Mental Model Management:

The effects of punishment and reinforcement re-using and re-configuring strategies within game spaces

By

Yajing Zhang

A thesis submitted in partial fulfillment of the requirements for the degree of

Master of Science

Department of Psychology

University of Alberta

Abstract

The development, storage and deployment of mental models are keystone cognitive processes central to successful operation in everyday life. We investigate the effects of punishment and reinforcement on people's ability to acquire, reuse, and reconfigure mental models. Across three experiments, 218 participants completed a competitive, binary-outcome dice game against a computerized opponent, where the goal was to defeat the opponent who played two different exploitable strategies. All participants played four blocks of game, and each block consists of *Pre*, *During*, and *Post* phases. In the *Pre* phase, a strategy was acquired. In the *During* phase, a fixed win rate manipulation was used to create conditions where the strategy learned in the *Pre* phase was either punished or reinforced. In the *Post* phase, to maximize wins participants had to either relearn the old strategy (where *Pre* and *Pos*t strategies were the same) or learn a new strategy (where *Pre* and *Post* strategies were different). The three experiments varied in their *During* phase design. Participants experienced a mild punishment (50%) for *Pre* strategy in Experiment 1, a severe punishment (25%) and a light reinforcement (75%) in Experiment 2, and an equal degree of punishment (44%) and reinforcement (88%) relative to the baseline *Pre* performance of 66% win rate in Experiment 3. Participants' proportions of win trials and optimal behaviours were analyzed in response to the reinforcement or punishment of old and new mental model strategies. Data revealed: (1) It is easier to relearn an old strategy relative to learning a new strategy; (2) The benefit of relearning old knowledge is weakened following a strong but not mild punishment; (3) Reinforcement of a learned strategy strengthens relearning old information but sabotages learning new information; and (4) Behaviours following a win are generally a stronger predictor in future performance than behaviours following a loss. These results indicate

that people have a tendency to stay at their previous strategies unless they are punished harshly. Additionally, wins produce more reliable and flexible behavior relative to losses, indicating the focus of future research should be on how individuals recover from loss rather than maintain success.

Keywords: Decision-making, Simple games, Mental model acquisition, Mental model update, Mental model reacquisition, Reinforcement, Punishment

Preface

This thesis is an original work by Yajing Zhang. No part of this thesis has been previously published.

Acknowledgments

The last two years have been sometimes tiring but always uplifting. I would first like to thank my supervisor, Dr. Ben Dyson, for his excellent support, from my first decision to defer my entry to the completion of this thesis, encouragement, especially in emphasizing the importance of study-life balance, and mentorship, prompt meetings helped a lot in solving confusions.

I would also like to thank my lab mate, Eunchan Na. Being the only other graduate in the lab, Eunchan guided me in studying and working and discussed interesting ideas with me, and was also the first person here with whom I could share my hobbies.

I am grateful to my partner, my family, and my friends, who always supported and inspired me. They were always by my side whenever I was depressed. They made me confirm that my journey in academia, although it may be tough and frustrating, would be full of hope and joviality. Additionally, there are special thanks to cats Kikino, Qiuqiu, and Piaoliang, the adorable creatures who have warmed me from the first day I met them.

Table of Contents

Chapter 1 ..... **1**
Introduction **1**
1.1 Core findings 2
1.2 Thesis organization 2
Chapter 2 ..... **3**
Background **3**
2.1 Mental models 3
2.2 Learning 3
2.3 Learning via experience 5
2.4 Mental model acquisition: building, updating, and retrieving 7
2.5 Relationship between new and old mental models 8
2.6 Operant conditioning and mental model acquisition 10
2.7 Aims 11
Chapter 3 ..... **12**
Experiment 1 ..... **12**
3.1 Introduction 12
3.1.1 Hypotheses 13
3.1.1.1 Predictions 13
3.2 Method 14
3.2.1 Participants 14
3.2.2 Stimuli 14
3.2.3 Design 15
3.2.4 Procedure 16
3.3 Results 17
3.3.1 Win rate distribution 17
3.3.2 Optimal behaviour rate (OBR) 19
3.3.3 Strategy suppression 22
3.4 Discussion 23
Chapter 4 ..... **25**
Experiment 2 ..... **25**
4.1 Introduction 25
4.1.1 Hypotheses 25

4.1.1.1 Predictions 25
4.2 Method 26
4.2.1 Participants 26
4.2.2 Stimuli 27
4.2.3 Design 27
4.2.4 Procedure 28
4.3 Results 28
4.3.1 Fixed win rate at 25% in During bins 29
4.3.1.1 Win rate distribution 29
4.3.1.2 Optimal behaviour rate (OBR). 30
4.3.1.3 Strategy suppression 32
4.3.2 Fixed win rate at 75% in During bins 33
4.3.2.1 Win rate distribution 33
4.3.2.2 Optimal behaviour rate (OBR) 35
4.3.2.3 Strategy enhancement 37
4.4 Discussion 39
Chapter 5 41
Experiment 3 41
5.1 Introduction 41
5.1.1 Hypotheses 41
5.1.1.1 Predictions 42
5.2 Method 43
5.2.1 Participants 43
5.2.2 Stimuli 43
5.2.3 Design 43
5.2.4 Procedure 44
5.3 Results 44
5.3.1 Fixed win rate at 44% in During bins 44
5.3.1.1 Win rate distribution 44
5.3.1.2 Optimal behaviour rate (OBR) 46
5.3.1.3 Strategy suppression 48
5.3.2 Fixed win rate at 88% in During bins 49
5.3.2.1 Win rate distribution 49
5.3.2.2 Optimal behaviour rate (OBR) 51
5.3.2.3 Strategy enhancement 53

5.4 Discussion 54
Chapter 6 55
General discussion 55
6.1 Summary of experiments. 55
6.2 Mental model acquisition and reacquisition following punishment and reinforcement 60
6.3 Behaviours following wins and losses 62
6.4 Contributions 63
6.5 Limitations and ideas for future studies 64
References 67
Appendix A: Cross analyses 75
A.1 Cross various fixed win rates analyses 75
A.2 Cross experiments analyses 78

List of Tables

Table 3.1 16
Table 3.2 18
Table 3.3 18
Table 3.4 19
Table 3.5 21
Table 4.1 28
Table 4.2 29
Table 4.3 30
Table 4.4 30
Table 4.5 31
Table 4.6 34
Table 4.7 35
Table 4.8 35
Table 4.9 37
Table 5.1 44
Table 5.2 45
Table 5.3 46
Table 5.4 46
Table 5.5 47
Table 5.6 50
Table 5.7 51
Table 5.8 51
Table 5.9 52
Table 6.1 58
Table 6.2 59
Table A.1 76
Table A.2 76
Table A.3 78
Table A.4 78
Table A.5 79

List of Figures

Figure 3.1 17
Figure 3.2 20
Figure 3.3 22
Figure 3.4 23
Figure 4.1 32
Figure 4.2 33
Figure 4.3 37
Figure 4.4 38
Figure 5.1 48
Figure 5.2 49
Figure 5.3 53
Figure 5.4 54

Chapter 1

Introduction

Decision-making is a basic daily task for everyone. By gathering and interpreting information from past events and current environments, people build mental models that are supposed to facilitate successful performance in the world (Craik, 1943; Brewer, 1987). However, the function of mental models is more intricate than it appears to be because of the tension in deciding whether to abandon an old representation of the world in favour of a new one. Based upon common sense and memory tasks, the time and effort consumed in relearning an old mental model is supposed to be less than learning a new one, and an old mental model may block the acquisition of a new one (e.g., Beda & Smith, 2018). However, there is a lack of direct comparison between the learning processes of old and new mental models. A further question is the degree of positively or negatively supporting evidence required to maintain an old mental model or create a new mental model, respectively.

With the success of using simple games studying decision-making, in this thesis, I investigate whether people are able to a) first acquire mental models to defeat an exploitable opponent in a competitive game, b) whether increasing or decreasing win rates as forms of reinforcement and punishment trigger the maintenance of the old model or development of a new model, and, c) whether there are overall performance differences in favour of building a new model and maintaining an old one.

1.1 Core findings

The core findings from the research provide evidence that:

- Compared to building a new mental model, people perform better in retrieving an old one.
- As the degree of punishment of the old mental model increases, people are more willing to build new mental models to achieve positive outcomes. In contrast, as the degree of reinforcement of an old mental model increases, new mental model acquisition is sabotaged.
- Behaviours following positive outcomes are generally more flexible and adaptive to new environments relative to behaviours following negative outcomes.

1.2 Thesis organization

In Chapter 2, previous studies in mental models, learning processes, and reinforcement and punishment are summarized to set the foundation for the current research. In Chapters 3 to 5, the details of experimental designs, data collection, results, and discussions of Experiments 1 - 3 are illustrated and reported separately. Finally, in Chapter 6, the thesis is summarized by a general discussion and some suggestions for future research direction.

Chapter 2

Background

2.1 Mental models

Humans are amazing in their capacity to condense finite, noisy, and ambiguous information and absorb what is useful (see Johnson-Laird, 2013, and, Tenenbaum, Kemp, Griffiths, & Goodman, 2011, for reviews). This capacity largely depends on past reactions or experiences (Bartlett, 1932; Wagoner, 2013), culminating in environmental representations, which will be consulted prior to future behaviour (Tolman, 1948). Craik (1943) describes such representations as “small-scale models,” which are able to use the knowledge of past events and select the best option from various alternatives. Brewer (1987) uses the term “mental model” for "all forms of mental representation, general or specific, from any domain, causal, intentional or spatial" (p. 193). The acquisition of a mental model is closely related to the process of learning.

2.2 Learning

Learning is an everyday activity in everyone’s life, with successful learning defined as improvement during future relative to current performance (McGeoch, 1942). From a cognitive standpoint, learning arises from the interaction between current stimulation and stored mental representation, and it results in a change in the learner's mental model (Vandenbosch & Higgins, 1996). A keystone element in the study of learning is the route via which individuals learn. Like a cook who learns to make a new meal for the first time, they may search for a readily-made recipe and follow it (*learning via instruction*), may copy other cooks who can cook the dish

(*learning via imitation*), or may adapt their prior knowledge of similar meals to the new one (*learning via experience*). Warnings and directions on drug-package inserts represent a clear case of *learning via instruction* as consequences (and the likelihood of those consequences) are specified, thereby providing the user the information they need to make the decision of taking the drug or not (Phillips, Fletcher, Marks, & Hine, 2016). *Learning via imitation* in daily life is exemplified by the way children mimic their parents' behaviours. Related theories arise from the seminal study on the Social Learning Theory (Bandura, Ross, & Ross, 1963), where Bandura and colleagues found that children's aggressive behaviours increased after they observed aggressive models, and would continue even when the models no longer existed. The method of *learning via experience* describes how people learn from their own interaction with environmental stimuli. Systematic studies are traced back to 20th-century scholars such as William James, John Dewey, and Jean Piaget, who emphasized experience in their theories of human learning and development (Passarelli & Kolb, 2011).

These three methods of learning are efficiently applied to decision-making situations like risky decisions (Hertwig, Barron, Weber, & Erev, 2004), prisoner's dilemma games (Kirchkamp & Nagel, 2007), and economic choices (Pingle, 1996). There appears to be no direct comparison among the three methods to suggest which one outperforms the others. Olson and Bruner (1974) state that the three methods are similar in the knowledge they specify, while different in the skills they develop. For example, *learning via instruction* highlights the skill of extracting information from language, *learning via imitation* highlights the skill of observing differences and imitating, while *learning via experience* represents the skill of obtaining information by perceptual input. Although varied, all three adaptive ways have been examined to be efficient toward successful learning (Norman, 1982). In the current study, participants played a game against the computer

and could only *learn via experience*, because there were no verbal or written hints telling them how to defeat the computerized opponent, and no chance to observe what other players were doing.

2.3 Learning via experience

Games are widely used in experiments to understand the mechanisms of decision-making because they are motivationally appealing (Ke, 2009). Although far less complex and complete compared to the real world, they parallel interactions within people and between people and environments (Schlenker & Bonoma, 1978). Additionally, using games to measure learning helps researchers to have control over variables that may affect players' behaviours (Lakkaraju et al. 2018). Performance outcomes via feedback are common stimuli provided to game players, which allow players to learn about game-related information. In a trial-by-trial game, stimuli are often presented to participants in each trial, and performance on the next trial of the participants modulates accordingly if learning by experience is taking place. An example is the zero-sum game rock-paper-scissors game (RPS; e.g., Cook et al., 2012). RPS is usually played between two players, reaching three outcomes: win, loss, or draw. For example, if player A chooses to play Paper, player B has three options: play Rock resulting in A-win-B-lose, Paper resulting in A-draw-B-draw, or Scissors resulting in A-lose-B-win. During the game, players are able to form strategies. For example, if both players choose rock, paper, and scissors randomly and of equal distribution (termed *mixed strategy*; Dyson, Wilbiks, Sandhu, Papanicolaou, & Lintag, 2016), their behaviours can realize a prediction called *mixed Nash equilibrium*, which is "a probabilistic distribution on the set of actions of each player. Each of the distributions should have the property that it is the best response to the other distributions; this means that each action assigned positive probability is among the actions that are best responses, in expectation, to the

distribution(s) chosen by the opponent(s)." (Daskalakis, Goldberg, & Papadimitriou, 2009; p. 89). In other words, in a multiple-trial two-player zero-sum game like RPS, the two players can come to a score of 0 on average by playing options stochastically.

However, in reality, people usually fail to play the *mixed strategy*, which makes their patterns to be predictable and exploitable for opponents (McNamara, Houston, & Leimar, 2021). Vice versa, if the opponent plays a pattern that can be noticed, people are capable of exploring and exploiting it. In the *exploration* state, people gather information about environments and interpret the information to build up mental models of the world. The ability to explore under uncertain environments is a function directed by the prefrontal cortex (Cavanagh, Figueroa, Cohen, & Frank, 2011; Badre, Long, & Frank, 2012). The process has also been observed in nature as animals try out different locations and various timings for collecting food (Jahn, 2023).

By exploring both directly (comparing known options and choosing the one with the highest payoff) and randomly (choosing uncertain options by chance), people aim to increase their reward in the long run (Wilson, Geana, White, Ludvig, & Cohen, 2014). After a few trials of exploration, both humans and monkeys can move into an *exploitation* state, where previously gathered information forms mental models for gaining more rewards. The transition from the *exploration* state to the *exploitation* state has been examined using a spatial selection task (e.g., Procyk & Goldman-Rakic, 2006), reward association task (e.g., Achterberg et al., 2022), and zero-sum games (e.g., Sun & Jia, 2023). However, because opponents' strategies may vary as time goes by, the *exploitation* state is not a final state. If an unfavourable outcome occurs, like a player losing a trial because of staying at their previous choice, one is likely to partially keep exploring until they feel certain of prospective rewards (Achterberg et al., 2022). The states transition is accompanied by the change of mental states as well. For example, in a game, wins

and rewards make people content with their current strategy, and they may be hopeful for more rewards; while they will become watchful and discontent when they start to lose, and may be inclined to explore new strategies for better outcomes (Young, 2009).

2.4 Mental model acquisition: building, updating, and retrieving

Mental model building happens in the exploration state where information is acquired and interpreted. Vandenbosch and Higgins (1996) propose that mental models can be established by simply gathering information without a specific goal. Within a zero-sum game, players learn to beat opponents without knowing what strategy the opponents are using, and players' mental models are built up in finite trials. For example, by using a two-player zero-sum game, Bakken (n.d.) finds that players can successfully learn the *mixed-strategy* and play optimally (i.e., randomly and equally-distributed choosing all options), and Brockbank and Vul (2021) find that players can recognize and counter-exploit their opponents' strategies. The two studies also suggest that people's ability to build such mental models is limited to opponents' relatively simple behavioural patterns. If opponents' regularity goes too complex, like consulting to the previous two actions of a player, the player is not able to recognize or counter-exploit it (Bakken, n.d.; Brockbank & Vul, 2021).

Because of the huge amount and variety of information people deal with every day, mental models must be flexible rather than static, and building up one mental model is not an end game in and of itself (Filipowicz, Anderson, & Danckert, 2016; Johnson-Laird, 2013). That is to say, mental models need to adapt to the changing environments. For example, imagine a student working on math problems. They solve the first problem with division. So "division can solve a math problem" has become their established mental model. When going to the second problem, they need to figure out whether it can still be solved by division (reacquisition of the old model),

or by division and subtraction together (accretion and tuning of the old model by integrating new information), or by subtraction solely (acquisition of a new model).

Filipowicz (2017) suggests there are three main stages for a mental model to be updated according to environments: the first is to build up and compare a mental model with the environments; the second is to detect if there is any mismatch between what the current mental model predicts and the real outcome; third is to think of alternative mental models if there are mismatches. Although there are individual differences, people are able to process such mental model updating. For instance, in a game called Plinko, in which players predict where exactly a ball will drop, Filipowicz, Anderson, and Danckert (2016) find that participants can learn the distribution of ball drops (which is relevant to information seeking and is determined by the inferior parietal lobule) and update their predictions accordingly (which is relevant to new mental model exploration and is determined by the medial prefrontal cortex).

2.5 Relationship between new and old mental models

Acquiring a new mental model (or learning a new strategy in a game) and reacquiring an old mental model (or relearning a readily learned / old strategy in a game) both require effort to explore. However, the degree of effort varies. Learning a new strategy consists of gathering information, figuring out regularities, and forming a novel mental model, while relearning an old strategy is about gathering information and retrieving an established mental model. The latter can be generally described as using fewer cognitive steps than the former. Therefore, relearning an old strategy is more efficient and less effortful than learning a new strategy, supported by Stöttinger et al. (2014). In their study, by employing computerized opponents playing two categories of strategies in two phases, they argue that if the strategy in phases 1 and 2 are of the

same category, participants choose more optimal options in phase 2, compared to the situation in which the strategies in phases 1 and 2 are of different categories.

There are many studies supporting that old knowledge has a large influence on learning new knowledge. Recall testing of old information has been suggested to have a positive effect on the learning of subsequent new information (see Pastötter & Bäuml, 2014, for review). On the other hand, many studies argue that relearning or retrieving old knowledge can block the study of new knowledge. For example, Beda and Smith (2018) find that retrieving a learned word pair hinders solving new word association problems. Finn and Roediger (2013) also find a similar impairment in associating faces-names-professions. In their experiments, participants learned face-name pairs first. And then they either saw the pairs again on screen or recalled the names with faces as clues before adding professions to the face-name pairs. The results suggest that those who recalled face-name pairs did worse in updating professions than those who re-saw the pairs. Such a block effect of old knowledge on new information has been discussed as *mental fixation*, which is defined as the proclivity to keep using old ideas, knowledge, and/or problem-solving attempts, regardless of knowing how unhelpful they are in new situations (Smith, 2003; Ditta, 2019). Another term to describe the negative effect is the *Einstellung (set) effect*, which indicates that when a piece of existing knowledge is triggered by a familiar feature, the exploration for a new strategy will be prevented (Bilalić, Mcleod, & Gobet, 2008). Davis and Chan (2015), more specifically, investigate both the positive and negative effects of relearning/retrieving old knowledge on new knowledge acquisition. By having participants tested (asking questions like "*What is the name of this face*") or restudied (simply showing old knowledge to participants), they find that with old knowledge testing, participants put more effort into old knowledge, which impairs new knowledge learning. However, Davis and Chan

also propose that testing-old only enhances learning-new when the two processes are clearly separated by blocks. Therefore, we may speculate that in a trial-by-trial task without block separation, relearning-old should only work as an inhibition to learning-new. What is also questionable is whether the abilities of relearning-old and learning-new are mutually exclusive: whether being good at relearning old knowledge infers being bad at learning new knowledge.

2.6 Operant conditioning and mental model acquisition

The process of operant conditioning (Skinner, 1963; Staddon & Cerutti, 2003) suggests that actions consistent with reinforcement (positive outcomes) are more likely to be repeated in the future, while actions consistent with punishment (negative outcomes) are less likely to be repeated in the future (the simplified logic of *Law of Effect*; Thorndike, 1911). These principles have been termed *win-stay* and *lose-shift* in a zero-sum game (Dyson, Wilbiks, Sandhu, & Papanicolaou, 2016). From a micro perspective (trial-by-trial) in such a game like RPS, for example, *win-stay* and *lose-shift* have been observed as a pattern of participants when against *mixed-strategy* opponents, suggesting that winning a trial encourages participants to perform that same action on the trial more in future trials (e.g., Wang, Xu, & Zhou, 2014) while losing discourages the previous action on the trial (e.g., Dyson et al., 2016). However, when playing against exploitable opponents (i.e., opponents playing in patterns that can be learned and exploited by participants), participants are able to jump out of the *win-stay* and *lose-shift* box. In other words, they learn to counter-exploit the opponents' strategies regardless of whether optimal choices conflict with *win-stay* / *lose-shift* or not (Sundvall & Dyson, 2022).Additionally, because *lose-shift* is examined to be inflexible in outcome magnitude while *win-stay* is flexible (Forder & Dyson, 2016), it is possible that behaviours after a win (regardless of staying or shifting) reflect future performance and the change of mental models better than behaviours after a loss. From a

macro perspective (block-by-block), a block of reinforcement for previous behaviours is assumed to encourage people to stay at them and not to try new ones, while a block of punishment may drive people away from previous behaviours and shift to new ones. From both micro and macro perspectives, reinforcement and punishment help understand mental model acquisition.

2.7 Aims

The experiments reported in this thesis aim to examine the ability of participants to build up new mental models of strategies and to retrieve old mental models in a zero-sum game, the effect of reinforcement and punishment on the acquisition of mental models, and if that effect is influenced by the degree of the reinforcement and punishment. Experiments 1, 2, and 3 addressed how well participants can acquire mental models (either new or old) by calculating how many trials they won and in how many trials they chose optimal actions to beat their opponents in different periods. The three experiments also compare the effect of different degrees of reinforcement and punishment on participants' performance by having varied manipulated win rates between two exploitable periods. The experiments aim to shed light on the influence of reinforcement and punishment on building a new mental model and retrieving an old one.

Chapter 3

Experiment 1

3.1 Introduction

Experiment 1 was designed to be an initial attempt to examine how well individuals buildup new mental models (represented by learning a new strategy in the experiment) in contrast to retrieving old mental models (represented by relearning an old strategy) when challenged with a chance win rate (50%) during learning. The current study used a zero-sum game called *Dice Dual* with binary outcomes (win and loss; after Hayes, 1975). The game has the same structure as the *Matching Pennies* game (e.g., Belot et al., 2013) but with 6 options (i.e., numbers 1 to 6) instead of 2 (i.e., head or tail).

All participants experienced four blocks in the experiment. Each block of the *Dice Dual* game was separated into three main parts (*Pre*, *During*, and *Post*; detailed in Method section), where exploitable strategies were available in *Pre* and *Post* bins, and fixed reinforcement/punishment mechanisms were established in *During* bins. The design allowed participants to relearn an old strategy (when the strategy in *Pre* and *Post* were the same) and learn a new strategy (when the strategies in *Pre* and *Post* were different).

In Experiment 1, a 50% win rate was fixed in *During* bins to influence participants' performance in *Post* bins. If participants' average win rate in *Pre* bins was higher than chance, then a fixed 50% win rate in *During* bins represented the punishment of their previous actions, potentially leading to suppression in relearning old strategies but facilitating the learning of new

strategies in *Post* bins. In other words, there are a readily-stored mental model A and a to-be-learned mental model B. The rejection of mental model A may negatively impact the re-use of mental model A in the future. This is in contrast to the new acquisition of mental model B. Therefore, the proportions of wins and optimal behaviour generalized by the novel learning of mental model B should be higher than the reuse of mental model A, following the punishment of mental model A.

3.1.1 Hypotheses

Experiment 1 examined the degree to which participants could successfully learn a new strategy and relearn an old strategy in response to a mild punishment (50% fixed win rate) delivered in the middle of learning. It is hypothesized that people can learn to exploit opponents' strategies in a simple game, and relearning an old strategy is easier than learning a new one. Additionally, according to the function of punishment and reinforcement in previous literature, it is assumed that the former suppresses people's early behaviours while the latter enhances the behaviours, in both trial-by-trial and block-by-block manners.

3.1.1.1 Predictions

1) Participants should learn to exploit the opponent's strategy during the initial stages of the game (*Pre* bins), leading to above-chance performance (the random chance of winning is 50% because the game has binary outcomes).
2) Participants should learn to exploit the opponent's strategy again during the final stages of the game (*Post* bins). With suppression of an old strategy in the *During* bins (fixed 50% win rate), performance should be better for the acquisition of a new strategy.

3) Optimal behaviour performance should be strongly linked to the results of the win rate performance.
4) Participants should play more optimal behaviours following wins than losses.
5) The degree of suppression of optimal behaviours exhibited between *Pre* and *During* bins should be positively correlated with the individual ability to learn a new strategy, but negatively correlated with the individual ability to relearn an old strategy.

3.2 Method

3.2.1 Participants

Data from a convenience sample of 76 participants (Mean = 18.76, SD = 2.13, 48 female, 58 right-handed) from the student population at the University of Alberta were analyzed. Two behavioural exclusion criteria were implemented: 1) item bias: where a participant selected the same item 100% of the time throughout at least one condition (1 participant excluded), and 2) ceiling win rate: where a participant won all trials in at least one condition (4 participants excluded). The second criterion was justified because a participant's behaviour preference after a loss could not be calculated if they achieved a 100% win rate. All participants gave their informed consent for inclusion and they were informed that the game took about 35 minutes to complete before they participated in the study. Participants received course credit and were only eligible to take part in one experiment in the following series. The protocol was approved at the University of Alberta under Research Ethics Board 2 (Pro00120832).

3.2.2 Stimuli

Game trials. Pictures of an ordinary six-sided die (white spots on a black background, participant; black spots on a white background, opponent) were displayed, with participants

sitting approximately 60 cm away from a 27" ViewSonic VX2757 Monitor. Participants chose one number using 6 linearly organized keys. Paradigms were controlled by Presentation 23.0 (build 10.27.21), and responses were recorded using a keyboard.

3.2.3 Design

The experiment was a 2x2 within-subject design with two strategies (*exploitable via repetition*, *exploitable via alternation*) and two learning contexts (learning a new strategy, relearning an old strategy) as factors.

Participants completed 504 trials of the Dice Dual game, divided into 4 counterbalanced blocks of 126 trials. Each block was divided into 7 bins of 18 trials (see Table 3.1). In each block, the first bin was always *unexploitable via mixed strategy* (M; i.e., the computer played all six numbers 3 times with equal proportion and randomly); and the fourth and fifth bins (*During*) were always *unexploitable via a fixed 50% win rate* ($F_{50;}$ i.e., no matter what number the participant chose, it was ensured that the participant's win rate of the two bins was fixed at 50%).

Two *exploitable* strategies were used in the other four bins: *exploitable via repetition* (R; i.e., choosing odd/even numbers repetitively; e.g., 264462), and *exploitable via alternation* (A; i.e., choosing odd and even numbers alternatingly; e.g., 523614). By using the same strategy, the second and third bins were combined and referred to as the *Pre* bins, and the sixth and seventh bins were combined and referred to as the *Post* bins. *Pre* and *Post* here referred to before or after exposure to the fixed 50% win rate (*During*).

In *relearning-old-strategy* conditions, the strategy used in *Pre* bins was the same as the one used in *Post* bins (e.g., *exploitable via alternation* strategy was used in *Pre* bins and *Post* bins). In *learning-new-strategy* conditions, the strategy used in *Pre* bins (i.e., the 2nd and 3rd bins) was different from the one used in *Post* bins (i.e., the 6th and 7th bins). For example, the

exploitable via alternation strategy was used in *Pre* bins, while the *exploitable via repetition* strategy was used in *Post* bins.

		Pre bins		*During* bins		*Post* bins	
	Bin 1	Bin 2	Bin 3	Bin 4	Bin 5	Bin 6	Bin 7
Relearning	M	R	R	F_{50}	F_{50}	R	R
Old	M	A	A	F_{50}	F_{50}	A	A
Learning	M	A	A	F_{50}	F_{50}	R	R
New	M	R	R	F_{50}	F_{50}	A	A

Table 3.1 Schematic depicting all four blocks of relearning old strategy and learning new strategy conditions in Experiment 1. Two unexploitable strategies: 1) M: computerized opponent using mixed-strategy; 2) F_{50}: participants' win rate was fixed-50%. Two exploitable strategies: 1) R: computerized opponent repetitively choosing odd/even numbers; 2) A: computerized opponent alternatingly choosing odd and even numbers. Each row represents one block, and each bin represents 18 trials. The sequence of the four blocks was counterbalanced.

3.2.4 Procedure

Participants were informed they would play 504 rounds of the Dice Dual game consisting of 4 blocks with 126 trials each with the goal to win as many trials as they could. At each trial, participants and the computerized opponent chose one number from the six sides of a die (i.e., 1, 2, 3, 4, 5, 6) prompted by a fixation cross. In the Even version of the Dice Dual, participants won the trial if the sum of the two sides was even and lost the trial if the sum was odd; in the Odd version, participants won the trial if the sum of the two sides was odd and lost the trial if the sum was even (versions were counterbalanced across participants). Instructions of the goal and the rules were given to participants in written words on computer screens before the game. There were no training trials. Responses from participants and opponents were then shown for 1000 ms for each trial. Participants' choice was shown on the right side of the screen by a black dice with white spots, and opponents' was shown on the left side by a white dice with black spots. The sum of the two dice was presented in the middle of the screen for 500 ms, after which the outcome of the trial was presented for 1000 ms in the form of "WIN +1" or "LOSE -1" (in green

or red font, respectively). Scores and trial numbers were updated, and the fixation cross returned ready for the next input (see Figure 3.1). Participants were instructed not to think too long and were encouraged to take breaks at will between any two blocks. Instructions and all other parameters were identical across all blocks. After completing all 4 blocks, participants completed the BIS/BAS scales (Carver & White, 1994), and then, participants were thanked for their time and debriefed.

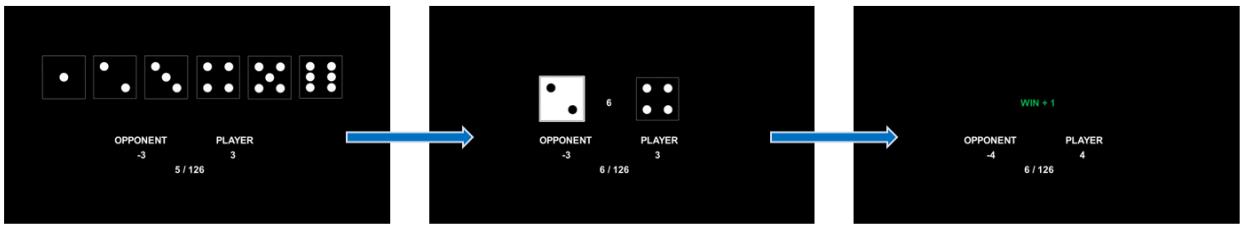

Figure 3.1 An example of a winning trial in the Even version of Dice Dual. When the sum of the two dice was even (in the second panel, 2 played by the opponent + 4 played by the player = 6), the player won 1 point and the opponent lost 1 point.

3.3 Results

3.3.1 Win rate distribution

The win rate represents a proportion of a participant's winning trials over all trials within a given condition (e.g., if the participant won 24 trials out of 36 trials, the win rate is 66.7%). See Table 3.2 for the distribution of player win rates across the two *Learning* conditions (relearning old vs. learning new). The average win rate in *Pre* bins (M = .637, SE = .015) collapsed across relearning-old and learning-new conditions was compared to chance performance (50%) via a one-sampled t-test: ($t(75) = 10.799$, $p < .001$). This confirms H1 in that, at a group level, participants learned to exploit opponent strategy during the initial stages of the game above chance. This observation further confirms that exposure to the fixed win rate in the *During* bins

will serve as a mild punishment mechanism (performance reduced on average by 13.7%). The average win rate in *Post* bins (M = .638, SE = .012) collapsed across relearning-old and learning-new conditions and was also compared to chance performance (50%) via a one-sampled t-test: ($t(75) = 11.055$, $p < .001$). This confirms the first half of H2 in that participants learned to exploit opponent strategy during the final stages of the game above chance at a group level.

Table 3.3 shows the inferential statistics of a two-way repeated measures ANOVA with *Learning* (Old, New), and *Period* (*Pre*, *Post* only; *During* bins were excluded in this analysis for the win rate in them was fixed at 50%). The result of the significant two-way interaction between *Learning* x *Period* suggested that the *Post* bins win rate in relearning-old-strategy condition (M = .662, SE = .015) was significantly higher than that in learning-new-strategy condition (M = .614, SE = .013; Tukey's HSD; $p < .05$). As expected, win rates between Old_{50} and New_{50} during *Pre* period were not significantly different from one another. Therefore, in contrast to the second-half prediction of H2, it was easier for participants to reacquire an old strategy, as opposed to adapting to a new strategy, following a period of mild punishment.

(win rate)	*Pre*	*During*	*Post*
Old_{50}	0.627 (.015)	0.50 (.000)	0.662 (.015)
New_{50}	0.647 (.015)	0.50 (.000)	0.614 (.013)

Table 3.2 Descriptive statistics of win rate in Experiments 1. Old: relearning old strategy; New: learning new strategy. $_{50}$: fixed 50% win rate in *During* bins. Note: Standard error in parenthesis.

	df	*F*	*MSE*	*p*	η_p^2
Learning (L)	1,75	1.486	.0094	.227	.019
Period (P)	1,75	.003	.0075	.956	.000
L x P	1,75	**10.415**	**.0173**	**.002**	**.122**

Table 3.3 Inferential statistics for win rate in Experiment 1. Two-way repeated measures ANOVA results. Learning (L): 2 levels (Old / New); Period (P): 2 levels (Pre / Post). Note: Significant effects in bold font.

3.3.2 Optimal behaviour rate (OBR)

The rate of optimal behaviour represents the degree to which participants initiated the correct action in response to the strategies presented. In the case of the *repetitive-strategy*, this was the proportion of *win-stay* trials over the total number of winning trials and the proportion of *lose-shift* trials over the total number of losing trials. For the *alternating-strategy*, this was the proportion of *win-shift* trials over the total number of winning trials and the proportion of *lose-stay* trials over the total number of losing trials. The optimal behaviour rate (OBR) is analysed on top of the win rate because the latter cannot indicate participants' performance in *During* bins for being fixed. Table 3.4 shows the distribution of player OBR across the two *Learning* conditions and the two *Outcome* conditions. Data in the current experiment and the subsequent experiments were collapsed across the nature of the opponent (*exploitable via repetition* and *exploitable via alternation*).

(optimal behaviour rate)		*Pre*	*During*	*Post*
After a win	Old_{50}	0.687 (.022)	0.596 (.018)	0.714 (.022)
	New_{50}	0.697 (.021)	0.636 (.017)	0.643 (.021)
After a loss	Old_{50}	0.608 (.017)	0.518 (.015)	0.660 (.016)
	New_{50}	0.644 (.019)	0.546 (.013)	0.636 (.016)

Table 3.4 Descriptive statistics of optimal behaviour rate (OBR) after a win and after a loss in Experiment 1. Old: relearning old strategy; New: learning new strategy. $_{50}$: fixed 50% win rate in *During* bins. Note: Standard error in parenthesis.

OBR was analyzed using a three-way repeated measures ANOVA with *Learning* (Old, New), *Period* (*Pre*, *During*, *Post*), and *Outcome* (after a win, after a loss) entered as factors (see

Table 3.5). There were significant main effects of *Period* as well as *Outcome* (*p*'s < .001), as well as significant two-way interactions between *Learning* x *Period* (*p* < .001), and between *Period* x *Outcome* (*p* = .014). There was no significant three-way interaction (*p* = .213; see Figure 3.2).

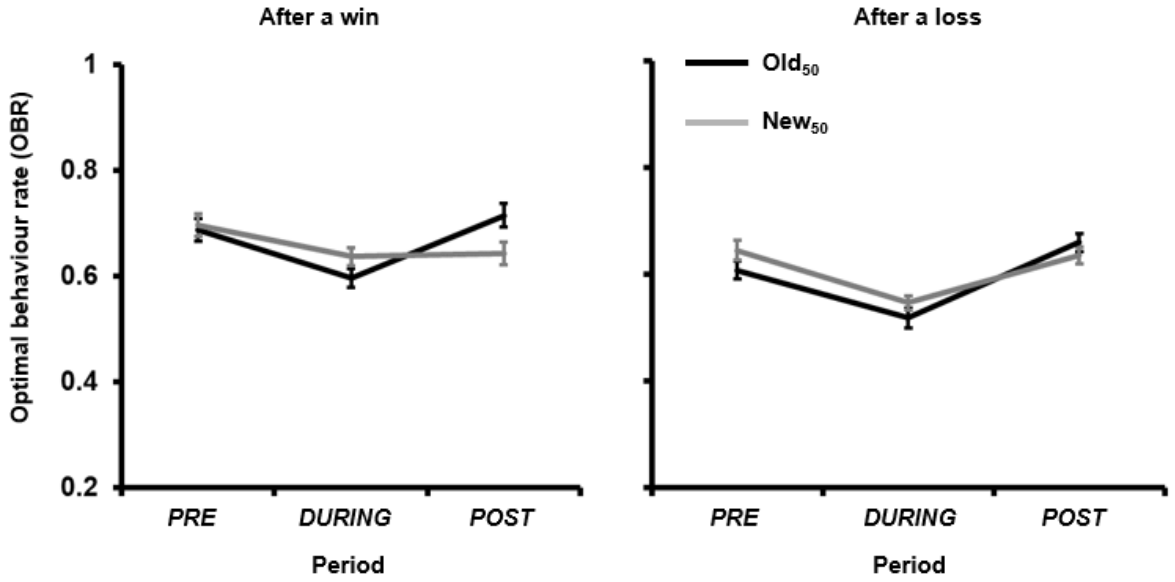

Figure 3.2 Distributions of participants' OBR in the two *Learning* conditions (relearning old strategy vs. learning new strategy) and two *Outcome* conditions (after a win vs. after a loss) collapsed across the *Exploitable* conditions (*exploitable via repetition* & *exploitable via alternation*) in Experiment 1. Note: Error bars represent standard errors.

In terms of the interaction between *Learning* and *Period* (2 x 3 interaction; see Figure 3.3.a), OBR was higher in *Post* bins in the relearning-old-strategy condition (M = .687, SE = .017), compared to the learning-new-strategy condition (M = .639, SE = .016; Tukey's HSD; *p* < .05), suggesting that participants were better at performing behaviour related to the retrieval of an old strategy than the learning of a new one. These OBR data are consistent with the two-way interaction between *Learning* x *Period* observed in win rates observed for post-relearning old trials and thus consistent with H3.

In terms of the interaction between *Period* and *Outcome* (see Figure 3.3.b), OBR after a win in *Pre* bins (M = .692, SE = .018) and *During* bins (M = .616, SE = .014) were higher than those after a loss in *Pre* bins (M =.626, SE = .015) and *During* bins (M = .532, SE = .012; Tukey's HSD; $p < .05$), respectively. However, the difference between rates of optimal behaviours after a win (M = .679, SE = .019) and after a loss (M = .648, SE = .013) was not significant in *Post* bins (Tukey's HSD; $p = .14$). Therefore, the data are broadly in support of H4 in that participants were generally more successful in expressing the correct mental model following wins relative to losses, although this difference was weaker during post-trials.

	df	*F*	*MSE*	*p*	η_p^2
Learning (L)	1,75	.096	.0240	.758	.001
Period (P)	2,150	**29.725**	**.0259**	**<.001**	**.284**
Outcome (O)	1,75	**41.442**	**.0200**	**<.001**	**.356**
L x P	2,150	**8.681**	**.0173**	**<.001**	**.104**
L x O	1,75	2.130	.0116	.149	.028
P x O	2,150	**4.417**	**.0124**	**.014**	**.056**
L x P x O	2,150	1.561	.0114	.213	.020

Table 3.5 Inferential statistics for OBR in Experiment 1. Three-way repeated measures ANOVA results. *Learning (L)*: 2 levels (Old / New); *Period (P)*: 3 levels (*Pre* / *During* / *Post*); *Outcome (O)*: 2 levels (Win / Loss). Note: Significant effects in bold font.

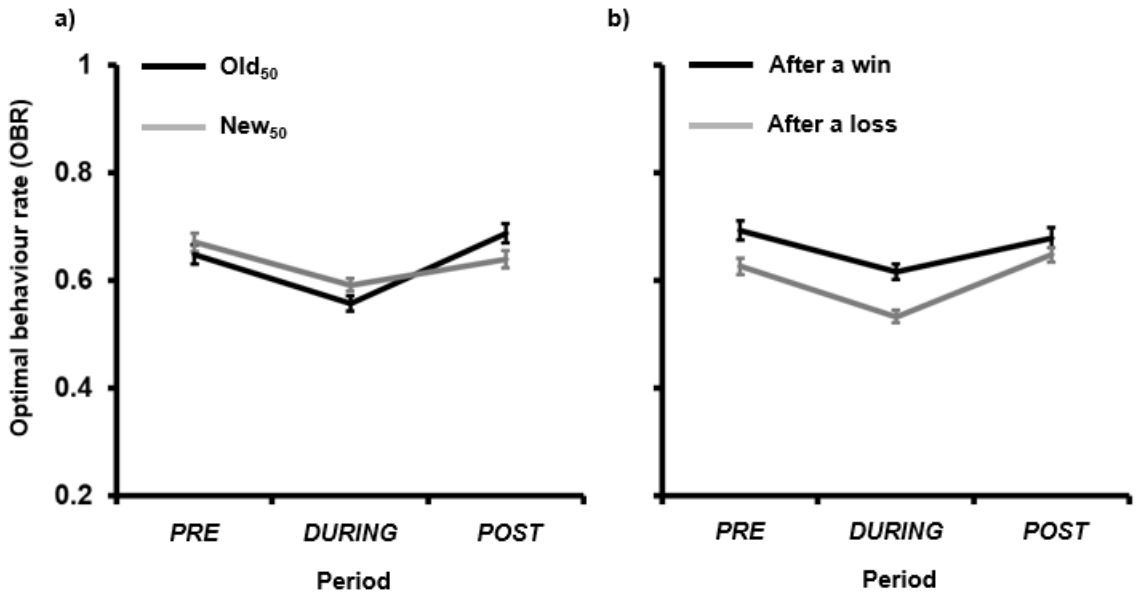

Figure 3.3 Distributions of participants' OBR in: a) the two *Learning* conditions (relearning old strategy vs. learning new strategy), and b) the two *Outcome* conditions (after a win vs. after a loss) collapsed across the *Exploitable* conditions (*exploitable via repetition* & *exploitable via alternation*) in Experiment 1. Note: Error bars represent standard errors.

3.3.3 Strategy suppression

To test individual sensitivity to mild punishment (in the form of decreasing win rates to a fixed 50% in the *During* period), the degree of OBR suppression was calculated as [OBR *Pre*] minus [OBR *During*] following both wins and losses. These were then correlated at an individual level with *Post*- win rates in the context of relearning an old strategy (Figure 3.4.a) and learning a new strategy (Figure 3.4.b). The change in the OBR after a win in *During* bins was positively correlated with the win rate in *Post* bins in both learning-new-strategy ($r = .359$, $p < .001$) and relearning-old-strategy conditions ($r = .361$, $p < .001$). However, the change of the OBR after a loss was not significantly correlated with later performance in either learning-new ($r = .219$, $p = .058$) or relearning-old conditions ($r = .065$, $p = .577$). This is in partial support of H5 in that participants who were more willing to stop their strategy when it ceased being effective performed better following mild punishment (fixed 50% win rate). However, this effect was

similar for learning a new strategy and for relearning an old strategy. Finally, the significant correlations after wins but not for losses showed that participants were able to exhibit more control over behaviour following positive relative to negative outcomes.

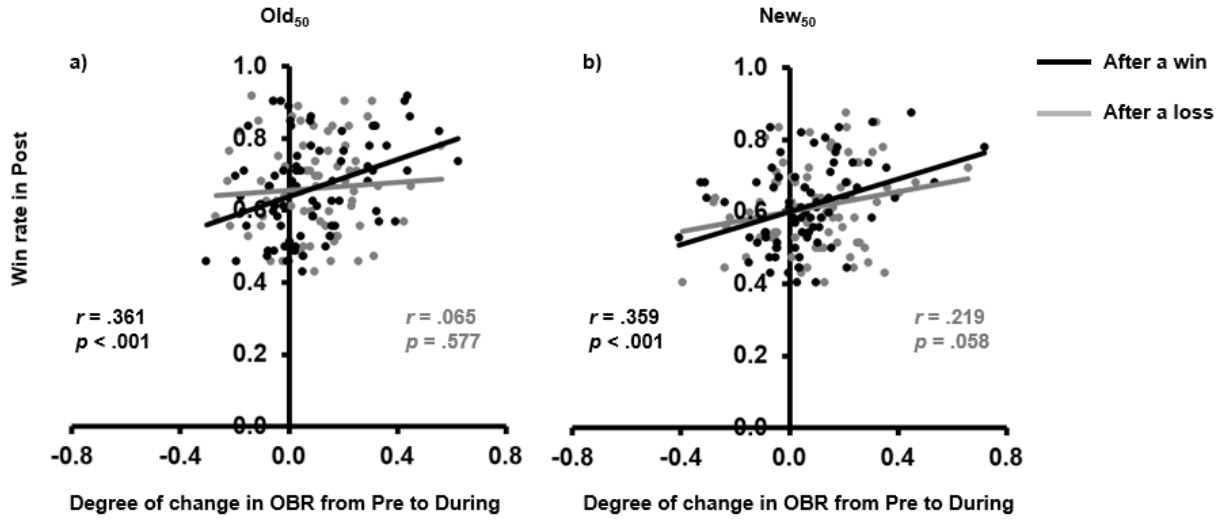

Figure 3.4 Correlation between win rate in *Post* bins and the degree of change in OBR from *Pre* bins to *During* bins in Experiment 1. A positive OBR difference indicates lower OBR in *During* bins than in *Pre* bins. Panel a) represents relearning-old-strategy condition; panel b) represents learning-new-strategy condition.

3.4 Discussion

Experiment 1 represents a successful framework for testing individual abilities to acquire, reject and either reactivate or change mental models as a function of punishment. The above chance *Pre* win rate suggested modest mental model acquisition before punishment. The OBR was in accordance with the win rate that it was suppressed during mild punishment (win rate reduced by 16% to a fixed 50%).

In contrast to the initial prediction, mild punishment in Experiment 1 facilitated the reinstatement of an old mental model, as opposed to the acquisition of a new mental model. The degree of OBR suppression was a significant predictor of both retrieving an old mental model

and building up a new mental model. Also importantly, the mental model acquisition was better expressed following positive relative to negative outcomes. Given that degree of OBR suppression was positively correlated with post-punishment win rates following wins but not losses, suggesting participants had a larger degree of behavioural flexibility following wins relative to losses. In other words, following punishment, participants were more willing to adjust their actions after a win to adapt to the new situation, compared to actions after a loss. This post-win flexibility has also been examined in speeding/slowing after wins and losses that reaction times after wins are more flexible than those after losses (Dyson, 2023) and that behaviours following wins are more flexible than those following losses (Forder & Dyson, 2016).

Reacquiring an old mental model was found to be easier than establishing a new mental model following punishment (Stöttinger et al., 2014). Experiment 1 showed that mild punishment of the old mental model affected both new mental model acquisition and old mental model reacquisition. Furthermore, even with such punishment, participants performed better in reacquisition. This suggests *mental fixation* (Smith, 2003) that old mental models are easy to retrieve but difficult to abandon.

Chapter 4

Experiment 2

4.1 Introduction

One reason why a failed mental model was relearned better than a new mental model may have been due to the mild degree of punishment the old model received. Therefore, in Experiment 2, the degree of punishment was increased to a fixed 25% win rate in *During* bins. Furthermore, Experiment 2 also employed a reinforcement condition where win rates were raised to 75% in *During* bins. This was to glean whether punishment and reinforcement act in similar, complementary ways with respect to mental model management.

4.1.1 Hypotheses

The fixed 25% win rate is assumed to serve as a punishment for old strategies, and following it, relearning old strategies is supposed to be suppressed while learning new ones is enhanced. The fixed 75% win rate is assumed to serve as a reinforcement for old strategies, and following it, relearning old strategies is supposed to be enhanced while learning new ones is suppressed.

4.1.1.1 Predictions

1) Participants should learn to exploit the opponent's strategy during the initial stages of the game (*Pre* bins), leading to above-chance performance (50%).
2) Participants should learn to exploit the opponent's strategy during the final stages of the game (*Post* bins). With stronger suppression of an old strategy in the *During* bins

(fixed 25% win rate), performance should be better for the acquisition of a new strategy.

3) OBR should be strongly linked to the results of the win rate performance.
4) Participants should exhibit more OBR following wins than losses.
5) The degree of suppression of optimal behaviours exhibited between *Pre* and *During* bins should be positively correlated with the individual ability to learn a new strategy, but negatively correlated with the individual ability to relearn an old strategy.
6) Participants should learn to exploit the opponent's strategy during the initial stages of the game (*Pre* bins), leading to above-chance performance (50%).
7) Participants should learn to exploit the opponent's strategy again during the final stages of the game (*Post* bins). With the enhancement of an old strategy in the *During* bins (fixed 75% win rate), performance should be better for the reacquisition of the old strategy.
8) OBR should be strongly linked to the results of the win rate performance.
9) Participants should exhibit more OBR following wins than losses.
10) The degree of enhancement in optimal behaviours exhibited between *Pre* and *During* bins should be positively correlated with the individual ability to relearn an old strategy, but negatively correlated with the individual ability to learn a new strategy.

4.2 Method

4.2.1 Participants

Data from a convenience sample of 72 participants (Mean = 19.19, SD = 2.52, 49 female, 53 right-handed) from the student population at the University of Alberta were analyzed. Eight participants were excluded from analyses according to Criterion 2 described in the Method

section of Experiment 1. Participants received course credit, and the protocol was approved at the University of Alberta under Research Ethics Board 2 (Pro00120832).

4.2.2 Stimuli

The stimuli and experimental set-up were identical to those of Experiment 1.

4.2.3 Design

In contrast to Experiment 1, Experiment 2 was a 2x2x2 within-subject design with two aimed strategies (*exploitable via repetition, exploitable via alternation*), two kinds of fixed win rates (25%, 75%) in *During* bins, and two learning contexts (learning a new strategy, relearning the old strategy) as factors. Participants encountered the same four types of computerized opponents (*unexploitable via mixed strategy, unexploitable via fixed win rate, exploitable via repetition,* and *exploitable via alternation*) as in Experiment 1. In *During* bins in Experiment 2, participants' win rate was designed to be 25% (F_{25}; punishment: encouraging participants to stop using strategies learned from previous trials) and 75% (F_{75}; reinforcement: encouraging participants to continue using strategies learned from previous trials) times of trials instead of 50% in Experiment 1.

Experiment 2 again had 4 blocks (126 trials each) with randomized order. The design of this experiment was generally identical to Experiment 1, apart from the two changes in *During* bins (see Table 4.1). First, participants' win rates were fixed at 25% (F_{25}) and 75% (F_{75}) instead of 50%. Second, the two bins of *During* bins were designed to consist of 20 trials and 16 trials (sequence fixed) instead of 18 trials each. This was to ensure the number of trials in both bins could be divisible by 4 (to have a 25%- and a 75%-win-rate).

		Pre bins		*During* bins		*Post* bins	
[Counterbalanced version 1]	Bin 1	Bin 2	Bin 3	Bin 4	Bin 5	Bin 6	Bin 7
Relearning	M	R	R	F_{25}	F_{25}	R	R
Old	M	A	A	F_{75}	F_{75}	A	A
Learning	M	R	R	F_{75}	F_{75}	A	A
New	M	A	A	F_{25}	F_{25}	R	R

		Pre bins		*During* bins		*Post* bins	
[Counterbalanced version 2]	Bin 1	Bin 2	Bin 3	Bin 4	Bin 5	Bin 6	Bin 7
Relearning	M	R	R	F_{75}	F_{75}	R	R
Old	M	A	A	F_{25}	F_{25}	A	A
Learning	M	R	R	F_{25}	F_{25}	A	A
New	M	A	A	F_{75}	F_{75}	R	R

Table 4.1 Schematic depicting two counterbalanced versions of all four blocks of relearning old strategy and learning new strategy conditions in Experiment 2. Two unexploitable strategies: 1) M: computerized opponent using mixed-strategy; 2) F_{25}: participants' win rate was fixed-25%, F_{75}: participants' win rate was fixed-75%. Two exploitable strategies: 1) R: computerized opponent repetitively choosing odd/even numbers; 2) A: computerized opponent alternatingly choosing odd and even numbers. Each row represents one block, and each bin represents 18 trials apart from Bin 4 (20 trials) and Bin 5 (16 trials). The sequence of the four blocks was counterbalanced.

4.2.4 Procedure

The procedure was identical to that of Experiment 1.

4.3 Results

To be consistent with the analyses in Experiment 1, instead of comparing the two conditions of *Fixed win rate* (25%, 75%, in *During* bins) in one analysis, the analyses were separated into *25%* and *75%* such that the Results section of Experiment 2 has the same structure as that of Experiment 1.

4.3.1 Fixed win rate at 25% in During bins

4.3.1.1 Win rate distribution

The win rate was calculated as in Experiment 1. See Table 4.2 for the distribution of player win rates across the two *Learning* conditions (relearning old vs. learning new). As in Experiment 1, the average win rate in *Pre* bins (M = .648, SE = .014) collapsed across relearning-old and learning-new conditions, was compared to chance performance (50%) via a one-sampled t-test: ($t(71) = 10.292$, $p < .001$). This confirms H1 in that at a group level, participants learned to exploit opponent strategy during the initial stages of the game above chance. This observation further confirms that exposure to the fixed win rate of 25% in the *During* bins will serve as a more severe punishment mechanism than the 50% in Experiment 1. The average win rate in *Post* bins (M = .671, SE = .016) collapsed across relearning-old and learning-new conditions, and was also compared to chance performance (50%) via a one-sampled t-test: ($t(71) = 10.955$, $p < .001$). This confirms the first half of H2 in that participants learned to exploit opponent strategy during the final stages of the game above chance at a group level, and participants were able to exploit opponents' strategies after the punishment of a 25%-win-rate in *During* bins.

(win rate)	*Pre*	*During*	*Post*
Old_{25}	0.662 (.017)	0.25 (.000)	0.660 (.020)
New_{25}	0.634 (.021)	0.25 (.000)	0.682 (.019)

Table 4.2 Descriptive statistics of win rate in Experiments 2. Old: relearning old strategy; New: learning new strategy. $_{25}$: fixed 25% win rate in *During* bins. Note: Standard error in parenthesis.

In contrast to Experiment 1, in the 25% during condition (Table 4.3), there were no significant main effects or interaction for win rates (all *p*s > .05). This suggests that the

advantage of relearning an old strategy following reduction to 50% win rate (Experiment 1) is abolished when win rate is reduced to 25% (Experiment 2), partially suggesting the second half of H2 that reacquiring old strategies did not outperform acquiring new strategies.

	df	*F*	*MSE*	*p*	η_p^2
Learning (L)	1,71	.030	.0232	.862	.000
Period (P)	1,71	2.464	.0157	.121	.034
L x P	1,71	2.900	.0157	.093	.039

Table 4.3 Inferential statistics for win rate with the fixed win rate at 25% in *During* bins in Experiment 2. Two-way repeated measures ANOVA results. *Learning (L)*: 2 levels (Old / New); *Period (P)*: 2 levels (*Pre* / *Post*). Note: Significant effects in bold font.

4.3.1.2 Optimal behaviour rate (OBR).

The rate of conducting optimal behaviours is defined in the same way as in Experiment 1. Table 4.4 shows the distribution of player OBR across the two *Learning* conditions and the two *Outcome* conditions.

(optimal behaviour rate)		*Pre*	*During*	*Post*
After a win	Old_{25}	0.718 (.028)	0.618 (.026)	0.728 (.029)
	New_{25}	0.654 (.034)	0.556 (.033)	0.755 (.027)
After a loss	Old_{25}	0.637 (.022)	0.519 (.018)	0.633 (.025)
	New_{25}	0.660 (.026)	0.552 (.019)	0.639 (.026)

Table 4.4 Descriptive statistics of optimal behaviour rate (OBR) after a win and after a loss in Experiment 2. Old: relearning old strategy; New: learning new strategy. $_{25}$: fixed 25% win rate in *During* bins. Note: Standard error in parenthesis.

OBR was again analyzed using a three-way repeated measures ANOVA with *Learning* (Old, New), *Period* (*Pre*, *During*, *Post*), and *Outcome* (after a win, after a loss) entered as factors (see Table 4.5 and Figure 4.1). Like Experiment 1, there were significant main effects of *Period*

as well as *Outcome* (*p*'s < .001); however, unlike Experiment 1, there were no significant two-way interactions between *Learning* x *Period* or *Period* x *Outcome* (*p*s > .05). The absence of significant interactions in OBR supports H3 that it is in accordance with win rate analyses.

For the main effect of *Outcome*, OBR was higher following wins (M = .671, SE = .015) than losses (M = .607, SE = .013), replicating Experiment 1 and supporting H4.

For the main effect of *Period*, OBR was the lowest in *During* bins (M = .561, SE = .011), compared to *Pre* bins (M = .667, SE = .017) and *Post* bins (M = .689, SE = .019), replicating Experiment 1.

	df	*F*	*MSE*	*p*	η_p^2
Learning (L)	1,71	.096	.0849	.758	.001
Period (P)	2,142	**27.439**	**.0488**	**<.001**	**.279**
Outcome (O)	1,71	**20.934**	**.0435**	**<.001**	**.228**
L x P	2,142	.831	.0358	.438	.012
L x O	1,71	3.262	.0483	.075	.044
P x O	2,142	2.830	.0324	.062	.038
L x P x O	2,142	2.662	.0289	.073	.036

Table 4.5 Inferential statistics for OBR in 25% condition Experiment 2. Three-way repeated measures ANOVA results. *Learning (L)*: 2 levels (Old / New); *Period (P)*: 3 levels (*Pre* / *During* / *Post*); *Outcome (O)*: 2 levels (Win / Loss). Note: Significant effects in bold font.

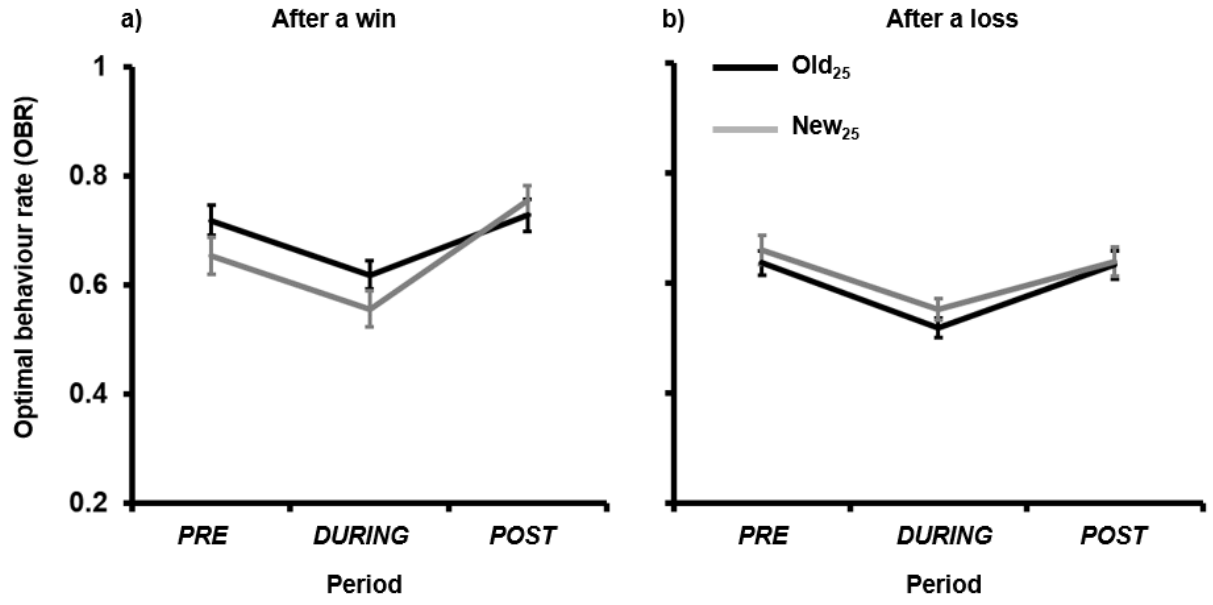

Figure 4.1 Distributions of participants' OBR in the two *Learning* conditions (relearning old strategy vs. learning new strategy) and the two *Outcome* conditions (after a win vs. after a loss) collapsed across the *Exploitable* conditions (*exploitable via repetitio*n & *exploitable via alternation*) in the 25%-fixed-win-rate condition in Experiment 2. Panel a) represents OBR after a win; panel b) represents OBR after a loss. Note: Error bars represent standard errors.

4.3.1.3 Strategy suppression

As in Experiment 1, the degree of OBR suppression was calculated as [OBR *Pre*] minus [OBR *During*] following both wins and losses. These were then correlated at an individual level with *Post*- win rates in the context of relearning an old strategy (Figure 4.2.a) and learning a new strategy (Figure 4.2.b). This represented a test of individual sensitivity to severe punishment (in the form of decreasing win rates to a fixed 25% in the *During* period).

There were positive correlations between win rate in *Post* bins and the degree of changed OBR from *Pre* to *During* bins both after a win and after a loss in *Old* (after a win: $r = .258$, $p = .029$; after a loss: $r = .288$, $p = .014$) and *New* conditions (after a win: $r = .312$, $p = .008$; after a loss: $r = .239$, $p = .043$). These positive correlations show that the higher degree an individual suppresses their OBR in response to the punishment mechanism of the 25%-fixed-win-rate, the

better performance they did in later bins, regardless of whether relearning old or learning new strategies. In other words, if a player successfully realized their previous strategies were less efficient and promptly changed the strategies, later they tended to achieve higher win rates. This is in partial support of H5 in that participants who were more willing to stop their strategy when it ceased being effective performed better following severe punishment (fixed 25% win rate). As in Experiment 1, this effect was similar for learning a new strategy and for relearning an old strategy. In contrast to Experiment 1, there was no evidence that participants had greater control over their behaviour following wins relative to losses.

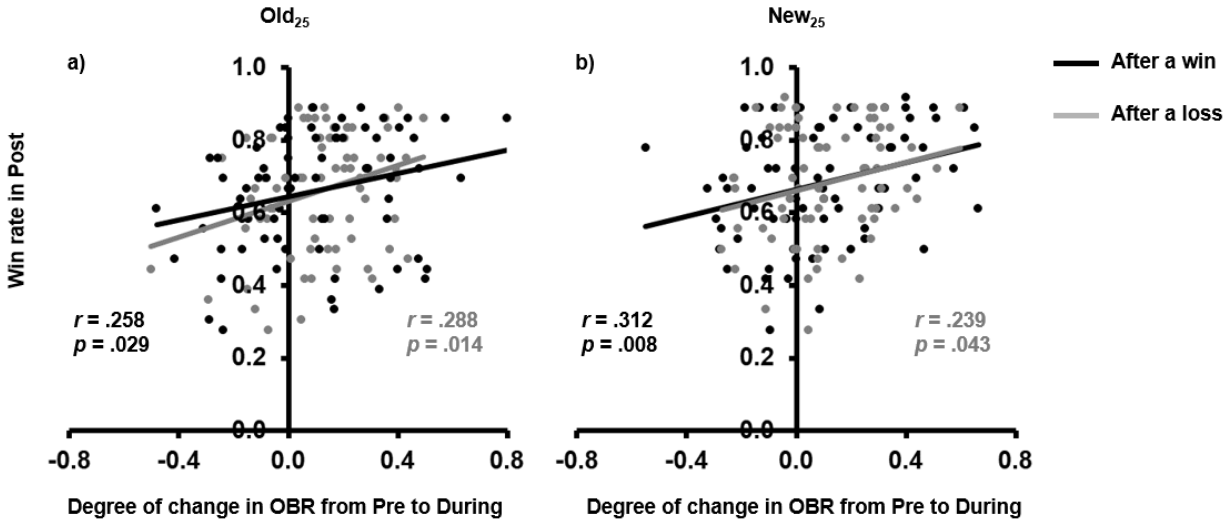

Figure 4.2 Correlation between win rate in *Post* bins and degree of change in OBR from *Pre* bins to *During* bins in 25%-fixed-win-rate condition Experiment 2. A positive OBR difference indicates lower OBR in *During* bins than in *Pre* bins. Panel a) represents OBR after a win and after a loss in the relearning-old condition; panel b) represents OBR after a win and a loss in the learning-new condition.

4.3.2 Fixed win rate at 75% in During bins

4.3.2.1 Win rate distribution

See Table 4.6 for the distribution of player win rates across the two *Learning* conditions (relearning old vs. learning new). The average win rate in *Pre* bins (M = .649, SE = .015)

collapsed across relearning-old and learning-new conditions, and was compared to chance performance (50%) via a one-sampled t-test: ($t(71) = 9.859$, $p < .001$). This again confirms H6 in that at a group level, participants learned to exploit opponent strategy during the initial stages of the game above chance. The average win rate in *Pre* bins was also significantly lower than the 75%-fixed-win-rate in *During* bins ($t(71) = -6.735$, $p < .001$). This observation confirms that exposure to the fixed win rate of 75% in the *During* bins will serve as an enhancement mechanism to learned strategies. The average win rate in *Post* bins (M = .643, SE = .018) collapsed across relearning-old and learning-new conditions, and was also compared to chance performance (50%) via a one-sampled t-test: ($t(71) = 7.912$, $p < .001$). This confirms the first half of H7 in that participants learned to exploit opponent strategy during the final stages of the game above chance at a group level, and participants were able to exploit opponents' strategies after the punishment of a 75%-win-rate in *During* bins.

(win rate)	*Pre*	*During*	*Post*
Old_{75}	0.610 (.021)	0.75 (.000)	0.694 (.022)
New_{75}	0.687 (.018)	0.75 (.000)	0.592 (.021)

Table 4.6 Descriptive statistics of win rate in Experiments 2. Old: relearning old strategy; New: learning new strategy. $_{75}$: fixed 75% win rate in *During* bins. Note: Standard error in parenthesis.

Table 4.7 shows the inferential statistics of a two-way repeated measures ANOVA with *Learning* (Old, New) and *Period* (*Pre*, *Post* only; *During* bins were excluded in this analysis for the win rate in them was fixed at 75%). The result of the significant two-way interaction between *Learning* x *Period* suggested that in relearning-old-strategy condition, the *Post* bins win rate (M = .694, SE = .022) was significantly higher than the *Pre* bins win rate (M = .610, SE = .021; Tukey's HSD; $p = .001$), while the opposite result was observed in learning-new-strategy

condition that *Post* win rate (M = .592, SE = .021) was significantly lower than *Pre* win rate (M = .687, SE = .018; Tukey's HSD; $p < .001$). Therefore, supporting H7, it was easier for participants to reacquire an old strategy, as opposed to adapting to a new strategy, following a period of enhancement. The unexpected difference between the relearning-old *Pre* win rate and learning-new *Pre* win rate was attributed to be out of chance because the sequence of blocks was counterbalanced and which version participants were in was randomized.

	df	*F*	*MSE*	*p*	η_p^2
Learning (L)	1,71	.547	.0220	.462	.008
Period (P)	1,71	.107	.0210	.744	.002
L x P	1,71	**34.108**	**.0168**	**<.001**	**.325**

Table 4.7 Inferential statistics for win rate with the fixed win rate at 75% in *During* bins in Experiment 2. Two-way repeated measures ANOVA results. *Learning (L)*: 2 levels (Old / New); *Period (P)*: 2 levels (*Pre* / *Post*). Note: Significant effects in bold font.

4.3.2.2 Optimal behaviour rate (OBR)

The rate of conducting optimal behaviours is defined in the same way as in Experiment 1. Table 4.8 shows the distribution of player OBR across the two *Learning* conditions and the two *Outcome* conditions.

(optimal behaviour rate)		*Pre*	*During*	*Post*
After a win	Old_{75}	0.641 (.033)	0.639 (.039)	0.735 (.038)
	New_{75}	0.766 (.024)	0.724 (.030)	0.596 (.035)
After a loss	Old_{75}	0.641 (.028)	0.657 (.032)	0.716 (.029)
	New_{75}	0.693 (.024)	0.656 (.032)	0.618 (.024)

Table 4.8 Descriptive statistics of optimal behaviour rate (OBR) after a win and after a loss in Experiment 1. Old: relearning old strategy; New: learning new strategy. $_{75}$: fixed 75% win rate in *During* bins. Note: Standard error in parenthesis.

OBR was again analyzed using a three-way repeated measures ANOVA with *Learning* (Old, New), *Period* (*Pre*, *During*, *Post*), and *Outcome* (after a win, after a loss) entered as factors (see Table 4.9 and Figure 4.3). In contrast to previous analyses with suppression (50% in Experiment 1 and 25% in Experiment 2), there was no significant main effect (p's > .05). However, there was a significant two-way interaction between *Learning* x *Period* ($p < .001$; 2 x 3 interaction; see Figure 2.3). In the interaction, participants played more optimal behaviours in *Post* bins in the relearning-old condition (M = .725, SE = .026) than in the learning-new condition (M = .601, SE = .025; Tukey's HSD; $p < .001$). In consistent with the difference in win rate, there was an unexpected significant difference between OBRs in *Pre* in learning-new (M = .729, SE = .020) and relearning-old conditions (M = .641, SE = .023; Tukey's HSD; $p = .009$. Despite that difference, there was a tendency that the OBR, in sequence of *Pre-During-Post*, dropped in learning-new condition while increased in relearning-old condition. This observation suggests that participants were better at performing behaviour related to the retrieval of an old strategy than the learning of a new one following the fixed 75% win rate in *During* bins. These OBR data are consistent with the two-way interaction between *Learning* x *Period* observed in win rates observed for post-relearning old trials and thus consistent with H8.

	df	*F*	*MSE*	*p*	η_p^2
Learning (L)	1,71	.027	.1206	.870	.000
Period (P)	2,142	.646	.0472	.526	.009
Outcome (O)	1,71	1.335	.0650	.252	.018
L x P	2,142	**17.279**	**.0491**	**<.001**	**.196**
L x O	1,71	1.041	.0806	.311	.014
P x O	2,142	.775	.0336	.462	.011
L x P x O	2,142	2.856	.0305	.061	.039

Table 4.9 Inferential statistics for OBR in 75% condition Experiment 2. Three-way repeated measures ANOVA results. *Learning (L)*: 2 levels (Old / New); *Period (P)*: 3 levels (*Pre* / *During* / *Post*); *Outcome (O)*: 2 levels (Win / Loss). Note: Significant effects in bold font.

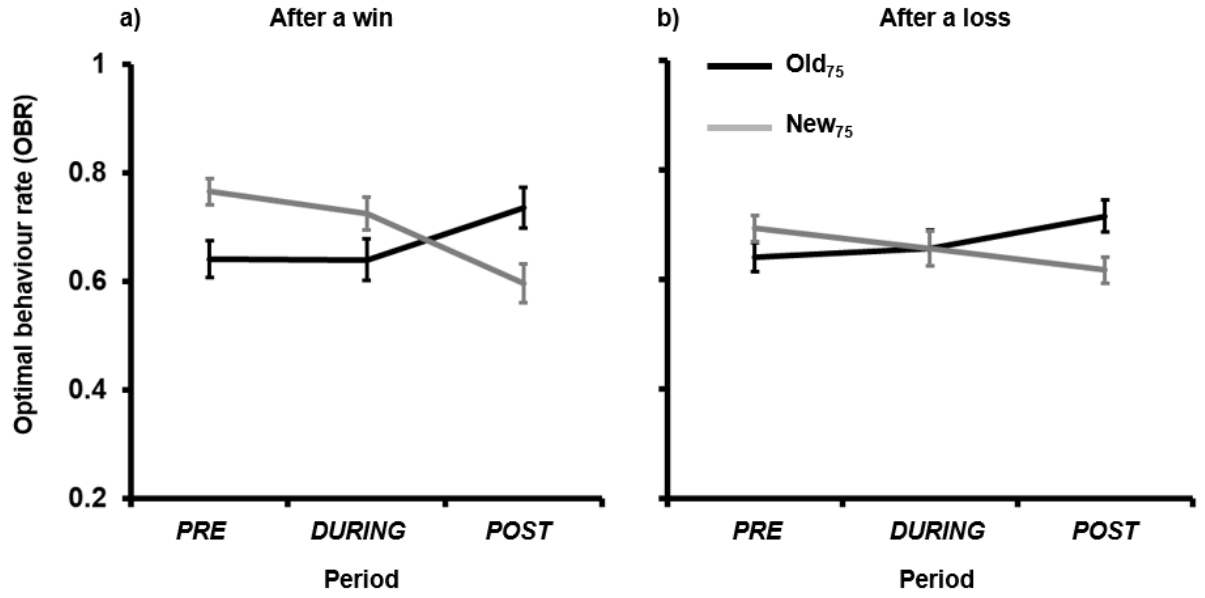

Figure 4.3 Distributions of participants' OBR in the two *Learning* conditions (relearning old strategy vs. learning new strategy) and the two *Outcome* conditions (after a win vs. after a loss) collapsed across the *Exploitable* conditions (*exploitable via repetitio*n & *exploitable via alternation*) in the 75%-fixed-win-rate condition in Experiment 2. Panel a) represents OBR after a win; panel b) represents OBR after a loss. Note: Error bars represent standard errors.

4.3.2.3 Strategy enhancement

As in Experiment 1, the degree of OBR suppression was calculated as [OBR *Pre*] minus [OBR *During*] following both wins and losses. These were then correlated at an individual level

with *Post*- win rates in the context of relearning an old strategy (Figure 4.4.a) and learning a new strategy (Figure 4.4.b). This represented a test of individual sensitivity to reinforcement (in the form of increasing win rates to a fixed 75% in the *During* period).

The change in the OBR in *During* bins was negatively correlated with the win rate in *Post* bins in the relearning-old condition, both following a win (r = -.372, p < .001) and following a loss (r = -.351, p = .003). In contrast, in the learning-new condition, the change in the OBR after a win in *During* bins was positively correlated with the win rate in the *Post* bins (r = .240, p = .042). This supports H10 that participants who were more willing to continue their strategy when it was more effective performed better following enhancement (fixed 75% win rate) when relearning the enhanced old strategy but performed worth when learning a new strategy. Additionally, the significant correlation after wins but not losses in the learning-new condition partially supports H9 that participants were able to exhibit more control over behaviour following positive relative to negative outcomes.

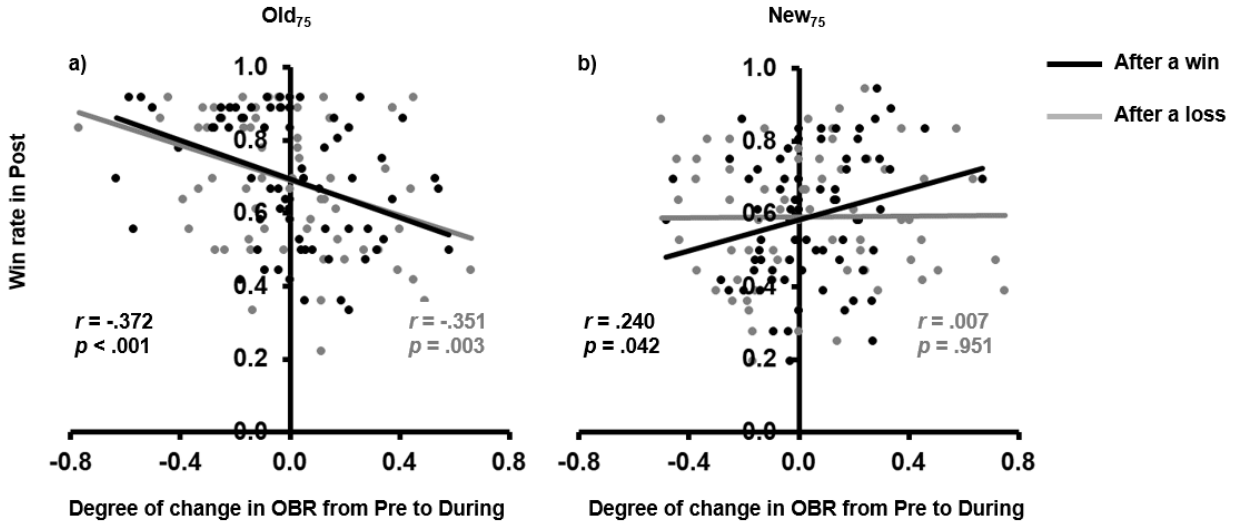

Figure 4.4 Correlation between win rate in *Post* bins and degree of change in OBR from *Pre* bins to *During* bins in 75%-fixed-win-rate condition Experiment 2. A positive OBR difference indicates lower OBR in *During* bins than in *Pre* bins. Panel a) represents OBR after a win and

after a loss in the relearning-old condition; panel b) represents OBR after a win and a loss in learning-new condition.

4.4 Discussion

Experiment 2, built on the base of Experiment 1, presents further insights into mental model acquisition and reacquisition as a function of punishment and reinforcement. The above chance *Pre* win rates in both punishment and reinforcement conditions suggested initial mental model acquisition. The OBR in punishment condition was in accordance with the win rate that it was suppressed during severe punishment (win rate dropped to fixed 25%), but not in reinforcement condition, which suggested that the fixed 75% win rate (around 10% higher than the average) was merely a light reinforcement.

The severe punishment (fixed 25% condition) in Experiment 2 facilitated the acquisition of a new mental model while disturbed the reinstatement of an old mental model more, compared to the mild punishment (fixed 50%) in Experiment 1. However, the above-chance *Post* OBR in the relearning-old condition showed that old mental models were temporarily suppressed but not permanently thrown away. The light reinforcement (fixed 75% condition), on the other hand, functioned the other way around that new model acquisition would be suppressed while old model reacquisition would be enhanced. The degree of OBR change was again a significant predictor of both retrieving an old mental model and building up a new mental model: participants who decreased OBR more from *Pre* to *During* bins did better in mental model acquisition and reacquisition following punishment, but worse in mental model reacquisition following reinforcement. This suggests that if a participant learned to adapt to a new situation better and faster, they may have achieved more wins in later trials. Mental model acquisition after reinforcement was better predicted by OBR change following positive compared to

negative outcomes, which may have again suggested a larger degree of behavioural flexibility following wins relative to losses.

Experiment 2 showed that punishment and reinforcement of old mental models influenced both reacquiring old models and acquiring new models. Unlike mild punishment in Experiment 1, severe punishment in Experiment 2 frustrated the advantage of old mental model reacquisition over new mental model acquisition, while light reinforcement would encourage the former and discourage the latter to an observable degree.

Chapter 5

Experiment 3

5.1 Introduction

One issue in the interpretation of Experiment 2 was that the degrees of punishment (25%) and reinforcement (75%) were unbalanced relative to an initial mental model acquisition (win) rate of 66%. Specifically, and on average, performance was only reinforced in the *During* bins by 9% but punished by 41%. Effects of losses should be stronger than effects of wins (loss aversion; Tversky & Kahneman, 1991); however, the effect in Experiment 2 could be due to the unbalanced win rate manipulation, rather than any fundamental difference between punishment and reinforcement. Therefore, Experiment 3 attempted to equate the degrees of punishment and reinforcement. Since the average *Pre* win rate in Experiments 1 and 2 was around 66%, the degree of punishment in *During* bins was set to a 44% win rate, and the degree of reinforcement in *During* bins was set to an 88% win rate. This was in the hope of providing a more balanced and comprehensive conclusion as to the effects of punishment and reinforcement in Experiment 3, presuming a *Pre* win rate across all participants of around 66%.

5.1.1 Hypotheses

The hypotheses in Experiment 3 are consistent with Experiment 2. The fixed 44% win rate is assumed to be a punishment for old strategies which is hypothesized to suppress old-strategy-relearning while enhance new-strategy-learning. The fixed 88% win rate is assumed to

be a reinforcement for old strategies which is hypothesized to enhance old-strategy-relearning while suppress new-strategy-learning.

5.1.1.1 Predictions

1) Participants should learn to exploit the opponent's strategy during the initial stages of the game (*Pre* bins), leading to above-chance performance (50%). The average win rate in *Pre* bins should not differ from 66%.
2) Participants should learn to exploit the opponent's strategy again during the final stages of the game (*Post* bins). With suppression of an old strategy in the *During* bins (fixed 44% win rate), performance should be better for the acquisition of a new strategy.
3) OBR should be strongly linked to the results of the win rate performance.
4) Participants should play more optimal behaviours following wins than losses.
5) The degree of suppression of optimal behaviours exhibited between *Pre* and *During* bins should be positively correlated with the individual ability to learn a new strategy, but negatively correlated with the individual ability to relearn an old strategy.
6) Participants should learn to exploit the opponent's strategy during the initial stages of the game (*Pre* bins), leading to above-chance performance (50%). The average win rate in *Pre* bins should not differ from 66%.
7) Participants should learn to exploit the opponent's strategy again during the final stages of the game (*Post* bins). With the stronger enhancement of an old strategy in the *During* bins (fixed 88% win rate), performance should be better for reacquisition of the old strategy.
8) OBR should be strongly linked to the results of the win rate performance.

9) Participants should play more optimal behaviours following wins than losses.
10) The degree of enhancement in optimal behaviours exhibited between *Pre* and *During* bins should be positively correlated with the individual ability to relearn an old strategy, but negatively correlated with the individual ability to learn a new strategy.

5.2 Method

5.2.1 Participants

Data from a convenience sample of 70 participants (Mean = 19.36, SD = 2.37, 33 female, 54 right-handed) from the student population at the University of Alberta were analyzed. Three participants were excluded from analyses according to criterion 1), and a further 7 participants were excluded according to Criterion 2. Criteria were described in the Method section of Experiment 1. Participants received course credit, and the protocol was approved at the University of Alberta under Research Ethics Board 2 (Pro00120832).

5.2.2 Stimuli

The stimuli and experimental set-up were identical to those of Experiments 1 and 2.

5.2.3 Design

Being generally identical to Experiment 2, the current experiment was a 2x2x2 within-subject design with two aimed strategies (*exploitable via repetition, exploitable via alternation*), two kinds of fixed win rates (44%, 88% here, instead of 25% and 75% in Experiment 2) in *During* bins, and two learning contexts (learning a new strategy, relearning the old strategy) as factors. The rest design was the same as Experiments 1 and 2 (see Table 5.1).

		Pre bins		*During* bins		*Post* bins	
[Counterbalanced version 1]	Bin 1	Bin 2	Bin 3	Bin 4	Bin 5	Bin 6	Bin 7
Relearning Old	M	R	R	F_{44}	F_{44}	R	R
	M	A	A	F_{88}	F_{88}	A	A
Learning New	M	R	R	F_{88}	F_{88}	A	A
	M	A	A	F_{44}	F_{44}	R	R

		Pre bins		*During* bins		*Post* bins	
[Counterbalanced version 2]	Bin 1	Bin 2	Bin 3	Bin 4	Bin 5	Bin 6	Bin 7
Relearning Old	M	R	R	F_{88}	F_{88}	R	R
	M	A	A	F_{44}	F_{44}	A	A
Learning New	M	R	R	F_{44}	F_{44}	A	A
	M	A	A	F_{88}	F_{88}	R	R

Table 5.1 Schematic depicting two counterbalanced versions of all four blocks of relearning old strategy and learning new strategy conditions in Experiment 3. Two unexploitable strategies: 1) M: computerized opponent using mixed-strategy; 2) F_{44}: participants' win rate was fixed-44%, F_{88}: participants' win rate was fixed-88%. Two exploitable strategies: 1) R: computerized opponent repetitively choosing odd/even numbers; 2) A: computerized opponent alternatingly choosing odd and even numbers. Each row represents one block, and each bin represents 18 trials. The sequence of the four blocks was counterbalanced.

5.2.4 Procedure

The procedure went was identical to that of Experiments 1 and 2.

5.3 Results

5.3.1 Fixed win rate at 44% in During bins

5.3.1.1 Win rate distribution

The win rate is calculated as in Experiments 1 and 2. See Table 5.2 for the distribution of player win rates across the two *Learning* conditions (relearning old vs. learning new). As in Experiments 1 and 2, the average win rate in *Pre* bins (M = .657, SE = .013) collapsed across relearning-old and learning-new conditions, was compared to chance performance (50%) via a

one-sampled t-test: ($t(69) = 11.788$, $p < .001$), and to the expected (66%) via a one-sample t-test: ($t(69) = -.229$, $p = .819$). This confirms H1 in that at a group level, participants learned to exploit opponent strategy during the initial stages of the game above chance (50%), and further confirms the degrees of punishment (44%) and reinforcement (88%) in this experiment are numerically equal, relate to the average *Pre* win rate of 65.7%.

The average win rate in *Post* bins (M = .661, SE = .019) collapsed across relearning-old and learning-new conditions, and was also compared to chance performance (50%) via a one-sampled t-test: ($t(69) = 8.306$, $p < .001$). Like Experiments 1 and 2, this confirms the first half of H2 in that participants learned to exploit opponent strategy during the final stages of the game above chance at a group level, and participants were able to exploit opponents' strategies after the punishment of 44%-win-rate in *During* bins.

(win rate)	*Pre*	*During*	*Post*
Old_{44}	0.662 (.021)	0.44 (.000)	0.682 (.022)
New_{44}	0.652 (.019)	0.44 (.000)	0.639 (.022)

Table 5.2 Descriptive statistics of win rate in Experiments 3. Old: relearning old strategy; New: learning new strategy. $_{44}$: fixed 44% win rate in *During* bins. Note: Standard error in parenthesis.

Replicating Experiment 2 but in contrast to Experiment 1, in the 44% *During* condition (Table 5.3), there were no significant main effects or interaction for win rates (all *p*s > .05). This suggests that the advantage of relearning an old strategy following reduction to 50% win rate (Experiment 1) is abolished when win rate is reduced to 44% (Experiment 3), similar to 25% (Experiment 2).

	df	*F*	*MSE*	*p*	η_p^2
Learning (L)	1,69	2.460	.0204	.121	.034
Period (P)	1,69	.060	.0165	.807	.001
L x P	1,69	.786	.0230	.379	.011

Table 5.3 Inferential statistics for win rate with the fixed win rate at 44% in *During* bins in Experiment 3. Two-way repeated measures ANOVA results. *Learning (L)*: 2 levels (Old / New); *Period (P)*: 2 levels (*Pre* / *Post*). Note: Significant effects in bold font.

5.3.1.2 Optimal behaviour rate (OBR)

The rate of conducting optimal behaviours was defined in the same way as in Experiments 1 and 2. Table 5.4 shows the distribution of player OBR across the two *Learning* conditions, two *Fixed-win-rate* conditions, and the two *Outcome* conditions.

(optimal behaviour rate)		*Pre*	*During*	*Post*
After a win	Old_{44}	0.688 (.033)	0.571 (.033)	0.721 (.035)
	New_{44}	0.714 (.029)	0.621 (.029)	0.681 (.034)
After a loss	Old_{44}	0.680 (.025)	0.565 (.022)	0.702 (.028)
	New_{44}	0.664 (.026)	0.531 (.023)	0.658 (.025)

Table 5.4 Descriptive statistics of optimal behaviour rate (OBR) after a win and after a loss in Experiment 3. Old: relearning old strategy; New: learning new strategy. $_{44}$: fixed 44% win rate in *During* bins. Note: Standard error in parenthesis.

OBRs were analyzed as in Experiments 1 and 2, using a three-way repeated measures ANOVA with *Learning* (Old, New), *Period* (*Pre*, *During*, *Post*), and *Outcome* (after a win, after a loss) entered as factors (see Table 5.5 and Figure 5.1). Replicating Experiment 2, there were significant main effects of *Period* ($p < .001$) as well as *Outcome* ($p = .047$), but no significant interactions ($ps > .05$).

For the main effect of *Outcome*, OBR was higher following wins (M = .666, SE = .020) than losses (M = .633, SE = .012), replicating Experiments 1 and 2 and supporting H4.

For the main effect of *Period*, OBR was the lowest in *During* bins (M = .572, SE = .012), compared to *Pre* bins (M = .687, SE = .017) and *Post* bins (M = .691, SE = .023), replicating Experiments 1 and 2.

	df	*F*	*MSE*	*p*	η_p^2
Learning (L)	1,69	.225	.0843	.637	.003
Period (P)	2,138	**27.734**	**.0457**	**<.001**	**.287**
Outcome (O)	1,69	**4.087**	**.0551**	**.047**	**.056**
L x P	2,138	1.005	.0544	.369	.014
L x O	1,69	1.485	.0647	.227	.021
P x O	2,138	.466	.0288	.628	.007
L x P x O	2,138	.853	.0326	.428	.012

Table 5.5 Inferential statistics for OBR in 44% condition Experiment 3. Three-way repeated measures ANOVA results. *Learning (L)*: 2 levels (Old / New); *Period (P)*: 3 levels (*Pre* / *During* / *Post*); *Outcome (O)*: 2 levels (Win / Loss). Note: Significant effects in bold font.

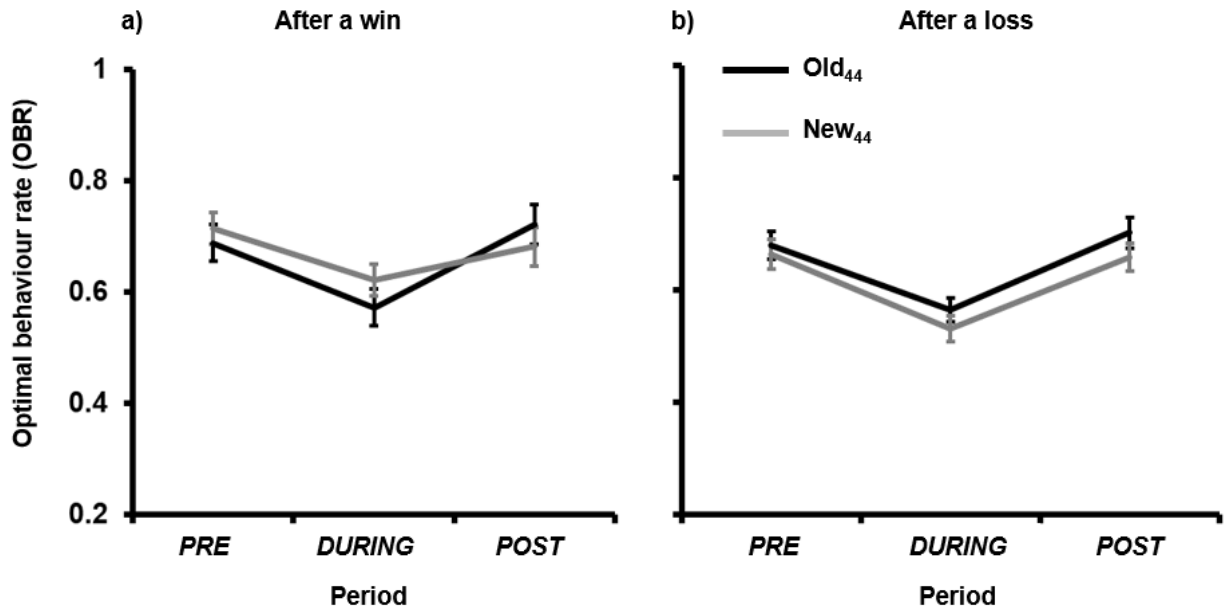

Figure 5.1 Distributions of participants' OBR in the two *Learning* conditions (relearning old strategy vs. learning new strategy) and the two *Outcome* conditions (after a win vs. after a loss) collapsed across the *Exploitable* conditions (*exploitable via repetitio*n & *exploitable via alternation*) in the 44%-fixed-win-rate condition in Experiment 3. Panel a) represents OBR after a win; panel b) represents OBR after a loss. Note: Error bars represent standard errors.

5.3.1.3 Strategy suppression

As in Experiments 1 and 2, the degree of OBR suppression was calculated as [OBR *Pre*] minus [OBR *During*] following both wins and losses. These were then correlated at an individual level with *Post* win rates in the context of relearning an old strategy (Figure 5.2.a) and learning a new strategy (Figure 5.2.b). This represented a test of individual sensitivity to punishment (in the form of decreasing win rates from around 66% to a fixed 44% in the *During* period).

No significant correlation between the *Post* win rate and the degree of changed OBR from *Pre* to *During* bins was shown in *Old* condition (after a win: $r = .029$, $p = .811$; after a loss: $r = .207$, $p = .085$). However, there was a positive correlation in *New* after a win ($r = .288$, $p = .016$) not after a loss ($r = .071$, $p = .559$). The positive correlation shows that the higher degree

an individual suppresses their OBR in response to the punishment mechanism of the 44%-fixed-win-rate, the better performance they did in later bins, only happening in behaviours following a win when learning a new strategy. This is in partial support of H5 in that participants who were more willing to stop their strategy when it ceased being effective performed better following punishment (fixed 44% win rate). As in Experiment 1, participants had greater control over their behaviour following wins relative to losses. However, unlike Experiments 1 and 2, this effect only worked for learning a new strategy.

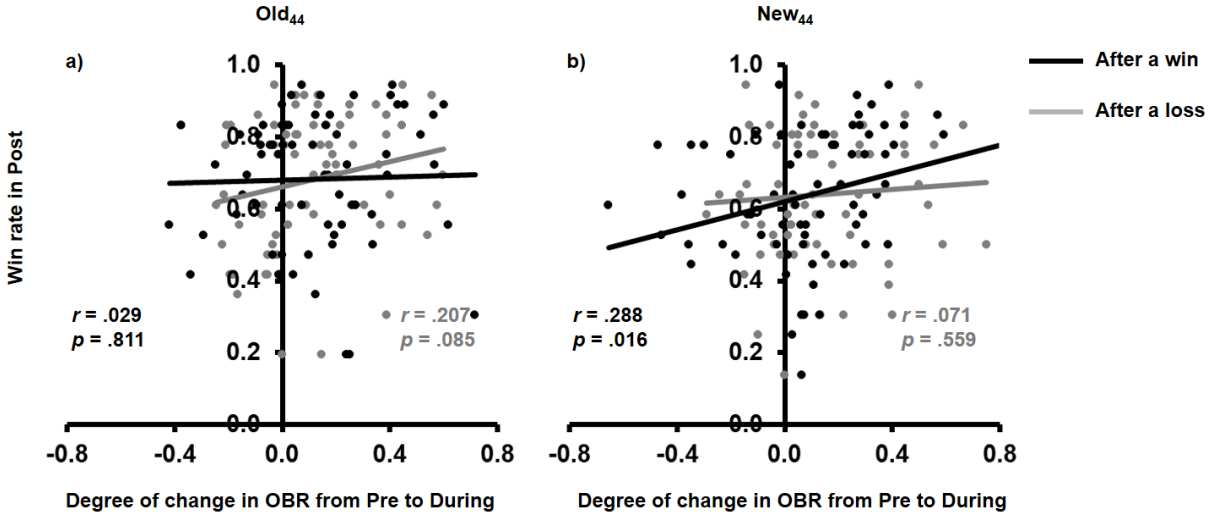

Figure 5.2 Correlation between win rate in *Post* bins and degree of change in OBR from *Pre* bins to *During* bins in 44%-fixed-win-rate condition Experiment 3. A positive OBR difference indicates lower OBR in *During* bins than in *Pre* bins. Panel a) represents OBR after a win and after a loss in the relearning-old condition; panel b) represents OBR after a win and a loss in the learning-new condition.

5.3.2 Fixed win rate at 88% in During bins

5.3.2.1 Win rate distribution

See Table 5.6 for the distribution of player win rates across the two *Learning* conditions (relearning old vs. learning new). The average win rate in *Pre* bins (M = .655, SE = .017)

collapsed across relearning-old and learning-new conditions, and was compared to chance performance (50%) via a one-sampled t-test: ($t(69) = 9.033$, $p < .001$). This again confirms H6 in that at a group level, participants learned to exploit opponent strategy during the initial stages of the game above chance. The average win rate in *Pre* bins was not significantly different from expected 66% ($t(69) = -.270$, $p = .788$), as in the fixed-44% condition. This observation, in line with H6, again confirms that the degrees of reinforcement (88%) and punishment (44%) distribute equally relate to the average *Pre* win rate of 65.5%. The average win rate in *Post* bins (M = .674, SE = .018) collapsed across relearning-old and learning-new conditions, and was also compared to chance performance (50%) via a one-sampled t-test: ($t(69) = 9.858$, $p < .001$). This confirms the first half of H7 in that participants learned to exploit opponent strategy during the final stages of the game above chance at a group level, and participants were able to exploit opponents' strategies after the reinforcement of 88%-win-rate in *During* bins.

(win rate)	*Pre*	*During*	*Post*
Old_{88}	0.657 (.023)	0.88 (.000)	0.725 (.021)
New_{88}	0.654 (.020)	0.88 (.000)	0.622 (.020)

Table 5.6 Descriptive statistics of win rate in Experiments 3. Old: relearning old strategy; New: learning new strategy. $_{88}$: fixed 88% win rate in *During* bins. Note: Standard error in parenthesis.

Table 5.7 shows the inferential statistics of a two-way repeated measures ANOVA with *Learning* (Old, New) and *Period* (*Pre*, *Post* only). The significant main effect of *Learning* showed that, collapsed across *Pre* and *Post* periods, the win rate in the relearning-old condition (M = .691, SE = .020) was higher than that in the learning-new condition (M = .638, SE = .014). This result suggests a better performance in relearning already learned strategies.

Replicating the 75% condition in Experiment 2, there was a significant two-way interaction between *Learning* x *Period* suggested that in relearning-old-strategy condition, the *Post* bins win rate (M = .725, SE = .021) was significantly higher than the *Pre* bins win rate (M = .657, SE = .023; Tukey's HSD; p = .015), while no significant difference was found in learning-new-strategy condition (Tukey's HSD; p = .481). Therefore, in support of the second half of H7, participants did well in reacquiring an old strategy but not in adapting to a new strategy following a period of enhancement.

	df	*F*	*MSE*	*p*	η_p^2
Learning (L)	1,69	**10.377**	**.0188**	**.002**	**.131**
Period (P)	1,69	1.196	.0195	.278	.017
L x P	1,69	**10.240**	**.0171**	**.002**	**.129**

Table 5.7 Inferential statistics for win rate with the fixed win rate at 88% in *During* bins in Experiment 2. Two-way repeated measures ANOVA results. *Learning (L)*: 2 levels (Old / New); *Period (P)*: 2 levels (*Pre* / *Post*). Note: Significant effects in bold font.

5.3.2.2 Optimal behaviour rate (OBR)

The rate of conducting optimal behaviours is defined in the same way as in Experiments 1 and 2. Table 5.8 shows the distribution of player OBR across the two *Learning* conditions and the two *Outcome* conditions.

(optimal behaviour rate)		*Pre*	*During*	*Post*
After a win	Old_{88}	0.708 (.031)	0.749 (.035)	0.775 (.034)
	New_{88}	0.679 (.031)	0.747 (.033)	0.655 (.031)
After a loss	Old_{88}	0.666 (.027)	0.718 (.039)	0.752 (.029)
	New_{88}	0.681 (.023)	0.665 (.039)	0.618 (.024)

Table 5.8 Descriptive statistics of optimal behaviour rate (OBR) after a win and after a loss in Experiment 3. Old: relearning old strategy; New: learning new strategy. $_{88}$: fixed 88% win rate in *During* bins. Note: Standard error in parenthesis.

OBR was again analyzed using a three-way repeated measures ANOVA with *Learning* (Old, New), *Period* (*Pre*, *During*, *Post*), and *Outcome* (after a win, after a loss) entered as factors (see Table 5.9 and Figure 5.3). There was a significant main effect of *Learning* (p = .021), and, as the 75% condition in Experiment 2, there was a significant two-way interaction between *Learning* x *Period* (p = .002). The significant main effect of *Learning* and the two-way interaction in OBR is consistent with the main effect and two-way interaction observed in win rates and thus supports H8.

For the main effect of *Learning*, OBR was higher in the relearning-old condition (M = .728, SE = .024) than in the learning-new condition (M = .674, SE = .018), suggesting optimal behaviours had been enacted more when participants were relearning an old strategy.

In terms of the interaction between *Learning* and *Period* (2 x 3 interaction; see Figure 5.3), replicating 75% condition in Experiment 2, OBR was higher in *Post* bins in the relearning-old condition (M = .763, SE = .026) compared to learning-new condition (M = .636, SE = .023; Tukey's HSD; $p < .001$). This observation again suggests that participants were better at performing behaviours related to the retrieval of an old strategy than the learning of a new one. These OBR data are consistent with the two-way interaction between *Learning* x *Period* observed in win rates observed for post-relearning old trials and thus consistent with H8.

	df	*F*	*MSE*	*p*	η_p^2
Learning (L)	1,69	**5.619**	**.1087**	**.021**	**.075**
Period (P)	2,138	1.639	.0567	.198	.023
Outcome (O)	1,69	3.405	.0769	.069	.047
L x P	2,138	**6.513**	**.0444**	**.002**	**.086**
L x O	1,69	.038	.0634	.846	.001
P x O	2,138	.771	.0328	.464	.011
L x P x O	2,138	1.263	.0315	.286	.018

Table 5.9 Inferential statistics for OBR in 88% condition Experiment 3. Three-way repeated measures ANOVA results. *Learning (L)*: 2 levels (Old / New); *Period (P)*: 3 levels (*Pre* / *During* / *Post*); *Outcome (O)*: 2 levels (Win / Loss). Note: Significant effects in bold font.

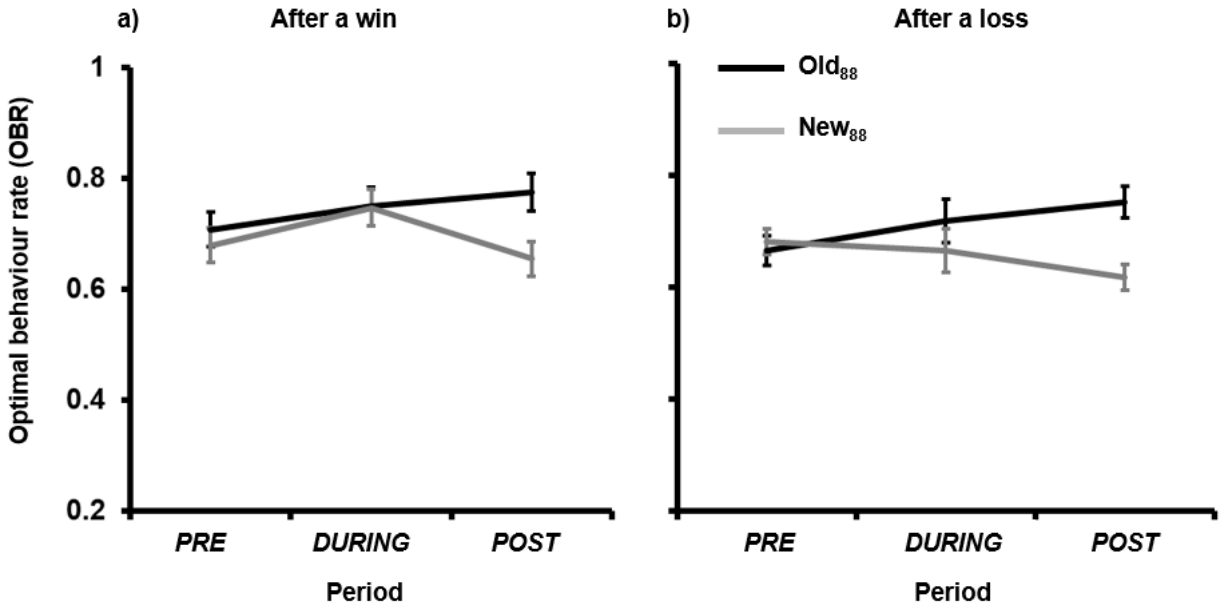

Figure 5.3 Distributions of participants' OBR in the two *Learning* conditions (relearning old strategy vs. learning new strategy) and the two *Outcome* conditions (after a win vs. after a loss) collapsed across the *Exploitable* conditions (*exploitable via repetitio*n & *exploitable via alternation*) in the 88%-fixed-win-rate condition in Experiment 3. Panel a) represents OBR after a win; panel b) represents OBR after a loss. Note: Error bars represent standard errors.

5.3.2.3 Strategy enhancement

Calculated as in Experiments 1 and 2, the change in the OBR after a win in *During* bins was negatively correlated with the win rate in *Post* bins in the relearning-old condition (r = -.268, p = .025). In contrast, there was no significant correlation in learning-new-strategy condition (ps > .05). This partially replicates 75% condition in Experiment 2 and supports H10 that participants who were more willing to continue their strategy when it was more effective performed better following enhancement (fixed 88% win rate) when relearning the enhanced old strategy. Additionally, the significant correlation after wins but not losses in the relearning-old

condition is in support of H9 that participants were able to exhibit more control over behaviour following positive relative to negative outcomes.

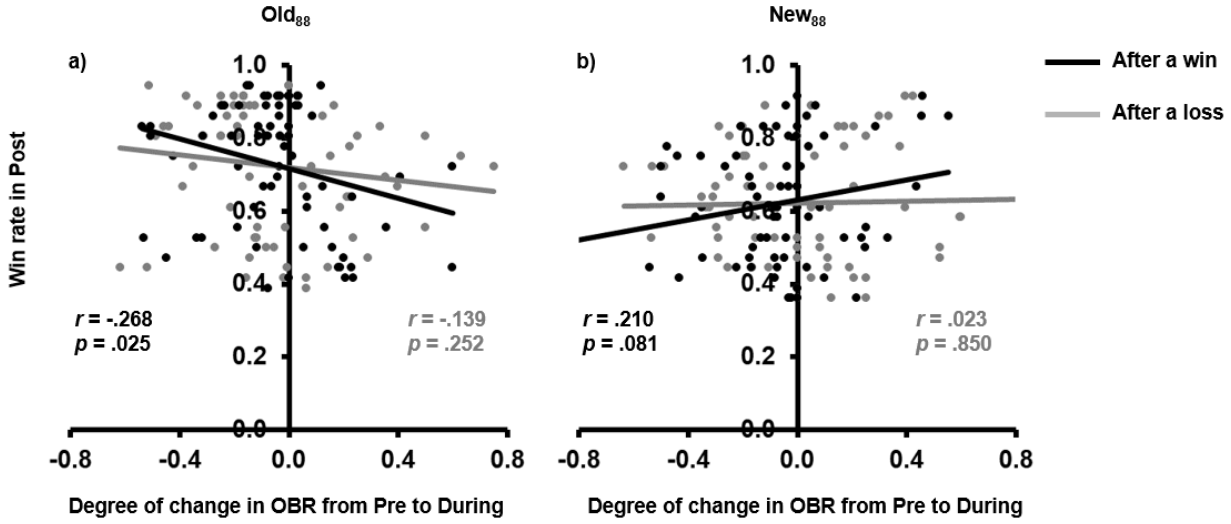

Figure 5.4 Correlation between win rate in *Post* bins and degree of change in OBR from *Pre* bins to *During* bins in 88%-fixed-win-rate condition Experiment 3. A positive OBR difference indicates lower OBR in *During* bins than in *Pre* bins. Panel a) represents OBR after a win and after a loss in the relearning-old condition; panel b) represents OBR after a win and a loss in the learning-new condition.

5.4 Discussion

Experiment 3 established a framework for mental model acquisition and reacquisition with a numerically equal distribution of punishment (44%) and reinforcement (88%), related to a baseline win rate of 66%. The results were mostly consistent with Experiment 2 that punishment suppressed old mental model reacquisition and enhanced new mental model acquisition, while reinforcement did the opposite way. In other words, the fixed win rate of 44% (Experiment 3) affected performance and behaviours similarly to 25% (Experiment 2), and 88% (Experiment 3) was similar to 75% (Experiment 2).

Chapter 6

General discussion

6.1 Summary of experiments.

How various degrees of punishment and reinforcement impacted on the development of new mental models or the maintenance of old mental model was examined across three experiments using a simple zero-sum competitive game. In Experiments 1-3, participants played the *Dice Dual* game against a computer opponent using two types of exploitable strategies, aiming to allow participants to adapt different optimal behaviours (*win-stay* / *lose-shift*, or, *win-shift* / *lose-stay*). In all three experiments, each game was separated into three periods. In the *Pre* bins, they learned a strategy; in the *During* bins, win rate was fixed to simulate punishment or reinforcement; in the *Post* bins, they either relearned the *Pre* strategy or learned a new one. The three experiments varied at the fixed win rate in *During* bins: 50% in Experiment 1, 25% and 75% in Experiment 2, and 44% and 88% in Experiment 3. Table 6.1 and Table 6.2 compare the main results across the three experiments.

In Experiment 1, the fixed win rate of 50% in *During* bins was interpreted as a mild punishment to previously learned strategy because participants' average *Pre* win rate was 63.7%. Following the fixed 50% win rate punishment to the *Pre* strategy, participants still did better in relearning the old strategy compared to learning a new strategy, supported by higher *Post* win rate and larger *Post* optimal behaviour rate (OBR) percentages (see Table 6.1). The changes of OBR from *Pre* to *During* bins were used to examine how flexible and adaptive participants were

when exposed to fixed win rates (see Table 6.2). The more participants suppressed their previous optimal behaviour, the better they did in learning a new strategy *and* relearning an old strategy. The OBR following wins were robustly higher than that following losses, indicating the former being more reliable. There was also higher flexibility in OBR following wins relative to losses because more variations were shown in the former from *Pre* to *Post* bins.

Experiment 2 attempted to increase the degree of punishment (from 50% in Experiment 1 to 25%) and to introduce an enhanced win rate condition (75%) to simulate the experience of reinforcement. The fixed win rate of 25% represented a severe punishment to the previously learned strategy because the participants' average *Pre* win rate was 65%, while the fixed 75% win rate was a light reinforcement. The severe punishment of a learned strategy abolished the advantage for old strategy as seen in Experiment 1, as participants performed similarly well during the *Post* phase when deploying both old and new strategies. These were again supported by the *Post* win rate and *Post* OBR percentages (see Table 6.1). Replicating the results of Experiment 1, the more participants suppressed OBR when the old strategy was punished, the better they did in learning new and in relearning old strategies. A novel contribution in Experiment 2 was the study of light reinforcement to a learned strategy by using a fixed 75% win rate in a separate condition. Here, participants did better in relearning the old strategy compared to learning a new strategy. In examining individual levels of strategy suppression (see Table 6.2), during reinforcement (75% win rate; *During*) the more participants (erroneously) suppressed that strategy, the worse they did in relearning the old strategy but the better they did in learning a new strategy. Consistent with the data from Experiment 1, mental model acquisition was better expressed following wins than following losses in the punishment (25% win rate) condition, but not in the novel reinforcement (75% win rate) condition.

Experiment 3 attempted to equate the degree of punishment and reinforcement by setting the fixed win rate in *During* bins at 44% and 88%, respectively. As hoped, the current punishment and reinforcement were objectively equated because participants' average *Pre* win rates were both 66%. Following the punishment to a learned strategy (44% win rate; *During*), participants again did similarly in relearning the old strategy and learning a new strategy, replicating Experiment 2. Following the reinforcement of a learned strategy (88% win rate; *During*), participants did better in relearning the old strategy compared to learning a new strategy, again replicating Experiment 2. As in all previous experiments, these conclusions were supported by both the *Post* win rate and *Post* OBR percentages (see Table 6.1). As in Experiment 2, the more participants suppressed their optimal behaviours when the old strategy was punished, the better they did in mental model new-acquisition *and* old-reacquisition. However, when the old strategy was reinforced, the more they suppressed, the worse they did in relearning while no effect in learning-new (see Table 6.2). Finally, and as in Experiment 2, mental model acquisition was better expressed following wins than following losses within punishment (44% win rate) but not reinforcement (88% win rate) conditions.

	Avg *Pre* Win Rate	Avg *Post* Win Rate	*Post* Win Rate	OBR Old / New	OBR Win / Lose
Experiment 1					
During 50	63.7% *	63.8% *	Old > New	Old > New	Win > Lose
Experiment 2					
During 25	64.8% *	67.1% *	Old = New	Old = New	Win > Lose
During 75	64.9% *	64.3% *	Old > New	Old > New	Win = Lose
Experiment 3					
During 44	65.7% *	66.1% *	Old = New	Old = New	Win > Lose
During 88	65.5% *	67.4% *	Old > New	Old > New	Win = Lose

Table 6.1 Main results across three experiments. Note: * = significant from 50%.

	Relearn-Old Strategy		Learn-New Strategy	
	Following a Win	Following a Loss	Following a Win	Following a Loss
Experiment 1				
During 50	+	×	+	×
Experiment 2				
During 25	+	+	+	+
During 75	−	−	+	×
Experiment 3				
During 44	×	×	+	×
During 88	−	×	×	×

Table 6.2 Correlation results across three experiments. A positive correlation between the changed OBR from Pre to During and Post performance indicates the more optimal behaviours suppressed in During bins the better performance in Post bins. Note: + = significant positive correlation, − = significant negative correlation, × = no significant correlation.

6.2 Mental model acquisition and reacquisition following punishment and reinforcement

In summary, old mental model reacquisition has an advantage over new model acquisition under conditions of mild punishment (fixed win rate of 50%; Experiment 1) and reinforcement (fixed win rates of 75% and 88%; Experiments 2 and 3). In these cases, win rate percentages in *Post* bins were higher for old models relative to new models, and relatedly, *Post* bins optimal behaviour rates (OBR) were higher for old models relative to new models. However, this advantage was gone under conditions of more severe punishment (fixed win rate of 25% and 44%; Experiments 2 and 3). In these cases, win rate percentages and OBRs in *Post* bins were equivalent between old and new models (for comparisons among reinforcement and punishment conditions, see Appendix A).

In all conditions, the *Post* win rate percentages and OBRs of old mental model reacquisition were never lower than those of new model acquisition. One explanation is that old models are triggered by familiar stimuli, which drives people away from the exploration of new models (*Einstellung effect;* Bilalić, Mcleod, & Gobet, 2008). Also, when known information occurs repeatedly, people may weigh the known information more than new information. Therefore, as a default position, individuals first reuse their learned strategies in the absence of any radical change in feedback (*self-reinforcement*; Wheeler, Bolton, & Sanquist, 1990).

However, once the nature of feedback reliably shifts into punishment, the initial benefit associated with the reuse of old information is lost. But what constitutes ‘reliable’ punishment? A mild punishment (50% from 66% baseline; -16% change), although detectable, does not influence people to the degree that they will abandon an old mental model in favour of a new

one, while a mild reinforcement (75% from 66% baseline; +9% change) allows participants to use the old model more continually. This can possibly be explained by *confirmation bias* in people believe in what they have learned and tend to overweight confirmatory evidence (Nickerson, 1998). As such, the degree of contrary evidence has to be greater than the degree of confirmatory evidence to make participants change their mental model. Although the 'mild' punishment (Experiment 1) was approximately twice the size of the 'light' reinforcement (Experiment 2), this percentage change may still not have been enough. Alternatively, the belief that failure is borne out of external chance but success results from internal control (*self-attribution bias*; Feather & Simon, 1971) may also further encourage people to stick to their current strategy. Third, people battle against the pressures of cognitive inertia (folks just carrying on doing the same thing; Messner & Vosgerau, 2010). If the punishment becomes more severe, people reckon the old model to be wrong and are more willing to learn a new model, but learning new knowledge takes time and effort and does not outperform relearning old knowledge. If the reinforcement of the prior strategy becomes stronger, on the other hand, people become faithful to the old model, which seriously frustrates new model learning. Fourth and finally, the 50% win rate could represent an objective 'chance' performance rather than the explicit punishment of previously learned information. As such, this may excuse the participant into thinking that there may be- in fact- *nothing* to learn in the current game environment. This is in contrast to our original intention that the reduction in win rate from 66% to 50% would index the failure of the current mental model. In the following reacquisition period, they may not be confident enough to retrieve their old mental model or deploy a newer mental model, as they may have believed the game was operating at chance thereby leading to a slow learning process. Therefore, although disliking losses (*loss aversion*; Tversky & Kahneman, 1991), people are still likely to retrieve old

mental models even if they are punished. Yet once behaviours appear to be correct, people are unlikely to shift away from their current thoughts and actions.

The correlation analyses (see Table 6.2) reveal a relationship between the flexibility in behaviour change and later performances. In punishment conditions (50%, 25%, and 44%), results are generally consistent and show that participants who successfully suppressed *During* OBR when being punished had higher *Post* win rates than those who did not suppress, regardless of mental model reacquisition or acquisition. In reinforcement conditions (75% and 88%), however, participants who wrongly suppressed *During* OBR when being reinforced did worse in old model reacquisition but better in new model acquisition (significant correlation in 75% but not 88% condition). The results suggest that when the old strategy is punished, because of the easiness of old mental model reacquisition, people who try other strategies perform well in both reusing and reconfiguring mental models. On the other hand, when the old strategy is reinforced, people who successfully detect the change and enhance the old strategy do well in old-relearning but fall short of new-learning. Therefore, the degree of reinforcement of old mental models might have a restriction that needs to be limited to a certain degree; otherwise, new mental model acquisition will be frustrated.

6.3 Behaviours following wins and losses

Another major finding is that behaviours after wins are generally more reliable and flexible relative to losses. In other words, the proportion of optimal behaviours after wins are robustly higher, better predict *Post* performance, and show more variation than optimal behaviours after losses. The advantage of this heightened cognitive state following wins showed in the current experiments is in line with previously examined post-win flexibility (e.g., Dyson, 2023; Forder & Dyson, 2016; Dixon, Larche, Stange, Graydon & Fugelsang, 2018). Such

discrepancy between higher-quality behaviours following positive outcomes and lower-quality behaviours following negative outcomes is probably due to a *lose-shift* inertia, which states an automatic tendency to choose other choices if people lose, regardless of what optimal behaviour should be used (Dyson, Wilbiks, Sandhu, Papanicolaou, & Lintag, 2016; Dyson, 2023; Alós-Ferrer, Hügelschäfer, & Li, 2016). In current studies, behaviours following losses, regardless of staying or shifting, appear more inflexible than behaviours following wins.

6.4 Contributions

Methodologically, the current thesis establishes a successful paradigm of using a simple game to test mental models, with adjustable degrees of punishment and reinforcement. The paradigm allows further discussions about relationships among mental models and factors which may influence mental model acquisition. Theoretically, the current thesis supports previous literature on the easiness of reacquiring old mental models and how they block new model acquisition (e.g., Stöttinger et al., 2014; Bilalić, Mcleod, & Gobet, 2008; Davis & Chan, 2015). Additionally, the current work emphasizes better performances following wins relative to following losses, and also indicates post-loss inertia regardless of stay or shift behaviour (e.g., Dyson, 2023; Dyson, Wilbiks, Sandhu, Papanicolaou, & Lintag, 2016). Practically, the current thesis serves as a hint in applied education fields that teachers and parents can evaluate the degree of punishment and reinforcement of students'/children's behaviours (e.g., Mayer, Sulzer, & Cody, 1968; Kazdin & Forsberg, 1974). A light reinforcement may be sufficiently effective for students to continue their behaviour, or to re-act a previous behaviour. A reinforcement of a high degree, on the other hand, will encourage a child's prior behaviours but may meanwhile impact future learning. As for punishment, it may require a higher degree for students/children to stop a behaviour when it is wrong or no longer adaptable.

6.5 Limitations and ideas for future studies

Although Experiments 1-3 represent a systematic manipulation of the degree of punishment and reinforcement experienced by participants, the current experiments lack a "true" baseline: a condition to test how participants perform in old mental model reacquisition and new mental model acquisition if their win rate was fixed and *continued* at their average win rate (around 66%). Based on the results of the current Experiments 1-3, in a fixed 66% win rate condition, people should be more likely to stay with their old strategy, relative to shifting to a new strategy. Furthermore, across three experiments, the highest degree of reinforcement (from 66% to 88%; +22% difference) was not as strong as the most severe punishment (from 66% to 25%; -41% difference). There is potentially a ceiling effect of both mechanisms such that further increases in punishment or reinforcement will not dramatically impact performance. In other words, the degree of punishment below 25% and the degree of reinforcement above 88% may not affect performance any further. However, it may be rigorous to examine higher reinforcement percentages.

In the current experiments, reinforcement and punishment are manipulated based on varied win rates. There are also possible ways to keep the win rate consistent while functioning as reinforcement or punishment. For example, in a reinforcement condition, the win rate can always be fixed at 50%, but participants gain more than 1 point for a winning trial (e.g., 2, 3, 4, etc. points for different degrees or reinforcement) while lose 1 point for a losing trial. In this way, the trials of winning and losing are equal, which guarantees participants will receive the same amount of trial information from positive and negative feedback. Participants' performance in relearning-old and learning-new may also be influenced by the way they learn about reinforcement and punishment. As experiencers, like in the current experiments, they experience

the reinforcement and punishment themselves. In future studies, participants can learn by observing others experiencing reinforcement and punishment. Learning via observation can also effectively influence relearning-old and learning-new, consistent with the current findings. However, the degree of influence is supposed to be weaken because participants may underrate the reinforcement or punishment as it happens on others (Clark, Lawrence, Astley-Jones, & Gray, 2008).

In addition to reinforcement and punishment designs, the current studies exclude data from participants who won all trials in *Pre* and *Post* phases because without losing any trial, their behaviours following losses could not be recorded. However, these participants may represent a group of people who performed the best in the current game. Their behaviours may be different from other participants, which requires further analyses.

Finally, the current experiments show that the punishment of learned strategy negatively affects old model reacquisition at a degree of 44% win rate rather than 50%. It is interesting to investigate whether the 44% win rate is the *starting point* for the change in the default preference for old mental models, or, if there is another spot between 44% and 50% that initiates this change. As previously mentioned, the 50% win rate may not have worked work as intended (as a punishment) because participants treated the win rate (50%) as an index of a game of chance rather than as a game of skill that is being performed badly. In other words, people perceive 50% win rate as "this game is now random" instead of "I am being punished". Additionally, the collapsing of strategy in the current thesis means that there are other tales to tell regarding the ease of moving into and out of specific strategies, like the two exploitable strategies (*repetition* and *alternation*) used here. Specifically, the *repetition* strategy used in the current studies meant that optimized behavior was in line with the operant conditioning principles of *win-stay* and *lose-*

shift. In contrast, the *alternation* strategy used in the current studies meant that optimized behavior was exactly in opposition to operant conditioning principle, requiring *win-shift* and *lose-stay*. Previous studies have shown that strategies that require "anti-operant conditioning" behaviours can be performed as well as strategies that align with *win-stay* and *lose-shift* (e.g., Sundvall & Dyson, 2022). However, moving to and from strategies that either align or misalign with operant conditioning may reveal constraints in our ability to recycle or dispose of current mental models. Besides, strategies with antagonistic optimal behaviours may increase the level of difficulty in learning-new. Therefore, future studies can focus on other degrees of punishment and reinforcement, detailing the effect of chance performance, and examining the influence of different strategies.

References

Achterberg, J., Kadohisa, M., Watanabe, K., Kusunoki, M., Buckley, M. J., & Duncan, J. (2022). A one-shot shift from explore to exploit in monkey prefrontal cortex. *Journal of Neuroscience, 42*(2), 276-287.

Badre, D., Doll, B. B., Long, N. M., & Frank, M. J. (2012). Rostrolateral prefrontal cortex and individual differences in uncertainty-driven exploration. *Neuron, 73*(3), 595-607.

Bakken, B. T. (n.d.). An empirical study of learning and decision making in a two-person zero-sum game with strategic uncertainty: Experimental design and preliminary results.

Bandura, A., Ross, D., & Ross, S. A. (1963). Imitation of film-mediated aggressive models. *The Journal of Abnormal and Social Psychology, 66*(1), 3.

Bartlett, F. C. (1932). *Remembering: A study in experimental and social psychology.* Cambridge, UK: Cambridge University Press.

Beda, Z., & Smith, S. M. (2018). Chasing red herrings: Memory of distractors causes fixation in creative problem solving. *Memory & Cognition, 46,* 671-684.

Belot, M., Crawford, V. P., & Heyes, C. (2013). Players of Matching Pennies automatically imitate opponents' gestures against strong incentives. *Proceedings of the National Academy of Sciences, 110*(8), 2763-2768.

Bilalić, M., McLeod, P., & Gobet, F. (2008). Why good thoughts block better ones: The mechanism of the pernicious Einstellung (set) effect. *Cognition, 108*(3), 652-661.

Brewer, W. F. (1987). Schemas versus mental models in human memory. in P. Morris (Ed.), *Modelling Cognition*, Wiley, New York.

Brockbank, E., & Vul, E. (2021). Humans fail to outwit adaptive rock, paper, scissors opponents. *In Proceedings of the Annual Meeting of the Cognitive Science Society* (Vol. *43*, No. 43).

Carver, C. S., & White, T. L. (1994). Behavioral inhibition, behavioral activation, and affective responses to impending reward and punishment: the BIS/BAS scales. *Journal of personality and social psychology, 67*(2), 319.

Cavanagh, J. F., Figueroa, C. M., Cohen, M. X., & Frank, M. J. (2012). Frontal theta reflects uncertainty and unexpectedness during exploration and exploitation. *Cerebral cortex, 22*(11), 2575-2586.

Clark, L., Lawrence, A. J., Astley-Jones, F., & Gray, N. (2009). Gambling near-misses enhance motivation to gamble and recruit win-related brain circuitry. *Neuron, 61*(3), 481-490.

Cook, R., Bird, G., Lünser, G., Huck, S., & Heyes, C. (2012). Automatic imitation in a strategic context: players of rock–paper–scissors imitate opponents' gestures. *Proceedings of the Royal Society B: Biological Sciences, 279*(1729), 780-786.

Craik, K. (1943). *The nature of explanation.* Cambridge, UK: Cambridge University Press.

Daskalakis, C., Goldberg, P. W., & Papadimitriou, C. H. (2009). The complexity of computing a Nash equilibrium. *Communications of the ACM, 52*(2), 89-97.

Davis, S. D., & Chan, J. C. (2015). Studying on borrowed time: How does testing impair new learning?. *Journal of Experimental Psychology: Learning, Memory, and Cognition, 41*(6), 1741.

Ditta, A. S. (2019). *An Investigation of the Effects of Retrieving and Being Re-exposed to Old Ideas on the Generation of New Ideas.* University of California, Santa Cruz.

Dixon, Larche, Stange, Graydon & Fugelsang, (2018). The frustrating effects of just missing the jackpot: Slot machine near-misses trigger large skin conductance responses, but no post-reinforcement pauses. *Journal of Gambling Studies, 29,* 661-674.

Dyson, B. (2023). Post-error speeding or post-win slowing? An empirical note on the interpretation of decision-making time as a function of previous outcome. PsyArXiv.

Dyson, B. J., Musgrave, C., Rowe, C., & Sandhur, R. (2020). Behavioural and neural interactions between objective and subjective performance in a Matching Pennies game. *International Journal of Psychophysiology, 147,* 128-136.

Dyson, B. J., Wilbiks, J. M. P., Sandhu, R., Papanicolaou, G., & Lintag, J. (2016). Negative outcomes evoke cyclic irrational decisions in Rock, Paper, Scissors. *Scientific reports, 6*(1), 20479.

Erev, I., & Roth, A. E. (1998). Predicting how people play games: Reinforcement learning in experimental games with unique, mixed strategy equilibria. *American economic review,* 848-881.

Feather, N. T., & Simon, J. G. (1971). Attribution of responsibility and valence of outcome in relation to initial confidence and success and failure of self and other. *Journal of Personality and Social Psychology, 18*(2), 173.

Filipowicz, A. (2017). *Adapting to Change: The Role of Priors, Surprise and Brain Damage on Mental Model Updating.*

Filipowicz, A., Anderson, B., & Danckert, J. (2016). Adapting to change: The role of the right hemisphere in mental model building and updating. *Canadian Journal of Experimental Psychology/Revue canadienne de psychologie expérimentale, 70*(3), 201.

Finn, B., & Roediger III, H. L. (2013). Interfering effects of retrieval in learning new information. *Journal of Experimental Psychology: Learning, Memory, and Cognition, 39*(6), 1665.

Forder, L., & Dyson, B. J. (2016). Behavioural and neural modulation of win-stay but not lose-shift strategies as a function of outcome value in Rock, Paper, Scissors. *Scientific reports, 6*(1), 33809.

Hayes, G. (1975). *Duell.* Beverly, MA, USA: Parker Brothers.

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.

Jahn, C. I., Grohn, J., Cuell, S., Emberton, A., Bouret, S., Walton, M. E., ... & Sallet, J. (2023). Neural responses in macaque prefrontal cortex are linked to strategic exploration. *Plos Biology, 21*(1), e3001985.

Johnson-Laird, P. N. (2013). Mental models and cognitive change. *Journal of Cognitive Psychology, 25*(2), 131-138.

Kazdin, A. E., & Forsberg, S. (1974). Effects of group reinforcement and punishment on classroom behavior. *Education and Training of the Mentally Retarded*, 50-55.

Ke, F. (2009). A qualitative meta-analysis of computer games as learning tools. *Handbook of Research on Effective Electronic Gaming in Education,* 1-32.

Kirchkamp, O., & Nagel, R. (2007). Naive learning and cooperation in network experiments. *Games and Economic Behavior, 58*(2), 269-292.

Lakkaraju, K., Epifanovskaya, L. W. E., Stites, M. C., Letchford, J., Reinhardt, J. C., & Whetzel, J. (2018). *Online Games for Studying Human Behavior (No. SAND2018-6296B).* Sandia National Lab.(SNL-NM), Albuquerque, NM (United States); Sandia National Lab.(SNL-CA), Livermore, CA (United States).

Mayer, G. R., Sulzer, B., & Cody, J. J. (1968). The use of punishment in modifying student behavior. *The Journal of Special Education, 2*(3), 323-328.

McGeoch, J. A. *The psychology of human learning.* New York: Longmans, Green and Co., 1942.

McNamara, J. M., Houston, A. I., & Leimar, O. (2021). Learning, exploitation and bias in games. *Plos one, 16*(2), e0246588.

Messner, C., & Vosgerau, J. (2010). Cognitive inertia and the implicit association test. *Journal of Marketing Research, 47*(2), 374-386.

Nickerson, R. S. (1998). Confirmation bias: A ubiquitous phenomenon in many guises. *Review of General Psychology, 2*(2), 175-220.

Olson, D. R., & Bruner, J. S. (1974). Learning through experience and learning through media. *Teachers College Record, 75*(5), 125-150.

Passarelli, A. M., & Kolb, D. A. (2011). The learning way: Learning from experience as the path to lifelong learning and development. *The Oxford Handbook of Lifelong Learning,* 70-90.

Pastötter, B., & Bäuml, K. H. T. (2014). Retrieval practice enhances new learning: the forward effect of testing. *Frontiers in Psychology,* 286.

Phillips, W. J., Fletcher, J. M., Marks, A. D., & Hine, D. W. (2016). Thinking styles and decision making: A meta-analysis. *Psychological Bulletin, 142*(3), 260.

Pingle, M., & Day, R. H. (1996). Modes of economizing behavior: Experimental evidence. *Journal of Economic Behavior & Organization, 29*(2), 191-209.

Procyk, E., & Goldman-Rakic, P. S. (2006). Modulation of dorsolateral prefrontal delay activity during self-organized behavior. *Journal of Neuroscience, 26*(44), 11313-11323.

Schlenker, B. R., & Bonoma, T. V. (1978). Fun and games: The validity of games for the study of conflict. *Journal of Conflict Resolution, 22*(1), 7-38.

Skinner, B. F. (1963). Operant behavior. *American Psychologist, 18*(8), 503.

Smith, S. M. (2003). The constraining effects of initial ideas. *Group Creativity: Innovation through Collaboration,* 15-31.

Staddon, J. E., & Cerutti, D. T. (2003). Operant conditioning. *Annual review of psychology, 54*(1), 115-144.

Stöttinger, E., Filipowicz, A., Danckert, J., & Anderson, B. (2014). The effects of prior learned strategies on updating an opponent's strategy in the rock, paper, scissors game. *Cognitive Science, 38*(7), 1482-1492.

Sun, Z., & Jia, G. (2023). Reinforcement learning for exploratory linear-quadratic two-person zero-sum stochastic differential games. *Applied Mathematics and Computation, 442,* 127763.

Sundvall, J., & Dyson, B. J. (2022). Breaking the bonds of reinforcement: Effects of trial outcome, rule consistency and rule complexity against exploitable and unexploitable opponents. *Plos one, 17*(2), e0262249.

Tenenbaum, J. B., Kemp, C., Griffiths, T. L., & Goodman, N. D. (2011). How to grow a mind: Statistics, structure, and abstraction. *Science, 331*(6022), 1279-1285.

Thorndike, E. L. (1911). *Animal Intelligence. Experimental studies.* New York: Macmillan Company.

Tolman, E. C. (1948). Cognitive maps in rats and men. *Psychological Review, 55*(4), 189.

Tversky, A., & Kahneman, D. (1991). Loss aversion in riskless choice: A reference-dependent model. *The quarterly journal of economics, 106*(4), 1039-1061.

Vandenbosch, B., & Higgins, C. (1996). Information acquisition and mental models: An investigation into the relationship between behaviour and learning. *Information Systems Research, 7*(2), 198-214.

Wagoner, B. (2013). Bartlett's concept of schema in reconstruction. *Theory & Psychology, 23*(5), 553-575.

Wang, Z., Xu, B., & Zhou, H. J. (2014). Social cycling and conditional responses in the Rock-Paper-Scissors game. *Scientific Reports, 4*(1), 5830.

Wheeler, W. A., Bolton, P. A., & Sanquist, T. F. (1990, October). Decision making in an emergency: When information is not enough. *In Proceedings of the Human Factors Society Annual Meeting* (Vol. *34*, No. 16, pp. 1137-1141). Sage CA: Los Angeles, CA: SAGE Publications.

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

Young, H. P. (2009). Learning by trial and error. *Games and economic behavior, 65*(2), 626-643.

Appendix A: Cross analyses

A.1 Cross various fixed win rates analyses

By comparing the results from *Fixed win rate* conditions in three experiments, it appeared that in 50%, 75%, and 88% conditions, *Post* OBR in the relearning-old condition was higher than that in the learning-new condition (significant *Learning* x *Period* interaction; *p*s < .05); however, in 25% and 44% conditions, *Post* OBR showed no difference between relearning-old and learning-new. It remained to be a question whether in 25% and 44% conditions, compared to 50%, 75%, and 88% conditions, *Post* OBR in relearning-old condition went down or *Post* OBR in learning-new condition went up. In other words, did fixed 25% and 44% win rates in *During* bins serve as a suppression mechanism for relearning old strategy or a reinforcement mechanism for learning new strategy? A 2x2 repeated measures ANOVA for *Post* OBR in each experiment between two fixed win rate conditions (25% vs. 75%; 44% vs. 88%) was thus conducted, with *Learning* (relearning-old, learning-new) and *Fixed win rate* (25%, 75%; 44%, 88%) entered as factors (Table A.1, A.2, respectively).

In Experiment 2 (Table A.1), when comparing 25% and 75% conditions, the significant two-way interaction between *Learning* x *Fixed win rate* ($p < .001$; 2 x 2 interaction) suggested no difference between *Post* OBR in relearning-old (M = .680, SE = .023) and learning-new (M = .697, SE = .022; Tukey's HSD; $p = .878$) conditions when the fixed win rate was 25%. In 75% condition, *Post* OBR was higher in relearning-old condition (M = .725, SE = .026) while lower in learning-new condition (M = .607, SE = .025; Tukey's HSD; $p < .001$).

In Experiment 3 (Table A.2), when comparing 44% and 88% conditions (Table 4.2), the significant two-way interaction between *Learning* x *Fixed win rate* (p = .010; 2 x 2 interaction) again suggested no difference between *Post* OBR in relearning-old (M = .711, SE = .027) and learning-new (M = .670, SE = .025; Tukey's HSD; p = .266) conditions when the fixed win rate was 44%; while higher *Post* OBR in relearning-old condition (M = .763, SE = .026) and lower *Post* OBR in learning-new condition (M = .636, SE = .023; Tukey's HSD; p < .001) when the fixed win rate was 88%.

The results from Experiments 2 and 3 were consistent, implying that when previous actions were punished (25% and 44% conditions), participants' ability to retrieve a learned strategy is suppressed while to obtain a new strategy is reinforced; when previous actions were encouraged (75% and 88% conditions), the opposite mechanism works that participants' ability to retrieve a learned strategy is reinforced while to obtain a new strategy is suppressed.

	df	*F*	*MSE*	*p*	η_p^2
Learning (L)	1,71	**6.438**	**.0288**	**.013**	**.083**
Fixed win rate (F)	1,71	.843	.0437	.362	.012
L x F	1,71	**17.496**	**.0189**	**<.001**	**.198**

Table A.1 Inferential statistics for OBR in Experiment 2. Two-way repeated measures ANOVA results. Learning (L): 2 levels (Old / New); Fixed win rate (F): 2 levels (25% / 75%). Note: Significant effects in bold font.

	df	*F*	*MSE*	*p*	η_p^2
Learning (L)	1,69	**20.409**	**.0245**	**<.001**	**.228**
Fixed win rate (F)	1,69	.122	.0489	.728	.002
L x F	1,69	**7.059**	**.0181**	**.010**	**.093**

Table A.2 Inferential statistics for OBR in Experiment 3. Two-way repeated measures ANOVA results. Learning (L): 2 levels (Old / New); Fixed win rate (F): 2 levels (44% / 88%). Note: Significant effects in bold font.

It was also noticeable that there was a difference between OBR following a win in 25% and 44% conditions (significant main effect of *Outcome*; $p < .05$) but not in 75% and 88% conditions. A 2x2 repeated measures ANOVA was therefore conducted for Experiments 2 and 3 separately to answer whether participants played more *Post* optimal behaviours after a loss or less optimal behaviours after a win following the fixed 75% / 88% win rate in *During* bins. *Fixed win rate* (25% vs. 75%; 44% vs. 88%) and *Outcome* (win, loss) were entered as factors (Table A.3, A.4, respectively).

In Experiment 2 (Table A.3), the significant two-way interaction between *Fixed win rate* x *Outcome* ($p = .016$; 2 x 2 interaction) showed that, in 75% condition, there was no difference between *Post* OBR following a win (M = .666, SE = .030) and following a loss (M = .667, SE = .021; Tukey's HSD; $p = .999$), while in 25% condition, *Post* OBR following a win (M = .741, SE = .024) was significantly higher than that following a loss (M = .636, SE = .022; Tukey's HSD; $p = .005$).

In Experiment 3 (Table A.4), there was no significant *Fixed win rate* x *Outcome* interaction ($p = .831$), suggesting the mechanism of 44% and 88% do not vary on *Post* OBR after a win and after a loss (Tukey's HSD; ps > .05).

The inconsistent results from Experiments 2 and 3 indicate that either the effect of suppression (25% and 44% conditions) and reinforcement (75% and 88% conditions) on behaviours following a win or a loss is unstable or there is no such effect and the observed difference between suppression and reinforcement conditions results from individual differences.

	df	*F*	*MSE*	*p*	η_p^2
Fixed win rate (F)	1,71	.843	.0437	.362	.012
Outcome (O)	1,71	**10.294**	**.0190**	**.002**	**.127**
F x O	1,71	**6.046**	**.0337**	**.016**	**.078**

Table A.3 Inferential statistics for OBR in Experiment 2. Two-way repeated measures ANOVA results. Fixed win rate (F): 2 levels (25% / 75%); Outcome (O): 2 levels (Win / Loss). Note: Significant effects in bold font.

	df	*F*	*MSE*	*p*	η_p^2
Fixed win rate (F)	1,69	.122	.0489	.728	.002
Outcome (O)	1,69	3.150	.0144	.080	.044
F x O	1,69	.046	.0289	.831	.001

Table A.4 Inferential statistics for OBR in Experiment 3. Two-way repeated measures ANOVA results. Fixed win rate (F): 2 levels (44% / 88%); Outcome (O): 2 levels (Win / Loss). Note: Significant effects in bold font.

A.2 Cross experiments analyses

Because there were different degrees of punishment and reinforcement conditions in Experiments 2 and 3, it was interesting to examine whether punishment and reinforcement affected optimal behaviours as general mechanisms. By grouping 25% (Experiment 2) and 44% (Experiment 3) fixed win rates in *During* bins into punishment group, and 75% (Experiment 2) and 88% (Experiment 3) fixed win rates in *During* bins into reinforcement group, the effects of punishment and reinforcement on *Post* OBR of relearning-old and learning-new were further analyzed via a mixed-subjects factorial ANOVA, with *Learning (L)* entered as within-subject factor and punishment and reinforcement in *During* bins (*During effect; PR)* entered as between-subject factor (Table A.5).

The significant two-way interaction between *During effect* x *Learning* ($p < .001$; 2 x 2 interaction) showed that, following punishment, there was no difference between *Post* OBR of relearning-old (M = .696, SE = .018) and learning new (M = .684, SE = .017; Tukey's HSD; $p = .910$). However, when following reinforcement, the Post OBR of relearning-old (M = .744, SE = .018) was significantly higher than that of learning-new (M = .621, SE = .017; Tukey's HSD; $p < .001$). In short, the punishment of a learned strategy weakened the advantage of relearning an old strategy over learning a new one, yet the latter did not outperform the former, while the reinforcement of a learned strategy strongly encouraged relearning-old but suppressed learning-new.

	df	*F*	*MSE*	*p*	η_p^2
During effect (PR)	1,282	.106	.0645	.745	.000
Learning (L)	1,282	**28.442**	**.0227**	**<.001**	**.092**
L x PR	1,282	**19.278**	**.0227**	**<.001**	**.064**

Table A.5 Inferential statistics for OBR in Experiments 2 and 3. Two-way mixed-subjects factorial ANOVA results. Between-subject factor: During effect (PR): 2 levels (Punishment / Reinforcement); Within-subject factor: Learning (L): 2 levels (Old / New). Note: Significant effects in bold font.