of RL models

Note: The effect of utility function was marginalized. The bold dots represented the mean G^2, and the black line indicated the simple effect of choice sensitivity (for original, not aligned-rank-transformed G^2). 试次依赖=trial-dependent，非试次依赖=trial-independent.

With model type (WSLS-1, WSLS-2, SS) as the independent variable, and G^2 as the dependent variable, an ART repeated measures analysis of variance was conducted. The results showed that the effect of model type was significant, $F(2, 384) = 8.28$, $p < .001$. The post hoc comparison showed that there was no significant difference in mean G^2 between the WSLS-1 and SS models, $t(384) = -1.08$, $p = .527$. But the mean G^2 of the WSLS-2 model was significantly higher than that of the WSLS-1 model, $t(384) = 3.94$, $p < .001$, and that of the SS model, $t(384) = 2.86$, $p = .013$.

Figure 4.2 was a violin plot of G^2 distributions of the heuristics models. Combining with the above tests and the mean grade, WSLS-2 was the best heuristics model.

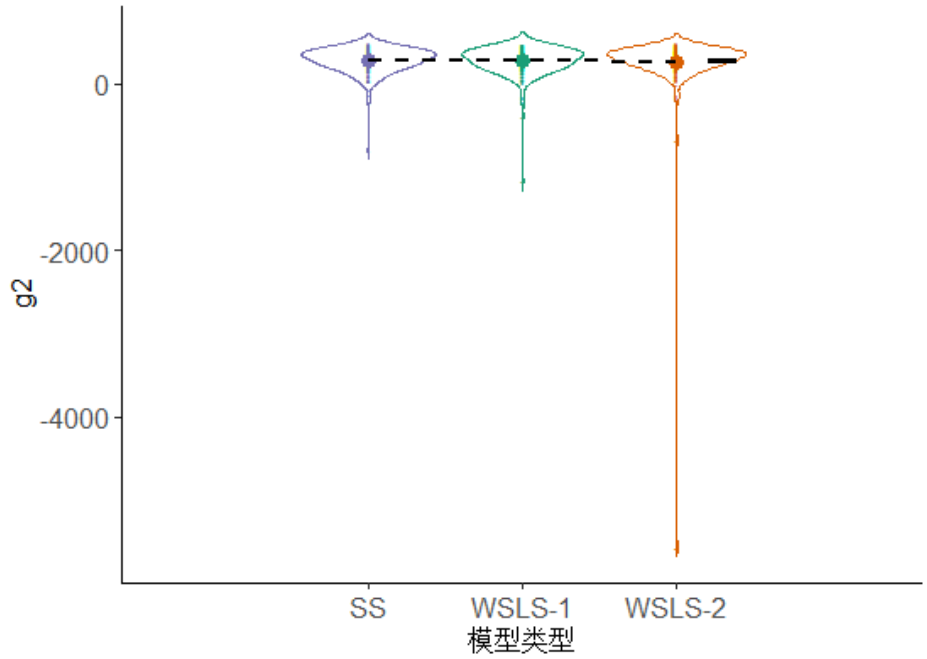

Figure 4.2 G^2 distributions of heuristics models

Note: The bold dots represented the mean G^2, and the black line indicated the difference between the WSLS-2 model mean G^2 and the merged mean G^2 of WSLS-1 model and SS model (for original, not aligned-rank-transformed G^2).

A paired sample Wilcoxon signed-rank test was conducted on the G^2 between the WSLS-2 model and the two best RL models. The results showed that the mean G^2 of the WSLS-2 model was significantly higher than those of the two best RL models, V_1=705, p_1 < .001, V_2 =743, p_2 < .001. Therefore, the best heuristics model performed better in generalization performance using G^2 as the evaluation metric. The heuristics model overall also outperformed the RL models in terms of the mean grade.

For all RL models and the WSLS-2 models, a fact was that they all contained few extremely low G^2, so the mean original G^2 was not a good evaluation of generalization performance. First, although the mean G^2 were smaller than 0 for all RL models, the median G^2 and the percentage of $G^2 > 0$ were still good supports. Second, although the mean G^2 of WSLS-2 was the lowest among the heuristics models, its median G^2 was higher. Since both the ART method and the mean grade could reduce the impact of outliers, the evaluation of the models was mainly based on the two methods.

4.3 Generalized *MSD C*omparison

The distributions of generalized *MSD* statistics of different models were shown in Table 4.2.

Table 4.2 Generalized *MSD* distributions of different models

Model			Generalized *MSD*						
Utility Function	Choice function	Choice sensitivity	10% quantile	Median	90% quantile	Mean	Standard deviation	Mean grade	SW test Significance
power utility	softmax	trial-independent	98.02	417.14	885.21	480.02	372.66	6.90	.001
		trial-dependent	90.38	433.40	968.06	536.32	379.79	7.85	.023
	probit	trial-independent	82.23	391.83	943.21	498.53	375.03	7.70	.007
		trial-dependent	88.91	430.68	941.91	526.33	389.87	7.68	.004
Logarithmic utility	softmax	trial-independent	77.40	319.98	886.05	433.95	346.33	5.58	.002
		trial-dependent	83.76	399.49	964.10	489.42	363.20	6.10	.008
	probit	trial-independent	101.63	366.53	956.58	463.78	354.25	6.85	.002
		trial-dependent	85.69	373.01	922.05	477.79	363.47	6.75	.003
WSLS-1			39.89	172.70	594.44	286.40	416.14	3.53	<.001
WSLS-2			31.58	103.50	640.14	255.82	441.32	3.00	<.001
SS			38.04	120.21	912.69	313.36	441.16	4.08	<.001

With utility function (power, logarithmic), choice function (softmax, probit), and choice sensitivity (trial-independent, trial-dependent) as the independent variables, and generalized *MSD* as the dependent variable, an ART repeated measures analysis of variance was conducted. The results showed that the main effect of the utility function was significant, $F(1, 273) = 13.63$, $p < .001$. The main effect of the choice sensitivity was significant, $F(1, 273) = 11.40$, $p = .001$.

A post hoc comparison on the utility function showed that the mean generalized *MSD* of power utility function models was significantly higher than that of the logarithmic utility function models, $t(273) = 3.69$, $p < .001$.

A post hoc comparison on the choice sensitivity showed that the mean generalized *MSD* of trial-dependent choice sensitivity models was significantly higher than that of the trial-independent choice sensitivity models, $t(273) = 3.38$, $p = .001$.

Figure 4.3 was a violin plot of the generalized *MSD* distributions of the RL models. Combining the above tests and the mean grade, the best RL model was the logarithmic + softmax + trial-independent model.

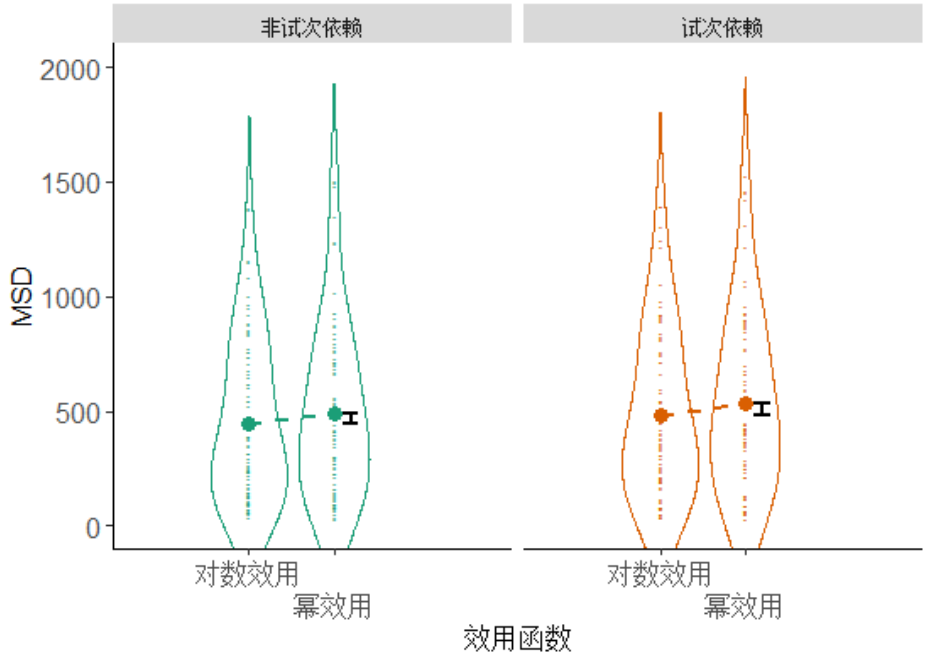

Figure 4.3 Generalized *MSD* distributions of RL models

Note: The effect of choice function was marginalized. The bold dots represented the mean generalized *MSD*, and the black line indicated the simple effect of utility function (for original, not aligned-rank-transformed generalized *MSD*). 对数效用=logarithmic utility, 幂效用=power utility

With model type (WSLS-1, WSLS-2, SS) as the independent variable, and generalized *MSD* as the dependent variable, an ART repeated measures analysis of variance was conducted. The results showed that the effect of model type was significant, $F(2, 78) = 7.61$, $p = .001$. The post hoc comparison showed that there was no significant difference in the mean generalized *MSD* between the WSLS-1 and SS models, $t(78) = -0.30$, $p = .951$. The mean generalized *MSD* of the WSLS-2 model was significantly smaller than that of the WSLS-1 model, $t(78) = -3.22$, $p = .005$, and that of the SS model, $t(78) = -3.52$, $p = .002$.

Figure 4.4 was a violin plot of the generalized *MSD* distributions of different heuristics models. Combining with the above tests and the mean grade, the best heuristics model was the WSLS-2 model.

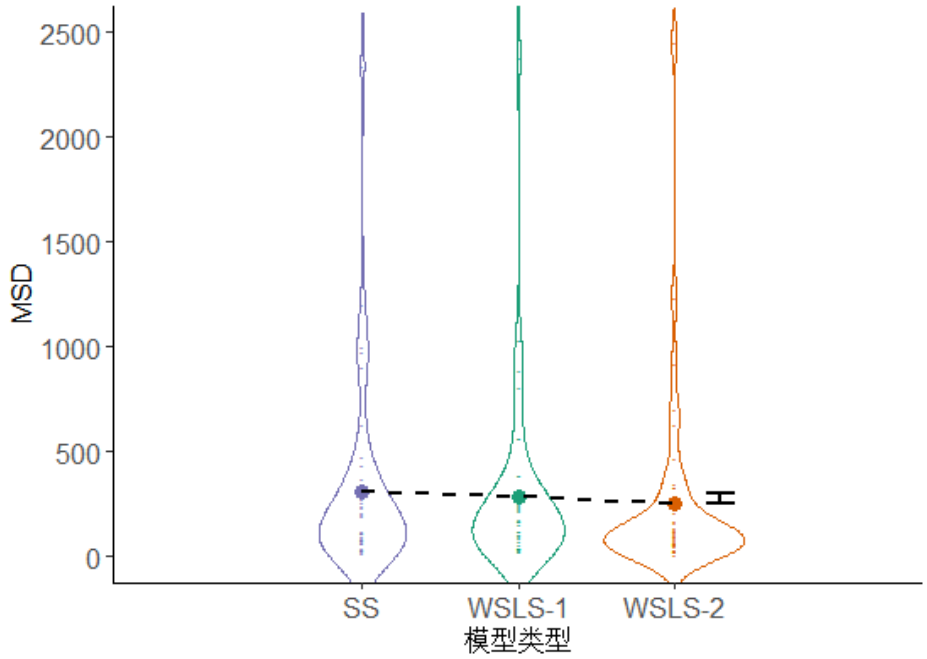

Figure 4.4 Generalized *MSD* distributions of heuristics models

Note: The bold dots represented the mean generalized *MSD*, and the black line indicated the difference between the WSLS-2 model mean generalized *MSD* and the merged mean generalized *MSD* of WSLS-1 model and SS model (for original, not aligned-rank-transformed generalized *MSD*).

A paired sample Wilcoxon signed-rank test was conducted on the generalized *MSD* between the WSLS-2 model and the logarithmic + softmax + trial-independent model. The results showed that the mean generalized *MSD* of the WSLS-2 model was significantly lower than that of the logarithmic + softmax + trial-independent model, V=679, $p < .001$. Therefore, the best heuristics outperformed the best RL model in terms of generalization performance when using generalized *MSD* as the metric. The heuristics models were overall better than the RL models in terms of the mean grade.

4.4 Discussion

Combining the results of the two metrics, the logarithmic utility function was better than the power utility function. The softmax choice function was slightly better than the probit choice function. The trial-independent choice sensitivity was better than the trial-dependent choice sensitivity.

For the heuristics model, WSLS-2 was always the best model. It supported the

assumptions of two conditional probabilities, and that a single trial could adjust both of them. In fact, the result that the WSLS-1 and the SS models did not beat each other also supported them. As mentioned in Section 1.4, the SS model was a restricted WSLS-2 model without the second assumption. The WSLS-1 model was a restricted WSLS-2 model without the first assumption. The result indicated that the two assumptions might be equally important.

As for the DMs' strategies, similar to the fitting performance comparison, the heuristics models showed an obvious advantage, suggesting a preference or larger weights of the heuristics.

5 Overall Discussion

5.1 Best RL Model

Based on the fitting and generalization performances, the logarithmic utility function showed an obvious advantage over the power utility function. Since this function was proposed to explain the magnitude effect in DFD (Scholten & Read, 2010), and has been used to fit the data relevant to this effect (Bouchouicha & Vieider, 2017; Scholten & Read, 2014), its advantage indicated the magnitude effect might also exist in DFE. In fact, although there is no study directly investigating this effect, Erev, Ert, and Yechiam (2008), and Konstantinidis, Taylor, and Newell (2018) did observe a trend of the magnitude effect in DFE experiments.

The softmax choice function was better than the probit choice function, while the potential of the ratio choice function remained to be explored. For the ratio choice function, although it seemed to be restricted, its potential has been observed in DFD. For example, Stott's (2006) comparison of PT models in DFD found that the ratio choice function achieved the best overall performance. In the study of Scholten and Read (2014) about the magnitude effect in DFD, they also found that the ratio choice function was better than the commonly used difference-based choice functions. In addition, the ratio choice function conformed to the Weber-Fechner law, indicating its psychophysical basis. The results of this study showed that it might be a good choice function in DFE too.

For the softmax choice function, it was the most widely used choice function in DFE (Ahn

et al., 2008; Lejarraga & Hertwig, 2017; Yechiam & Busemeyer, 2005). This study demonstrated its good performance again. Although it was intuitively an LCA choice function, it could also be derived from RUT. Yellott (1977) has proved that when the random components in RUT followed an independent and identically distributed Gumbel distribution, the DM's choice followed the softmax choice function. This conclusion could be extended to problems with more than two options. It showed that the softmax choice function had good mathematical properties. In fact, the advantage of it over the probit choice function might also come from its advantage in functional form. Although the standard logistic and normal distributions had similar shapes, the former had a higher kurtosis coefficient and a thicker tail. When the difference in expectation was large and the choice sensitivities were the same (i.e., the $\theta(t)$ in the softmax choice function and the $1/\theta(t)$ in the probit choice function), then the probability of choosing the option with smaller expectation was larger when predicted by the softmax choice function. Its insensitivity to extreme differences could better adapt to the phenomenon of sequential dependence in DFE (Erev & Barron, 2005). This phenomenon meant that even in the late stage of choices, DMs still alternated between different options. In other words, exploration still existed. Since the softmax choice function was insensitive to extreme differences, relevant models allowed this exploration. But the probit choice function could only adapt to this phenomenon by reducing the choice sensitivity (it could be found in parameter estimates), which pushed the predictions to the random level, resulting in poor performances. In addition, Daw et al. (2006) used functional magnetic resonance imaging to record the cortical activities of DMs in a DFE task and found that they were significantly correlated with the predicted probabilities by the softmax choice model, indicating the function's neurophysiological basis.

The trial-independent choice sensitivity was better when using *BIC*, G^2, and generalized *MSD* as the evaluation metrics, which was consistent with previous research (Ahn et al., 2008; Beitz, Salthouse, & Davis, 2014). However, it was still premature to make a conclusion. First, the trial-dependent choice sensitivity performed better at both the best and overall levels when using fitted *MSD* as the metric. Second, each DM completed multiple problems in TPT, causing difficulties in capturing the dynamic changes in choice sensitivity. For example, the DM's choice sensitivity might be mainly affected by the practice effect in the early presented problems, but by the fatigue effect in the later presented problems. When these problems were

jointly used to estimate parameters, the performance of trial-independent choice sensitivity might be better on the contrary. Third, the choice sensitivity might not change monotonically even in a single problem. For example, it might be affected by both practice and fatigue, showing an inverted U-shaped change, which could not be captured by the functional form considered in this study.

5.2 Best Heuristics Model

Based on the fitting and generalization performances, the WSLS-2 model was the best heuristics model. It indicated that its assumptions of two conditional probabilities, and that a single trial can adjust both were necessary. The generalization performance comparison even showed that they were equally important.

The DMs tended to stay after win trials and shift after loss trials. This was an intuitive inference and was also the pattern predicted by the RL models (Erev & Barron, 2005). This study showed that there were differences in the initial stay tendencies.

As for the latter assumption, although it seemed reasonable to only adjust the conditional probability related to the next trial after a single win or loss trial, it could meet some difficulties explaining the choices in the late stage. Due to the underweighting of small probabilities in DFE, people generally preferred the certain options or the risky options that were most likely to produce higher outcomes. In the late stage, the occasional loss trials did not have a strong impact on the stay probability of the next trial. It could not be captured by the WSLS-1 model. For the WSLS-2 model, both stay probabilities approached 1 in the late stage, which could better describe the participants' stable preferences.

5.3 Dominant Strategy

The heuristics models always outperformed the RL models, no matter at the best or overall levels. This advantage could be analyzed from two perspectives: single strategy and mixed strategy.

If people used a single strategy in DFE, it indicated that the dominant strategy was the WSLS strategy. Otto, Taylor, and Markman (2011) used probabilistic learning tasks and found

that under no cognitive loads, a WSLS model better explained the DM's behavior. But under cognitive loads, a RL model performed better. Their explanation was that WSLS was an explicit, rule-based strategy, while RL was an implicit, information-integrated strategy. Cognitive loads made the DMs rely more on the implicit learning process, so that the RL model could better capture their choices. As evidence, they found that participants' subjective preferences for the WSLS strategy were positively correlated with the performances of the WSLS model, but there was no significant correlation between the subjective preferences and the performances of the RL model. This was consistent with the assumption that explicit learning was conscious, but implicit learning was not. Similarly, Worthy et al. (2012) designed a DFE task in which the outcome distributions changed dynamically. The change was influenced by the DM's choices. The results still showed that the WSLS model performed better under no cognitive loads, while the RL model performed better under cognitive loads. It indicated that the differences between WSLS and RL models in different DFE tasks might be universal. From this perspective, the DMs in TPT completed the problems without cognitive loads, so the explicit learning system, i.e., WSLS, dominated.

However, Estes (1997, 2002) believed that an important feature of human cognition was the existence of multiple concurrent processes. That was, people might use mixed strategies when making decisions. In fact, some studies have used weighted models of RL and heuristics to capture behavioral patterns in DFE, and have achieved better results than single models (Ahn et al., 2014; Worthy & Maddox, 2014; Worthy Pang & Byrne, 2013). From this perspective, the advantage of the heuristics model was due to its larger weight in the mixed strategy. This larger weight could be caused by a variety of reasons. For example, in TPT, the DMs had to complete multiple problems. They might prefer to use simple strategies to complete the task faster but maintain the consistency of choices.

Although the heuristics model showed obvious advantages, it did not mean that the RL models were less important. First, if the DMs used a mixture of strategies, a single model only captured the characteristics of the corresponding strategy. When the task context changed (for example, the task changed to IGT), leading to a larger weight of the RL strategy, the investigation of the RL models could be useful for model selection. Second, compared to the heuristics models, the RL models were normative. It could adapt to different paradigms of DFE

tasks (Ahn et al., 2008;Sutton & Barto, 2018; Yechiam & Busemeyer, 2005). Moreover, each parameter in the RL models had theoretical meaning. For example, the parameter α in the utility function could represent the degree of risk aversion, and the parameter λ could represent the degree of loss aversion. In fact, in the field of computational psychiatry, some studies have compared the differences in estimated parameters of RL models of different populations, and attempted to use this difference as an auxiliary diagnostic metric (Lane, Yechiam, & Busemeyer, 2006; Fridberget al., 2010). Third, some assumptions of heuristics models were too strong. The WSLS models assumed that the DMs made choices based solely on the relative magnitude of outcomes. But imagine two DFE problems. The first problem was to choose between option (3, 1) and option (4, 0.8). The second problem was to choose between option (3, 1) and option (400, 0.8). If people only used the WSLS strategy, the choices should be similar at least at the group level. But this conclusion was obviously counter-intuitive. Second, the heuristics models assumed that when the result of the t^{th} trial is equal to that of the $(t - 1)^{th}$ trial, it was still a win trial. But imagine that a DM chose option (-4, 0.8) three times continuously, with feedback 0, -4, -4. Then the third trial was defined as a win trial, and the DM was believed to stay on the next trial. This was also counter-intuitive, since the DM knew that this option had a higher outcome of 0. Therefore, heuristics models might not be intuitive in more complex problems.

5.4 Limitations and Future Directions

On the technical aspect, this study conducted parameter estimation using the traditional maximum likelihood method at the individual level. However, Ahn, Krawitz, Kim, Busemeyer, and Brown (2013) compared the parameter recovery capabilities of different estimation methods on simulated data and found that the hierarchical Bayesian method considering the group characteristics performed better. They also pointed out that the parameter estimates by traditional methods were usually close to the boundary of the specified range and deviated greatly from the true values. This phenomenon was also found in this experiment. In Table 2.1 and Table 2.2, for some parameters whose values were restricted to be larger than 0, the standard deviation was even greater than the mean. This reflected that most estimates were close to the lower bound. This extreme distribution of parameter estimates indicated the

insufficient information at the individual level. Incorporating the information at the group level could reduce the amount of information required, and help obtain more stable and accurate estimates. In addition, the hierarchical Bayesian method provided a natural way to compare models. It was conducted by creating a discrete uniform prior over models. The posterior probabilities could be directly used to measure model performances. At the same time, this method could automatically penalize complex models since the probabilities were distributed to a broader space (Kruschke, 2014). Considering the above advantages, future research could try hierarchical Bayesian estimation to reproduce the results of this study.

For the components of the RL models, the ratio choice function was worthy of further exploration. Although this study had extended it into the loss domain problems, it was still unsuitable for the mixed domain problems. And this was what future research could do. Another function to be explored was the trial-dependent choice sensitivity. Intuitively, choice sensitivity should change dynamically during learning. But the monotonic functional form might be too restricted, and might produce extreme sensitivity when the trial number was too small or large . Future research could try more flexible functional forms.

Regarding the dataset, although TPT was currently the largest public DFE dataset, the range of outcomes in the problems was small (i.e., from -30 to 30), and the problem structure was simple. In fact, Glöckner, Hilbig, Henninger and Fiedler (2016) found a reversed DEG in problems where both options contained multiple outcomes, indicating that problem structures might affect the psychological processes of DFE. As discussed in Section 5.3, the heuristics might not be advantageous when the options had very different expectations. Future research could try more diverse problem structures to investigate how the models could adapt to complex environments. At the same time, as mentioned in Section 5.1, completing many problems might lead to changes in the DM's strategy and made it difficult to capture the change of choice sensitivity. Therefore, future research should still control the number of problems.

Finally, this study focused on single strategies. But the choices might be the results of multiple concurrent processes. Future research could try to construct weighted models of RL and heuristics strategies, and compare them with single strategies to test whether the weighted models can better capture the behavioral patterns in DFE.

6 Conclusions

This study compared 12 RL models and 3 heuristics models on binary choice DFE data, and arrived at the following conclusions:

(1) For the RL models, the logarithmic utility function was better than the power utility function. The softmax choice function was better than the probit choice function, and the potential of the ratio choice function remained to be explored. Although the trial-independent choice sensitivity was better than the trial-dependent choice sensitivity based on most metrics, the trial-dependence of choice sensitivity remained to be explored.

(2) For the heuristics model, the WSLS-2 model was always better than the WSLS-1 and SS models, which supported the assumptions of two conditional probabilities, and that a single trial would adjust both of them.

(3) The heuristics model showed obvious advantages at both the overall and best levels, but the relationship between the heuristics and RL models needed to be further explored.

7 Appendix

Since the magnitude effect is more stable in the gain domain, the following proof is limited to the gain domain problems (the following proof is for DED, where the expectation of an option is the expected utility with the probabilities transformed by a monotonic function). The low-magnitude problem is a choice between option A(m, p) and option B(n, q), and the high-magnitude problem is a choice between option C(am, p) and option D(an, q). where $0 < m < n$, $0 < q < p \leq 1$, $a > 1$ and $pm = qn$. The magnitude effect refers to the probability that a DM chooses C in the second problem being higher than that of choosing A in the first problem. The difference increases with a larger a.

First, I will use the elasticity of the utility function to prove that when using the ratio choice function, the power utility function does not support the magnitude effect, but the logarithmic utility function does.

The elasticity of the power utility function and its derivative are

$$\varepsilon_u(x) = \frac{xu'(x)}{u(x)} = \frac{\alpha x^{\alpha}}{x^{\alpha}} = \alpha$$

$$\varepsilon'_u(x) = 0$$

(5-1)

The elasticity of the logarithmic utility function and its derivative are

$$\varepsilon_u(x) = \frac{xu'(x)}{u(x)} = \frac{\frac{\alpha x}{1+\alpha x}}{\ln(1+\alpha x)} = \frac{\alpha x}{(1+\alpha x)\ln(1+\alpha x)}$$

$$\varepsilon'_u(x) = \frac{\alpha(1+\alpha x)\ln(1+\alpha x) - \alpha x\left(\alpha \ln(1+\alpha x) + \frac{\alpha(1+\alpha x)}{1+\alpha x}\right)}{[(1+\alpha x)\ln(1+\alpha x)]^2}$$

$$= \frac{\alpha[\ln(1+\alpha x) - \alpha x]}{[(1+\alpha x)\ln(1+\alpha x)]^2} < 0$$

(5-2)

Therefore, the elasticity of the power utility function is constant, and that of the logarithmic utility function decreases as the outcome increases.

If treating the ratio of the expectation between the safe and risky options in the high-magnitude problem as a function of *a*, this function and its derivative are:

$$f(a) = \frac{\pi(p)u(am)}{\pi(q)u(an)}$$

$$f'(a) = \frac{\pi(p)}{\pi(q)}\left(\frac{mu'(am)u(an) - nu'(an)u(am)}{u^2(an)}\right)$$

$$= \frac{\pi(p)u(am)u(an)}{a\pi(q)u^2(an)}(\varepsilon_u(am) - \varepsilon_u(an))$$

(5-3)

For the power utility function, because the elasticity is constant, the ratio does not change as *a* changes. Therefore, when using the ratio choice function, the DM's probability of choosing the safe option is the same in the two problems. For the logarithmic utility function, because the elasticity decreases, the ratio increases as *a* increases. Therefore, when using a ratio choice function, the DM has a higher probability of choosing the safe option in the second problem than in the first problem.. And the difference increases with a larger *a*.

Second, when using a difference-based choice function like the softmax or probit choice

function, the power utility function somewhat supports the magnitude effect, but does not allow preference reversal. In other words, it does not allow a preference for the risky option in the first problem, but a preference for the safe option in the second problem.

If treating the difference of the expectation between the safe and risky options in the high-magnitude problem as a function of a, with the power utility function, this function and its derivative are:

$$f(a) = \pi(p)u(am) - \pi(q)u(an) = \pi(p)(am)^{\alpha} - \pi(q)(an)^{\alpha}$$
$$= a^{\alpha}\big(\pi(p)u(m) - \pi(q)u(n)\big)$$
$$f'(a) = a^{\alpha-1}\big(\pi(p)u(m) - \pi(q)u(n)\big) \tag{5-4}$$

The difference of expectation increases with a larger a. To explain the magnitude effect, it is required that the DM always prefers the safe option.

Finally, the logarithmic utility function always supports the magnitude effect and allows the occurrence of preference reversal..

The support of the magnitude effect has been proved for the ratio choice function. Here I will prove it also holds for the difference-based choice functions.

When the utility function is a logarithmic utility function, the function defined above and its derivative are

$$f(a) = \pi(p)u(am) - \pi(q)u(an) = \frac{\pi(p)\ln(1+\alpha am) - \pi(q)\ln(1+\alpha an)}{\alpha}$$
$$f'(a) = \frac{m\pi(p)}{1+\alpha am} - \frac{n\pi(q)}{1+\alpha an} = \frac{m\pi(p) - n\pi(q) + \alpha amn\pi(p) - \alpha amn\pi(q)}{(1+\alpha am)(1+\alpha an)} \tag{5-5}$$

Since objective probabilities are mapped to decision weights, $f'(a)$ is not always positive. But when a is relatively large, the sign of $f'(a)$ is mainly determined by $\alpha amn\pi(p)$ - $\alpha amn\pi(q)$, which is strictly greater than 0. In the magnitude effect studies, a can range from tens to tens of thousands (Scholten, & Read, 2014; Weber & Chapman, 2005). Under this condition, $f(a)$ is an increasing function of a. When a increases, the difference decreases even reverses. Therefore, the logarithmic utility function can support the magnitude effect in this context.

If the logarithmic utility function supports the phenomenon of preference reversal, it

should require that the risky option has a larger expectation in the first problem, but a lower expectation in the second problem.Since the ratio of expectations is an increasing function of a ($a > 1$ is not required), we only need to prove that when $a \to 0$, the expectation of the risky option is higher, and when $a \to \infty$, the expectation of the risky option is lower.

For the former, the ratio between expectations is:

$$\lim_{a\to 0}\frac{\pi(p)u(am)}{\pi(q)u(an)} = \lim_{a\to 0}\frac{\pi(p)\ln(1+\alpha am)}{\pi(q)\ln(1+\alpha an)} = \lim_{a\to 0}\frac{\pi(p)\dfrac{\alpha m}{1+\alpha am}}{\pi(q)\dfrac{\alpha n}{1+\alpha an}} = \frac{m\pi(p)}{n\pi(q)} \tag{5-6}$$

Since $pm = qn$, and the weighting function is subadditive, $\pi(p)/\pi(q) < n/m$, the risky option has a higher expectation.

For the latter, the ratio between expectations is

$$\lim_{a\to \infty}\frac{\pi(p)u(am)}{\pi(q)u(an)} = \lim_{a\to \infty}\frac{\pi(p)\ln(1+\alpha am)}{\pi(q)\ln(1+\alpha an)} = \lim_{a\to \infty}\frac{\pi(p)\dfrac{\alpha m}{1+\alpha am}}{\pi(q)\dfrac{\alpha n}{1+\alpha an}} = \frac{\pi(p)}{\pi(q)} \tag{5-7}$$

Since the weighting function is an increasing function of probability, the risky option has a lower expectation..

In summary, neither utility function can strictly explain the magnitude effect when using the difference-based choice functions. However, considering that the DMs tend to prefer the risky option when the magnitude is very low, the logarithmic utility function seems more appropriate In addition, among the combinations of the utility and choice functions, the only one that can strictly explain the magnitude effect is the logarithmic + ratio, while the only one that cannot explain the magnitude effect at all is the power + ratio.

References

Ahn, W., Busemeyer, J., Wagenmakers, E., & Stout, J. (2008). Comparison of decision learning models using the generalization criterion method. *Cognitive Science: A Multidisciplinary Journal*, *32*(8), 1376-1402. doi:10.1080/03640210802352992

Ahn, W., Krawitz, A., Kim, W., Busemeyer, J. R., & Brown, J. W. (2013). A model-based fMRI analysis with hierarchical Bayesian parameter estimation. *Decision*, *1*(S), 8-23. doi:10.1037/2325-9965.1.s.8

Ahn, W., Vasilev, G., Lee, S., Busemeyer, J. R., Kruschke, J. K., Bechara, A., & Vassileva, J. (2014). Decision-making in stimulant and opiate addicts in protracted abstinence: Evidence from computational modeling with pure users. *Frontiers in Psychology*, *5*, 849-863. doi:10.3389/fpsyg.2014.00849

Audibert, J., Munos, R., & Szepesvári, C. (2009). Exploration–exploitation tradeoff using variance estimates in multi-armed bandits. *Theoretical Computer Science*, *410*(19), 1876-1902. doi:10.1016/j.tcs.2009.01.016

Barron, G., & Erev, I. (2003). Small feedback-based decisions and their limited correspondence to description-based decisions. *Journal of Behavioral Decision Making*, *16*(3), 215-233. doi:10.1002/bdm.443

Beitz, K. M., Salthouse, T. A., & Davis, H. P. (2014). Performance on the Iowa gambling task: From 5 to 89 years of age. *Journal of Experimental Psychology: General*, *143*(4), 1677-1689. doi:10.1037/a0035823

Bouchouicha, R., & Vieider, F. M. (2017). Accommodating stake effects under prospect theory. *Journal of Risk and Uncertainty*, *55*(1), 1-28. doi:10.1007/s11166-017-9266-y

Busemeyer, J. R., & Wang, Y. (2000). Model comparisons and model selections based on generalization criterion methodology. *Journal of Mathematical Psychology*, *44*(1), 171-189. doi:10.1006/jmps.1999.1282

Camilleri, A. R., & Newell, B. R. (2011). When and why rare events are underweighted: A direct comparison of the sampling, partial feedback, full feedback and description

choice paradigms. *Psychonomic Bulletin & Review*, *18*(2), 377-384. doi:10.3758/s13423-010-0040-2

Collins, A. G., & Frank, M. J. (2014). Opponent actor learning (Opal): Modeling interactive effects of striatal dopamine on reinforcement learning and choice incentive. *Psychological Review*, *121*(3), 337-366. doi:10.1037/a0037015

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. doi:10.1038/nature04766

Erev, I., & Barron, G. (2005). On adaptation, maximization, and reinforcement learning among cognitive strategies. *Psychological Review*, *112*(4), 912-931. doi:10.1037/0033-295x.112.4.912

Erev, I., Ert, E., & Yechiam, E. (2008). Loss aversion, diminishing sensitivity, and the effect of experience on repeated decisions. *Journal of Behavioral Decision Making*, *21*(5), 575-597. doi:10.1002/bdm.602

Erev, I., Ert, E., Roth, A. E., Haruvy, E., Herzog, S. M., Hau, R., Lebiere, C. (2010). A choice prediction competition: Choices from experience and from description. *Journal of Behavioral Decision Making*, *23*(1), 15-47. doi:10.1002/bdm.683

Estes, W. K. (1950). Toward a statistical theory of learning. *Psychological Review*, *57*(2), 94-107. doi:10.1037/h0058559

Estes, W. K. (1997). Processes of memory loss, recovery, and distortion. *Psychological Review*, *104*(1), 148-169. doi:10.1037/0033-295x.104.1.148

Estes, W. K. (2002). Traps in the route to models of memory and decision. *Psychonomic Bulletin & Review*, *9*(1), 3-25. doi:10.3758/bf03196254

Fridberg, D. J., Queller, S., Ahn, W., Kim, W., Bishara, A. J., Busemeyer, J. R., Stout, J. C. (2010). Cognitive mechanisms underlying risky decision-making in chronic cannabis users. *Journal of Mathematical Psychology*, *54*(1), 28-38. doi:10.1016/j.jmp.2009.10.002

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

Glöckner, A., Hilbig, B. E., Henninger, F., & Fiedler, S. (2016). The reversed description-experience gap: Disentangling sources of presentation format effects in risky choice. *Journal of Experimental Psychology: General*, *145*(4), 486-508. doi:10.1037/a0040103

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. doi:10.1111/j.0956-7976.2004.00715.x

Kahneman, D., & Tversky, A. (1979). Prospect theory: An analysis of decision making under risk. *Econometrica*, *47*(2), 263-291. doi:10.21236/ada045771

Kahneman, D., & Tversky, A. (1982). The psychology of preferences. *Scientific American*, *246*(1), 160-173. doi:10.1038/scientificamerican0182-160

Konstantinidis, E., Taylor, R. T., & Newell, B. R. (2018). Magnitude and incentives: Revisiting the overweighting of extreme events in risky decisions from experience. *Psychonomic Bulletin & Review*, *25*(5), 1925-1933. doi:10.3758/s13423-017-1383-8

Kruschke, J. (2014). *Doing Bayesian data analysis: A tutorial with R, JAGS, and Stan* (2nd ed.). Boston, MA: Academic Press.

Lane, S. D., Yechiam, E., & Busemeyer, J. R. (2006). Application of a computational decision model to examine acute drug effects on human risk taking. *Experimental and Clinical Psychopharmacology*, *14*(2), 254-264. doi:10.1037/1064-1297.14.2.254

Lejarraga, T., & Hertwig, R. (2017). How the threat of losses makes people explore more than the promise of gains. *Psychonomic Bulletin & Review*, *24*(3), 708-720. doi:10.3758/s13423-016-1158-7

Luce, R. D. (1959). *Individual choice behavior*. New York, NY: Wiley.

Luce, R. D., & Suppes, P. P. (1965). Preference, utility, and subjective probability. In R. D. Luce, R. R. Bush, & E. Galanter (Eds.), *Handbook of mathematical psychology* (pp. 252-410). New York, NY: Wiley.

Markowitz, H. (1952). The utility of wealth. *Journal of Political Economy*, *60*(2), 151-158. doi:10.1086/257177

Nelder, J. A., & Mead, R. (1965). A simplex method for function minimization. *The Computer Journal, 7*(4), 308-313. https://doi.org/10.1093/comjnl/7.4.308

Otto, A. R., Taylor, E. G., & Markman, A. B. (2011). There are at least two kinds of probability matching: Evidence from a secondary task. *Cognition*, *118*(2), 274-279. doi:10.1016/j.cognition.2010.11.009

Prelec, D., & Loewenstein, G. (1991). Decision making over time and under uncertainty: A common approach. *Management Science*, *37*(7), 770-786. doi:10.1287/mnsc.37.7.770

Rescorla, R. A., & Wagner, A. R. (1972). A theory of Pavlovian conditioning: Variations in the effectiveness of reinforcement and nonreinforcement. In A. H. Black & W. F. Prokasy (Eds.), *Classical conditioning II: Current research and theory* (pp. 64-99). New York, NY: Appleton-Century-Crofts.

Scholten, M., & Read, D. (2010). The psychology of intertemporal tradeoffs. *Psychological Review*, *117*(3), 925-944. doi:10.1037/a0019619

Scholten, M., & Read, D. (2014). Prospect theory and the "forgotten" fourfold pattern of risk preferences. *Journal of Risk and Uncertainty*, *48*(1), 67-83. doi:10.1007/s11166-014-9183-2

Schulz, E., & Gershman, S. J. (2019). The algorithmic architecture of exploration in the human brain. *Current Opinion in Neurobiology*, *55*, 7-14. doi:10.1016/j.conb.2018.11.003

Schwarz, G. (1978). Estimating the dimension of a model. *The Annals of Statistics, 6*(2), 461-464. doi:10.1214/aos/1176344136

Shapiro, S. S., & Wilk, M. B. (1965). An analysis of variance test for normality (complete samples). *Biometrika, 52*(3-4), 591-611. doi:10.1093/biomet/52.3-4.591

Stott, H. P. (2006). Cumulative prospect theory's functional menagerie. *Journal of Risk and Uncertainty*, *32*(2), 101-130. doi:10.1007/s11166-006-8289-6

Sutton, R. S., & Barto, A. G. (2018). *Reinforcement learning: An introduction* (2nd ed.). Cambridge, MA: MIT Press.

Thurstone, L. L. (1927). A law of comparative judgment. *Psychological Review*, *34*(4), 273-286. doi:10.1037/h0070288

Tversky, A., & Kahneman, D. (1981). The framing of decisions and the psychology of choice. *Science*, *211*(4481), 453-458. doi:10.1126/science.7455683

Weber, B. J., & Chapman, G. B. (2005). Playing for peanuts: Why is risk seeking more common for low-stakes gambles? *Organizational Behavior and Human Decision Processes*, *97*(1), 31-46. doi:10.1016/j.obhdp.2005.03.001

Wobbrock, J. O., Findlater, L., Gergle, D., & Higgins, J. J. (2011). The aligned rank transform for nonparametric factorial analyses using only anova procedures. *Proceedings of the Annual Conference on Human Factors in Computing Systems 11*, 143-146. doi:10.1145/1978942.1978963

Worthy, D. A., Hawthorne, M. J., & Otto, A. R. (2013). Heterogeneity of strategy use in the Iowa gambling task: A comparison of win-stay/lose-shift and reinforcement learning models. *Psychonomic Bulletin & Review*, *20*(2), 364-371. doi:10.3758/s13423-012-0324-9

Worthy, D. A., Maddox, W. T., & Markman, A. B. (2008). Ratio and difference comparisons of expected reward in decision-making tasks. *Memory & Cognition*, *36*(8), 1460-1469. doi:10.3758/mc.36.8.1460

Worthy, D. A., & Maddox, W. T. (2014). A comparison model of reinforcement-learning and win-stay-lose-shift decision-making processes: A tribute to W.K. Estes. *Journal of Mathematical Psychology*, *59*, 41-49. doi:10.1016/j.jmp.2013.10.001

Worthy, D. A., Otto, A. R., & Maddox, W. T. (2012). Working-memory load and temporal myopia in dynamic decision making. *Journal of Experimental Psychology: Learning, Memory, and Cognition*, *38*(6), 1640-1658. doi:10.1037/a0028146

Worthy, D. A., Pang, B., & Byrne, K. A. (2013). Decomposing the roles of perseveration and expected value representation in models of the Iowa gambling task. *Frontiers in Psychology*, *4*, 640-648. doi:10.3389/fpsyg.2013.00640

Wulff, D. U., Mergenthaler-Canseco, M., & Hertwig, R. (2018). A meta-analytic review of two modes of learning and the description-experience gap. *Psychological Bulletin*, *144*(2), 140-176. doi:10.1037/bul0000115

Yechiam, E., & Busemeyer, J. R. (2005). Comparison of basic assumptions embedded in learning models for experience-based decision making. *Psychonomic Bulletin & Review*, *12*(3), 387-402. doi:10.3758/bf03193783

Yellott, J. I. (1977). The relationship between Luce's choice axiom, Thurstone's theory of comparative judgment, and the double exponential distribution. *Journal of Mathematical Psychology*, *15*(2), 109-144. doi:10.1016/0022-2496(77)90026-8

Yoon, S., Vo, K., & Venkatraman, V. (2017). Variability in decision strategies across description-based and experience-based decision making. *Journal of Behavioral Decision Making*, *30*(4), 951-963. doi:10.1002/bdm.2009