LEARNING IN NON-STATIONARY ENVIRONMENTS

by

Cameron Dale Hassall

Submitted in partial fulfilment of the requirements
for the degree of Master of Science

at

Dalhousie University
Halifax, Nova Scotia
August 2013

TABLE OF CONTENTS

LIST OF TABLES v

LIST OF FIGURES vi

ABSTRACT xi

LIST OF ABBREVIATIONS USED xii

ACKNOWLEDGEMENTS xiii

CHAPTER 1: INTRODUCTION 1

1.1 Reinforcement Learning 1

1.2 Dopamine and Medial-Frontal Reward Processing 2

1.3 Expected and Unexpected Uncertainty 5

1.4 Probabilistic Reversal Learning 7

1.5 Neural Correlates of Uncertainty Detection 9

1.6 The P300 and the LC-NE system 10

1.7 Summary 11

CHAPTER 2: EXPERIMENT 1 13

2.1 Introduction 13

2.2 Method 16

2.2.1 Participants 16

2.2.2 Apparatus and Procedure 16

2.2.3 Data collection 19

2.2.4 Data analysis 20

2.3 Results 24

2.3.1 Accuracy 24

2.3.2 Response Time 25
2.3.3 The fERN 26
2.3.4 The P300 27
2.4 Discussion 28
CHAPTER 3: EXPERIMENT 2 32
3.1 Introduction 32
3.2 Method 34
3.2.1 Participants 34
3.2.2 Apparatus and Procedure 34
3.2.3 Data Collection 37
3.2.4 Data Analysis 37
3.3 Results 41
3.3.1 Accuracy 41
3.3.2 Response Time 43
3.3.3 Environment Ratings 43
3.3.4 The fERN 45
3.3.5 The P300 46
3.4 Discussion 47
CHAPTER 4: SIMULATION 50
4.1 Introduction 50
4.2 Design 53
4.2.1 Action Selection: Reinforcement Learning 54
4.2.2 Uncertainty Detection: ACh and NE 56

4.3 Data Analysis 59
4.3.1 Simulating Experiment 1 59
4.3.2 Simulating Experiment 2 60
4.4 Simulation Results 61
4.4.1 Behavioural 61
4.4.2 ACh and NE 61
4.4.3 Prediction Errors 64
4.4.4 Experiment 2 versus Experiment 1 65
4.5 Discussion 67
CHAPTER 5: GENERAL DISCUSSION 71
5.1 Overview of Current Results 71
5.2 Connection to Current Research and Theory 73
5.3 Conclusion 75
REFERENCES 77
APPENDIX 86

LIST OF TABLES

Table 2.1. Experiment 1: Behavioural means and standard errors. 26

Table 2.2. Experiment 1: Means and standard errors for the fERN. 27

Table 2.3. Experiment 1: Means and standard errors for the P300..... 28

Table 3.1. Experiment 2: Behavioural means and standard errors. 43

Table 3.2. Experiment 2: Means and standard errors for the fERN. 46

Table 3.3. Experiment 2: Means and standard errors for the P300..... 47

LIST OF FIGURES

Figure 1.1. Dopaminergic activity encodes a reinforcement-learning prediction error. Prior to learning, midbrain dopaminergic activity increases at the time of reward. Following learning, dopaminergic activity decreases at the first indication that events are better than expected. Image from Schultz, Dayan, and Montague, 1997..... 3

Figure 2.1. Experimental design, with timing details. Over the course of a block (20 trials), participants learned that selecting one of the squares resulted in more wins, on average, compared to selecting the other coloured square..... 19

Figure 2.2. Average medial-frontal response to feedback (a) with scalp topographies (b). Context shifts only occurred in late non-stationary blocks. The dashed rectangle shows the interval of analysis (200 – 400 ms post feedback)..... 22

Figure 2.3. P300 response to feedback early in learning (a and b) and late in learning (c and d) for each environment type. The grand average (all conditions pooled) was maximal at Pz. A dashed rectangle indicates the region of analysis (300 – 500 ms post feedback)..... 23

Figure 2.4. P300 scalp topography for the average response to all feedback, maximal at Pz..... 24

Figure 2.5. Performance Curve. In the non-stationary environment, the optimal choice switched on trial 12, on average. Dashed lines are shown at 50% (chance) and 80%, for reference. 25

Figure 2.6. Behavioural results. (a) Mean performance early and late in a block for each environment. (b) Mean response time early and late in a block for each environment. 26

Figure 2.7. Mean fERN amplitudes, early and late in learning, for each environment. 27

Figure 2.8. P300 amplitudes in response to feedback. When context shifts occurred, there was an enhanced P300 response to wins. 28

Figure 3.1. Experimental design for (a) blocks and (b) trials, with timing details. At the beginning of each block, participants were shown the casino in which they were gambling. After 20 trials, they were asked to rate the honesty of the current casino. 36

Figure 3.2. Grand average (a) waveforms and (b) scalp topographies in response to feedback. Context shifts only occurred late in non-stationary blocks. The dashed rectangle shows the interval of analysis (200 – 400 ms post feedback). Note that the scales for the scalp topographies are identical (black = -4 μV, white = 1 μV). 39

Figure 3.3. P300 response to feedback early in learning (a and b) and late in learning (c and d) for each environment type. A dashed rectangle shows the region of analysis: 300 – 500 ms post feedback. 40

Figure 3.4. P300 scalp topography for the average response to all feedback, maximal at Pz. 41

Figure 3.5. Performance curve in (a) Experiment 1, and (b) the present study. In the non-stationary environment, the optimal choice switched on trial 12, on average. Dashed lines are shown at 50% (chance) and 80%, for reference..... 42

Figure 3.6. Mean accuracies, for each environment, for the current experiment and Experiment 1. Adding environment cues improved performance regardless of environment type..... 42

Figure 3.7. Behavioural results. Mean performance (a) and response time (b) in Experiment 1 and (c, d) Experiment 2. 44

Figure 3.8. Casino ratings (a) over all trials and (b) grouped by environment type. Participants were able to discriminate the environments based on the feedback they received throughout a block. 44

Figure 3.9. Mean fERN amplitudes, early and late in learning, for each environment in (a) Experiment 1, and (b) the present study. In the present experiment, the fERN was enhanced over time in the non-stationary environment, but reduced in the stationary environment..... 45

Figure 3.10. P300 amplitudes in response to feedback in (a) Experiment 1, and (b) the current experiment. When context shifts occurred, there was an enhanced P300 response to wins and losses. 47

Figure 4.1. Model design. Action selection (right) was made via reinforcement learning. Node weights for each option [w_1 w_2] were used by a softmax function (P) to generate a response. Option weights were also used to compute the current value of the system at action selection, depending on which unit was active (i.e. which option was chosen). A prediction error unit

(PE) compared this predicted value with the actual reward value (r) to generate a prediction error (δ). Feedback was also used to detect uncertainty in the prefrontal cortex (PFC) working memory (left). Here, feedback history within a certain context was used to estimate expected uncertainty (γ), which was used to determine the likelihood of sticking with the current belief about which option was best (λ). If a context shift was detected based on these values, then the weights associated with each option were reset. 56

Figure 4.2. Model performance (b) compared to actual performance in Experiment 1 (a). Accuracy was defined as the proportion of times that a virtual participant made the optimal response (i.e. made the response most likely to result in a win). Note that a context shift occurred around trial 12. Dashed lines are shown at accuracies of 80% and 50%. 61

Figure 4.3. ACh and NE levels, over time. Following the context shift, ACh and NE levels increase to indicate rising expected and unexpected uncertainty, respectively. 62

Figure 4.4. Mean values for P*, the model estimate of the probability that the current context belief remained the same in the (a) stationary and (b) non-stationary environment. For the sake of this illustration, μ=1 was assumed to be the correct context in both the stationary environment, and in the non-stationary environment prior to the context switch. If P*(μ=1) < P*(μ=2) then a context update occurred. 63

Figure 4.5. Mean number of context shift detections ("context updates") for each trial in the (a) stationary and (b) non-stationary environment. A context shift occurred on trial 12, on average. 64

Figure 4.6. Model prediction error means for (a) all trials, and (b) grouped by early/late in learning. Prediction errors were enhanced in the non-stationary environment, reflecting unexpected losses. The y-axis is reversed here, mirroring the ERP convention for plotting the fERN. 64

Figure 4.7. Model performance when NE was modulated by environment (b) compared to actual performance in Experiment 2 (a). Overall NE levels were reduced in the stationary environment and enhanced in the non-stationary environment. Dashed lines are shown at 80% and 50% accuracy. 65

Figure 4.8. Total number of context updates in the (a) stationary environment, where NE was reduced, and (b) non-stationary environment, where NE was enhanced. 66

Figure 4.9. Overall performance in (a) the stationary environment and (b) the non-stationary environment. Compared here are the reinforcement learning (RL) model alone, the RL model augmented by uncertainty detection (RL+UD), and the augmented model when overall neurotransmitter levels were modulated (RL+UD*). 66

Figure 4.10. Simulated prediction errors (fERNs) for Experiment 1 and Experiment 2 (a). An arrow highlights the difference: a positive prediction error later in the stationary blocks in the Experiment 2 simulation. Compare this with the actual fERN amplitudes measured in Experiments 1 and 2 (b). 67

ABSTRACT

Real-world decision making is challenging due, in part, to changes in the underlying reward structure: the best option last week may be less rewarding today. Determining the best response is even more challenging when feedback validity is low. Presented here are the results of two experiments designed to determine the degree to which midbrain reward processing is responsible for detecting reward contingency changes when feedback validity is low. These results suggest that while midbrain reward systems may be involved in detecting unexpected uncertainty in non-stationary environments, other systems are likely involved when feedback validity is low – namely, the locus-coeruleus-norepinephrine system. Finally, a computational model that combines these systems is described and tested. Taken together, these results downplay the role of the midbrain reward system when feedback validity is low, and highlight the importance of the locus-coeruleus-norepinephrine system in detecting reward contingency changes.

LIST OF ABBREVIATIONS USED

ACh	acetylcholine
ANOVA	analysis of variance
DA	dopamine
EEG	electroencephalography
ERN	error-related negativity
ERP	event-related potential
fERN	feedback error-related negativity
LC	locus coeruleus
mPFC	medial prefrontal cortex
NE	norepinephrine
OFC	orbitofrontal cortex
PFC	prefrontal cortex
rERN	response error-related negativity
RL	reinforcement learning
SD	standard deviation
SE	standard error
SRO	stimulus-response-outcome
WM	working memory

ACKNOWLEDGEMENTS

I would like to acknowledge the help and kind support of my supervisor, Dr. Olave Krigolson. I would also like to thank the other members of my committee, Dr. Aaron Newman and Dr. Thomas Trappenburg, for their time and effort.

This work would not have been possible without the financial support of both the Natural Sciences and Engineering Research Council of Canada and the Nova Scotia Health Research Foundation.

Thank you to my parents for not only encouraging me to do what I loved to do, but for lending me support when I needed it.

Finally, I would like to thank my wife, Aisling. I could not have done it without you.

CHAPTER 1: INTRODUCTION

In order to maximize their utility, value-driven decision makers must first learn about their world, and then optimally exploit this information by selecting those actions that are most likely to lead to a reward. However, the world is an uncertain place. With that said, we have evolved systems to adapt to varying levels and types of uncertainty in the world around us. Successfully detecting that something about the world is different is important in uncertain environments, since the most rewarding action may change. The neural mechanisms behind the detection of and adaptation to uncertainty remain unclear due, in part, to the different forms of uncertainty that have been identified. However, before this issue can be properly addressed, a brief review of reinforcement learning (RL) theory and mechanisms is warranted.

1.1 Reinforcement Learning

Detecting shifts in reward payouts is an important characteristic of RL models. Here, RL refers to a class of models through which optimal actions may be learned: using feedback to learn to do the right thing at the right time. According to Sutton and Barto (1998) these models map situations to actions – learning what to do in a given situation – in order to maximize long-term rewards. An essential component of RL models, and one relevant to uncertainty detection, is the prediction error. A prediction error is a comparison between expected and actual outcomes. An unexpected reward or an unexpected cue that predicts a reward will elicit a positive prediction error. Likewise, an unexpected punishment, or a lack of reward when one is expected, will elicit a negative prediction error. For example, under RL theory, a positive prediction error is generated when a wandering rat encounters unexpected food. Over time, the rat may learn that food

can always be found at the same location. Thus, on subsequent explorations, predictive cues (e.g. familiar junctions in a maze) may elicit positive prediction errors, while no prediction error is generated when – as expected – the food is actually located. In RL models, prediction errors are used to alter the strength of the associations between situations and actions such that rewarded actions, but not punished actions, are more likely to be repeated.

Implicit in the calculation of a prediction error is a representation of uncertainty, since expected rewards and punishments do not lead to prediction errors, and thus do not change behaviour. However, as others have pointed out (e.g. Alexander & Brown, 2011; Payzan-LeNestour & Bossaerts, 2011), the lack of an explicit representation of uncertainty in RL models (i.e. model-free RL) may be a limiting factor when modelling human behavioural and neural data, since humans are clearly able to detect and adapt to different forms of uncertainty in appropriate ways (see Sections 1.4 and 1.5 for examples).

1.2 Dopamine and Medial-Frontal Reward Processing

Thorndike's Law of Effect (1911) states that an action that leads to a reward becomes connected to the situation that the action arose from. While Thorndike's Law of Effect describes the goal of RL, and RL itself describes the computations behind how actions and situations become connected, the neural mechanisms behind a RL system in humans (if one exists) remain unclear. One popular view is that dopamine (DA) signals that events are better or worse than expected (a RL prediction error: Montague, Dayan, & Sejnowski, 1996). While the exact role of DA is controversial (see Beeler, 2012, for a review), Schultz, Dayan, and Montague (1997) showed that rather than signalling all

positive outcomes, dopaminergic activity instead marks when outcomes are unexpected. They did this by observing that monkey dopaminergic neurons were activated by rewarding stimuli (in their case, a drop of juice) before, but not after learning that a tone predicted the reward. Instead, with learning, dopaminergic activity shifted earlier in time to the first predictor that a reward was forthcoming, i.e. the tone. Furthermore, dopaminergic activity at the expected time of reward following the tone decreased if the predicted reward was withheld (i.e. a negative prediction error: Figure 1.1).

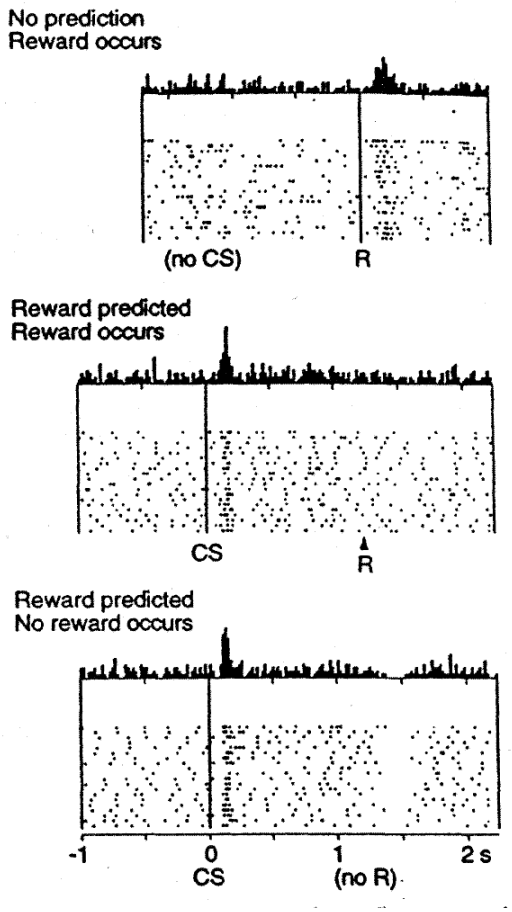

Figure 1.1. Dopaminergic activity encodes a reinforcement-learning prediction error. Prior to learning, midbrain dopaminergic activity increases at the time of reward. Following learning, dopaminergic activity decreases at the first indication that events are better than expected. Image from Schultz, Dayan, and Montague, 1997.

While Schultz et al. (1997) studied reward-related dopaminergic activity in monkeys, it is possible to evaluate high-level reward processing in humans by examining electroencephalographic (EEG) recordings. Indeed, the high temporal resolution of EEG makes it particularly suited to measuring the fast processing of rewards and punishments that occurs in a range of learning tasks. In particular, there is a medial-frontal negative deflection in the event-related potential (ERP) – averaged EEG in response to a particular event – when humans make errors. This deflection, termed the error-related-negativity (ERN), is thought to be generated in anterior cingulate cortex (ACC: Dehaene, Posner, & Tucker, 1994; Holroyd, Yeung, Nystrom, Mars, Coles, & Cohen, 2004). In 2002, Holroyd and Coles linked what was known about midbrain DA activity – namely, that it encodes a RL prediction error – with the observation that the ERN also appears to be sensitive to whether ongoing events are better or worse than expected. Holroyd and Coles (2002) proposed that the ERN is elicited when midbrain DA conveys a RL prediction error to ACC. Furthermore, the role of ACC has recently been interpreted as a region devoted not only to learning, but to predicting the most likely outcomes of actions (Alexander & Brown, 2011), making it a good candidate region for uncertainty detection. In particular, Alexander and Brown (2011) modelled activity in medial prefrontal cortex (mPFC), including ACC, under the assumption that these areas detect and signal surprising outcomes. Thus, both neuroimaging and modelling evidence (Holroyd & Coles, 2002; Alexander & Brown, 2011) suggest that ACC may be involved in detecting unexpected uncertainty.

Two types of ERN have been identified: the response ERN (rERN) and the feedback ERN (fERN). The rERN occurs when participants make mistakes in speeded

response-time tasks (e.g. Gehring, Goss, Coles, Meyer, & Donchin, 1993), and is generated at the time an erroneous response is made. Specifically, the rERN is thought to reflect evaluation of an efference copy of the motor command (Allain, Hasbroucq, Burle, Grapperon, & Vidal, 2004). The fERN, in contrast, is generated upon receiving external feedback indicating an error (additionally, external feedback must be the first indicator that an error has occurred). For example, Miltner, Braun, and Coles (1997) had participants estimate the length of one second and, sometime after each response, provided feedback to indicate either a correct or incorrect estimation. Importantly, participants were unaware they had committed an error until the feedback was received. Miltner et al. (1997) observed that the ERP response to correct feedback differed from the ERP response to incorrect feedback. The difference was maximal over medial-frontal cortex, and was localized to ACC. In summary, the difference between the rERN and the fERN is, arguably, the timing of the information that signals whether events are better or worse than expected.

1.3 Expected and Unexpected Uncertainty

Not all uncertainty is the same. Yu and Dayan (2005) distinguished between expected and unexpected uncertainty (also see Bach & Dolan, 2012 or Bland & Schaeffer, 2012, for recent reviews). Expected uncertainty – the most commonly studied form – arises when actions are sometimes (but not always) rewarded. Expected uncertainty is usually expressed as feedback validity: the likelihood of receiving specific feedback for an action. For example, the best thing to do when making a decision about whether or not to wear a raincoat is to check the weather forecast. Weather forecasts are not always correct, nor do we expect them to be. A keen observer of weather forecasts

may even develop a sense of the accuracy of these predictions. If the forecast accuracy is known to be 90% (the feedback validity), then it is not unexpected when one day out of ten the forecast is incorrect. In fact, it would be unusual for the weather forecast to be correct 100% of the time. The particular day on which the forecast will fail is unknown, but it fails with a known or expected uncertainty.

In contrast, unexpected uncertainty is characterized by a shift in the underlying rule structure determining which actions are most likely to be rewarded. If the weather forecasts mentioned above were to suddenly be highly inaccurate for several days in a row, perhaps due to an equipment malfunction, users of the forecast would experience unexpected uncertainty. In this new environment, the weather forecast may no longer be the best predictor compared to looking out the window, or even flipping a coin. Thus, actions that were optimal in the past may become suboptimal in an environment with unexpected uncertainty (a non-stationary environment), requiring a decision maker to seek out new information about the world. In a stationary environment, on the other hand, optimal actions remain stable over time. Complicating matters slightly, expected uncertainty and unexpected uncertainty interact with one another. In environments with both forms of uncertainty, a lack of reward may or may not signal a shift in the underlying reward structure. If expected uncertainty is high (i.e. rewards are seldom expected) then it can be very difficult for decision makers to detect and adapt to unexpected uncertainty (Behrens, Woolrich, Walton, & Rushworth, 2007; Bland and Schaefer, 2011).

Consider, for example, the challenge of determining which of two slot machines is most likely to result in a win when played. A highly rewarding slot machine may pay

out 50% of the time, while another slot machine may pay out only 30% of the time. If the better slot machine happens to lose, as it will about half the time, an optimal player should not change his or her belief that the current choice is best. Likewise, if playing the worse slot machine results in an occasional win, an optimal player should not mistake it for the better option. Thus, single plays are not that informative when unexpected uncertainty is high, and several plays are required in order to estimate the likelihood of each option winning. With enough examples, however, an astute player should notice that one slot machine is better. Over time, an optimal player learns to select the best slot machine, and comes to expect a reward about half the time (the expected uncertainty). Since the expected uncertainty of the better choice is only 50%, compared to 30% for the worse choice, it would be difficult for a player in this example to determine that the win probabilities of the two slot machines had reversed. In fact, several plays would likely be required in order to detect a reversal, i.e. to notice that something had changed.

1.4 Probabilistic Reversal Learning

Human performance in non-stationary environments is usually studied using probabilistic reversal learning, two examples of which were given in the previous section. In these learning tasks, a previously rewarded response will begin to be punished, requiring participants to inhibit that response in favour of a new choice (Rolls, Hornak, Wade, & McGrath, 1997; Swainson, Rogers, Sahakian, Summers, Polkey, & Robbins, 2000; Cools, Clark, Owen, & Robbins, 2002; Behrens et al., 2007; Bland & Schaeffer, 2011). By far, the bulk of the work on probabilistic reversal learning has focused on pharmacological variables (e.g. Bari, Theobald, Caprioli, Mar, Aidoo-Micah, Dalley, & Robbins, 2010) and/or brain damage or dysfunction (e.g. Fellows & Farah, 2003; Waltz

& Gold, 2007; Adleman, Kayser, Dickstein, Blair, Pine, & Leibenluft, 2011). For example, humans with orbitofrontal cortex (OFC) lesions tend to perseverate in their responses in probabilistic reversal tasks (Hornak, O'Doherty, Bramham, Rolls, Morris, Bullock, & Polkey, 2004). In particular, Hornak et al. (2004) had participants learn which of two fractal images was more likely to result in a reward when selected (70% versus 40%). Participants with OFC lesions tended to stick to previously rewarded responses, even after a reversal, reducing their overall performance relative to controls. Studies like these, as well as neuroimaging work with healthy participants (Cools et al., 2002), have implicated both OFC and ACC (Paulus, Hozack, Frank, & Brown, 2002) in detecting and adapting to changes in reward probabilities.

Two issues arise from existing research on probabilistic reversal learning. First, while feedback validity may be manipulated, it is seldom reduced below 70% (70%: O'Doherty, Critchley, Deichmann, & Dolan, 2003; 70%: Hornak et al., 2004; 75% or 80 %: Behrens et al., 2007; 80%: Chase et al., 2010; 83.3% or 73.3%: Bland and Schaefer, 2011). Thus, the degree to which OFC and ACC are involved in detecting probability shifts in non-stationary environments when feedback validity is very low (e.g. below 70%) is unclear.

Furthermore, in most experiments, reversals may occur at any time (O'Doherty et al., 2003; Hornak et al., 2004; Chase et al., 2010), although they may occur more or less frequently, depending on the research question (Behrens et al., 2007; Bland and Schaefer, 2011). In particular, while experimenters may define certain blocks as stationary or non-stationary, participants are typically told that a reversal may occur at any time. The implication of this is that, from the perspective of participants, most experimental tasks

for studying probabilistic reversal learning occur in non-stationary environments, albeit with periods of greater or lesser unexpected uncertainty. This issue – participants' expectations about when reversals may occur – should be considered when making any stationary/non-stationary comparisons, especially given recent work highlighting the role of cognitive control in uncertainty detection (Mushtaq, Bland, & Schaeffer, 2011).

1.5 Neural Correlates of Uncertainty Detection

To investigate the difference between expected and unexpected uncertainty, Yu and Dayan (2005) modelled probabilistic reversal learning by simulating levels of two neurotransmitters: acetylcholine (ACh) and norepinephrine (NE). According to Yu and Dayan (2005), ACh levels signal expected uncertainty (Sarter & Bruno, 1997), and NE levels signal unexpected uncertainty (Bouret and Sara, 2005; Doya, 2008). For example, in a cueing task, ACh levels rise and fall depending on the likelihood of a cue being valid (Sarter & Parikh, 2005). Specifically, ACh levels are enhanced when cue validity is low, and reduced when cue validity is high. Bouret and Sara (2005) reviewed evidence from recordings in rat and monkey locus coeruleus (LC), the main source of NE in the brain. These electrophysiological results suggest that LC activation occurs when ongoing behaviour is interrupted due to new information about the environment, such as when a previously rewarded behaviour is no longer rewarding. Thus, tonic ACh activity appears to be linked to expected uncertainty, while phasic NE activity appears to signal the detection of unexpected uncertainty (Doya, 2008). One possible benefit of ACh and NE modulation in response to uncertainty may be to allocate attention in optimal ways, e.g. in a target detection task (Avery, Nitz, Chiba, & Krichmar, 2012). Another possible

benefit is that the successful detection of context shifts may improve performance in a learning task (e.g. Hornak et al., 2004).

1.6 The P300 and the LC-NE system

ERP methodology provides a means for indirectly measuring the output of the LC-NE system in humans. The P300 is a positive-deflection in the ERP that typically peaks 300-500 ms post stimulus (Sutton, Braren, Zubin, & John, 1965; Polich, 2007). Although it has been linked to several different cognitive functions, perhaps the most influential account of the P300 is the context-updating hypothesis (Donchin, 1981; Donchin & Coles, 1988). Based on early observations that the P300 is sensitive to stimulus frequency (i.e. it is enhanced by rare events), the idea behind context updating is that new information sometimes requires an update to one's internal model of the world. The amplitude of the P300 is thought to reflect the degree to which the internal representation changes. In other words, the P300 is enhanced under uncertain conditions (Polich, 1990). Consistent with this interpretation, it has been suggested that the amplitude of the P300 is sensitive to the extent of the locus coeruleus-norepinephrine (LC-NE) system's modulation of information processing (Nieuwenhuis, Aston-Jones, & Cohen, 2005). In particular, the role of the LC-NE system is believed to involve the outcome of internal decision-making processes regarding task-relevant stimuli (such as rewards and punishments). This is based on earlier observations that LC neurons in monkeys respond exclusively to target stimuli in a visual discrimination task, even after a reversal occurs, but before behavioural adjustment occurs (Aston-Jones, Rajkowski, & Kubiak, 1997).

1.7 SUMMARY

The exact neural mechanisms behind how humans deal with unexpected uncertainty remain unclear. Although there is evidence implicating a medial-frontal RL system in uncertainty detection, methodological issues make it difficult to determine its exact role. This thesis consists of three experiments designed to determine the extent to which medial-frontal reward processing is used to detect and adapt to unexpected uncertainty. The goal of Experiment 1 was to show that low feedback validity reduces the impact of unexpected uncertainty on the fERN. The results of Experiment 1 supported this hypothesis: unlike previous work using high-validity feedback (Bland and Schaeffer, 2011) the magnitude of the fERN was identical in both stationary and non-stationary environments, even though participants were able to detect context shifts when they occurred. This suggests that while medial-frontal reward processing may play a role in detecting unexpected uncertainty, it is not the only system involved. This assertion was supported by the observation that while there was no fERN difference between the stationary and non-stationary environments in Experiment 1, there was a feedback-locked P300 enhancement in the non-stationary environments.

In Experiment 1, as with most other experiments designed to compare stationary and non-stationary environments, participants were given no cues as to the identity (stationary or non-stationary) of the environment, other than the trial-to-trial feedback that they received. In order to further distinguish between stationary and non-stationary environments, Experiment 2 used environmental cues to indicate to participants that a context shift could or could not occur in the near future. Otherwise, Experiment 2 used exactly the same task as Experiment 1. As predicted, participants performed better in

Experiment 2 compared to Experiment 1, and also displayed enhanced fERN and P300 components in non-stationary environments. Comparisons between Experiment 2 and Experiment 1 suggest that a system other than medial-frontal reward processing is at least partially responsible for detecting unexpected uncertainty.

Finally, in Chapter Four, the behavioural and EEG data observed in Experiments 1 and 2 were simulated. The model combined a RL component, which generated prediction errors and selected actions, with an uncertainty detection component developed by Yu and Dayan (2005). The modelling results suggest that while a RL system plays a strong role in learning in non-stationary environments, it is likely that this system is augmented by an ACh-NE system that explicitly computes various forms of uncertainty.

CHAPTER 2: EXPERIMENT 1

2.1 Introduction

Our choices are informed by the past; a history of rewards and punishments often drives our actions. Thus, the validity of those rewards and punishments is of critical importance for learning systems. When even the best decisions often result in a loss, learning is difficult. Feedback validity describes the degree to which rewards and punishments can be trusted. When feedback validity is high, choosing the best option almost always results in a reward. Likewise, a poor option almost always results in a lack of reward. When feedback validity is low, however, it is difficult to know if one's current actions are optimal.

Besides feedback validity (i.e. expected uncertainty), unexpected uncertainty also impacts learning. The implication of unexpected uncertainty – occasional shifts in the underlying rule structure of a task – is that an optimal response in one instance may no longer be so in the next. Bland and Schaefer (2011) recently used EEG to investigate the interaction between expected uncertainty (which they called feedback validity) and unexpected uncertainty. To do this they had participants learn several stimulus-response-outcome (SRO) rules. SRO rules link an outcome to an action made in a certain context (de Witt & Dickinson, 2009). Following a correct response, a reward was given with a certain likelihood (the feedback validity), and withheld some proportion of the time (the expected uncertainty). Likewise, incorrect responses also lead to a reward with some small probability. Every so often, the SRO rules would switch – the previously rewarded response was now punished, and vice-versa. Bland and Schaefer (2011) manipulated both expected and unexpected uncertainty across several blocks. In some blocks, the SRO

rules changed frequently (a non-stationary environment), and in some they remained stable (a stationary environment). They observed a modulation of the fERN, an ERP component sensitive to rewards and punishments (Miltner et al., 1997), which depended on unexpected uncertainty. In particular, the amplitude of the fERN was enhanced during times when the SRO rules were shifting compared to when they were stable. In addition, they noted that performance (i.e. how often people made the optimal response) suffered when feedback validity was low (i.e. when expected uncertainty was high). Bland and Schaefer (2011) also analyzed the P300 in response to feedback, and noted an enhanced P300 during times of unexpected uncertainty.

Several issues arose in the Bland and Schaefer (2011) study. One issue was that, for optimal responses, high feedback validity was defined as times when the probability of the optimal choice winning was 83.3%, i.e. p(win) = 83.3%; low feedback validity was defined as p(win) = 73.3%. This meant that even in the low feedback validity condition, correct responses still resulted in a reward 73.3% of the time. Likewise, for incorrect responses, p(loss) = 83.3% (high validity) or 73.3% (low validity). A second minor issue was that, based on these probabilities, punishments would be much less frequent than rewards. This could cause the P300, an ERP component sensitive to rare events, to contaminate the fERN by deflecting it in the positive direction (Falkenstein, 2004; Holroyd & Krigolson, 2007). In addition, for their P300 analysis, Bland and Schaefer (2011) grouped together both win and loss feedback responses. Thus, it was not clear if the enhanced P300 was due to the response to wins, losses, or both. Finally, Bland and Schaefer (2011) grouped all ERP responses within each block, which is problematic since the amplitude of the fERN is known to change over time (i.e. with learning: Krigolson,

Pierce, Holroyd, & Tanaka, 2009). In order to claim that any observed effect on the fERN amplitude is due to a difference in uncertainty, it is therefore advisable to track the fERN amplitude within a block (e.g. early and late within a block). Given these issues, it is not clear if the fERN is enhanced in non-stationary environments compared to stationary environments when feedback validity is greatly reduced (i.e. below 70%). Reducing feedback validity should also result in fewer wins compared to losses, mitigating the contamination of the fERN by the P300.

The goal of this experiment was to extend Bland and Schaefer's (2011) work, employing feedback validity below what they considered low. To do this, EEG were recorded while participants played a gambling game in which selecting one of two coloured squares resulted in either a win or a loss. Importantly, selecting one of the coloured squares was more likely to result in a win compared to selecting the other coloured square. Occasionally, and without warning, a context shift would occur (i.e. the colours on the squares would switch), requiring participants to adapt their behaviour in order to maximize their overall wins. Recall that the fERN is thought to index a RL prediction error – the difference between expected and actual rewards (Holroyd & Coles, 2002). Based on this definition, and on other work on the fERN (Holroyd, Larsen, & Cohen, 2004; Amiez, Joseph, & Procyk, 2005; Hajcak, Moser, Holroyd, & Simons, 2006; Holroyd, Hajcak, & Larsen, 2006), RL theory would predict that the fERN should be enhanced in non-stationary environments following a context shift, replicating Bland and Schaefer's (2011) results. However, given that feedback validity here was greatly reduced, making it difficult to tell when a context shift occurred, it was predicted that

there would be little or no difference in the fERN between the stationary environment and the non-stationary environment.

Unlike the fERN, which signals that feedback is different than expected, the P300 marks a context update – the modification of an internal model of the environment (Donchin & Coles, 1988). Thus, it was predicted that the P300 response to both wins and losses would be enhanced in the non-stationary environment following a context shift, since in this environment unexpected losses signalled that a context shift had occurred, and wins signalled that the correct context had been redetermined. Implicit in this prediction is the assumption that participants would be able to detect the context shifts in the non-stationary environment when they occurred and adapt their behaviour accordingly, despite reduced feedback validity.

2.2 Method

2.2.1 Participants

Twenty university-aged participants (4 male, mean age: 20+/- 0.4) with no known neurological impairments and with normal or corrected-to-normal vision took part in the experiment. All of the participants were volunteers who were recruited through an online signup system, and they all received two credit points in an undergraduate psychology course for their participation. The participants provided informed consent approved by the Health Sciences Research Ethics Board at Dalhousie University.

2.2.2 Apparatus and Procedure

Participants were seated 75 cm in front of a computer display and used a Logitech USB game controller to perform a gambling task (written in MATLAB [Version 7.14, Mathworks, Natick, USA] using the Psychophysics Toolbox Extension, Brainard, 1997).

Participants received both verbal and written instructions. We encouraged participants to minimize head and eye movements, and to maintain fixation on the centre of the display throughout the experiment. Participants were told that the goal of the task was to win as many points as possible by maximizing the number of wins they received (two points each) and minimizing their losses (minus one point each).

The gambling task was a two-armed bandit (Sutton and Barto, 1998). On each trial, participants chose between two coloured squares, which represented slot machines. Square colours were chosen randomly at the beginning of each block. Each choice resulted in either a win or a loss. Participants were told that choosing one of the coloured squares was more likely to result in a win compared to choosing the other coloured square. Unknown to participants, choosing the higher probability square initially resulted in a win with probability p(win) = 0.6, and choosing the lower probability square resulted in a win with p(win) = 0.1. Similar to the task used by Miltner et al. (1997), these parameters were adjusted throughout the experiment in order to balance the overall number of wins and losses. Specifically, if the overall ratio of one feedback type to the other (either wins:losses or losses:wins) exceeded 3:2, then p(win) was adjusted to make the task either easier or more challenging by increasing/decreasing the higher p(win) by 5%, and decreasing/increasing the lower p(win) by 2.5%.

In half of the blocks, chosen at random, the p(win) values of the squares would swap partway through the block. Within these blocks – called non-stationary blocks – the probability swap, or context shift, occurred randomly between trials 11 and 20. The exact switch trial was chosen from a type of folded normal distribution, SD = 1, centred on trial 12, such that a switch was never allowed to occur on trials 1-10.

Participants were given the following verbal and written instructions:

> In this experiment, you will be playing several slot machine games. During each game, you will pull the arm on one of two coloured slot machines. One of the coloured slot machines is better - selecting it will result in more wins and fewer losses, in the long run. Your goal in this task is maximize the number of points you receive over 24 games. Sometimes, which slot machine is best will change. If this happens, you should change your response in order to maximize the number of wins you receive. Try to keep your eyes on the + in the centre of the screen throughout the experiment. Wait for the + to change colour before responding. Click the left gamepad button to select the left slot machine. Click the right gamepad button to select the right slot machine. Press LEFT or RIGHT to see a summary.

Participants were then shown a summary as follows:

- When the + changes colour, choose a slot machine
- One of the coloured slot machines is the better choice
- Sometimes, which slot machine is best may change
- Your goal is maximize your total number of points
- Press LEFT or RIGHT to begin

Each trial began with the presentation of a white 1.1 cm central fixation cross subtending 0.84 degrees of visual angle for 400-600 ms. Next, two coloured squares appeared, each 2.8 cm (2.14 degrees) across, equidistant to the left and right of the fixation cross. The squares were 11.3 cm (8.62 degrees) apart centre-to-centre. After another 400-600 ms, the fixation cross changed colour to light gray to cue participants to choose a square. Participants made their selection by pressing the left or right button on the gamepad with the index finger of their left or right hand, respectively. If participants responded too early (i.e. before the "go" cue) or too late (after 2000 ms) the trial resulted in a loss so that all valid responses occurred within a 0-2000 ms window following the go cue.

Following a valid response, the squares were occluded, leaving only a light gray fixation cross on the display for 400-600 ms. Next, participants were shown feedback

indicating the outcome of the trial ("WIN" or "LOSE") for 1000 ms. Wins resulted in a gain of two points, while losses resulted in a loss of one point. Participants were shown their total score at the end of each block. See Figure 2.1 for an overview of the experimental design, with timing details.

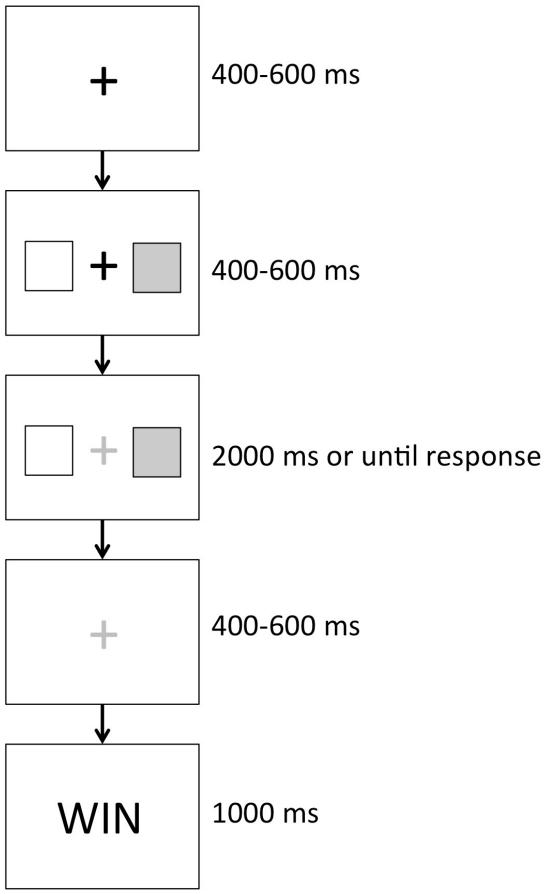

Figure 2.1. Experimental design, with timing details. Over the course of a block (20 trials), participants learned that selecting one of the squares resulted in more wins, on average, compared to selecting the other coloured square.

2.2.3 Data collection

The experimental program recorded participant responses and response times. The EEG was recorded from 16 electrode locations using Brain Vision Recorder software (Version 1.20, Brain Products, GmbH, Munich, Germany). The electrodes were mounted

in a fitted cap with a standard 10-20 layout and were recorded with an average reference built into the amplifier (see Figure A1 in the Appendix for the exact electrode configuration). The vertical electrooculogram was recorded from an electrode placed above the right eye. Electrode impedances were kept below 20 kΩ. The EEG data were sampled at 1000 Hz and amplified (V-Amp, Brainproducts, GmbH, Munich, Germany).

2.2.4 Data analysis

For each trial (1-20), accuracy, defined as the proportion of times across all blocks that the option most likely to result in a win was chosen, was computed for each condition (stationary/non-stationary) for each participant. Means and standard errors of accuracies and response times were also computed for each condition and participant both early (trials 1-10) and late (trials 11-20) in learning, and were compared using a 2 (environment: stationary, non-stationary) by 2 (learning phase: early, late) repeated-measures analysis of variance (ANOVA). Means and standard deviations of adjusted p(win) values were computed, and the total number of wins and losses were compared using a paired-samples t-test. An alpha level of .05 was assumed for all statistical tests, and all error measures represented one standard error.

EEG data were filtered through a (0.1 Hz – 25 Hz pass band) phase shift-free Butterworth filter and rereferenced to the average of the two mastoid channels. Next, ocular artifacts were corrected using the algorithm described by Gratton, Coles, and Donchin (1983). Subsequent to this, all trials were baseline corrected using a 200 ms epoch prior to stimulus onset. Finally, trials in which the change in voltage in any channel exceeded 10 μV per sampling point or the change in voltage across the epoch was greater than 100 μV were discarded. In total, 2% of the data were discarded.

In order to evaluate participant responses to visual feedback (i.e. when the word "WIN" or "LOSE" was displayed), 800 ms epochs of data (from 200 ms before feedback onset to 600 ms after feedback onset) were extracted from the continuous EEG for each trial, channel, and participant, for all feedback presentations. These epochs were grouped based on their feedback type (win or loss), time within a block (early: trials 1-10 or late: trials 11-20), and environment type (stationary or non-stationary). ERPs were created by averaging within these groupings for each participant (early win, early loss, late win, late loss). Two grand average waveforms were also created by averaging over either all wins or all losses, in order to determine likely scalp locations and timing windows for analysis. Finally, all wins and losses were pooled into a third grand average, i.e. the average response to feedback, regardless of valence.

To analyze the fERN, difference waves were created by subtracting the average waveform for win trials from the average waveform for loss trials, early and late in learning, for each environment type. This produced four difference waves for each participant: early stationary, early non-stationary, late stationary, and late non-stationary (see Figure 2.2). A grand difference waveform was also created by subtracting the average for all wins from the average for all losses. Based on the peak of the grand difference wave (Appendix, Figure A2), and in line with previous work (Holroyd & Coles, 2002; Krigolson & Holroyd, 2007; Krigolson et al., 2008), the fERN was defined as the maximum negative deflection of the difference waveform 200–400 ms post feedback at electrode site FCz, where the grand difference wave was maximal (i.e. the greatest difference).

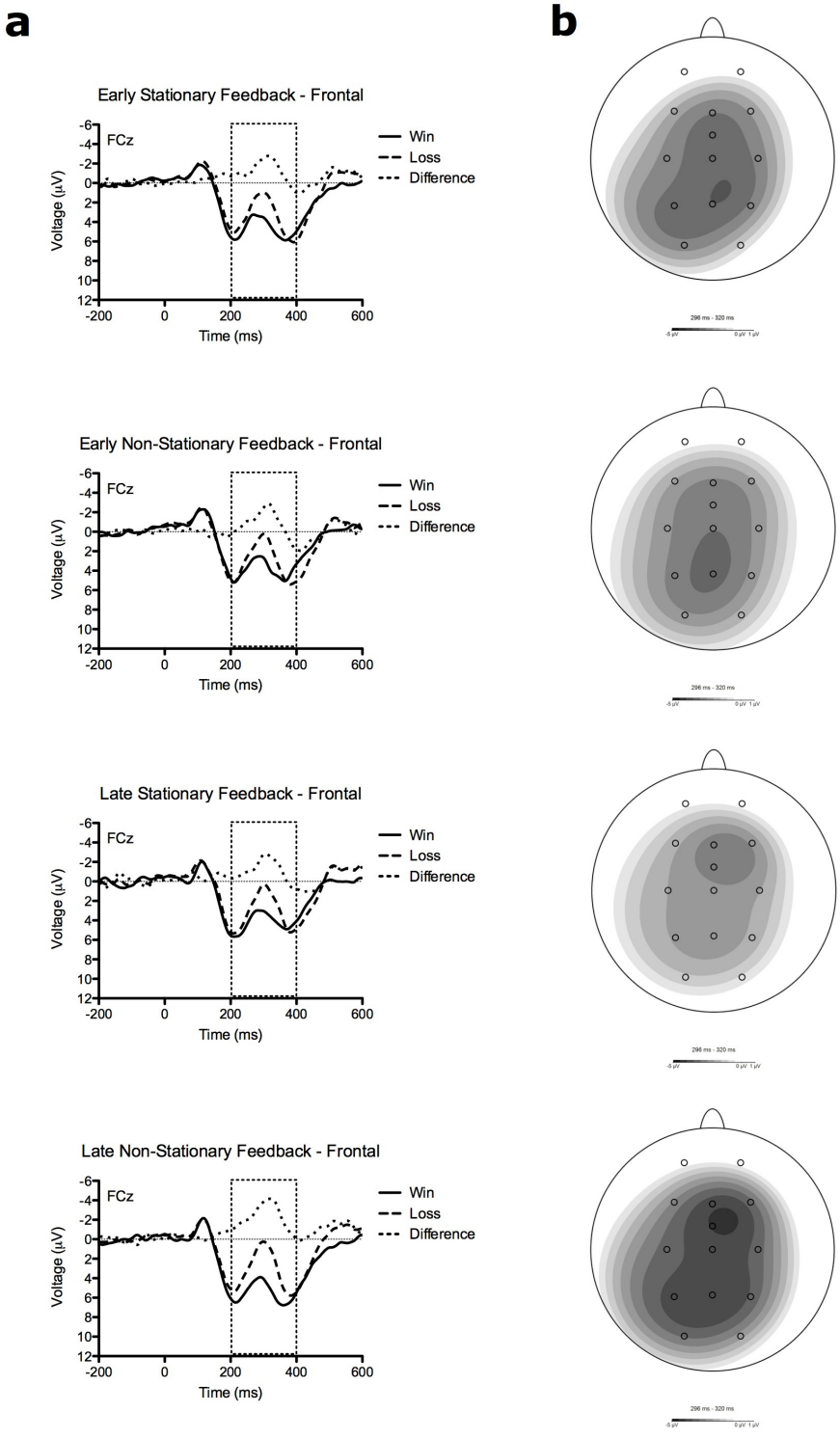

Figure 2.2. Average medial-frontal response to feedback (a) with scalp topographies (b). Context shifts only occurred in late non-stationary blocks. The dashed rectangle shows the interval of analysis (200 – 400 ms post feedback).

Based on an observation of the average combined response to all wins and losses at parietal electrode sites (Appendix, Figure A3), the P300 was defined as the average voltage 300-500 ms post feedback. In line with previous work, the analysis was done at electrode site Pz because this is where the P300 of the average of all wins and losses was maximal (Polich & Margala, 1997; Nieuwenhuis et al., 2005; Wu & Zhou, 2009). See Figure 2.4 for the P300 scalp topography. A P300 was defined for each feedback type (win/loss), learning phase (early/late), and environment (stationary/non-stationary). Figure 2.3 shows the average waveforms for these conditions.

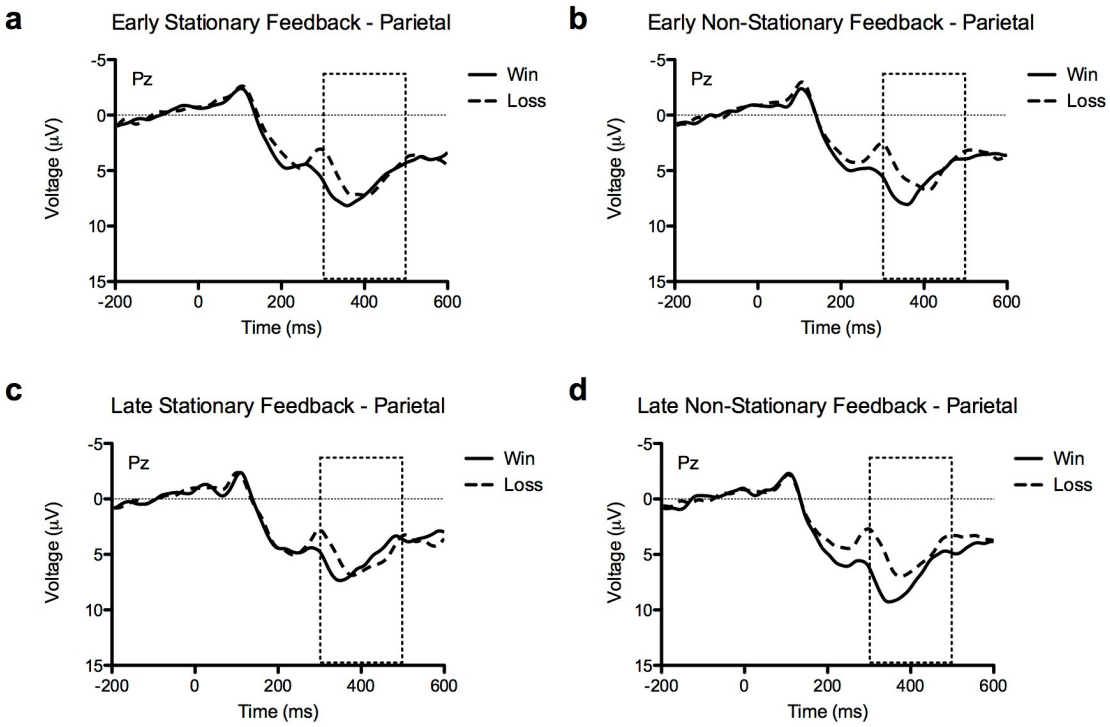

Figure 2.3. P300 response to feedback early in learning (a and b) and late in learning (c and d) for each environment type. The grand average (all conditions pooled) was maximal at Pz. A dashed rectangle indicates the region of analysis (300 – 500 ms post feedback).

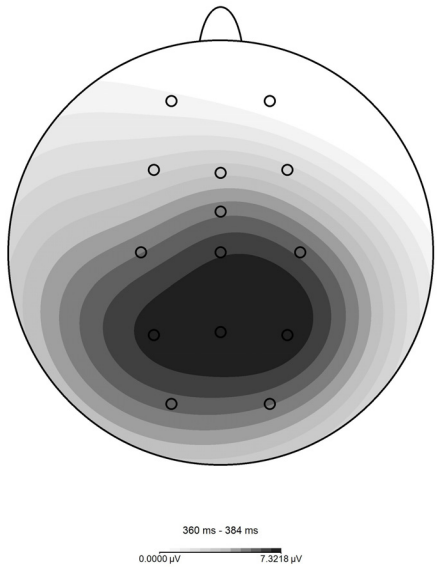

Figure 2.4. P300 scalp topography for the average response to all feedback, maximal at Pz.

The fERN peaks were compared using a 2 (environment: stationary, non-stationary) by 2 (learning phase: early, late) repeated-measures ANOVA. P300 peaks were compared using a 2 (environment: stationary, non-stationary) by 2 (learning phase: early, late) by 2 (feedback: win, loss) repeated-measures ANOVA. For the P300 ANOVA, only significant effects were reported. As with the behavioural data, an alpha level of .05 was assumed for all statistical tests, and all error measures represented one standard error.

2.3 Results

2.3.1 Accuracy

Mean accuracies with standard errors are presented in Table 2.1. There was an effect of time ($F(1,19) = 9.536$, $p = .006$): mean accuracies increased later in a block compared to earlier in a block as participants learned the correct response. Accuracies

were also worse in the non-stationary environment ($F(1,19) = 28.156$, $p < .001$) compared to the stationary environment. Finally, there was an interaction between time and environment ($F(1,19) = 24.396$, $p < .001$): accuracies became worse in the non-stationary environment, following a context shift, but better in the stationary environment (See Figure 2.6a). The mean adjusted p(win) values were .67 +/- .01 and .14 +/- .01. Due to these adjustments to the p(win) values, there was no difference in the total number of wins and losses $t(19) = 0.3951$, $p = .3486$.

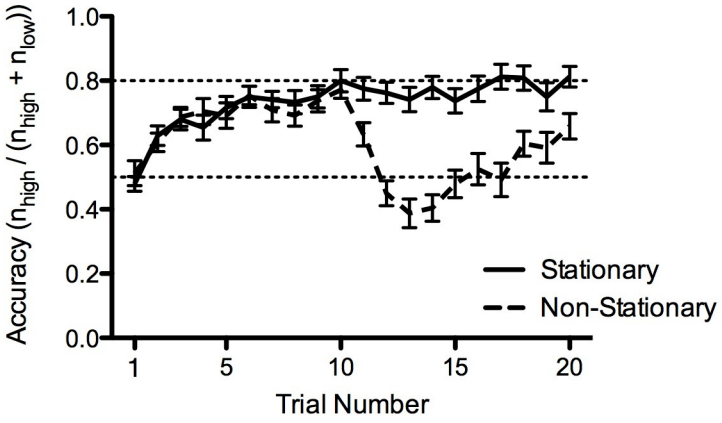

Figure 2.5. Performance Curve. In the non-stationary environment, the optimal choice switched on trial 12, on average. Dashed lines are shown at 50% (chance) and 80%, for reference.

2.3.2 Response Time

Mean response times with standard errors are presented in Table 2.1. There was a significant effect of time ($F(1,19) = 11.240$, $p = .016$): participants were slower to respond in the second half of a block. There was no effect of environment ($F(1,19) = 1.552$, $p = .228$) and no time/environment interaction ($F(1,19) = 2.566$, $p = .126$. See Figure 2.6b.

Table 2.1. Experiment 1: Behavioural means and standard errors.

		Early		Late	
Variable	Environment	Mean	SE	Mean	SE
Accuracy (%)	Stationary	70	2	77	3
	Non-Stationary	70	2	53	0.7
Response Time (ms)	Stationary	447	32	473	36
	Non-Stationary	444	31	458	35

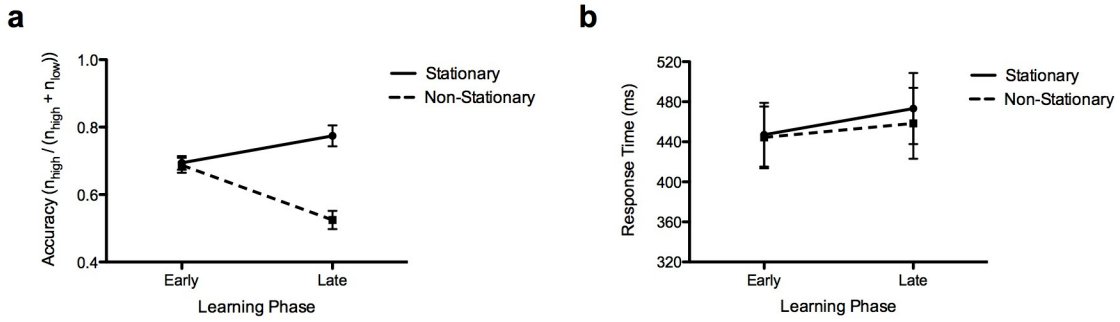

Figure 2.6. Behavioural results. (a) Mean performance early and late in a block for each environment. (b) Mean response time early and late in a block for each environment.

2.3.3 The fERN

An analysis of difference waves (losses minus wins) locked to the onset of feedback revealed ERP components with latencies and scalp distributions (maximal at FCz) consistent with a fERN, both early and late in learning, and for each environment type (see Figure 2.2).

A repeated-measures ANOVA of the fERN peaks revealed a main effect of time ($F(1,19) = 7.967$, $p = .011$): the fERN was enhanced later in a block compared to earlier

in a block. There was no effect of environment ($F(1,19) = 0.241$, $p = .629$) and no time/environment interaction ($F(1,19) = 1.011$, $p = .327$). See Figure 2.7 and Table 2.2.

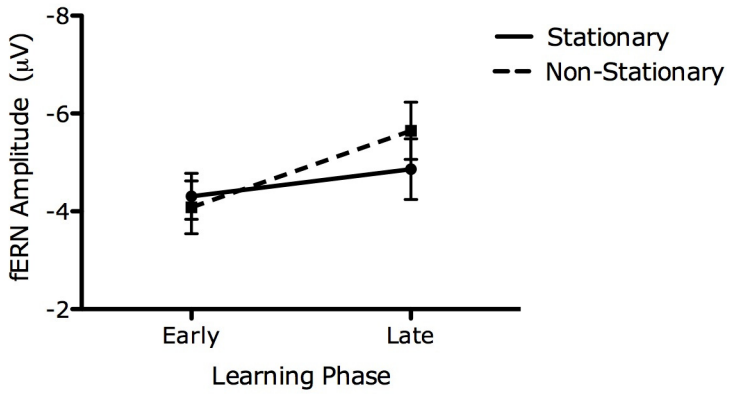

Figure 2.7. Mean fERN amplitudes, early and late in learning, for each environment.

Table 2.2. Experiment 1: Means and standard errors for the fERN.

	Early		Late	
Environment	Mean	SE	Mean	SE
Stationary (μV)	-4.3	0.5	-4.9	0.6
Non-Stationary (μV)	-4.1	0.5	-5.6	0.6

2.3.4 The P300

An analysis of mean amplitude in the P300 time range (300 – 500 ms) at a posterior electrode site (Pz) revealed interactions between time and environment ($F(1,19) = 7.173$, $p = .015$) and between feedback and environment ($F(1,19) = 5.922$, $p = .025$). As shown in Figure 2.8, the P300 was enhanced later in learning for non-stationary blocks compared to stationary blocks. Furthermore, the P300 for win trials was enhanced

relative to loss trials in the non-stationary blocks, but not the stationary blocks. See Table 2.3 for exact P300 amplitudes.

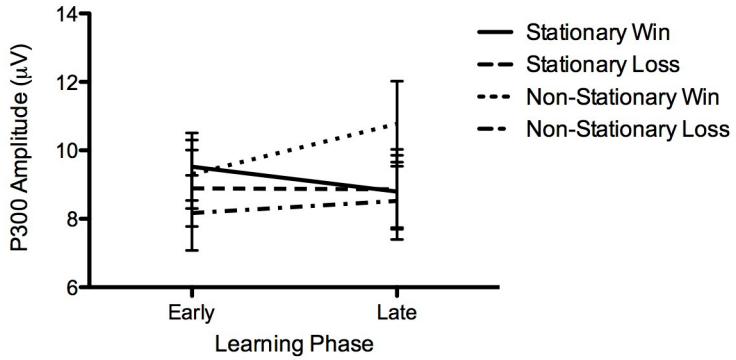

Figure 2.8. P300 amplitudes in response to feedback. When context shifts occurred, there was an enhanced P300 response to wins.

Table 2.3. Experiment 1: Means and standard errors for the P300.

		Early		Late	
Feedback Type	Environment	Mean	SE	Mean	SE
Win (μV)	Stationary	9.5	1.0	8.8	1.1
	Non-Stationary	9.3	1.0	10.8	1.3
Loss (μV)	Stationary	8.9	1.1	8.9	1.2
	Non-Stationary	8.1	1.1	8.5	1.1

2.4 Discussion

The goal of this experiment was to determine whether or not low feedback validity (high expected uncertainty) during learning leads to an enhancement of RL prediction errors generated within medial-frontal cortex. Participants in this study were able to learn optimal responses in a simple gambling game in which the optimal response

could change. Two different ERP responses to feedback were measured and analyzed: the P300 and the fERN.

As predicted, using reduced feedback validity resulted in a failure to reproduce Bland and Schaefer's (2011) observation of an enhanced fERN following context shifts. Thus, even though participants here were able to learn optimal responses, they did not show an enhanced fERN in the non-stationary environment relative to the stationary environment. However, there was a main effect of time such that the fERN was enhanced later in each block, regardless of environment. In the non-stationary environment, this is unsurprising since the context shifts that occurred later in a block were characterized by unexpected losses, and should therefore have elicited a larger prediction error relative to earlier in a block. That an enhanced fERN was mirrored in the stationary environment, as predicted, may be surprising to some given that the fERN is typically reduced with learning (Krigolson et al., 2008). More specifically, as feedback became expected – as the wins and losses became less surprising – there should have been less of a prediction error later in learning (Sutton & Barto, 1998). Thus, assuming that the fERN is a RL prediction error (Holroyd & Coles, 2002), the fERN should have diminished over time in the stationary environment, but not the non-stationary environment.

One possible explanation for this surprising fERN result is that due to the high expected uncertainty in this task, participants were prone to falsely detecting context shifts in the stationary environment. If this were the case here, then one would expect the fERN in the stationary environment to resemble the fERN in the non-stationary environment (which it did). Thus, not only did the fERN in the non-stationary environment fail to decrease with learning, as predicted, but it also closely matched the

fERN in the non-stationary environment, making any comparison between these quantities problematic in this design, as well as many others (O'Doherty et al., 2003; Hornak et al., 2004; Behrens et al., 2007; Chase et al., 2010; Bland and Schaefer, 2011).

This hypothesis – that low feedback validity made stationary environments more like non-stationary environments – is consistent with the response time data: participants were slower to respond later in a block compared to earlier, in both environments. This slowing of response times suggests that participants came to deliberate over their responses in both environments, i.e. that even in the stationary environment they were unsure about which response was optimal later in a block.

As the performance curve for this task showed, however, participants were able to detect and adapt to the change in outcome probabilities. Thus, while midbrain reward processing as indexed by the fERN may be important for action selection, and may also contribute to uncertainty detection in general, there is most likely some other mechanism helping to detect unexpected uncertainty. A clue about this other mechanism may be that, while there was no fERN difference between environments, there was a P300 difference. Specifically, the P300 in response to wins (but not losses) was enhanced in the non-stationary environment following context shifts. A context-updating interpretation (Donchin, 1981; Donchin & Coles, 1988) of this enhancement would suggest that feedback following a context shift was used to both signal that a change had occurred (in the case of losses), and that the new context had been determined (in the case of wins). Consistent with this suggestion, the P300 enhancement could also reflect an increase in phasic NE in response to the detection of context shifts in the non-stationary environment (Nieuwenhuis et al., 2005).

It is unclear why the P300 to losses would not also be enhanced in the non-stationary environment following a context shift, especially given that a sequence of losses should signal the shift itself. One speculation is that, for some reason, losses were less salient compared to wins in this task. A possible explanation for this is that just as stationary environments became more like non-stationary environments, participants tended to confuse non-stationary environments for stationary environments. That is, participants may have confused the actual context shifts in non-stationary environments for the random feedback variations that occurred in stationary environments. If this was the case here, then it is reasonable to suggest that participants may have tended to ignore (initially) the losses that signalled a context shift, and instead focus more on wins, i.e. that wins were more salient because they were more informative (Nieuwenhuis et al., 2005).

In conclusion, the results presented here suggest that when feedback validity is low, midbrain reward processing may not be solely responsible for detecting context shifts. In particular, uncertainty about the identity of the environment may make the accurate detection of context shifts challenging. Experiment 2 will attempt to validate this claim by providing cues to participants as to which environment they are in.

CHAPTER 3: EXPERIMENT 2

3.1 INTRODUCTION

The results from Experiment 1 suggest that low feedback validity impacts midbrain reward processing such that the fERN in stationary environments closely resembles the fERN in non-stationary environments. This surprising outcome may have been due to participant difficulty in detecting context shifts when they occurred in the non-stationary environment, or to mistakenly believing that context shifts had occurred in the stationary environment (i.e. mistaking expected uncertainty for unexpected uncertainty). Put another way: when a context shift can occur at any time (in any block of an experiment), then one could argue that the entire experiment is a non-stationary environment. This could pose serious challenges to any attempt to make stationary/non-stationary comparisons within such a design, yet to date this is precisely what most research on unexpected uncertainty has done (Yu & Dayan, 2005; Behrens et al., 2007; Chase et al., 2010; Bland & Schaefer, 2011), and it was what was done in Experiment 1.

Thus, in order to fully investigate the neural basis of learning in non-stationary environments, one should contrast a known non-stationary environment with a known stationary environment, i.e. environments in which participants are aware that a context shift either could or could not occur. The reasoning behind this claim is based on the idea of NE modulation in the ACh-NE system of uncertainty detection (Yu and Dayan, 2005; Bouret and Sara, 2005). NE, according to Yu and Dayan (2005), indexes unexpected uncertainty. In other words, NE increases at times when the current belief about the environment (e.g. which option is best) is in doubt, and decreases when that belief is firm – an inverse measure of confidence. Thus, in Yu and Dayan's (2005) model, context

updates are more likely when NE levels are raised (confidence is low) and less likely when NE levels are lowered (confidence is high). This modelling result is also seen in the laboratory. For example, monkeys whose overall level of NE is increased through pharmacological intervention are more likely to switch their response strategy compared to controls (Steere & Arnsten, 1997). Specifically, Steere and Arnsten (1997) observed that monkeys injected with guanfacine, which increases the firing rate of LC neurons and raises NE levels, were more likely than controls to detect reward contingency reversals in a reward-based learning task (also see Devauges & Sara, 1990). Thus, increasing NE via the LC-NE system may increase performance in uncertain environments by improving reward reversal detection. The idea that NE modulation may play a role in different environments will be explored further in Chapter Four, in which human performance in the present experiment and in Experiment 1 will be simulated.

The goal of this experiment was therefore to examine learning in known stationary and non-stationary environments. To do this, participants played the same gambling game as in Experiment 1, with cues added at the beginning of each block that indicated whether a block was stationary or non-stationary. Similar to how experienced sailors know that the weather is more unpredictable at certain times of the year, participants here could learn to identify those blocks in which a reward reversal could occur. It was predicted that the addition of environmental cues would not only lead to an overall improvement in performance (in both environments), but also an observed difference in medial-frontal reward processing in each environment, as indexed by the fERN. Specifically, it was hypothesized that the fERN would be enhanced later in a block in the non-stationary environment (i.e. after the context shift occurred, when losses were

unexpected), and reduced later in a block in the stationary environment (as losses became expected). This prediction was based on previous work implicating ACC in uncertainty detection (Behrens et al., 2007; Chase et al., 2010; Bland & Schaefer, 2011). Finally, based on the context-updating hypothesis (Donchin & Coles, 1988) and on previous work on uncertainty (Bland & Schaeffer, 2011), it was predicted that the P300 would be enhanced for both wins and losses later in a non-stationary block, as these events provided the information that a context shift had occurred, and that the correct context had been relearned.

3.2 Method

3.2.1 Participants

Twenty university-aged participants (9 male, mean age: 21+/- 0.6) with no known neurological impairments and with normal or corrected-to-normal vision took part in the experiment. None of the participants had been tested in Experiment 1. All of the participants were volunteers who received credit in an undergraduate course for their participation. The participants provided informed consent approved by the Health Sciences Research Ethics Board at Dalhousie University.

3.2.2 Apparatus and Procedure

The apparatus and procedure here were identical to Experiment 1, except that prior to each block the participants were shown a cue to indicate the environment they were in. Participants were told that they would be playing a gambling game in two different casinos – a dishonest casino, in which the best choice could change partway through a block, and an honest casino, in which the best choice never changed. The

environmental cue consisted of a picture and name unique to that casino. The exact picture and name for each casino type (honest/dishonest) was randomized between participants.

Since participants were not told which casino was the dishonest one, a way to determine if they knew which casino was which was needed. At the end of each block, participants were asked to rate, with a button press, the honesty of the casino they had just played in on a scale from 1 (dishonest) to 5 (honest). Similar to Experiment 1, participants were given the following verbal and written instructions (differences from Experiment 1 have been italicized here):

> In this experiment, you will be playing several slot machine games. During each game, you will pull the arm on one of two coloured slot machines. One of the coloured slot machines is better - selecting it will result in more wins and fewer losses, in the long run. Your goal in this task is maximize the number of points you receive over 24 games. *You will be playing this game in one of two casinos. You will be shown a picture of the casino you are playing in before each set of 20 trials. One of the casinos is dishonest, and may switch which coloured slot machine is the better choice partway through a game. The other casino is honest, and never switches the payouts within a game. At the end of several of the games, you will be asked to rate the casino you just finished playing in. The rating scale goes from 1 (dishonest) to 5 (honest). You will indicate your rating by pressing one of buttons 1-5.* Try to keep your eyes on the + in the centre of the screen throughout the experiment. Wait for the + to change colour before responding. Click the left gamepad button to select the left slot machine. Click the right gamepad button to select the right slot machine. Press LEFT or RIGHT to see a summary.

Participants were also shown the following summary (difference from Experiment 1 italicized):

> - When the + changes colour, choose a slot machine
> - One of the coloured slot machines is the better choice
> - *In one of the casinos, which slot machine is best may change multiple times*
> - Your goal is maximize your total number of points
> - Press LEFT or RIGHT to begin' - Press LEFT or RIGHT to begin

See Figure 3.1 for an overview of the present experiment.

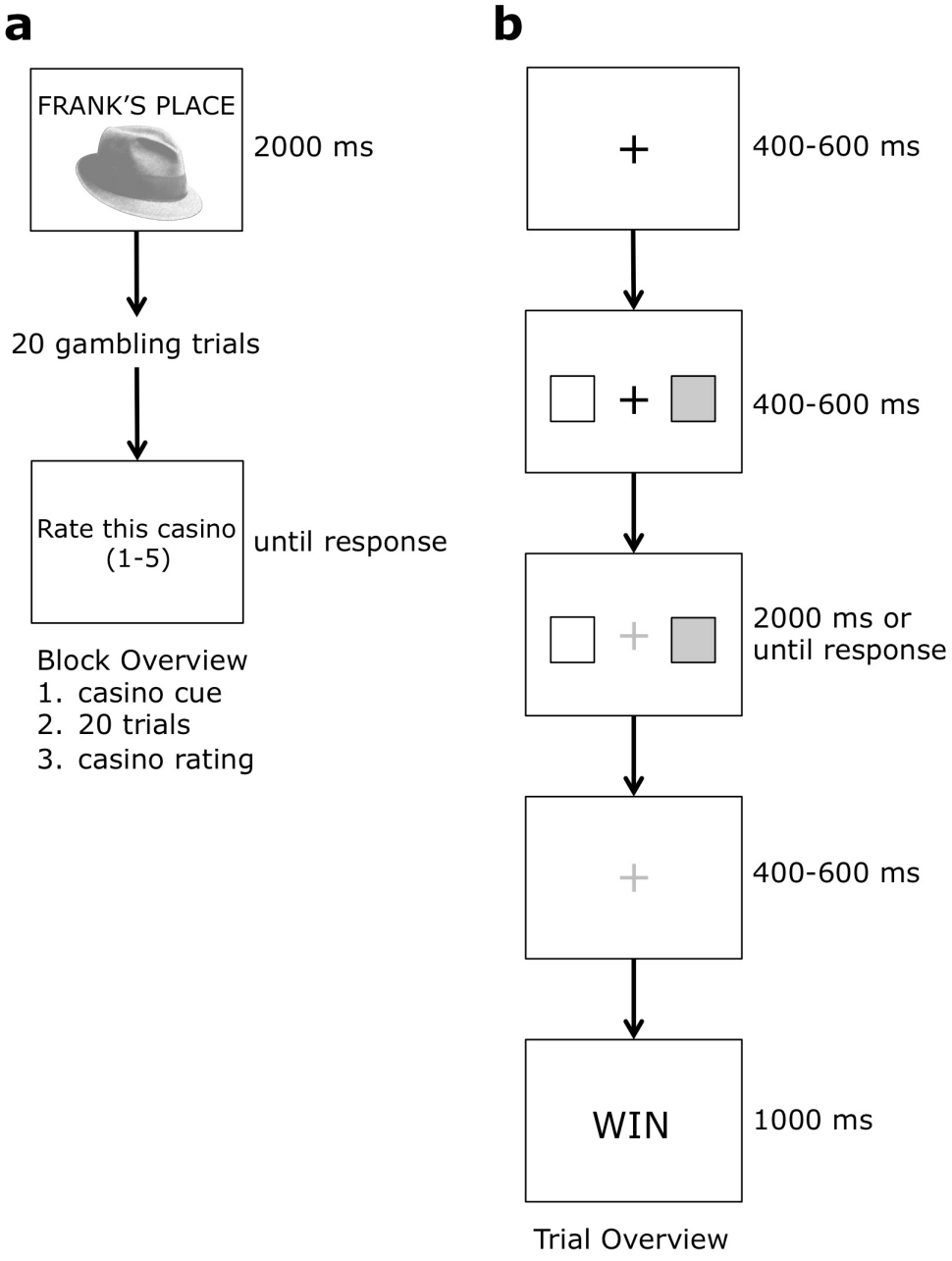

Figure 3.1. Experimental design for (a) blocks and (b) trials, with timing details. At the beginning of each block, participants were shown the casino in which they were gambling. After 20 trials, they were asked to rate the honesty of the current casino.

3.2.3 Data Collection

The experimental program recorded participant responses (slot machine selections and casino ratings) and response times. The EEG was recorded from 64 electrode locations using Brain Vision Recorder software (Version 1.20, Brain Products, GmbH, Munich, Germany). The electrodes were mounted in a fitted cap with a standard 10-20 layout and were recorded with an average reference built into the amplifier (see Figure A4 in the Appendix for the exact electrode configuration). The vertical and horizontal electrooculograms were recorded from electrodes placed above and below the right eye, and on the outer canthi of the left and right eyes. Electrode impedances were kept below 20 kΩ and the EEG data were sampled at 1000 Hz and amplified (Quick Amp, Brainproducts, GmbH, Munich, Germany).

3.2.4 Data Analysis

Behavioural data were analyzed as in Experiment 1. Additionally, means and standard errors of casino ratings were computed for each trial, and mean casino ratings for each environment (stationary/non-stationary) were computed for each participant and compared using a paired-samples t-test in order to determine if participants were aware of which casino was the dishonest casino (i.e. the non-stationary environment). Finally, mean performance in Experiment 1 was compared to mean performance in Experiment 2 using a 2 (environment: stationary, non-stationary) by 2 (Experiment: One, Two) mixed ANOVA. An alpha level of .05 was assumed for all statistical tests, and all error measures represent one standard error.

EEG data were also analyzed and epoched as in Experiment 1. In total, 3% of the data were discarded after artifact rejection. Previous research (fERN: Holroyd & Coles, 2002; Krigolson & Holroyd, 2007; Krigolson et al., 2008 and P300: Polich & Margala, 1997; Nieuwenhuis et al., 2005; Wu & Zhou, 2009), and an examination of the grand average waveforms for both frontal and parietal responses (Appendix, Figures A5 and A6) suggested that the same analyses windows would be appropriate here as in Experiment 1. In particular, as in Experiment 1, the difference wave (losses – wins) of the grand average response to feedback in the ERN time range (200 – 400 ms post feedback) was maximal at FCz (see Figure 3.2). Finally, the overall average (all wins and losses combined) was maximal in the P300 time range (300 -500 ms post feedback) at electrode site Pz (See Figures 3.3 and 3.4).

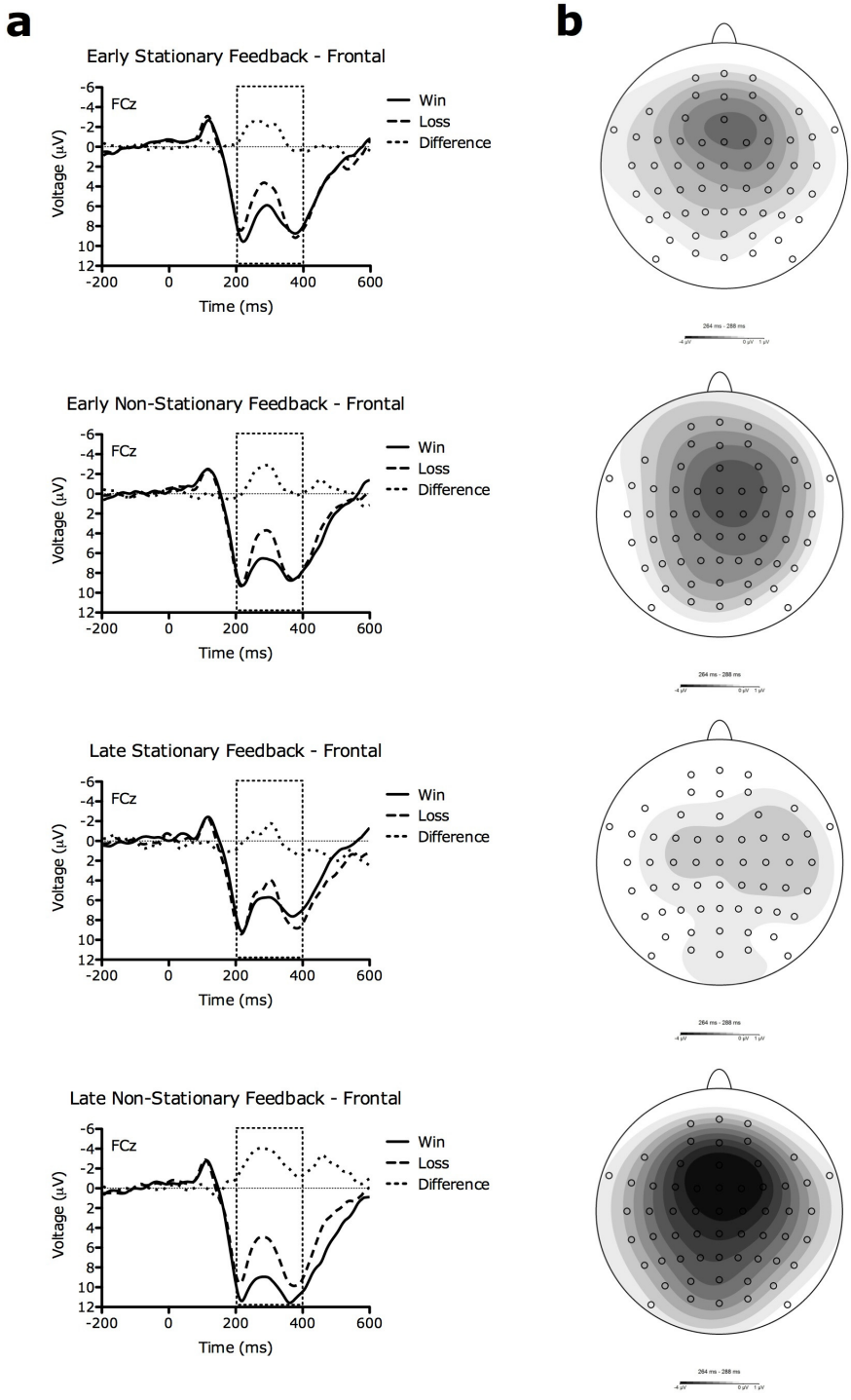

Figure 3.2. Grand average (a) waveforms and (b) scalp topographies in response to feedback. Context shifts only occurred late in non-stationary blocks. The dashed rectangle shows the interval of analysis (200 – 400 ms post feedback). Note that the scales for the scalp topographies are identical (black = -4 μV, white = 1 μV).

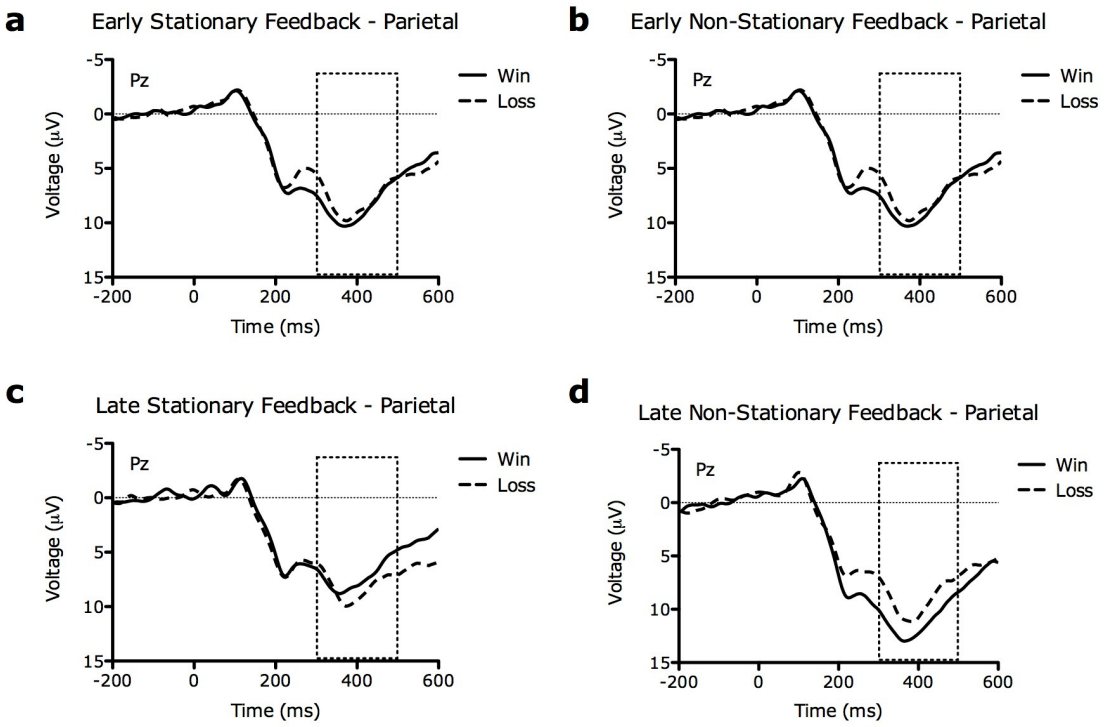

Figure 3.3. P300 response to feedback early in learning (a and b) and late in learning (c and d) for each environment type. A dashed rectangle shows the region of analysis: 300 – 500 ms post feedback.

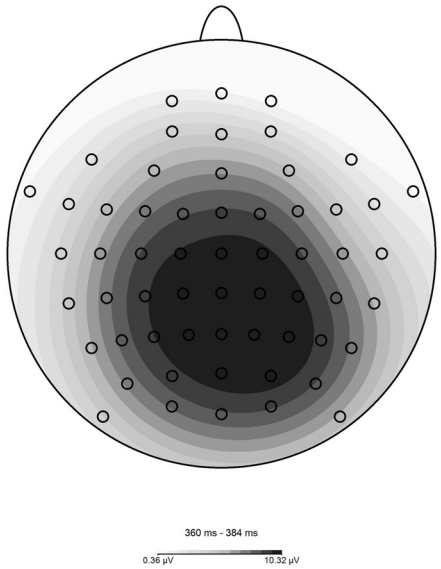

Figure 3.4. P300 scalp topography for the average response to all feedback, maximal at Pz.

3.3 Results

3.3.1 Accuracy

The mean accuracies are presented in Table 3.1. There was a significant effect of environment ($F(1,19) = 51.879$, $p < .001$). Performance was worse in the non-stationary environment compared to the stationary environment. There was also a significant interaction between time and environment ($F(1,19) = 106.463$, $p < .001$): accuracies became worse in the non-stationary environment, but better in the stationary environment (See Figure 3.6c). The mean adjusted p(win) values were .63 +/- .01 and .12 +/- .004. As in Experiment 1, due to these adjustments to the p(win) values, there was no difference in the total number of wins and losses, $t(19) = 0.4451$, $p = .6612$. Finally, there was a significant main effect of both environment ($F(1,38) = 76.72$, $p < .001$) and Experiment

($F(1,38) = 19.13$, $p < .001$): performance was better in the present experiment, regardless of environment type (Figures 3.5 and 3.6)

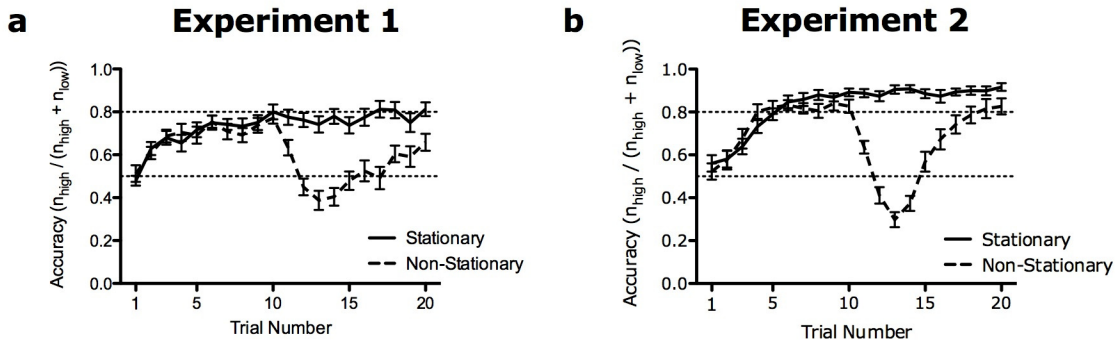

Figure 3.5. Performance curve in (a) Experiment 1, and (b) the present study. In the non-stationary environment, the optimal choice switched on trial 12, on average. Dashed lines are shown at 50% (chance) and 80%, for reference.

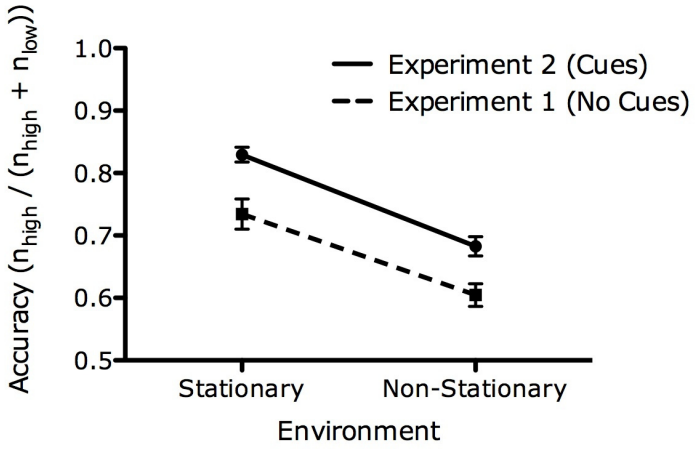

Figure 3.6. Mean accuracies, for each environment, for the current experiment and Experiment 1. Adding environment cues improved performance regardless of environment type

3.3.2 Response Time

Mean response times are presented in Table 3.1. There was no significant effect of time ($F(1,19) = 0.220$, $p = .540$). There was also no significant effect of environment ($F(1,19) = 0.389$, $p = .228$) and no significant time/environment interaction ($F(1,19) = 0.155$, $p = ..699$. See Figure 3.6d.

Table 3.1. Experiment 2: Behavioural means and standard errors.

		Early		Late	
Variable	Environment	Mean	SE	Mean	SE
Accuracy (%)	Stationary	80	1.8	89	0.2
	Non-Stationary	75	1.8	61	2.5
Response Time (ms)	Stationary	373	25.7	377	26.4
	Non-Stationary	372	23.9	373	28.4

3.3.3 Environment Ratings

See Figure 3.7b. Mean ratings for each trial suggested that participants were able to distinguish the casinos from trial one. Participants rated the non-stationary casino significantly lower (more dishonest) than the stationary casino, $t(1,19) = 7.88$, $p < .001$ (Figure 3.7a).

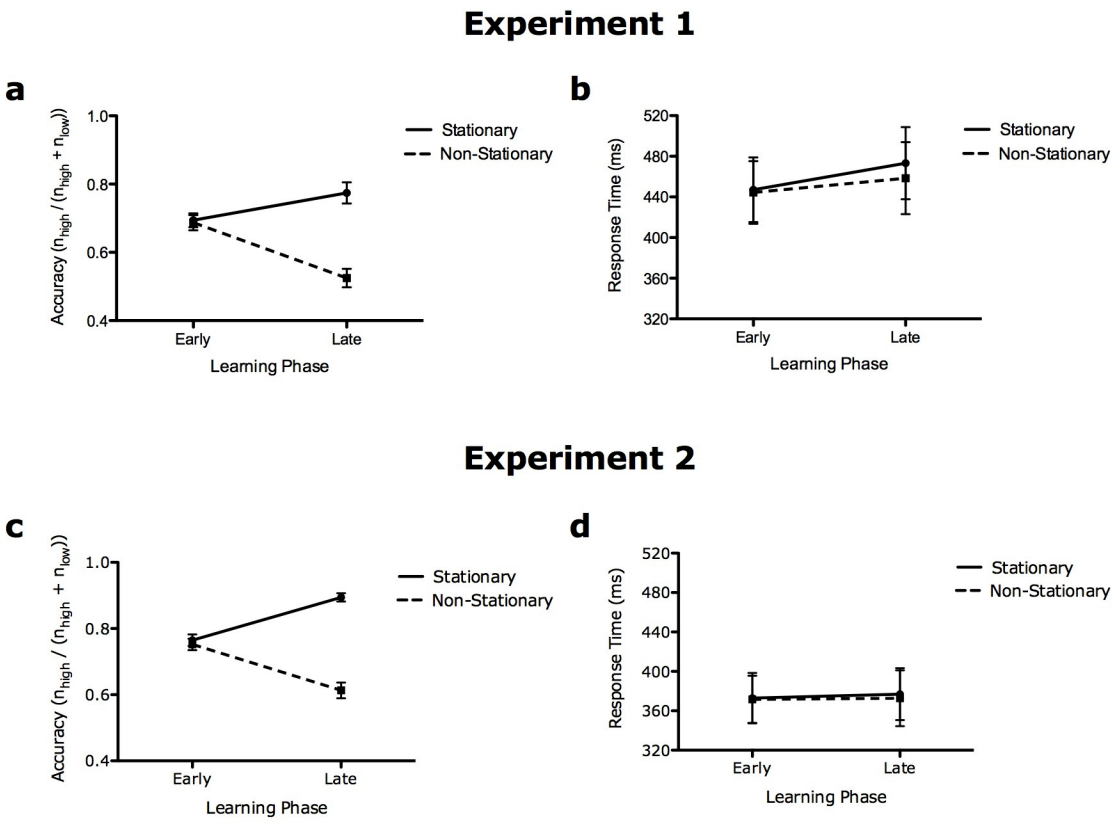

Figure 3.7. Behavioural results. Mean performance (a) and response time (b) in Experiment 1 and (c, d) Experiment 2.

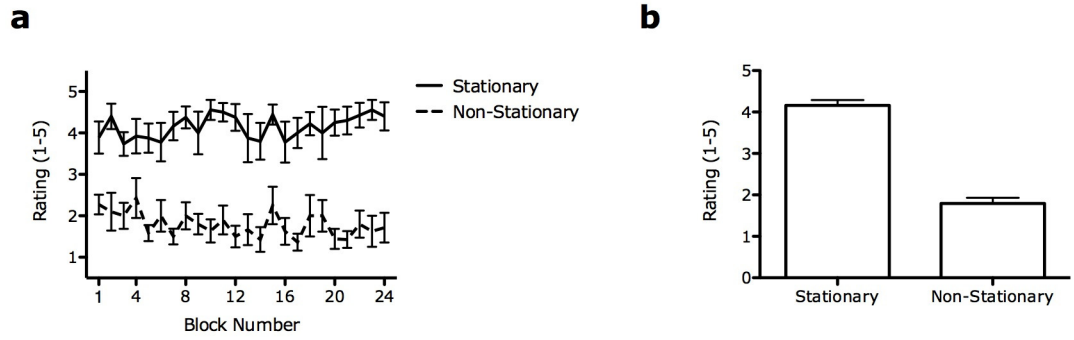

Figure 3.8. Casino ratings (a) over all trials and (b) grouped by environment type. Participants were able to discriminate the environments based on the feedback they received throughout a block.

3.3.4 The fERN

An analysis of difference waves (losses minus wins) locked to the onset of feedback revealed ERP components with latencies and scalp distributions (maximal at FCz) consistent with a fERN, both early and late in learning, and for each environment type (see Figure 3.2).

There was a main effect of environment on the magnitude of the fERN ($F(1,19) = 6.145$, $p = .023$) and a significant time/environment interaction ($F(1,19) = 7.489$, $p = .013$). The fERN was greater in the stationary environment compared to the non-stationary environment. Furthermore, the fERN increased later in a block in the non-stationary environment, but decreased later in a block in the stationary environment. There was no main effect of time ($F(1,19) = 0.004$, $p = .949$). See Figure 3.8 and Table 3.2.

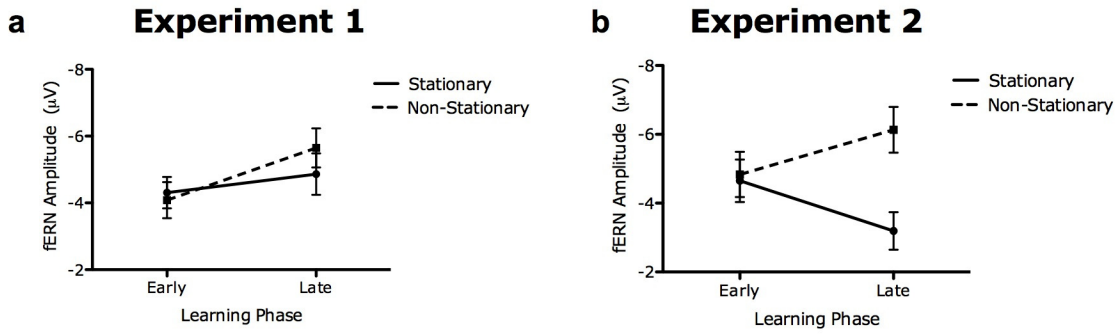

Figure 3.9. Mean fERN amplitudes, early and late in learning, for each environment in (a) Experiment 1, and (b) the present study. In the present experiment, the fERN was enhanced over time in the non-stationary environment, but reduced in the stationary environment.

Table 3.2. Experiment 2: Means and standard errors for the fERN.

	Early		Late	
Environment	Mean	SE	Mean	SE
Stationary (μV)	-4.4	0.6	-3.1	0.6
Non-Stationary (μV)	-4.7	0.6	-5.6	0.6

3.3.5 The P300

There was an interaction between time and environment ($F(1,19) = 19.583$, $p < .001$) on P300 amplitude: the P300 was enhanced later for non-stationary, but not stationary, environments. There was also an interaction between feedback and environment ($F(1,19) = 13.401$, $p = .002$): wins elicited a larger P300 than losses in the non-stationary, but not stationary, environments. Additionally, there were main effects of time ($F(1,19) = 8.027$, $p = .01$) and environment ($F(1,19) = 26.999$, $p > .001$). As in Experiment 1, later blocks elicited a larger P300, and the P300 was enhanced for non-stationary environments compared to stationary environments (Figure 3.10 and Table 3.3).

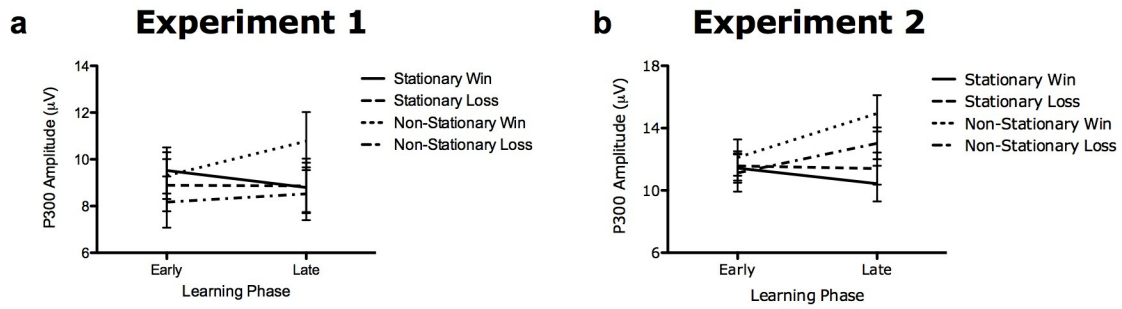

Figure 3.10. P300 amplitudes in response to feedback in (a) Experiment 1, and (b) the current experiment. When context shifts occurred, there was an enhanced P300 response to wins and losses.

Table 3.3. Experiment 2: Means and standard errors for the P300.

		Early		Late	
Feedback Type	Environment	Mean	SD	Mean	SD
Win (μV)	Stationary	11.4	0.9	10.4	1.1
	Non-Stationary	12.1	1.2	14.9	1.2
Loss (μV)	Stationary	11.6	0.9	11.4	1.0
	Non-Stationary	11.1	1.2	13.0	1.0

3.4 Discussion

The goal of this study was to try to determine the effect of cues on learning under uncertainty. Participants learned, over 20 trials, which of two slot machines was more likely to lead to a win when selected. Prior to each block, participants were shown a cue indicating that the casino they were playing in was either non-stationary (the best response could change) or stationary (the best response never changed). The behavioural

data indicated that participants were able to discern which casino was the non-stationary casino, and were also able to determine the optimal response in both environments.

The hypothesis regarding the fERN was supported. The fERN is thought to reflect a RL prediction error (Holroyd and Coles, 2002). Therefore, unexpected wins and losses should elicit larger prediction errors compared to expected wins and losses. In the known stationary environment, with learning, the fERN diminished as wins and losses became more expected. In the known non-stationary environment, the fERN was enhanced later in a block – the losses (and wins, eventually) experienced following a context shift were unexpected, and therefore resulted in a greater prediction error.

As with the fERN results, an environment-dependent difference in the amplitude of the P300 component was also observed. The feedback-locked P300 for both wins and losses was enhanced later in non-stationary blocks compared to stationary blocks. This is consistent with the context-updating account of the P300 (Donchin, 1981; Donchin & Coles, 1988), which holds that a P300 is observed whenever new information causes an update to one's internal model of the world. This includes information about the probabilistic structure of a task (Donchin & Coles, 1988). In order to maximize their rewards, participants here had to detect and adapt to probabilistic reversals. Presumably, the observed P300 enhancement marks this detection.

Consistent with a context-updating account is the explanation by Nieuwenhuis et al. (2005) that the P300 is modulated by phasic activity of the LC-NE system. Nieuwenhuis et al. (2005) proposed that the LC regulates exploratory behaviour in humans, i.e. breaking away from one behaviour in favour of another, through the release of NE. The change from one mode of control (automatic) to another (exploratory) is

marked by an increase in the P300. Thus, phasic NE and the P300 serve as a "neural interrupt signal" (Dayan & Yu, 2006; Doya, 2008), breaking away from previously rewarding behaviour so that new actions may be explored. In Chapter Four, the idea of NE modulation will be investigated further by modelling the human behavioural and EEG results presented here and in Experiment 1.

CHAPTER 4: SIMULATION

4.1 INTRODUCTION

Simulations of human behaviour have provided insight into the mechanisms behind how we learn and how we adapt to uncertain environments. One of the major benefits of computational modelling as a research tool is that it forces us to be explicit about not only the mechanisms, but also the exact computations, behind cognitive theories. Furthermore, modelling helps us to explain existing results and to make predictions for new experiments.

One explanation for how we detect uncertainty in the world is via Bayesian inference. In this view, new evidence is observed (e.g. wins and losses), and beliefs about outcomes are updated (e.g., that a win is 80% likely). Under unexpected uncertainty, outcome probabilities can change, and there may be a performance benefit to detecting these changes. Yu and Dayan's (2005) model of uncertainty detection estimated both expected and unexpected uncertainty over time and used these quantities to determine the likelihood that a probability change, or context shift, had occurred (also see Payzan-LeNestour & Bossaerts, 2011). In particular, Yu and Dayan (2005) hypothesized that expected uncertainty and unexpected uncertainty are tracked, over time, via the neurotransmitters acetylcholine (ACh) and norepinephrine (NE), respectively (also see: Doya, 2008). ACh seems to be specifically modulated in tasks that manipulate expected uncertainty (e.g. Sarter & Bruno, 1997). Likewise, NE seems to be specifically modulated in tasks involving unexpected uncertainty (e.g. Bouret and Sara, 2005). Based on evidence implicating different roles for ACh and NE in dealing with uncertainty, Yu and Dayan (2005) described an ACh-NE system for explicit uncertainty detection.

In contrast, another type of model has uncertainty detection implicitly built in: reinforcement learning (RL: Sutton & Barto, 1998). In RL models, rewards and punishments are used to strengthen (or weaken) associations between situations and actions. An action that previously lead to a reward is more likely to be repeated because the association between it and a given situation is stronger relative to the association between it and other possible actions. The strength of these associations, or connections, can be interpreted as a belief that taking a particular action will lead to a particular outcome – in other words, association strength may describe expected uncertainty. Unexpected uncertainty, however – the likelihood of a context shift from one reward rule to another – is not explicitly captured by RL models. Instead, standard RL models use all rewards and punishments to modify the weights associated with actions, including feedback signalling a context switch – for example, a series of losses. While unexpected uncertainty is not computed separately from expected uncertainty in RL models, if enough punishments are received for taking a certain action, then that action will be less likely to be repeated in the future. Thus, the preferred action may change within RL models, achieving the (presumed) goal of unexpected uncertainty detection in action selection. Taking this notion even further, others (Alexander and Brown, 2011) have augmented RL models with explicit uncertainty detection, e.g. by modulating learning rate based on the current level of uncertainty.

In summary, there are two possible models to account for uncertainty detection. Explicit models of uncertainty detection (Yu & Dayan, 2005; Payzan-LeNestour & Bossaerts, 2011) compute separate values for unexpected and expected uncertainty. RL models, in contrast, combine all uncertainties into a set of weights used to guide actions.

Systems implementing these two models are hypothesized to exist within humans (Holroyd & Coles, 2002; Yu & Dayan, 2005), which raises questions about both redundancy and interaction. While early work on unexpected uncertainty (Yu & Dayan, 2005) did not address the role of uncertainty detection in action selection, recent models (Payzan-LeNestour & Bossaerts, 2011) have begun to compare the actions predicted by different uncertainty detection models with the actions predicted by RL-only models. In particular, Payzan-LeNestour & Bossaerts (2011) modelled human responses in a 6-armed bandit task (determining which of 6 options was most likely to yield a reward) and found that, in some cases, the output of a Bayesian model matched human responses more closely than a RL model. The Payzan-LeNestour & Bossaerts (2011) model included representations for both expected and unexpected uncertainty but, unlike the Yu and Dayan (2005) model, included an action selection component (softmax: Equation 4.1). While a Bayesian model seemed to produce the best fit for their human participants' responses, Payzan-LeNestour & Bossaerts (2011) also noted that responses became less Bayesian under increasing levels of uncertainty.

Less work has been done to model human data by combining a RL model with a Bayesian component for detecting context shifts. Yu and Dayan's (2005) model included no component for action selection (they modelled cue validity), and Payzan-LeNestour and Bossaerts (2011) were interested in comparing a Bayesian learner to a RL model alone. Also, although they did not include an explicit Bayesian component, Alexander and Brown (2011) simulated ACC activity with a RL model in which the learning rate of the model was adjusted based on the current level of surprise. Thus, unlike Yu and Dayan's (2005) model, recent RL models (Alexander & Brown, 2011; Payzan-LeNestour

& Bossaerts, 2011) do not explicitly detect context shifts. The goal of this experiment, therefore, was to model the human behavioural and EEG data observed in Experiments 1 and 2 by combining a standard RL model (e.g. Sutton & Barto, 1998) with Yu and Dayan's (2005) model for detecting expected and unexpected uncertainty. A combined model offered two advantages over a standard RL model alone. First, a combined model was able to better account for the human performance in Experiments 1 and 2 compared to a model without the ability to detect environmental uncertainty. Second, the model was able to account for the observed differences between Experiments 1 and 2. Specifically, the performance advantage observed in Experiment 2 compared to Experiment 1 may be due, in part, to improved detection of context shifts (more hits, and fewer false alarms). This was tested in the simulation by raising the overall level of NE in the non-stationary environment, and lowering NE in the stationary environment. Finally, The prediction errors generated by the RL component of the model mirrored the fERN measured in Experiments 1 and 2.

4.2 Design

The current model was written in MATLAB by adapting the same experimental code that was used in Experiments 1 and 2. Twenty virtual subjects were created, and each subject performed the same number of trials and blocks as the human participants in Experiments 1 and 2 (24 blocks of 12 trials each). The model incorporated two modules: one for action selection, and one for uncertainty detection. See Figure 4.1 for an overview of the model.

4.2.1 Action Selection: Reinforcement Learning

The model used RL to make decisions in the gambling task. The specifics of the model were based on earlier work by Holroyd and Coles (2002, 2008) as well as general guidelines described by Sutton and Barto (1998). In particular, the model used rewards to guide its actions such that previously rewarded actions were more likely to reoccur. Action selection was based on the relative strength, or weights, of two nodes, $\boldsymbol{w} = [w_1 w_2]$, where w_1 and w_2 are the weights associated with selecting Option 1 and Option 2, respectively.

Similar to the Holroyd and Coles (2008) model, the weights were each initialized to small random values between 0 and 0.01 at the beginning of each block. A softmax activation function was used to determine the probability of selecting each action, i:

$$p_i = \frac{e^{w_i/\tau}}{e^{w_1/\tau} + e^{w_2/\tau}} \tag{4.1}$$

Thus, the softmax function converted weights into action probabilities (Sutton & Barto, 1998) - the likelihood of taking each action. The parameter τ was the temperature, and described the degree to which options with lower weights are explored. In general, a higher temperature results in more exploration, and a lower temperature biases action selection toward heavily weighted options. At action selection, a value layer became active to describe which option was selected $\boldsymbol{v} = [v_1 v_2]$ such that $v_i = 1$ if option i was selected. This allowed a prediction to be generated regarding the total value ($\hat{V}$) of the system (e.g. if a reward was forthcoming) according to $\hat{V} = \boldsymbol{w} \cdot \boldsymbol{v}$.

Following each action, the model received feedback with probabilities similar to what human participants experienced – namely, that selecting the better option resulted in a win 60% of the time, and selecting the worse option resulted in a win only 10% of the

time. Following feedback, a prediction error was computed by comparing the actual reward received, *r*, with the predicted value of the system:

$$\delta = r - \hat{V} \tag{4.2}$$

The prediction error on trial t (δ_t) was used to update the weights according to:

$$\boldsymbol{w}_{t+1} = \boldsymbol{w}_t + \alpha\delta_t\boldsymbol{v}_t \tag{4.3}$$

In other words, only the weight of the currently active unit was updated, e.g. $\boldsymbol{v}_t = [1\ 0]$ if Option 1 was selected and $\boldsymbol{v}_t = [0\ 1]$ if Option 2 was selected. Here, the parameter α represents the learning rate: the degree to which prediction errors alter the weights. The learning rate and temperature parameters were each set to 0.01, which yielded a performance curve that was similar to what was seen in Experiments 1 and 2.

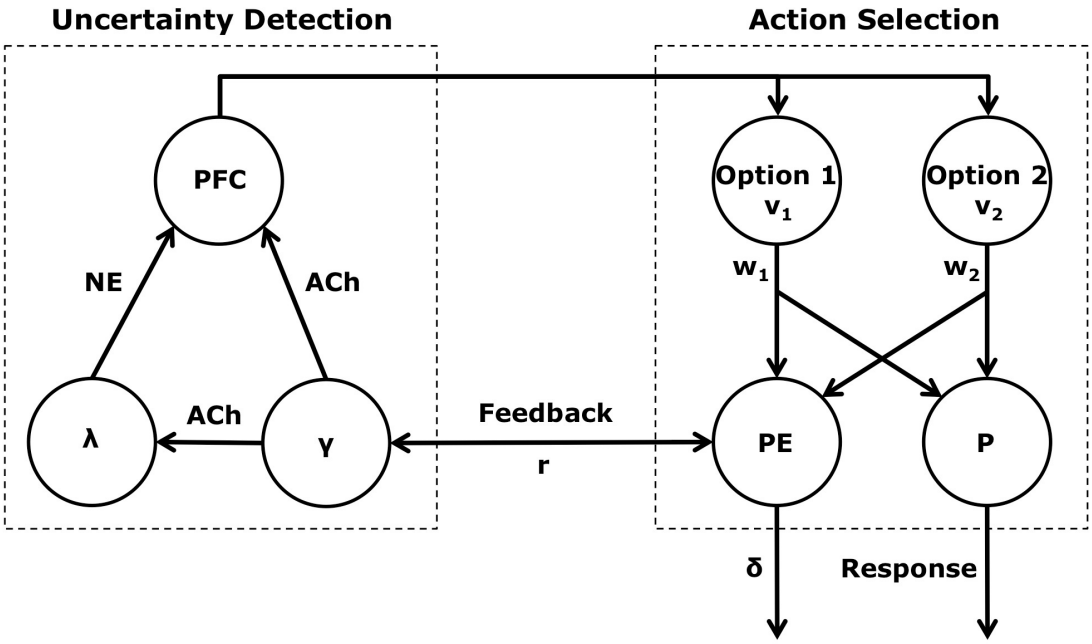

Figure 4.1. Model design. Action selection (right) was made via reinforcement learning. Node weights for each option [w_1 w_2] were used by a softmax function (P) to generate a response. Option weights were also used to compute the current value of the system at action selection, depending on which unit was active (i.e. which option was chosen). A prediction error unit (PE) compared this predicted value with the actual reward value (r) to generate a prediction error (δ). Feedback was also used to detect uncertainty in the prefrontal cortex (PFC) working memory (left). Here, feedback history within a certain context was used to estimate expected uncertainty (γ), which was used to determine the likelihood of sticking with the current belief about which option was best (λ). If a context shift was detected based on these values, then the weights associated with each option were reset.

4.2.2 Uncertainty Detection: ACh and NE

The model also used feedback to estimate both expected and unexpected uncertainty, based on work by Yu and Dayan (2005). Using a history of rewards and punishments, the model computed, on a trial-to-trial basis, the likelihood that the belief about the current context (i.e. which option was best) was correct. Each win and loss either confirmed or disconfirmed this belief. If enough disconfirming information was

received, the model switched its belief to the alternative hypothesis that the other option was better.

Besides requiring a decision (i.e. which option to select), the task that was modelled here differed from that of Yu and Dayan's (2005) in an important way. In the present task there were only two competing hypotheses (that either Option 1 or Option 2 was best), unlike Yu and Dayan's (2005) task, to determine which of several cues was most likely to predict the appearance of a stimulus. Yu and Dayan's (2005) model did not assess each hypothesis simultaneously due to the complexity of such a computation, but it was not unreasonable for the current model to do so. Specifically, Yu and Dayan (2005) considered two competing hypotheses: that the best-known context (belief about which of several cues was valid) was either correct or incorrect. The alternative hypothesis for their model always involved all other possibilities. When a context shift was detected, the Yu and Dayan (2005) model's current hypothesis switched randomly to one of the other possible cues. In the present model, however, there was only ever one other choice, so the current hypothesis could only ever switch from one option to the other.

The present model used information from the environment - in this case, feedback - to compute, for each trial, the likelihood of two competing hypotheses: that a belief about the current context (i.e. the environment) was either correct or incorrect. Since this task involved only two options, the competing hypotheses were that either Option 1 was best ($\mu_t = 1$) or that Option 2 was best ($\mu_t = 2$). Here, μ_t refers to the current hypothesis on trial t. On every trial where the best-known option was chosen (i.e. an exploitation trial) the following estimate of expected uncertainty was computed:

$$\gamma_t^* = \frac{Number\ of\ Wins}{Number\ of\ Trials\ in\ Current\ Context} \tag{4.4}$$

The current context was defined as those trials over which the current belief was held (i.e. since the last context shift). According to Yu and Dayan (2005), this quantity is the inverse of the current level of ACh, i.e. $ACh = 1 - \gamma_t^*$.

Next, the quantity from Equation 4.4 was used to compute the strength of the current belief – an estimate of how likely the current belief was to be true. This was done according to:

$$\lambda_i^* = \frac{P^*(\mu_t=i)}{P^*(\mu_t=i) + P^*(\mu_t \neq i)} + \phi \quad (4.5)$$

The tuning parameter Φ was used to bias the model's current belief, making it more or less certain that the current belief was correct (see section 4.3.2). Here, P^* is an estimate of the likelihood of each hypothesis – in this case, each option. Specifically,

$$P^*(\mu_t = i) = \gamma_t^*(\lambda_{t-1}^*\beta + (1 - \lambda_{t-1}^*)(1 - \beta)) \quad (4.6)$$

$$P^*(\mu_t \neq i) \approx 0.5(\lambda_{t-1}^*(1 - \beta) + (1 - \lambda_{t-1}^*)\beta) \quad (4.7)$$

The parameter β is the probability of sticking with the current hypothesis, and was set to 0.9 in the simulation, reflecting that for the most part we expect contexts to remain the same (Yu & Dayan. 2005). The factor of 0.5 in Equation 4.7 represents an estimate of the validity of the alternative option. In this particular task, the alternative option actually had a validity characterized by p(win) = 0.1 and p(loss) = 0.9, but this was assumed to be unknown to the model.

Thus, following wins on exploitation trials (trials for which the best known option was selected) the model's estimate of the likelihood of the current hypothesis (λ^*) increased, while its estimate of unexpected uncertainty ($NE = 1 - \lambda^*$) decreased. Conversely, following exploitation losses – losses in which the best-known option was selected – belief in the current hypothesis decreased, the estimate of unexpected

uncertainty (NE) increased, and a decision had to be made: whether to stick to the current hypothesis, or abandon it for the alternative option. Similar to Yu and Dayan's (2005) model, this involved comparing $P^*(\mu_t = i)$ to $P^*(\mu_t \neq i)$, which they hypothesized happens in prefrontal working memory (PFC node in Figure 4.1). If $P^*(\mu_t = i) > P^*(\mu_t \neq i)$, then no context shift was detected and the current belief was held. If $P^*(\mu_t = i) < P^*(\mu_t \neq i)$, then the current belief ($\mu_t = i$) was abandoned for the alternative belief ($\mu_t \neq i$). Following context updates, the weights were reset to random values between 0 and 0.01, and the context probability estimate (γ_t^*) was set to 0, as at the beginning of each block.

4.3 Data Analysis

4.3.1 Simulating Experiment 1

In order to assess the benefit of explicit uncertainty detection in a RL model, mean accuracies were computed for 20 simulated participants for each environment type (stationary/non-stationary) using both a RL model alone, and a RL model augmented by uncertainty detection (RL+UD). All other measurements taken when simulating Experiment 1 were done on the combined RL+UD model. As in Experiment 1, accuracy was defined as the proportion of times that the optimal response was chosen. In order to examine the overall performance curve, mean accuracy was also computed for each trial and environment type (stationary/non-stationary) across all participants. Note that error bars on all plots represent one standard error.

To determine the possible effect of ACh and NE levels on performance, mean simulated levels of these neurotransmitters were computed for each trial and environment

type. Furthermore, mean P* was computed for each trial, environment, and hypothesis (i.e. that either one option or the other was best). Recall that P* is the probability that a particular guess about the environment (e.g. which response is optimal) is correct. It was assumed here, without loss of generality, that Option 1 was optimal, and that in the non-stationary environment Option 2 became optimal following the context shift. In order to illustrate the number of context updates that occurred – the number of times a context shift was detected – the total mean numbers of context shifts were computed, by trial, for each environment.

To simulate the neural data from Experiment 1 (specifically, the fERN) mean prediction errors (δ) were computed for each trial and environment type. Similar to the analysis done in Experiment 1, means and standard errors were computed both early (trials 1-10) and late (trials 11-20) in a block, for each environment type.

4.3.2 Simulating Experiment 2

To simulate the results from Experiment 2 – i.e. the effect of environmental cues – the overall level of NE was lowered in the non-stationary environment and raised in the stationary environment. This was done by modulating Φ in Equation 4.5, e.g. raising λ^*, thereby lowering NE, making context shift detections less likely. The resulting mean accuracies for each trial and environment type were computed, as well as overall mean accuracies for each environment type. To illustrate the effect that NE modulation had on uncertainty detection, mean total numbers of context updates (i.e. context shift detections) were computed, for each trial and environment type.

4.4 Simulation Results

4.4.1 Behavioural

Model performance is shown in Figure 4.2b, which can be compared to human performance for Experiment 1 in Figure 4.2a. The model was able to detect and adapt to the context shifts that occurred around trial 12 in the non-stationary environment, although it was never able to fully recover within 20 trials. For later comparison, a version of the model without any uncertainty detection was run (RL alone: Figure 4.9).

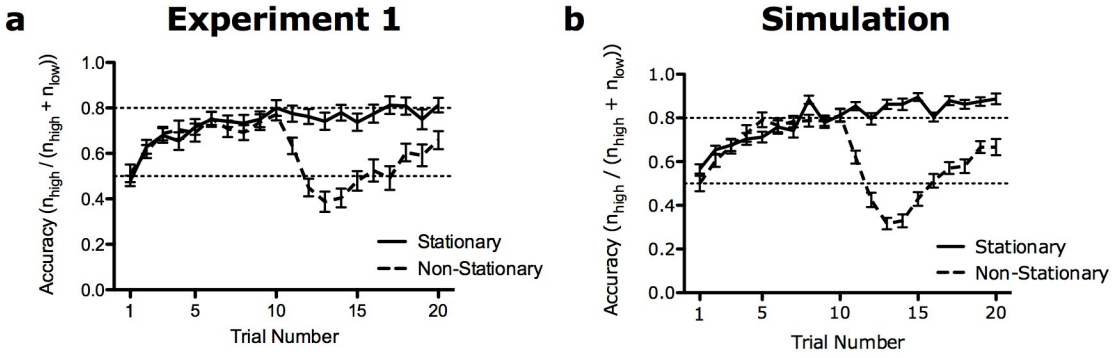

Figure 4.2. Model performance (b) compared to actual performance in Experiment 1 (a). Accuracy was defined as the proportion of times that a virtual participant made the optimal response (i.e. made the response most likely to result in a win). Note that a context shift occurred around trial 12. Dashed lines are shown at accuracies of 80% and 50%.

4.4.2 ACh and NE

In the stationary environment, as the model was exposed to the expected uncertainty of selecting the best option, the parameter γ came to reflect the probability of the best option winning. In the non-stationary environment, however, ACh levels increased following a context shift (~ trial 12) as more and more losses were detected,

then settled down as the context was updated to reflect the shift (Figure 4.3a). Likewise, NE (1-λ) decreased as the belief that the current context was the correct one increased. In the non-stationary environment, following a context shift, NE levels increased to indicate doubt in the current context belief (Figure 4.3b). These values (γ and λ) were used to determine P*, the likelihood that a context update occurred (see Equation 4.5). A context shift was accompanied by a decrease in the likelihood of the current hypothesis being true, and an increase in the likelihood of the alternative hypothesis (Figure 4.4b). If these values changed enough, then a context update was triggered, and the RL weights were reset, as described in Section 4.2.2.

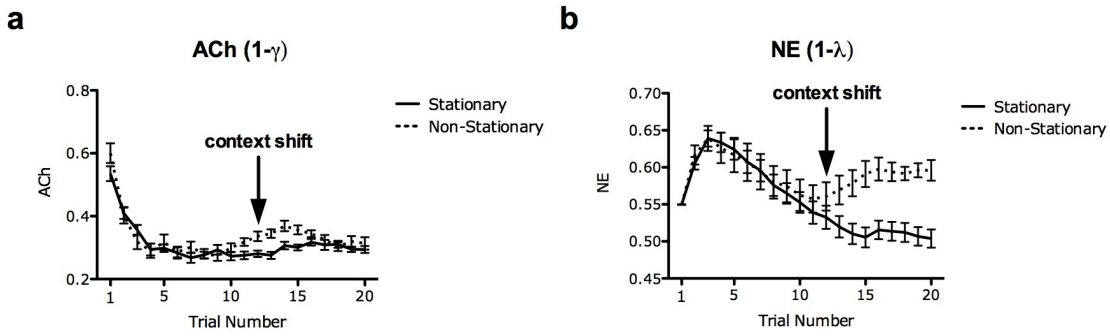

Figure 4.3. ACh and NE levels, over time. Following the context shift, ACh and NE levels increase to indicate rising expected and unexpected uncertainty, respectively.

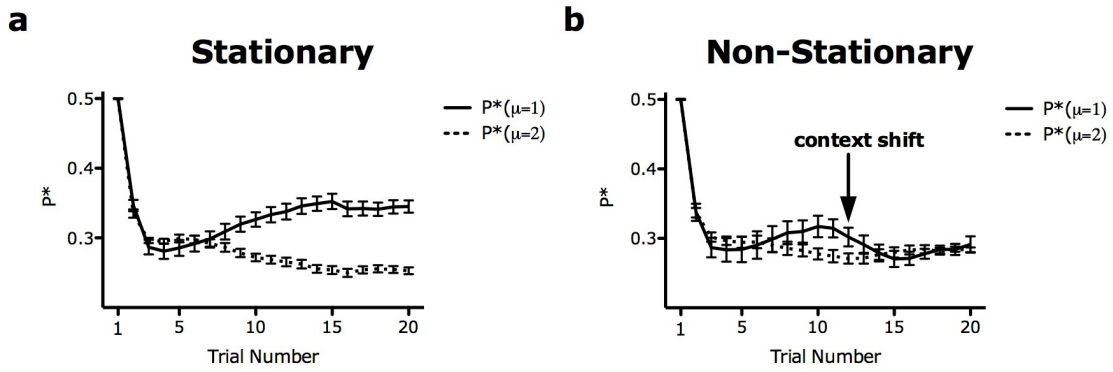

Figure 4.4. Mean values for P*, the model estimate of the probability that the current context belief remained the same in the (a) stationary and (b) non-stationary environment. For the sake of this illustration, μ=1 was assumed to be the correct context in both the stationary environment, and in the non-stationary environment prior to the context switch. If P*(μ=1) < P*(μ=2) then a context update occurred.

The model correctly identified more context shifts in the non-stationary compared to the stationary environment (Figure 4.5b versus 4.5a). There were many false alarms (detecting a context shift when none existed) in both the stationary environment and the non-stationary environment in trials 1-10, before the levels of expected and unexpected uncertainty could stabilize.

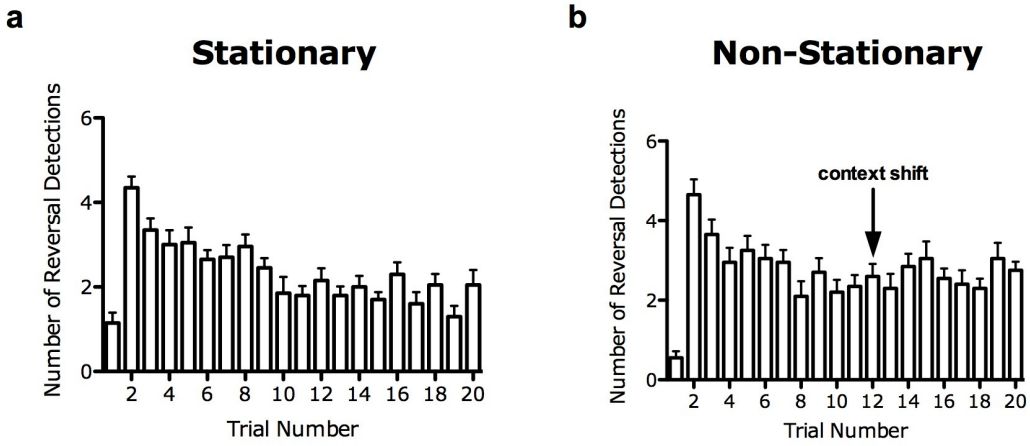

Figure 4.5. Mean number of context shift detections ("context updates") for each trial in the (a) stationary and (b) non-stationary environment. A context shift occurred on trial 12, on average.

4.4.3 Prediction Errors

The RL portion of the model generated prediction errors, shown in Figure 4.6. Later in learning, there was an enhanced prediction error, indicating that events were worse than anticipated.

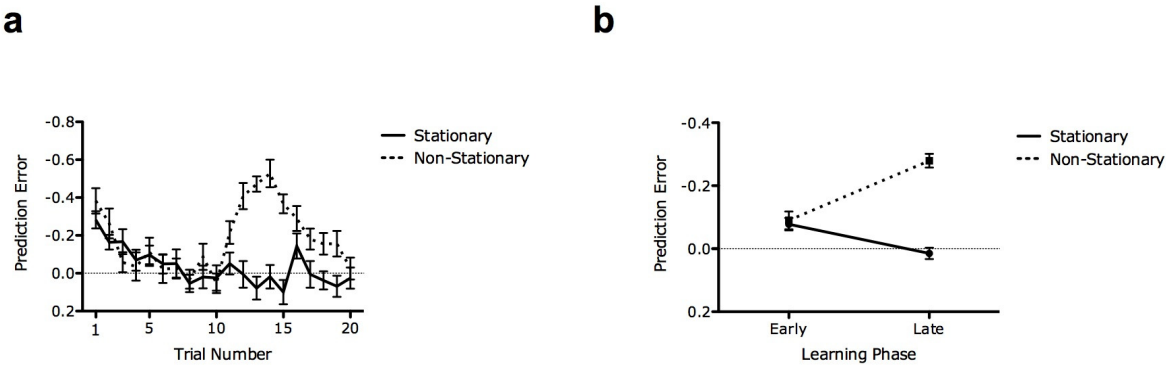

Figure 4.6. Model prediction error means for (a) all trials, and (b) grouped by early/late in learning. Prediction errors were enhanced in the non-stationary environment, reflecting unexpected losses. The y-axis is reversed here, mirroring the ERP convention for plotting the fERN.

4.4.4 Experiment 2 versus Experiment 1

Reducing overall levels of NE in the stationary environment, and raising overall levels of NE in the non-stationary environment resulted in a performance increase relative to no NE intervention (Experiment 2: Figure 4.7).

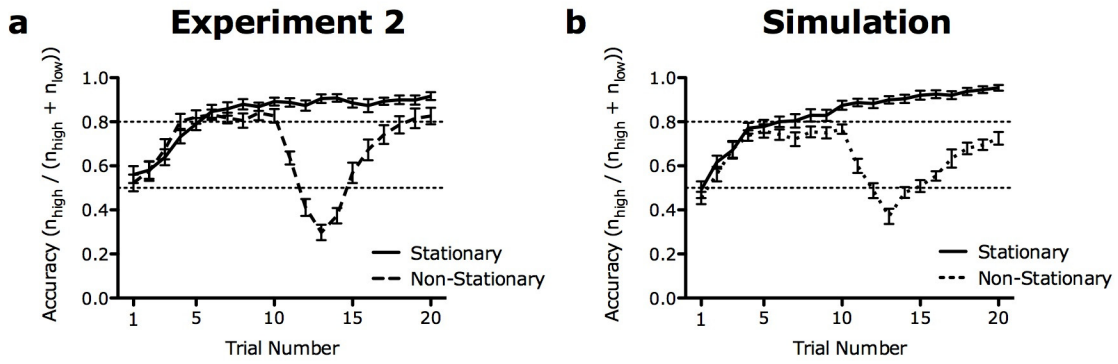

Figure 4.7. Model performance when NE was modulated by environment (b) compared to actual performance in Experiment 2 (a). Overall NE levels were reduced in the stationary environment and enhanced in the non-stationary environment. Dashed lines are shown at 80% and 50% accuracy.

The enhanced performance when NE levels were modulated based on the environment was due to an improved ability to detect context shifts. In particular, there were fewer false alarms in the stationary environment, and more hits (correct detections of context shifts) in the non-stationary environment (Experiment 2: Figure 4.8, compared to Figure 4.5).

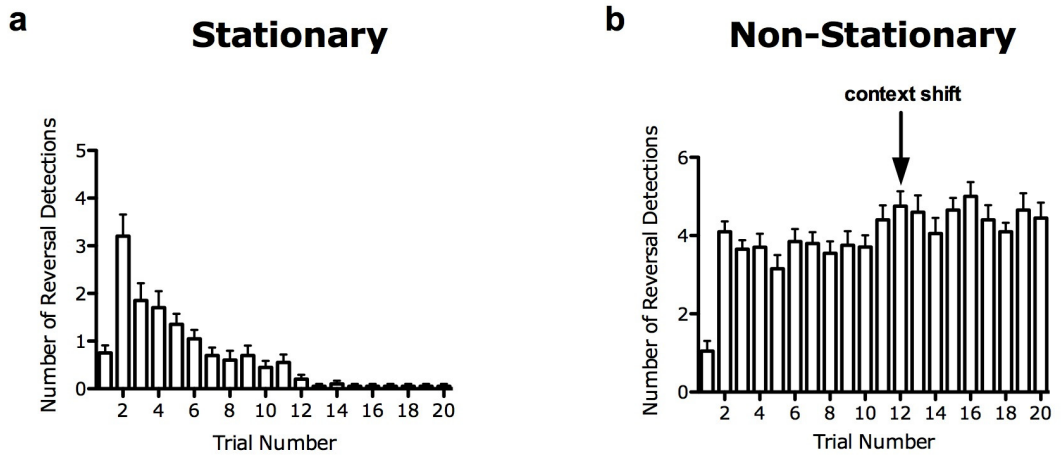

Figure 4.8. Total number of context updates in the (a) stationary environment, where NE was reduced, and (b) non-stationary environment, where NE was enhanced.

Finally, in order to assess the benefit of including uncertainty detection in the model, averages and standard deviations were computed for the RL model alone, the RL model with uncertainty detection added (RL+UD), and the augmented model with NE modulation (RL+UD*). See Figure 4.9.

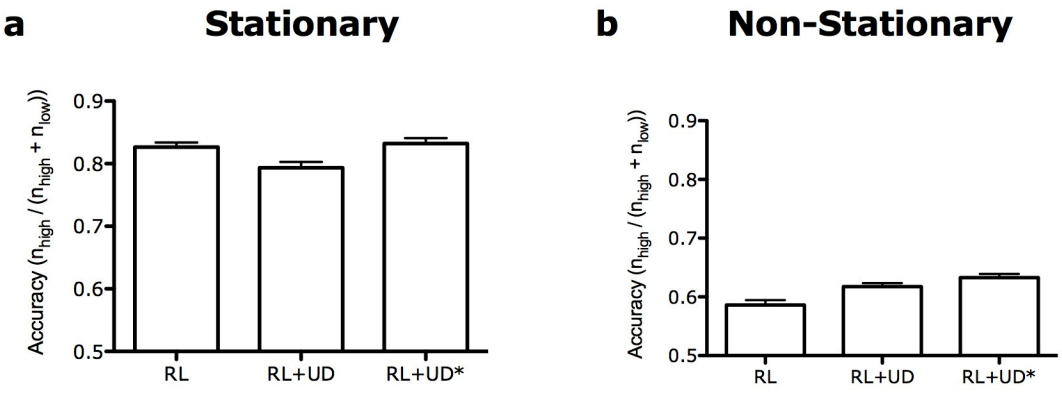

Figure 4.9. Overall performance in (a) the stationary environment and (b) the non-stationary environment. Compared here are the reinforcement learning (RL) model alone, the RL model augmented by uncertainty detection (RL+UD), and the augmented model when overall neurotransmitter levels were modulated (RL+UD*).

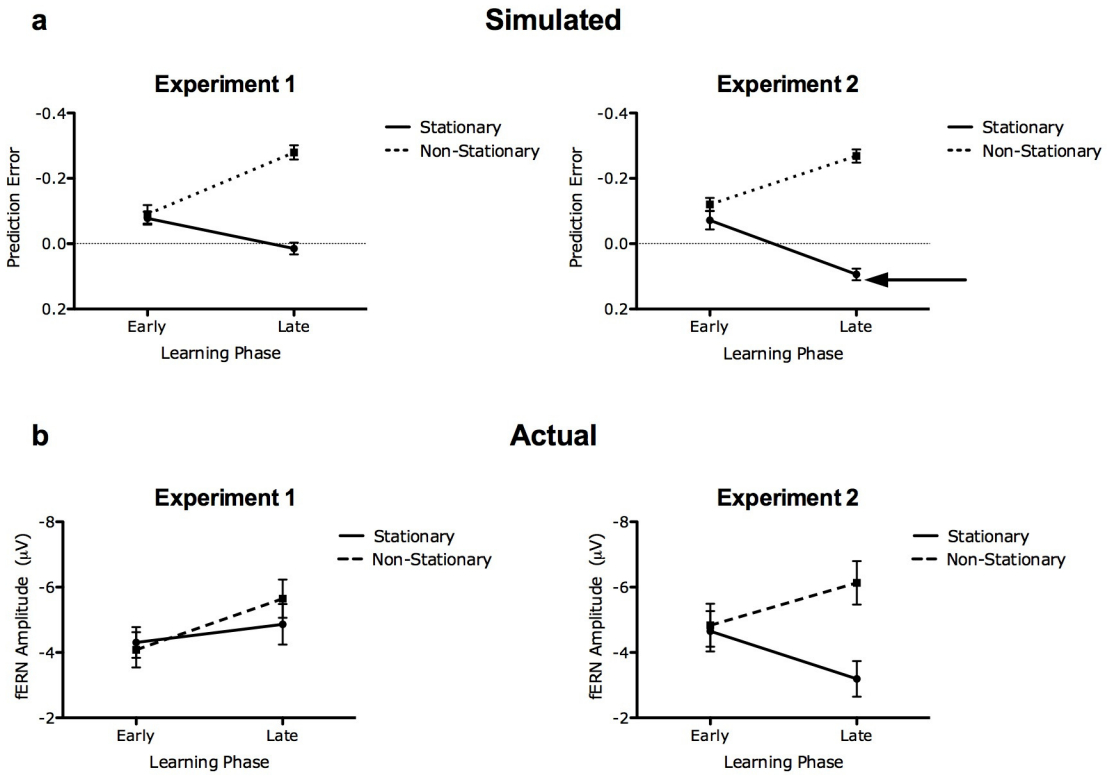

Figure 4.10. Simulated prediction errors (fERNs) for Experiment 1 and Experiment 2 (a). An arrow highlights the difference: a positive prediction error later in the stationary blocks in the Experiment 2 simulation. Compare this with the actual fERN amplitudes measured in Experiments 1 and 2 (b).

4.5 Discussion

The goal of this experiment was to simulate human behavioural and neural data for a two-armed bandit task played in a stationary environment and a non-stationary environment. Each bandit selection lead to a reward with differing probabilities; these probabilities could swap in the non-stationary environment, but not the stationary environment. The results presented here suggest that an ACh-NE system for uncertainty detection can improve the performance of a RL model in non-stationary environments, but hinder performance in stationary environments. However, with appropriate

modulation of overall NE levels, performance in both stationary and non-stationary environments can be improved over and above a RL model alone.

In particular, the ACh-NE system of uncertainty detection modelled here kept an ongoing record of past wins and losses, which it used to infer the probability that the reward probabilities had switched (that a context shift had occurred). By informing the RL component of this shift, the model was able to adapt more quickly than it did with a RL model alone. This, of course, improved performance in the non-stationary environment, but not the stationary environment, since context shifts only occurred in the non-stationary environment. In fact, as shown here, any context updates that did occur in the stationary environment could only hurt performance. Thus, without proper modulation of NE levels, adding explicit uncertainty detection to a RL model may improve performance in a non-stationary environment, but not a stationary environment.

These simulations also offer one possible explanation for the observed difference in behavioural results between Experiment 1 and Experiment 2. In particular, participants performed better in Experiment 2, in which they were presented with cues about the identity of the current environment, compared to Experiment 1, in which no cues were given. Here, it was proposed that the non-stationary environmental cue elevated overall NE levels in an ACh-NE uncertainty detection system, while the stationary environmental cue lowered NE levels. The effect of this modulation was that random feedback fluctuations in the stationary environment were less likely to be mistaken for context shifts, and actual context shifts in the non-stationary environment were more likely to be correctly detected.

The model presented here was also able to generate prediction errors, which according to Holroyd and Coles (2002) are reflected by the fERN in human ERP studies. While the prediction errors generated by the model nicely mirrored the results in Experiment 2, they were somewhat incongruent with the results from Experiment 1. In particular, the model predicted that the fERN should be the same between environments initially, and then decrease with learning in the stationary environment (as feedback became predictable) and increase in the non-stationary environment (following context shifts). However, in Experiment 1 the fERN increased in magnitude in both the stationary environment and the non-stationary environment. Thus, while the model produced results in line with previous research (e.g. Krigolson et al., 2008), the fERN results observed in Experiment 1 remain somewhat unexplained other than the speculations offered in Section 2.4, i.e. that participants treated the stationary environment as a non-stationary environment. Consistent with this explanation, examining Figure 4.10 suggests that one difference that NE modulation made was that, due to the improved performance in this condition, there was a positive overall prediction error (i.e. things were better than expected). In the Experiment 1 simulation, in contrast, the prediction error was more negative from the additional errors that were made due to mistakenly believing that a context shift had occurred. While this difference may not totally account for the fERN difference between Experiment 1 and Experiment 2, it seems like a reasonable starting point.

In conclusion, there is convincing evidence that the ACh-NE system is responsible for the context updates that occur when unexpected shifts in feedback rules are detected (Bouret and Sara, 2005; Yu & Dayan, 2005; Doya, 2008). This information

can be used to improve the performance of a RL model in non-stationary environments, at the expense of performance in stationary environments. Cue-driven modulation of NE (enhancement of NE in non-stationary environment and reduction of NE in stationary environments) may lead to further improvements, and could explain the observed performance benefit in Experiment 2 over Experiment 1.

CHAPTER 5: GENERAL DISCUSSION

Chapters Two and Three reported the behavioural and ERP results of two experiments designed to determine the degree to which medial-frontal RL plays a role in adapting to non-stationary environments – environments in which the optimal response may change. These experiments were motivated by the lack of research on this topic, and by the open issues that were identified in the work that does exist (e.g. Bland and Schaefer, 2011). While most work on non-stationary environments thus far has focused on tasks with high feedback validity (Yu & Dayan, 2005; Behrens et al., 2007; Chase et al., 2010; Bland & Schaefer, 2011), little has been said about detecting context shifts when feedback validity is low. Thus, to adequately describe the role of the medial-frontal cortex in adapting to non-stationary environments, it was necessary to test medial-frontal activity when there was little chance of a reward (Experiment 1). To further distinguish between human performance in stationary and non-stationary environments, Experiment 2 provided participants with cues such that the identity of the current environment (stationary or non-stationary) could be known. Finally, in Chapter Four, both behavioural and neural data were simulated for the tasks described in Experiments 1 and 2 using a model that combined RL with an ACh-NE model of uncertainty detection (Yu & Dayan, 2005). This chapter will summarize the results of Chapters Two, Three, and Four, as well as the contributions that these findings make to the study of learning in non-stationary environments.

5.1 Overview of Current Results

The task used in both Experiments 1 and 2 was a two-armed bandit in which participants had to learn, over 20 trials, which of two coloured squares was most likely to

lead to a reward when selected. This was repeated over several blocks, with new colours selected randomly at the beginning of each block. Importantly, in half of the blocks, outcome probabilities would switch partway through, requiring participants to detect this change and adapt their responses. In Experiment 1, participants were given no indication as to which type of environment (stationary or non-stationary) they were in, other than the trial-to-trial feedback they received. In contrast, Experiment 2 provided participants with unique cues as to the nature of each environment (i.e. the two-armed bandit was played in two different casinos – one honest, and one dishonest.)

In Experiment 1 there was no difference in medial-frontal reward processing, as indexed by the fERN component of the ERP, between the stationary environment and the non-stationary environment. There was, however, an enhanced P300 component in response to wins during/following context shifts. Also, participant responses slowed over time in both environments as participants grew uniformly unsure about their responses. Adding environmental cues in Experiment 2 resulted in an increase in the magnitude of the fERN over time in the non-stationary environment, and a decrease over time in the stationary environment. Thus, in this case, the RL processes underlying the fERN appeared to play a role in detecting and adapting to unexpected uncertainty. Furthermore, unlike in Experiment 1, participant responses did not slow down over time. Finally, the Experiment 2 P300 in response to both wins and losses was enhanced in the non-stationary environment, but not in the stationary environment.

The contrasting results of Experiments 1 and 2 suggest that while midbrain reward processing may play a role in detecting and adapting to unexpected uncertainty, it cannot be the whole story. Both Experiments saw an increase in the feedback-locked

P300, though – a result that was explored more with a simulation in Chapter Four. In particular, changes in feedback following a context shift triggered an increase in simulated phasic NE levels which, according to the simulation, may be used to improve the performance of a RL model. Furthermore, the simulated results for Experiment 2 suggest that the effect of the non-stationary environmental cue may have been to raise overall NE levels, making reversal detections more likely, while the stationary cue may have lowered overall NE levels, making reversal detections less likely. When NE was altered in this way in the simulation, there was a performance improvement in the cued task over the uncued task, another observed contrast between Experiments 1 and 2. Additionally, the fERN differences between the stationary and non-stationary environments in Experiment 2 were adequately captured by the prediction errors generated by the simulation. Furthermore, the simulation also offered a plausible explanation for the surprising fERN results seen in Experiment 1. In particular, the enhanced fERN observed later in stationary blocks in Experiment 1 may have been due to the performance cost associated with mistaking random feedback fluctuations for context shifts requiring behavioural adaptation.

5.2 Connection to Current Research and Theory

The most popular theories about detecting unexpected uncertainty involve Bayesian inference (Yu & Dayan, 2005; Rushworth & Behrens, 2008). Yu and Dayan (2005) linked a Bayesian approach with neurotransmitter modulation (NE for unexpected uncertainty, ACh for expected uncertainty). Several studies have also linked P300 enhancement with phasic increases in NE via the LC-NE system (see Nieuwenhuis et al., 2005, for a review). Given these theories on NE and unexpected uncertainty (Doya,

2008), it is not surprising that an increase in the P300 was observed in response to feedback following the context shifts in Experiments 1 and 2. Surprisingly, however, while this effect was observed for both wins and losses in Experiment 2, it was only observed for wins in Experiment 1, while in both Experiments 1 and 2 the P300 for wins was enhanced relative to the P300 for losses. The P300 is not typically sensitive to reward valence (Yeung & Sanfey, 2004; Holroyd et al., 2004), although it is sensitive to reward magnitude (Yeung & Sanfey, 2004), so this difference likely has more to do with saliency and/or context updating than valence. As speculated earlier, while no particular loss feedback following a context shift necessarily signalled a context update, there is reason to believe that a win following several losses would be highly salient, and thus elicit a larger P300 (Nieuwenhuis et al., 2005).

The fERN in Experiment 1 was not enhanced in the non-stationary environment, contrary to what others have found (Bland & Schaeffer, 2011). It is argued here that this is due to the low feedback validity that was used in the present work. In particular, participants may have mistaken losses signalling a context shift for losses due to expected uncertainty, i.e. that selecting the best option should often result in a loss. The implication of this result is that while medial-frontal cortex may play a role in detecting unexpected uncertainty, other systems are likely at work as well – in particular, the LC-NE system, as indexed by the P300. In Experiment 2, in contrast, the fERN became enhanced in the non-stationary environment only. This is consistent with both existing EEG (Chase et al., 2010; Bland & Schaeffer, 2011) and MRI work (Behrens et al., 2007).

According to Yu and Dayan's (2005) model, NE and ACh levels interact, so more work could be done manipulating both expected and unexpected uncertainty in a

paradigm similar to what was used here. As the results presented here indicate, the level of feedback validity that is used in experiments on uncertainty may greatly impact the degree to which medial-frontal reward processing is recruited for uncertainty detection. Furthermore, one unexplored issue here (and elsewhere) is how the feedback validity of competing options may interact. In Experiments 1 and 2, the less optimal choice always had a feedback validity characterized by p(win) = 0.1. Technically speaking, this describes a highly valid feedback situation (i.e., the outcome of choosing this option could be known with 90% accuracy). While the outcome probabilities in the present work were chosen via pilot testing in order to equalize the total number of win and loss trials, manipulation of the win likelihoods of each option (or several options) may yield similarly fruitful results to what was found here.

5.3 Conclusion

When feedback cannot be completely trusted – when a lack of reward doesn't necessarily mean that a wrong choice was made – then learning becomes challenging. This challenge is compounded when the best option may change over time. Just as different forms of uncertainty have been identified (Bland & Schaefer, 2012), different mechanisms have been uncovered behind how uncertainty is detected and dealt with (Payzan-LeNestour & Bossaerts, 2011). The topic of the present work was unexpected uncertainty – shifts in the underlying probabilities determining which action is optimal. The goal here was to investigate the role of two possible mechanisms behind detecting and adapting to unexpected uncertainty: midbrain reward processing, as indexed by the fERN, and the LC-NE system, as indexed by the P300.

By reducing feedback validity, a non-stationary environment was created (Experiment 1) where the LC-NE system appeared to play more of role in detecting and responding to unexpected uncertainty compared to the midbrain RL system. This suggests two separate systems for unexpected uncertainty detection: a model-free RL system, located in the midbrain, and some other system (possibly Bayesian: Payzan-LeNestour & Bossaerts, 2011; Wilson & Niv, 2012), related to LC-NE functioning. In Experiment 2, a stationary and a non-stationary environment were created where, despite low feedback validity, both midbrain reward processing and the LC-NE system appeared to be engaged for uncertainty detection. Finally, behavioural and neural data were simulated to show that it is possible to explain the difference between stationary and non-stationary performance through modulation of NE. In particular, NE may be enhanced in non-stationary environments, facilitating the correct detection of context shifts, and reduced in stationary environments, leading to fewer false context shift detections. Taken together, these results tend to downplay (but not exclude) the role of medial-frontal reward processing in unexpected uncertainty detection, and highlight the role of the LC-NE system

REFERENCES

Adleman, N. E., Kayser, R., Dickstein, D., Blair, R. J. R., Pine, D., & Leibenluft, E. (2011). Neural Correlates of Reversal Learning in Severe Mood Dysregulation and Pediatric Bipolar Disorder. *Journal of the American Academy of Child & Adolescent Psychiatry*, *50*(11), 1173–1185.e2. doi:10.1016/j.jaac.2011.07.011

Alexander, W. H., & Brown, J. W. (2011). Medial prefrontal cortex as an action-outcome predictor. *Nature Neuroscience*, *14*(10), 1338–1344. doi:10.1038/nn.2921

Allain, S., Hasbroucq, T., Burle, B., Grapperon, J., & Vidal, F. (2004). Response monitoring without sensory feedback. *Clinical Neurophysiology*, *115*(9), 2014–2020. doi:10.1016/j.clinph.2004.04.013

Amiez, C., Joseph, J.-P., & Procyk, E. (2005). Anterior cingulate error-related activity is modulated by predicted reward. *European Journal of Neuroscience*, *21*(12), 3447–3452. doi:10.1111/j.1460-9568.2005.04170.x

Aston-Jones, G, Rajkowski, J., & Kubiak, P. (1997). Conditioned responses of monkey locus coeruleus neurons anticipate acquisition of discriminative behavior in a vigilance task. *Neuroscience*, *80*(3), 697–715. doi:10.1016/S0306-4522(97)00060-2

Aston-Jones, Gary, & Cohen, J. D. (2005). An Integrative Theory of Locus Coeruleus-Norepinephrine Function: Adaptive Gain and Optimal Performance. *Annual Review of Neuroscience*, *28*(1), 403–450. doi:10.1146/annurev.neuro.28.061604.135709

Avery, M. C., Nitz, D. A., Chiba, A. A., & Krichmar, J. L. (2012). Simulation of cholinergic and noradrenergic modulation of behavior in uncertain environments. *Frontiers in Computational Neuroscience*, *6*. doi:10.3389/fncom.2012.00005

Bach, D. R., & Dolan, R. J. (2012). Knowing how much you don't know: a neural organization of uncertainty estimates. *Nature Reviews Neuroscience*, *13*(8), 572–586. doi:10.1038/nrn3289

Bari, A., Theobald, D. E., Caprioli, D., Mar, A. C., Aidoo-Micah, A., Dalley, J. W., & Robbins, T. W. (2010). Serotonin Modulates Sensitivity to Reward and Negative Feedback in a Probabilistic Reversal Learning Task in Rats. *Neuropsychopharmacology*, *35*(6), 1290–1301. doi:10.1038/npp.2009.233

Beeler, J. A. (2012). Thorndike's law 2.0: dopamine and the regulation of thrift. *Frontiers in Neuroscience*, *6*, 116. doi:10.3389/fnins.2012.00116

Behrens, T. E. J., Woolrich, M. W., Walton, M. E., & Rushworth, M. F. S. (2007). Learning the value of information in an uncertain world. *Nature Neuroscience*, *10*(9), 1214–1221. doi:10.1038/nn1954

Bland, A. R., & Schaefer, A. (2011). Electrophysiological correlates of decision making under varying levels of uncertainty. *Brain Research*, *1417*, 55–66. doi:10.1016/j.brainres.2011.08.031

Bland, A. R., & Schaefer, A. (2012). Different Varieties of Uncertainty in Human Decision-Making. *Frontiers in Neuroscience*, *6*. doi:10.3389/fnins.2012.00085

Bouret, S., & Sara, S. J. (2005). Network reset: a simplified overarching theory of locus coeruleus noradrenaline function. *Trends in Neurosciences*, *28*(11), 574–582. doi:10.1016/j.tins.2005.09.002

Cavanagh, J. F., Figueroa, C. M., Cohen, M. X., & Frank, M. J. (2011). Frontal Theta Reflects Uncertainty and Unexpectedness during Exploration and Exploitation. *Cerebral Cortex*. doi:10.1093/cercor/bhr332

Chase, H. W., Swainson, R., Durham, L., Benham, L., & Cools, R. (2010). Feedback-related Negativity Codes Prediction Error but Not Behavioral Adjustment during Probabilistic Reversal Learning. *Journal of Cognitive Neuroscience*, *23*(4), 936–946. doi:10.1162/jocn.2010.21456

Cools, R., Clark, L., Owen, A. M., & Robbins, T. W. (2002). Defining the Neural Mechanisms of Probabilistic Reversal Learning Using Event-Related Functional Magnetic Resonance Imaging. *The Journal of Neuroscience*, *22*(11), 4563–4567.

Dayan, P., & Yu, A. J. (2006). Phasic norepinephrine: A neural interrupt signal for unexpected events. *Network: Computation in Neural Systems*, *17*(4), 335–350. doi:10.1080/09548980601004024

De Wit, S., & Dickinson, A. (2009). Associative theories of goal-directed behaviour: a case for animal–human translational models. *Psychological Research PRPF*, *73*(4), 463–476. doi:10.1007/s00426-009-0230-6

Dehaene, S., Posner, M. I., & Tucker, D. M. (1994). Localization of a neural system for error detection and compensation. Psychological Science, 5(5), 303–305.

Devauges, V., & Sara, S. J. (1990). Activation of the noradrenergic system facilitates an attentional shift in the rat. *Behavioural Brain Research*, *39*(1), 19–28. doi:10.1016/0166-4328(90)90118-X

Donchin, E. (1981). Surprise!… Surprise? *Psychophysiology*, *18*(5), 493–513. doi:10.1111/j.1469-8986.1981.tb01815.x

Donchin, E., & Coles, M. G. H. (1988). Is the P300 component a manifestation of context updating? *Behavioral and Brain Sciences*, *11*(03), 357–374. doi:10.1017/S0140525X00058027

Doya, K. (2008). Modulators of decision making. *Nature Neuroscience*, *11*(4), 410–416. doi:10.1038/nn2077

Falkenstein, M. (2004). Errors, Conflicts, and the Brain: A Review of the Contributions to the Error Conference, Dortmund 2003. *Journal of Psychophysiology*, *18*(4), 153–163. doi:10.1027/0269-8803.18.4.153

Fellows, L. K., & Farah, M. J. (2003). Ventromedial frontal cortex mediates affective shifting in humans: evidence from a reversal learning paradigm. *Brain*, *126*(8), 1830–1837. doi:10.1093/brain/awg180

Gehring, W. J., Goss, B., Coles, M. G. H., Meyer, D. E., & Donchin, E. (1993). A Neural System for Error Detection and Compensation. *Psychological Science*, *4*(6), 385–390. doi:10.1111/j.1467-9280.1993.tb00586.x

Gratton, G., Coles, M. G. ., & Donchin, E. (1983). A new method for off-line removal of ocular artifact. *Electroencephalography and Clinical Neurophysiology*, *55*(4), 468–484. doi:10.1016/0013-4694(83)90135-9

Hajcak, G., Moser, J. S., Holroyd, C. B., & Simons, R. F. (2006). The feedback-related negativity reflects the binary evaluation of good versus bad outcomes. *Biological Psychology*, *71*(2), 148–154. doi:10.1016/j.biopsycho.2005.04.001

Holroyd, C. B, Nieuwenhuis, S., Yeung, N., & Cohen, J. D. (2003). Errors in reward prediction are reflected in the event-related brain potential. *Neuroreport*, *14*(18), 2481.

Holroyd, C. B., Nieuwenhuis, S., Yeung, N., Nystrom, L., Mars, R. B., Coles, M. G. H., & Cohen, J. D. (2004). Dorsal anterior cingulate cortex shows fMRI response to internal and external error signals. *Nat Neurosci*, *7*(5), 497–498. doi:10.1038/nn1238

Holroyd, Clay B., & Coles, M. G. H. (2002). The neural basis of human error processing: Reinforcement learning, dopamine, and the error-related negativity. *Psychological Review*, *109*(4), 679–709. doi:10.1037/0033-295X.109.4.679

Holroyd, Clay B., & Coles, M. G. H. (2008). Dorsal anterior cingulate cortex integrates reinforcement history to guide voluntary behavior. *Cortex*, *44*(5), 548–559. doi:10.1016/j.cortex.2007.08.013

Holroyd, Clay B., Hajcak, G., & Larsen, J. T. (2006). The good, the bad and the neutral: Electrophysiological responses to feedback stimuli. *Brain Research*, *1105*(1), 93–101. doi:10.1016/j.brainres.2005.12.015

Holroyd, Clay B., & Krigolson, O. E. (2007). Reward prediction error signals associated with a modified time estimation task. *Psychophysiology*, *44*(6), 913–917. doi:10.1111/j.1469-8986.2007.00561.x

Holroyd, Clay B., Larsen, J. T., & Cohen, J. D. (2004). Context dependence of the event-related brain potential associated with reward and punishment. *Psychophysiology*, *41*(2), 245–253. doi:10.1111/j.1469-8986.2004.00152.x

Hornak, J., O'Doherty, J., Bramham, J., Rolls, E. T., Morris, R. G., Bullock, P. R., & Polkey, C. E. (2004). Reward-related Reversal Learning after Surgical Excisions in Orbito-frontal or Dorsolateral Prefrontal Cortex in Humans. *Journal of Cognitive Neuroscience*, *16*(3), 463–478. doi:10.1162/089892904322926791

Krigolson, O. E., & Holroyd, C. B. (2007). Predictive information and error processing: The role of medial-frontal cortex during motor control. *Psychophysiology*, *44*(4), 586–595. doi:10.1111/j.1469-8986.2007.00523.x

Krigolson, O. E., Pierce, L. J., Holroyd, C. B., & Tanaka, J. W. (2009). Learning to Become an Expert: Reinforcement Learning and the Acquisition of Perceptual Expertise. *Journal of Cognitive Neuroscience*, *21*(9), 1833–1840. doi:10.1162/jocn.2009.21128

Miltner, W. H. R., Braun, C. H., & Coles, M. G. H. (1997). Event-Related Brain Potentials Following Incorrect Feedback in a Time-Estimation Task: Evidence for a "Generic" Neural System for Error Detection. *Journal of Cognitive Neuroscience*, *9*(6), 788–798. doi:10.1162/jocn.1997.9.6.788

Montague, P. R., Dayan, P., & Sejnowski, T. J. (1996). A framework for mesencephalic dopamine systems based on predictive Hebbian learning. *The Journal of Neuroscience*, *16*(5), 1936–1947.

Mushtaq, F., Bland, A. R., & Schaefer, A. (2011). Uncertainty and Cognitive Control. *Frontiers in Psychology*, *2*. doi:10.3389/fpsyg.2011.00249

Nieuwenhuis, S., Aston-Jones, G., & Cohen, J. D. (2005). Decision making, the P3, and the locus coeruleus--norepinephrine system. *Psychological Bulletin*, *131*(4), 510–532. doi:10.1037/0033-2909.131.4.510

O'Doherty, J., Critchley, H., Deichmann, R., & Dolan, R. J. (2003). Dissociating Valence of Outcome from Behavioral Control in Human Orbital and Ventral Prefrontal Cortices. *The Journal of Neuroscience*, *23*(21), 7931–7939.

Paulus, M. P., Hozack, N., Frank, L., & Brown, G. G. (2002). Error Rate and Outcome Predictability Affect Neural Activation in Prefrontal Cortex and Anterior Cingulate during Decision-Making. *NeuroImage*, *15*(4), 836–846. doi:10.1006/nimg.2001.1031

Payzan-LeNestour, E., & Bossaerts, P. (2011). Risk, Unexpected Uncertainty, and Estimation Uncertainty: Bayesian Learning in Unstable Settings. *PLoS Comput Biol*, *7*(1), e1001048. doi:10.1371/journal.pcbi.1001048

Polich, J., & Margala, C. (1997). P300 and probability: comparison of oddball and single-stimulus paradigms. *International Journal of Psychophysiology*, *25*(2), 169–176.

Polich, John. (1990). Probability and inter-stimulus interval effects on the P300 from auditory stimuli. *International Journal of Psychophysiology*, *10*(2), 163–170. doi:10.1016/0167-8760(90)90030-H

Polich, John. (2007). Updating P300: An integrative theory of P3a and P3b. *Clinical Neurophysiology*, *118*(10), 2128–2148. doi:10.1016/j.clinph.2007.04.019

Polich, John, & Margala, C. (1997). P300 and probability: comparison of oddball and single-stimulus paradigms. *International Journal of Psychophysiology*, *25*(2), 169–176. doi:10.1016/S0167-8760(96)00742-8

Rolls, E. T., Hornak, J., Wade, D., & McGrath, J. (1994). Emotion-related learning in patients with social and emotional changes associated with frontal lobe damage. *Journal of Neurology, Neurosurgery & Psychiatry*, *57*(12), 1518–1524. doi:10.1136/jnnp.57.12.1518

Rushworth, M. F. S., & Behrens, T. E. J. (2008). Choice, uncertainty and value in prefrontal and cingulate cortex. *Nature Neuroscience*, *11*(4), 389–397. doi:10.1038/nn2066

Sarter, M., & Bruno, J. P. (1997). Cognitive functions of cortical acetylcholine: toward a unifying hypothesis. *Brain Research Reviews*, *23*(1–2), 28–46. doi:10.1016/S0165-0173(96)00009-4

Sarter, M., & Parikh, V. (2005). Choline transporters, cholinergic transmission and cognition. *Nature Reviews Neuroscience*, *6*(1), 48–56. doi:10.1038/nrn1588

Schultz, W., Dayan, P., & Montague, P. R. (1997). A Neural Substrate of Prediction and Reward. *Science*, *275*(5306), 1593–1599. doi:10.1126/science.275.5306.1593

Steere, J. C., & Arnsten, A. F. T. (1997). The α-2A noradrenergic receptor agonist guanfacine improves visual object discrimination reversal performance in aged rhesus monkeys. *Behavioral Neuroscience*, *111*(5), 883–891. doi:10.1037/0735-7044.111.5.883

Sutton, R. S., & Barto, A. G. (1998). *Reinforcement Learning: An Introduction*. Cambridge, MA: MIT Press.

Sutton, S., Braren, M., Zubin, J., & John, E. R. (1965). Evoked-Potential Correlates of Stimulus Uncertainty. *Science*, *150*(3700), 1187–1188. doi:10.1126/science.150.3700.1187

Swainson, R., Rogers, R. D., Sahakian, B. J., Summers, B. A., Polkey, C. E., & Robbins, T. W. (2000). Probabilistic learning and reversal deficits in patients with Parkinson's disease or frontal or temporal lobe lesions: possible adverse effects of

dopaminergic medication. *Neuropsychologia*, *38*(5), 596–612. doi:10.1016/S0028-3932(99)00103-7

Thorndike, E. (1911). Animal Intelligence: Experimental Studies. New York: Macmillan.

Waltz, J. A., & Gold, J. M. (2007). Probabilistic reversal learning impairments in schizophrenia: Further evidence of orbitofrontal dysfunction. *Schizophrenia Research*, *93*(1–3), 296–303. doi:10.1016/j.schres.2007.03.010

Wilson, R. C., & Niv, Y. (2012). Inferring Relevance in a Changing World. *Frontiers in Human Neuroscience*, *5*. doi:10.3389/fnhum.2011.00189

Wu, Y., & Zhou, X. (2009). The P300 and reward valence, magnitude, and expectancy in outcome evaluation. *Brain Research*, *1286*, 114–122. doi:10.1016/j.brainres.2009.06.032

Yeung, N., & Sanfey, A. G. (2004). Independent Coding of Reward Magnitude and Valence in the Human Brain. *The Journal of Neuroscience*, *24*(28), 6258–6264. doi:10.1523/JNEUROSCI.4537-03.2004

Yu, A. J., & Dayan, P. (2005). Uncertainty, Neuromodulation, and Attention. *Neuron*, *46*(4), 681–692. doi:10.1016/j.neuron.2005.04.026

APPENDIX

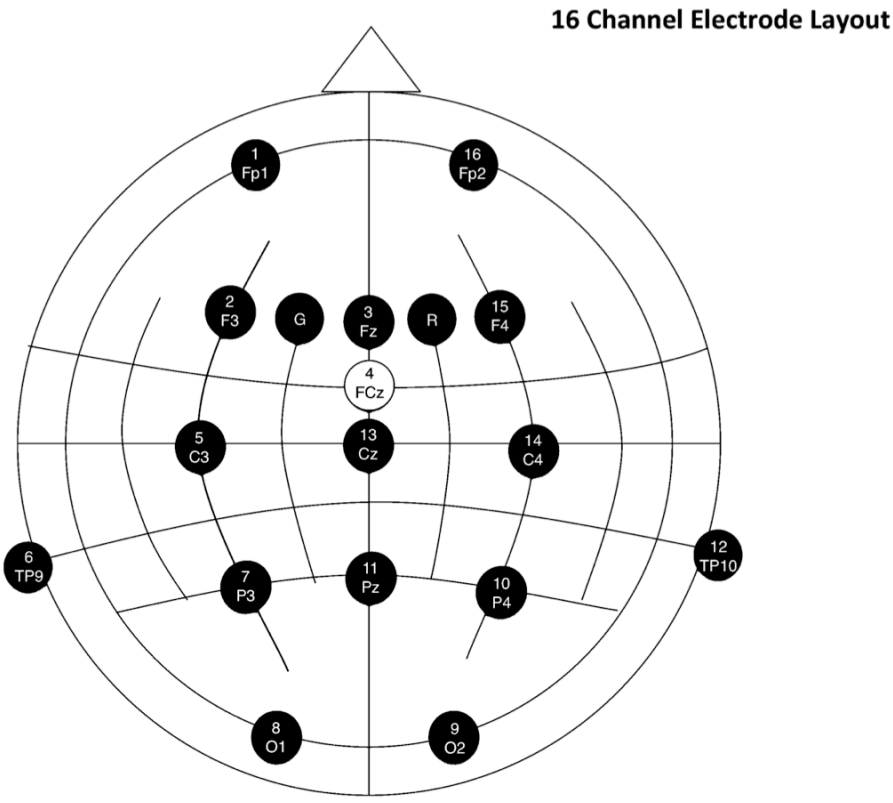

Figure A1: Electrode layout for Experiment 1.

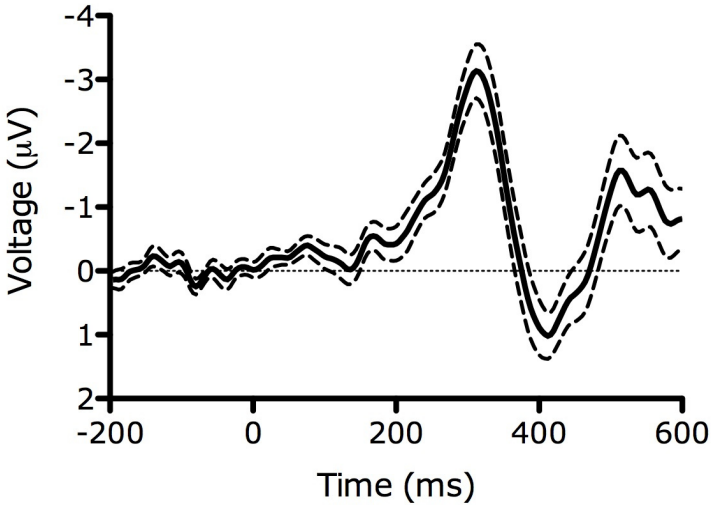

Figure A2. Grand average difference waveform (losses – wins) at electrode site FCz for Experiment 1. Dashed lines indicate one standard error.

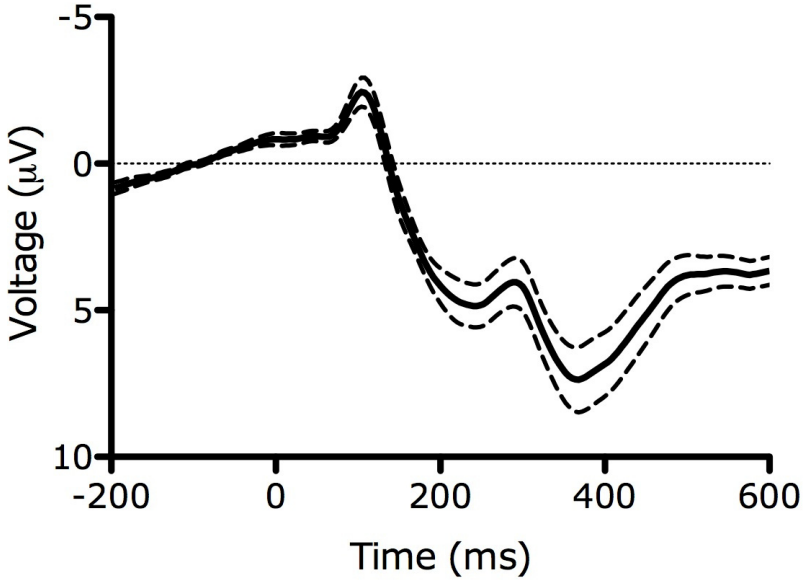

Figure A3. Grand average response to all wins at electrode site Pz for Experiment 1. Dashed lines indicate one standard error.

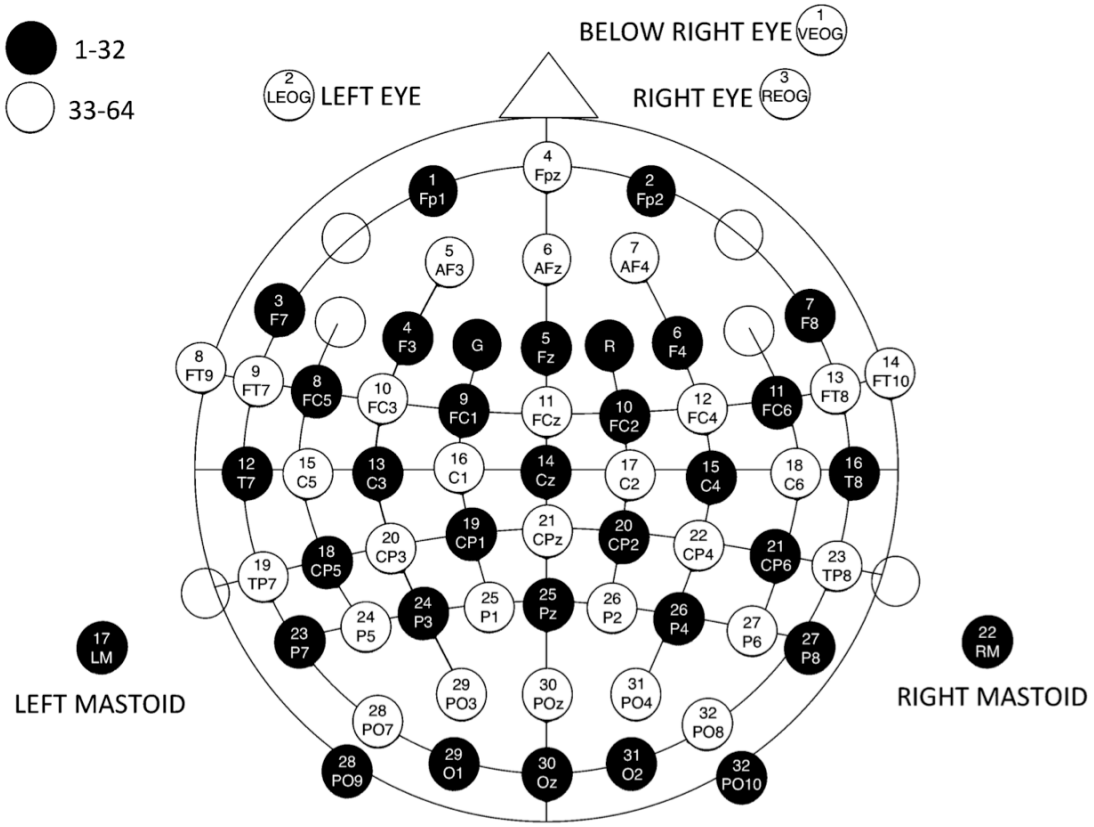

Figure A4: Electrode layout for Experiment 2.

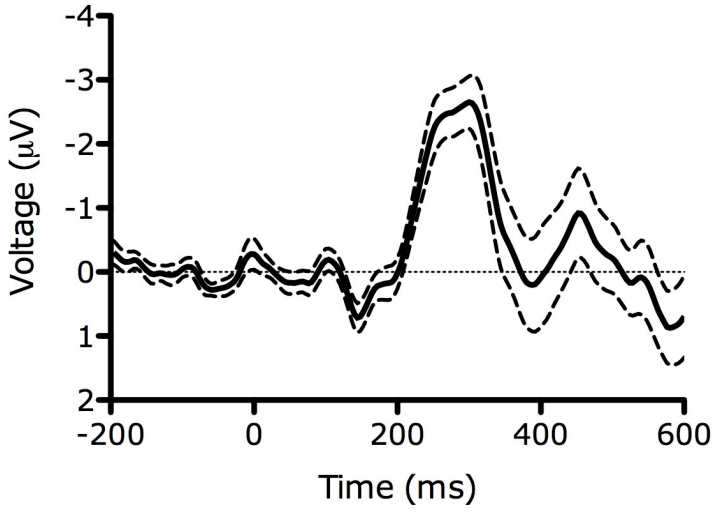

Figure A5. Grand average difference waveform (losses – wins) at electrode site FCz for Experiment 2. Dashed lines shows indicate one standard error.

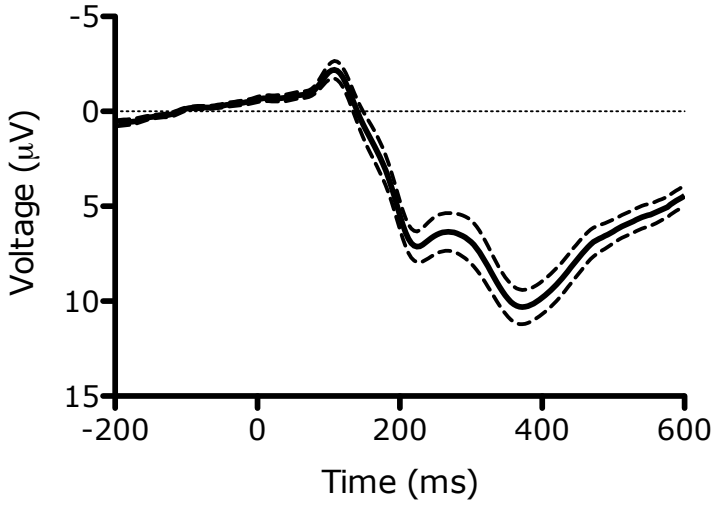

Figure A6. Grand average response to all wins at electrode site Pz for Experiment 2. Dashed lines indicate one standard error.