PROBING THE REPRESENTATION OF DECISION VARIABLES USING EEG AND EYE TRACKING

by

PABLO J. MORALES

A DISSERTATION

Presented to the Department of Psychology
and the Graduate School of the University of Oregon
in partial fulfillment of the requirements
for the degree of
Doctor of Philosophy

June 2018

DISSERTATION APPROVAL PAGE

Student: Pablo J. Morales

Title: Probing the Representation of Decision Variables Using EEG and Eye Tracking

This dissertation has been accepted and approved in partial fulfillment of the requirements for the Doctor of Philosophy degree in the Department of Psychology by:

Ulrich Mayr	Chairperson
Elliot Berkman	Core Member
Brice Kuhl	Core Member
William Harbaugh	Institutional Representative

and

Sara D. Hodges	Interim Vice Provost and Dean of the Graduate School

Original approval signatures are on file with the University of Oregon Graduate School.

Degree awarded June 2018

DISSERTATION ABSTRACT

Pablo J. Morales

Doctor of Philosophy

Department of Psychology

June 2018

Title: Probing the Representation of Decision Variables Using EEG and Eye Tracking

Value based decisions are among the most common types of decisions made by humans. A considerable body of work has investigated how different types of information guide such decisions, as well as how evaluations of their outcomes retroactively inform the parameters that were used to inform them. Several open questions remain regarding the nature of the underlying representations of decision-relevant information. Of particular relevance is whether or not positive and negative information (i.e. rewards/gains vs. punishments/losses/costs) are treated as categorically distinct, or whether they are represented on a common scale. This question was examined across three different studies utilizing a variety of methods (traditional event-related potentials, multivariate pattern classification, and eye tracking) to obtain a more comprehensive picture of how decision-relevant information is represented A common theme among the three studies was that positive and negative types of information seems to be, at least initially, represented as categorically distinct (whether it be information about gains vs. losses, or value vs. effort). Additionally, integration of different types of information appears to take place during the later phases of the decision period, which may also be when distortions in the representation of value information (ex. loss aversion) may occur. Overall, this body of work advances our understanding of the

underpinnings of value based decisions by providing additional insight about how decision-relevant information is represented in a dynamic and flexible manner.

CURRICULUM VITAE

NAME OF AUTHOR: Pablo J. Morales

GRADUATE AND UNDERGRADUATE SCHOOLS ATTENDED:

University of Oregon, Eugene
California State University, Fresno

DEGREES AWARDED:

Doctor of Philosophy, Psychology, 2018, University of Oregon
Master of Science, Psychology, 2013, University of Oregon
Master of Arts, Psychology, 2011, California State University, Fresno
Bachelor of Arts, Psychology, 2009, California State University, Fresno

AREAS OF SPECIAL INTEREST:

Neuroeconomics

PROFESSIONAL EXPERIENCE:

Graduate Teaching Fellow, Department of Psychology, 2011-2018

GRANTS, AWARDS, AND HONORS:

Graduate Student Award, Cognitive Neuroscience Society, 2017

Promising Scholar Award, University of Oregon, 2011-2012

PUBLICATIONS:

Knight, E., Christian, C.B., **Morales, P.J.,** Harbaugh, W., Mayr, U. & Mehta, P. (2017). Exogenous testosterone enhances cortisol and affective responses to social-evaluative stress in dominant men. *Psychoneuroendocrinology*, *85*, 151-157.

ACKNOWLEDGMENTS

First and foremost, I would like to thank my mother and father for supporting me throughout my time at the University of Oregon. They were an invaluable source of empathy and understanding, especially since I was born around the time my father was completing the qualifying exams for *his* PhD. I would like to thank my advisor, Dr. Ulrich Mayr, whose guidance, support, and openness to letting me pursue my own research interests made this journey possible. I would also like to express my gratitude to my labmates Jason Hubbard, Atsushi Kikumoto, and Melissa Moss, as well as the members of my committee, colleagues, research assistants, and friends that have supported me over the years. Finally, I would like to thank my girlfriend, Jessica Hardwicke, who has proved to be an invaluable source of strength and motivation.

TABLE OF CONTENTS

Chapter Page

Chapter Page

Chapter Page

LIST OF FIGURES

Figure Page

LIST OF TABLES

Table Page

CHAPTER I

INTRODUCTION

The ability to evaluate information in the environment in order to inform decision making and goal-directed behavior is ubiquitous to all organisms. Generally, this process can be described as the sampling of evidence from to-be-decided-upon options until some criterion is met that directly promotes a decision among them. Critically, the accumulated evidence is converted into a *decision variable* (DV) on which some sort of criterion can be imposed (Shadlen & Kiani, 2013). The purpose of the DV is to map accumulated evidence onto the appropriate action for the organism. Such a general mechanism for information evaluation is a core component of cognition, and is flexibly applicable across domains. For example, both perceptual decision making based on objective stimulus properties and value based/economic decision making based on subjective attributions of options based upon the organism's preferences or internal states utilize this common mechanism. Thus, decisions to determine which stimulus is brighter, which food item is healthier, whether or not to risk an uncertain gamble or choose a safe alternative, and whether or not to expend effort to obtain a reward follow this same fundamental process.

Integral to advancing our understanding of decision making across organisms is to probe the underlying representations of behaviorally relevant DVs. Indeed, the advent of neuroscience has allowed for a more in-depth examination of where and when important DVs are represented and manipulated in the brain, and has led to the development of the multidisciplinary field of *neuroeconomics* (Platt & Glimcher, 1999; Rangel, Camerer, & Montague, 2008; Wyart, de Gardelle, Scholl, & Summerfield, 2012). Of particular interest for both economists and psychologists is understanding how information about

losses and gains is represented. Previous evidence suggests that the two are represented asymmetrically, with more psychological impact attributed to losses (Kahneman & Tversky, 1979). This often leads can lead to attitudes like loss aversion, which favor avoiding losses, rather than pursuing equivalent gains. Similarly, it is unclear how value information for both intrinsic and extrinsic attributes of an object are processed. For example, are intrinsic benefits (ex. taste, appearance) of an object processed in the same manner as extrinsic costs (ex. effort)? Are these different types of information represented in a continuous, parallel manner, or a discrete, serial manner?

The current work will address these topics in three separate studies, using behavioral, eye tracking, and neuroimaging methods. Generally, studying the decision process can be approached in several different ways; one approach is to focus on the decision phase itself (where DVs are generated) and probe how decision relevant information is being represented prior to choice. Alternatively, another approach is to focus on the outcomes of decisions to make inferences about how decision relevant information was represented prior to choice. The first study adopts the latter approach to clarify what information is represented in the feedback-related negativity (FRN), a well-characterized ERP component believed to be associated with outcome processing. A wide body of literature has demonstrated that the FRN is sensitive to the valence of an outcome (i.e. losses/negative outcomes vs. gains/positive outcomes)(Gehring & Willoughby, 2002; Nieuwenhuis, Holroyd, Mol, & Coles, 2004; Sambrook & Goslin, 2015), but the literature remains mixed regarding to what degree the FRN represents information about the probability and magnitude of outcomes (Alexander & Brown, 2011; Hajcak, Holroyd, Moser, & Simons, 2005; Holroyd & Coles, 2002; San Martín,

René, Manes, Hurtado, Isla, & Ibañez, 2010). This might be due to experimental designs that don't allow for each of these decision-relevant variables to be studied simultaneously. To address this, we designed a modified two-armed bandit task to assess how information about valence, probability, and magnitude of monetary losses and gains might be represented in the FRN.

Although attitudes such as loss aversion, which may reflect underlying distortions in valuation, have been widely studied (Fox & Poldrack, 2008; Kahneman & Tversky, 1979; Tom, Fox, Trepel, & Poldrack, 2007; Yechiam & Hochman, 2013) it is unclear at what stage of processing the biases manifest. Is information about losses and gains processed in a manner reflective of loss aversion during initial valuation, or are these biases imposed only after information about gains and losses has been integrated? Furthermore, does loss aversion reflect a shift in the categorical boundary that separates losses and gains, or does it instead reflect a selective sensitivity to the magnitude of gains, but not losses (Kahneman & Tversky, 1979)? The second study of the current work addressed this by employing a novel gambling task where optimal performance requires choices to be based on the integration of multiple pieces of evidence. This task, in conjunction with multivariate pattern classification methods inspired by previous neuroimaging work (Haxby et al., 2001; Kriegeskorte, 2008) offers an opportunity to map the representational space of value-related information during the decision phase in an attempt to directly address these questions.

The value ascribed to an item under consideration is often times related to an intrinsic characteristic of that item (ex. taste, color, etc.). However, value can also be affected by characteristics that are *extrinsic* to the item (ex. the cost or effort required to

obtain the item). The degree to which information for both intrinsic and extrinsic attributes of an object are processed remains unclear. Few studies have examined whether these different types of information attributed to items under consideration (i.e. value vs. effort) are represented and ultimately integrated via parallel or serial processes (Harris, Adolphs, Camerer, & Rangel, 2011). The third study of the current work probed the dynamics of the representation of value and effort related information by utilizing eye tracking while subjects performed a value based decision task where in some instances, they were given foreknowledge about the presence of an effortful task that would be associated with one of the upcoming decision options. Such a design allowed us to independently track the online prioritization of effort or value information, as well as track when (if at all) the two types of information were integrated into something resembling a DV via fixation behavior. Furthermore, this gave us the opportunity to examine whether or not integration as indexed through visual fixations were actually predictive of choice. Such an approach provides insight into the temporal generation of behaviorally relevant DVs. Overall, this work implemented a multi-method approach to obtain a more comprehensive understanding of the underlying representations of decision-relevant information that are critical for flexible decision making behavior.

CHAPTER II

THE FEEDBACK RELATED NEGATIVTY IS MODULATED BY VALENCE AND MAGNITUDE, BUT NOT PROBABILITY, OF OUTCOMES

Introduction

In order to survive, it is fundamental for organisms to be able to monitor and evaluate the outcomes of their actions. Being able to distinguish whether an action resulted in a rewarding or non-rewarding/punishing outcome provides an opportunity to guide future behavior. Indeed, over the past several decades, advances in neuroscience have led to a search for whether there exist neural indicators of a flexible outcome monitoring system. A candidate signal that has been reliably established to evaluate outcomes across many studies is the so-called feedback-related negativity (FRN), an event-related potential believed to be elicited by the anterior cingulate cortex (ACC) (Hauser et al., 2014; Miltner, Braun, & Coles, 1997). In virtually all studies examining the FRN, it has been shown to differentiate between feedback indicating correct and incorrect task performance, or (i.e. feedback valence). Specifically, a large FRN (negative deflection in the event-related potential) is typically present anywhere between 200-400 milliseconds following the onset of incorrect –but not correct– feedback (Gehring & Willoughby, 2002; Hajcak, Moser, Holroyd, & Simons, 2007; Holroyd & Coles, 2002; Holroyd, Larsen, & Cohen, 2004; Holroyd, Nieuwenhuis, Yeung, & Cohen, 2003; Miltner et al., 1997; San Martín, 2012; San Martín, Kwak, Pearson, Woldorff, & Huettel, 2016). For example, during gambling tasks, it is common to see an FRN elicited following a gamble that results in a loss compared to a gamble that results in a gain. It should be noted that while an FRN following losses is the commonly reported effect, it

can respond flexibly to whatever the "worst" outcome is (see Holroyd, Larsen, and Cohen (2004)). This effect has been shown to occur irrespective of the modality of feedback (visual, auditory, tactile, etc.) (Miltner et al., 1997; Talmi, Atkinson, & El-Deredy, 2013) and across a variety of tasks (time estimation, reinforcement learning, gambling, etc.) (Ferdinand, Mecklinger, Kray, & Gehring, 2012; Gehring, Goss, Coles, Meyer, & Donchin, 1993; Hajcak, Moser, Holroyd, & Simons, 2006; San Martín, René, Appelbaum, Pearson, Huettel, & Woldorff, 2013).

Although the FRN seems to serve as a robust categorical indicator for "good" or "bad" outcomes, a large body of research has also examined whether or not it carries additional decision-relevant parameters. Perhaps the most frequently investigated of these is whether or not the FRN carries information about the perceived *probability* of an outcome occurring. This is often discussed in terms of a reward prediction error (RPE), which is formally defined as the difference between the anticipated reward outcome (often derived by prior experience) and the delivered reward outcome, and is a common topic in reinforcement learning (Botvinick, Weinstein, Solway, & Barto, 2015; Cavanagh, Frank, Klein, & Allen, 2010; Nieuwenhuis et al., 2004; Sambrook & Goslin, 2014; Silvetti, Castellar, Roger, & Verguts, 2014; Zani & Proverbio, 2003). One of the first major theories to propose the functional significance of the FRN was the Reinforcement-Learning (RL) theory of the FRN (Holroyd & Coles, 2002). This theory posits that the FRN reflected a RPE elicited by the ACC. More specifically, it explicitly proposed that the FRN represented a *negative* RPE (-RPE) such that FRNs are elicited when feedback is indicative of outcomes that are *worse* than expected, rather than *better* than expected. RL theory was thus claimed that FRNs are typically the result of

unexpected losses, and that the size of the FRN should scale monotonically with the degree of unexpectedness. Thus, according to the RL theory, the FRN represents both the valence of the outcome (i.e. good vs. bad), and also the relative expectedness or probability of the outcome (as would be required to classify as a prediction error signal).

Following the proposal of RL theory, a number of studies sought to explicitly test the claim that FRNs are the result of unexpected losses by designing studies where they vary the probability of a certain outcomes. At the time of publishing, there was little evidence aside from the study included in Holroyd and Coles' (2002) paper proposing the RL theory of the FRN that suggested that it scaled with unexpected losses. Since then, a number of additional studies have described effects consistent with this theory, suggesting that the FRN reflects a –RPE. (Bellebaum, Polezzi, & Daum, 2010; M. X. Cohen, Elger, & Ranganath, 2007; Hewig et al., 2007; Martin & Potts, 2011; Walsh & Anderson, 2012). However, not all studies have produced effects consistent with RL theory; in several cases, FRNs produced by unexpected losses were equivalent to those produced by expected losses (Hajcak et al., 2005; 2006). However, an additional study conducted by the same group found that FRNs behaved in a manner consistent with RL theory when subjects were asked to predict the outcome of their choice prior to receiving feedback; here, unpredicted losses resulted in larger FRNs than predicted losses (Hajcak et al., 2007). These inconsistences suggest that this specific effect may be sensitive to subtle changes in task design, and thus the replicability of the effects proposed by the RL theory merit further investigation.

In addition to the RL theory proposed by Holroyd and Coles (2004), more recently, Alexander and Brown (2011) have provided an alternate account as to how the

FRN may represent the probability of outcomes. They propose a prediction-response-outcome (PRO) model, which posits that the FRN produced by the ACC instead reflects an *unsigned* prediction error (UPE). In contrast to the RL model, which claims the FRN is a signed prediction error (specifically, a *negative* prediction error), the PRO model claims that the FRN responds to the *absolute* deviation from expectation. As such, the PRO model predicts that equivalent FRNs should be elicited by *both* improbable gains and improbable losses; the model explicitly states that these responses should be agnostic to the valence of the outcome. This implies that the FRN reflects more of a generalized "surprise" signal, rather than one that is specific to –RPEs.

Even prior to the publication of Alexander and Brown's (2011) paper proposing the PRO model, there was some evidence to suggest the FRN reflected a UPE. One study found that during a difficult time estimation task where subjects were asked to rate their performance prior to feedback, a mismatch between their rating and the feedback would result in an FRN (Oliveira, McDonald, & Goodman, 2007). Importantly, this FRN would be elicited following both unpredicted positive and negative feedback, and was equivalent in both conditions. Following the proposal of the PRO model, several other studies found similar results using paradigms that involved time estimation (Ferdinand et al., 2012) as well as delivery/omission of monetary rewards and electric shocks (Talmi et al., 2013). However, most previous studies that have examined the FRN and have manipulated outcome probability have not found effects consistent with the PRO model. In some cases, unexpected events were actually observed to result in more positive waveforms (the FRN is a negative deflection) (Walsh & Anderson, 2011). The relative dearth of evidence suggesting the FRN behaves as a UPE as proposed by the PRO model,

as well as the inconsistencies with previous findings suggest that further investigation is required to better clarify the functional significance of the FRN.

In addition to sensitivity to the probability of outcomes, a number of studies have examined if and how the FRN responds to the *magnitude* of outcomes. The RL theory proposed by Holroyd and Coles (2002) asserts that the FRN should behave in a manner that codes the magnitude of an outcome (Holroyd & Coles, 2002; Nieuwenhuis et al., 2004; Walsh & Anderson, 2012). Specifically, waveforms should be more negative following larger negative outcomes when compared to smaller negative outcomes, and should be more positive following larger positive outcomes when compared to smaller positive outcomes. In addition to the probability sensitivity proposed by RL theory, magnitude sensitivity would make the FRN sensitive the expected value of an outcome; that is, the value (magnitude) of the outcome weighted by the probability of the outcome.

As with the numerous studies examining the degree to which the FRN carries information about the probability of an outcome, the evidence regarding whether it does the same for the magnitude of an outcome is mixed. Several studies have (to certain degrees) supported the idea that the FRN codes magnitude sensitivity in a manner consistent with RL theory (Bellebaum et al., 2010; Gu et al., 2011; Sambrook & Goslin, 2014; San Martín, René et al., 2010). Interestingly, in some cases, magnitude sensitivity of the FRN is only found following either gains or losses, but not both. For example, Bellebeaum et al. (2010) showed that FRNs displayed magnitude sensitivity to only losses, whereas San Martin et al. (2010) showed the same effect but only for gains. The relative abundance of findings like these may be indicative of some sort asymmetry in loss/gain magnitude coding for the FRN. To make the picture even more unclear, several

studies have actually found *the opposite* of what is predicted by RL theory with regards to magnitude coding of the FRN, where larger outcomes resulted in smaller FRNs compared to smaller ones (Banis & Lorist, 2012; Kamarajan et al., 2009; Santesso, Dzyundzyak, & Segalowitz, 2011).

Clearly, the literature is mixed regarding what sort of information is represented by the FRN. This may be due in part to experimental designs that emphasize specific parameters while neglecting others. At the very least, the FRN is a robust indicator of the relative "goodness" or "badness" of an outcome (Nieuwenhuis et al., 2004; Walsh & Anderson, 2012). However, the probability and magnitude effects remain wildly inconsistent, especially in the former case with two mutually exclusive theories (RL and PRO) of how the FRN codes the probability of outcomes. Furthermore, in many instances, probability and magnitude effects are studied separately; either a study only manipulates magnitude, but hold probability constant, or vice versa (Bellebaum et al., 2010; Gehring & Willoughby, 2002; Hajcak et al., 2005; 2006; Holroyd & Coles, 2002; Sambrook & Goslin, 2014). With theories such as RL theory proposing that the FRN may be modulated as a function of *both* probability and magnitude, it is necessary to develop a paradigm in which both probability and magnitude can be examined. In order to address these inconsistencies and clarify the informational content being represented by the FRN, the current work designed a gambling task specifically tailored to tease out the effects of outcome valence (loss vs gain), outcome expectancy (expected vs. unexpected), and outcome magnitude (large vs small) on the FRN. To further maximize the discriminability of effects, we performed trial by trial analyses using hierarchical linear modeling on the individual waveforms, rather than difference waves between the

averaged gain/loss waveforms, which allows us to tease out specific gain/loss related effects (e.g. magnitude effects). This approach will hopefully clarify the functional significance of the FRN in a manner that will further refine theory moving forward.

Methods

Participants

A total of 32 subjects **(**17 male**)** were recruited in return for monetary compensation ranging between $25 and $45**.** All subjects had normal or corrected-to-normal vision and gave written informed consent according to procedures approved by the University of Oregon Institutional Review Board.

Stimulus Displays

Stimuli for the task were generated in Matlab using the Psychophysics toolbox extension (Brainard, 1997). They were presented on a 17-inch flat cathode ray tube computer screen. Subjects viewed the screen from a distance of approximately 100 cm.

Experimental Task

Subjects were paid a base rate of $25 to perform 800 trials (8 blocks; 100 trials per block) of a modified 2-armed bandit task specifically designed to assess the responsiveness of the feedback-related brain activity to monetary outcomes of different valences (gain vs. loss), levels of expectancy, and magnitudes (high vs. low). Subjects were instructed that on each trial they had $10 of house money to gamble with, and that outcomes of each trial were entirely independent from one another (i.e. one could not accumulate money across subsequent trials). They were also told that in addition to their base pay, at the end of the experiment, one trial would be drawn at random, and any remaining money they earned through gambling ($0-20) would be theirs to keep. Each

trial began with the presentation of a central fixation cross in between two empty square frames for 500 ms. Following this, a probability cue was presented above fixation for 200 ms that explicitly cued the probability of the current trial resulting in a *gain*. This information was communicated as a percentage that could be either a 30%, 40%, 50%, 60%, or 70% chance of gain on the current trial. Each probability cue occurred an equal number of times per block. It is important to note that the probability of incurring a monetary *loss* on any given trial was equal to 1 minus the probability of reward (e.g. if the initial cue states that probability of gain is 40%, probability of loss is 60%). 1000 ms after the offset of the probability cue, a gambling option appeared inside each of the two previously empty frames; one was always a low option ($1-$5) and the other was always a high option ($6-10). The position of the low and high gambling options (left or right) was randomized, and each combination of low and high gambles occurred an equal number of times within each probability. Subjects had to indicate whether they would rather accept the left or right gamble via a key press using either the left ("z" key) or right ("?/" key) hand, respectively. Upon generating a response, the frame surrounding their chosen gamble was bolded for 800ms. Feedback for the outcome of their gamble was presented for 1500 milliseconds by having the entire chosen gamble change color (orange or blue) to indicate whether they had just gained or lost the amount of money they had gambled on (the colors indicating gains or losses were counterbalanced across subjects). See Figure 2.1 for a timeline of a trial of this gambling task. Critically, gambling outcomes played out exactly to their probabilities; exactly 30% of trials with the 30% cue were wins, while the other 70% were losses, etc. Each trial was separated by a 500 millisecond intertrial interval. At the end of the experiment, one trial was drawn at

random, and the monetary outcome was applied to an endowment given to the subject prior to the study. Any remaining money was theirs to keep afterwards.

Electrophysiological Recording, Processing, and Analysis

EEG was recorded from a cap with 22 embedded Ag/AgCL electrodes specifically designed for EEG recording (Electrocap International) using the 10/20 system of electrode placement. Electrode sites included F3, FZ, F4, T3, C3, CZ, C4, T4, P3, PZ, P4, P03, P04, P0Z, T5, T6, O1, O2, OL, and OR. All electrode sites were referenced to the left mastoid, and all data was re-referenced off-line to the algebraic average of the left and right mastoids. Horizontal electrooculogram (EOG) were recorded from electrodes placed approximately 1 cm to the left

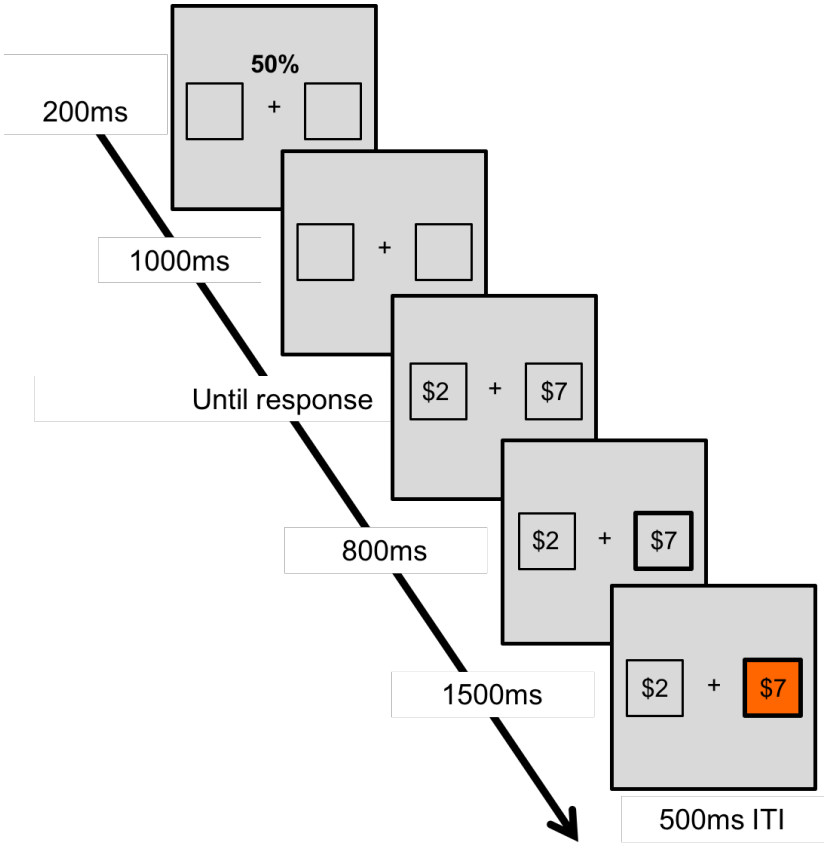

Figure 2.1. Timeline of a trial for the gambling task.

and right of the external canthi of the eyes of each subject to measure horizontal eye movements. Vertical EOG were recorded to detect subject eye blinks from an electrode placed beneath the left eye and referenced to the left mastoid. Impedances of all channels were kept below 5 kΩ. EEG and EOG signals were amplified with an SA Instrumentation amplifier with a bandpass filter of 0.01-80 Hz and were digitized at 250 Hz in LabView 6.1 running on a PC.

EEG data were preprocessed offline in MATLAB (MathWorks) using the EEGLAB toolbox (Delorme & Makeig, 2004) and custom scripts. For each trial, raw EEG was separated into 1200 millisecond epochs time-locked to the onset of the feedback stimulus (200 milliseconds pre-stimulus, 1000 milliseconds post-stimulus). The 200 milliseconds pre-stimulus portion served as a baseline. Trials with vertical eye movements in excess of 300 μV across a 200 millisecond sliding window, or horizontal eye movements in excess of 35μV across a 250 millisecond were excluded from analyses.

A large body of previous work has suggested that the FRN is fronto-centrally localized and peaks anywhere between 200-400 milliseconds post-feedback. (Gehring & Willoughby, 2002; Miltner et al., 1997; Nieuwenhuis et al., 2004; Sambrook & Goslin, 2015). Thus, the FRN was operationalized as the mean voltage between 260-360 milliseconds post-feedback at channel FZ (this time window was chosen based upon a visual inspection of the grand average loss vs. gain waveforms).

Results

Choice Behavior Reveals Probability Matching

Binary logistic hierarchical linear models were constructed in R (R Core Team, 2016) using the lme4 package (Bates, Mächler, Bolker, & Walker, 2015) to assess whether subject's trialwise choice behavior (bet low or bet high) was sensitive to the probability cue indicating the explicit probability of experiencing a gain on that trial following their choice. A full model was constructed which specified both random intercepts (subject ID) and random slopes (probability) where trials (level 1) were nested within subjects (level 2). There was a significant linear effect of probability (b = 4.52), z = 10.24, p < .001 demonstrating that as the probability of experiencing a gain increased, the proportion of high bets likewise increased in a linear fashion (see Figure 2.2**)**. This supports evidence for probability matching during the task, rather than adopting a more optimized strategy (i.e. always betting low when probability of gain < 50%, always betting high when probability of gain > 50%, and betting randomly when probability of gain = 50%). Additionally, there was significant quadratic effect of probability (b = 4.52), z = 10.24, p < .001, suggesting that the propensity to choose the high bet decreased as probabilities approached maximal uncertainty (i.e. the 50% probability cue).

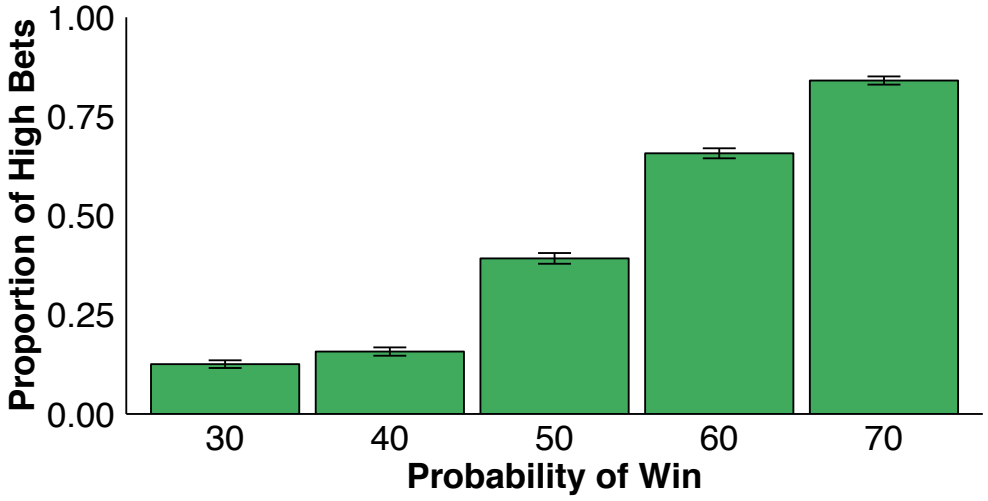

Figure 2.2. Proportion of high bets across different probabilities of winning in the gambling task.

FRN Responds Primarily to Valence and Magnitude of Outcomes

Hierarchical linear models were constructed in R (R Core Team, 2016) using the lme4 package (Bates et al., 2015) to assess which of the three experimental factors best explained modulations in FRN amplitude: outcome valence (gain/loss), magnitude of outcome (high/low), of outcome, and probability. A full model was constructed which specified both random intercepts and random slopes for each factor where trials (level 1) were nested within subjects (level 2). Consistent with previous literature (Gehring & Willoughby, 2002; Miltner et al., 1997; Nieuwenhuis et al., 2004), there was a main effect of outcome valence (b = 2.06), t = 6.59, p < .001; FRNs were larger (i.e. more negative) following losses (M = 4.18, SE = 0.09) when compared to gains (M = 6.60, SE = 0.10). With regards to the PRO model proposed by Alexander and Brown (2011), since the predictor used in this model reflected the probability of *gain,* a main effect of probability as described by the PRO model would best be captured by a quadratic function, rather than a linear one in order to represent a symmetric effect of the FRN on each side of the range of probabilities (where either wins or losses are more expected). In contrast with the PRO model, there was no quadratic effect of probability (i.e. UPE) observed with the FRN, (b = 0.53), t = 1.42, p = *ns*. However, there was also a main effect of outcome magnitude (b = 0.97), t = 2.36, p < .05; FRNs were smaller following high magnitude outcomes (M = 6.28, SE = 0.10) when compared to low magnitude outcomes (M = 4.72, SE = 0.09).

In addition to the main effects, there were several significant interactions that informed how information about valence, probability, and magnitude might be carried in the FRN. There was a significant two-way interaction between outcome valence and

magnitude (b = 1.03), t = 2.72, p < .01. This interaction was primarily driven by the fact that FRNs were significantly diminished (i.e. the waveforms were more positive) following *high magnitude gains*. Thus, it appeared that FRNs were only sensitive to the magnitude of gains, but not losses. Interestingly, a significant three-way interaction (b = -2.54), t = -2.80, p < .05 indicated that high magnitude gains that are *more expected* actually result in *larger* FRNs. This specific result is difficult to interpret a point that will be revisited in the discussion. Figure 2.3 shows waveforms for all possible valence, probability and magnitude conditions. See Figure 2.4 for mean FRN amplitudes in the 260-360 millisecond time window.

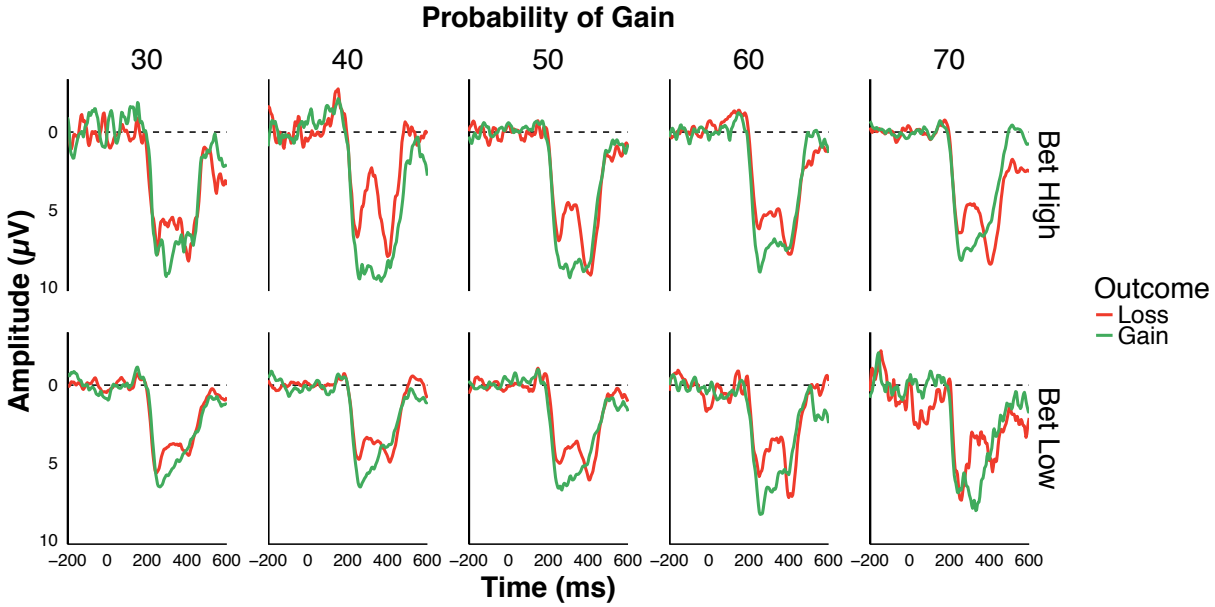

Figure 2.3. Grand average waveforms time-locked to the onset of the feedback for every combination of valence, probability, and magnitude condition.

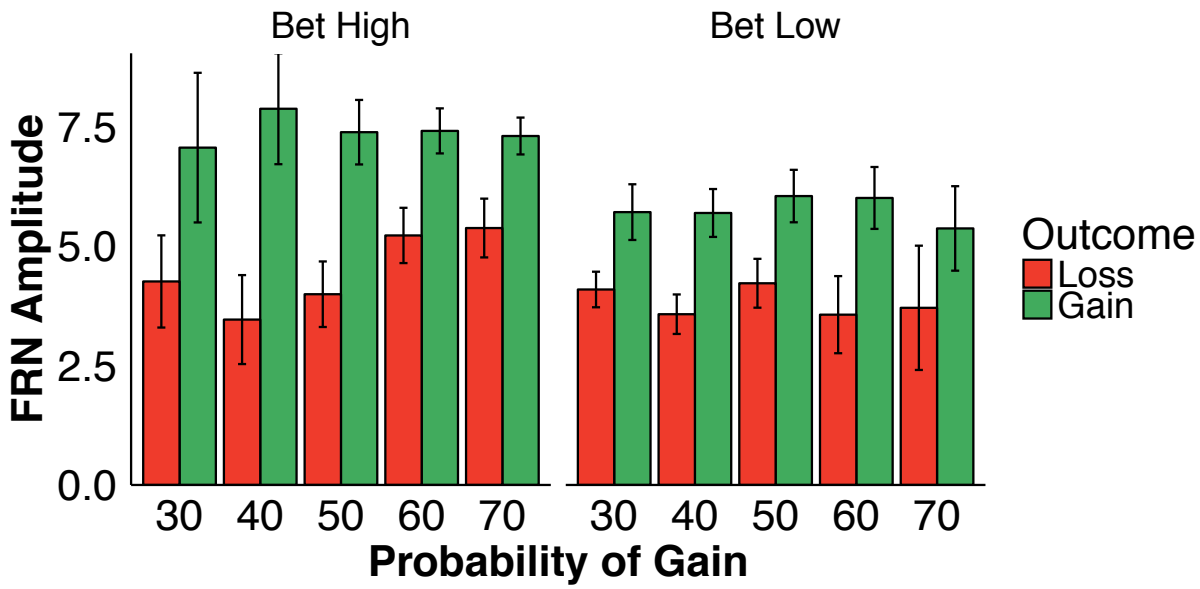

Figure 2.4. Mean FRN amplitudes across the 260-360 millisecond time window for all conditions (higher bars indicate smaller FRNs).

Discussion

The ability to accurately monitor the outcomes of our actions is essential for survival. The so-called feedback-related negativity (FRN) is an event-related potential that is widely believed to be a neural indicator of outcome monitoring. Aside from coding the relative "goodness" or "badness" of an outcome, the degree to which the FRN is sensitive to additional choice-relevant parameters is unclear, despite several theoretical models which attempt to establish the functional role of the FRN in carrying information about the probability and magnitude of the outcomes experienced (Alexander & Brown, 2011; Holroyd & Coles, 2002). The current work sought to clarify these inconsistencies by designing a paradigm that simultaneously tests effects of valence, probability, and magnitude of outcomes on the amplitude of the FRN. Overall, it was found that the FRN was primarily sensitive to the valence of the outcome, consistent with the overwhelming majority of previous literature examining the FRN. Additionally, there was no evidence

to suggest that the FRN was selectively sensitive to the probability of an outcome in either a manner consistent with a –RPE or UPE. Finally, there was some degree of modulation of the FRN by the magnitude of outcomes, but this was primarily driven by the outcome of gains, not losses. Each of these results will be discussed further in the following sections.

The current work found a very large effect of outcome valence on the amplitude of the FRN; FRNs were overall larger (i.e. more negative) following feedback for losses, and smaller following feedback for gains. This result is unsurprising, since it is the result that helped bring the FRN into the forefront of ERPs putatively associated with outcome monitoring via feedback (Miltner et al., 1997). The current work adds to an already large body of literature to suggest the FRN likely represents the output of an outcome monitoring systems that to some extent bins outcomes along a binary good/bad dimension (Hajcak et al., 2006). This result further reinforces a large body of literature that the FRN in part represents the motivational value of ongoing outcomes.

The FRN Does Not Respond to Outcome Probability

The current work did not find any strong evidence to indicate that the FRN is sensitive to the probability of an outcome. This contrasts with both the RL and PRO models, which propose that the FRN behaves like a –RPE, and UPE, respectively. Our results are largely consistent with studies like that of Hajcak et. al (2005), who found that probability of an outcome did not modulate the amplitude of the FRN, but inconsistent with numerous other studies that have found such an effect. This may be fundamentally due to the nature of the task employed in the current work; we utilized a gambling task that explicitly cued the probability that the current trial would result in a gain, and then

allowed subjects to freely choose what kind of bet they wanted to place. In this task, the relevant probability information was provided explicitly, and subjects did not have to (and technically could not) learn the optimal response strategy via reinforcement learning. Interestingly, this feature of explicit instruction of the probability of reward was also found in Hajcak et al. (2005), who failed to find an effect of outcome probability on the amplitude of the FRN.

To clarify some of the instances in which the FRN did not respond like a –RPE, Holroyd (2009) designed a series of experiments to probe necessary conditions that need to be met in order for the FRN to behave in such a manner. The first of these experiments was a replication of the original Hajcak (2005) study but with more extreme probability conditions; rather than conditions that reflected 25%, 50%, and 75% chance of reward, Holroyd et al. (2009) used 5%, 50%, and 95% (in response to the suggestion by Hajcak et al. (2005) that the FRN may only respond to extreme expectancy violations). Here, using an experiment where probability of reward was explicitly cued prior to choice, they found at least partial support for RL theory; unexpected losses resulted in larger FRNs than expected losses, but the relationship between unexpected, expected, and equiprobable losses was not monotonic. This effect though, was considerably weaker than previous studies that provided support for RL theory. In experiments two and three, subjects were required to actually learn the optimal response strategy via trial-and-error. However, in experiment two, despite subjects believing that they had learned the optimal strategy, feedback was random. In contrast, in experiment three, feedback was truthful and directly the result of having learned a response strategy. The results of experiment two mirrored those of experiment one, but experiment three demonstrated a monotonic sensitivity of

the FRN to outcome expectedness, consistent with what is predicted by RL theory (Holroyd & Coles, 2002).

The work of Holroyd (2009) thus strongly suggests that the experimental factors that need to be present in order for the FRN to behave in the way predicted by RL theory are that subjects need to actually learn a response strategy via trial-and-error, and that feedback needs to be truthful to their responses. In contrast, the current work, as well as Hajcak et al. (2005) did not utilize a task in which subjects could learn an optimal response strategy, due to explicitly providing subjects with the probabilities of reward prior to having to make a choice. When discussing why Hajcak et al. (2005) failed to find –RPE effects of the FRN, Walsh and Anderson (2012) suggested that subjects may not have had sufficient experience in the three probability conditions present in their experiment to develop strong enough expectations that could subsequently be violated in a manner that would result in a –RPE, or that the explicit probability cue did not directly map on to subjects' expectations (Hajcak et al., 2007). Although the current work did not allow for subjects to traditionally learn a strategy (like Hajcak et al. (2005)), one advantage of the current design is that we are able to indirectly assess subjects' expectations of the outcomes in each of the five probability conditions through their choice behavior. Here, subjects behaved in a way that indicated an attempt to maximize their earnings and minimize losses (i.e. betting high when the probability of winning is high, and betting low when the probability of winning is low). Furthermore, outcome probabilities within each probability condition were not manipulated in any way; participants very likely experienced violations of expectations. It is very unlikely that subjects in the current experiment did not have any expectations about the trial outcomes

in the different probability conditions. Whether or not this was insufficient to create an expectation whose violation would be manifested in the amplitude of the FRN is something that needs to be more carefully studied. This is further reinforced by the fact that other work has demonstrated that learning via instruction vs. trial-and-error does not result in quantitatively different FRN responses (Walsh & Anderson, 2011).

In addition to the FRN not behaving like a –RPE as predicted by RL theory, the current work found no evidence that the FRN behaves like a UPE, as predicted by the PRO model. This was the case even when the current task included losses and gains that equally unexpected, and in principle, should have both resulted in an FRN if the PRO model is correct. Evidence in support of the PRO model is fairly limited, with only a few studies demonstrating that the FRN can behave in a manner consistent with an UPE (Ferdinand et al., 2012; Oliveira et al., 2007; Talmi et al., 2013). Additionally, these experiments employ tasks that are quite different from the traditional reinforcement learning or gambling tasks often used when studying the FRN; Talmi et al. (2013) used a completely passive task where subjects were cued to receive either rewards or painful electric shocks that were either delivered or omitted, Oliveira et al. (2007) used a difficult motion estimation task where subjects had to rate their performance prior to receiving feedback about their performance, and Ferdinand et al. (2012) used difficult time estimation task. Ferdinand et al. (2012) criticized the use of gambling tasks to evaluate models like the PRO model for several reasons. First, they state that since gambling tasks do not typically offer opportunities for behavioral adjustment (similar to the findings of Holroyd et al. (2009)) and thus likely do not result in participants developing expectations of the outcomes, and that gambling tasks do commonly deliver gains and

losses in an equiprobable manner (see Gehring and Willoughby (2002)). Neither of these criticisms apply to the current task when you consider subject choice behavior, and the fact that we explicitly manipulated the probability of outcomes to not be purely equiprobable. Thus, it is unclear why there was no overall main effect of outcome expectancy on FRN amplitudes. One possibility is that the current task was not engaging enough to result in equally salient expectancy violations for both the loss and gain domains. The tasks that have shown the FRN behaving like a UPE have been either very difficult (Ferdinand et al., 2012; Talmi et al., 2013) or incredibly salient by using stimuli like painful electric shocks (Oliveira et al., 2007).

The FRN Responds to Outcome Magnitude, but Only for Gains

Despite the mixed literature regarding whether or not the FRN carries information about the magnitude of the outcome, the current work found a reliable main effect for outcome magnitude on the amplitude of the FRN. However, this main effect was qualified by an interaction that demonstrated that this magnitude was largely present for just gains, and not losses. Specifically, higher gains, as opposed to smaller gains, resulted in smaller FRNs/more positive waveforms overall. This is consistent with the results of San Martin et al. (2010), and only partially consistent with the RL theory prediction that larger losses should produce larger FRNs, and larger gains should produce smaller FRNs/more positive waveforms (Holroyd & Coles, 2002; Sambrook & Goslin, 2015). San Martin et al. (2010) explain their results by suggesting that the fact that FRNs were smaller following larger gains might have been due to a general loss avoidance strategy, whereby a greater value is placed on avoiding losses, which is reflected in more positive waveforms following large gains. Additionally, it has been suggested that these ERPs

following gains actually reflect subcortical reward-related processing (e.g. from the ventral striatum) (Holroyd, Pakzad-Vaezi, & Krigolson, 2008; Proudfit, 2015), and as such, it is reasonable that these ERPs would scale with the magnitude of gains. A complimentary explanation for the magnitude effect present only for gains but not losses is that the FRN to some extent reflects the asymmetric valuation of gains and losses as proposed by Prospect Theory (Fox & Poldrack, 2008; Kahneman & Tversky, 1979; Tversky & Kahneman, 1981; 1992). Such an explanation would explain why the FRN would not scale with magnitude of losses, since the proposed utility function for losses is markedly steeper, allowing less discriminability between losses of different magnitudes (i.e. similarly sized FRNs). In contrast the comparatively more gradual utility function for gains might allow for finer discriminability for gains of different magnitudes to produce the effects observed in the current study.

In the current work we also observed a significant three-way interaction between valence, the linear effect of probability of gain, and magnitude. This interaction indicated that high, expected gains actually resulted in *larger* FRNs (i.e. more negative waveforms). This is a somewhat anomalous finding that is difficult to interpret; if the FRN behaves as a true expected value signal, according to RL theory, one would expect that it FRNs would be *smaller* following high, expected gains. However, our findings are not consistent with that notion. Furthermore, this is the only significant effect that included an effect of probability of the outcome. In any case, it does not offer clear support for either the RL theory or the PRO model.

Limitations

One limitation of this study were the low bin counts for certain conditions. This was particularly apparent in the high gain probability/bet low/lose and low gain probability/bet high/win. The design of this study provided subjects with free choice, and thus actually required them to make choices to populate the various condition bins. This was compounded by the fact that the probability conditions resulted in outcomes that were true to the cued probabilities and not manipulated in any way. However, this was to a certain degree accounted for by relying on trial by trial analysis of EEG using hierarchical linear modeling, which can help mitigate the effects of low bin counts. Additionally, in the current task, it is possible that magnitude and probability are confounded, since the probability cues so strongly influenced choice behavior. This further justifies our analytical approach by independently teasing out the effects of probability and magnitude on a trial-by-trial basis. These experimental design choices were necessary in order to simultaneously test the effects of valence, probability, and magnitude on the FRN. If instead we only manipulated magnitude or probability across blocks, the task would become less engaging. This is problematic because it has been demonstrated that FRN effects are actually reduced the more passive a task becomes, and this reduction is correlated with subject ratings of task engagement (Yeung, Holroyd, & Cohen, 2005).

Conclusions

The current work sought to clarify what type of information is carried by the feedback-related negativity, a purported index of active outcome monitoring. Our findings helped further reinforce that at the very least, the FRN is a powerful indictor of the valence of an outcome, and correctly and coarsely distinguishes between them along

an abstract “good” or “bad” dimension. The lack of strong support for either the RL theory and PRO model with regards to the sensitivity of the FRN to –RPEs/magnitude and UPEs, respectively, in conjunction with the mixed support for these models in the literature, suggests that either they do not completely capture the true nature of the type of information carried in the FRN, or that the behavior of the FRN is highly contingent upon the type of task one is using to examine it. With regards to the lack of –RPE effects predicted by RL theory, claims of subjects not having clear expectations of the likely outcome of a trial (Walsh & Anderson, 2012) seem dubious considering how subjects made gambling choices in a way that indicated an expectation of the likely outcome of a trial given the initially cued probability. Thus, future work should better characterize whether or not behaviorally revealed expectations in tasks where no optimal response strategy can be learned are insufficient in contributing to expectancy violation/RPE effects that would manifested in the amplitude of the FRN. Finally, the fact that magnitude effects were present only for gains and not losses suggests that the FRN may to some degree represent value in an asymmetric manner, similar to the asymmetric utility functions proposed in Prospect Theory. Overall, the current work provides the impetus to more critically assess a number of proposed theories of the functional significance of the FRN.

CHAPTER III

WHEN DO DISTORTIONS IN VALUATION MANIFEST?

Study one demonstrated that signals like the FRN robustly distinguish between losses and gains in a categorical manner. Despite the fact that the FRN did not seem to be sensitive to the probability of an outcome, it did seem to carry information about the magnitude of an outcome. Specifically, the FRN distinguished between the high and low magnitude gains, but not losses. This result is consistent with an asymmetric representation of gains and losses (Fox & Poldrack, 2008; Kahneman & Tversky, 1979), whereby losses are processed in a more "all or none" manner regardless of magnitude, while preserving the sensitivity to the magnitude of gains. This is the hallmark of the valuation distortion known as loss aversion. Since the FRN largely represents the evaluation of an outcome presumably after all decision-relevant information has been computed to guide choice, it is unclear *when* these biases manifest during the decision process itself. This question, among others, was explored in study two.

CHAPTER IV

MULTVARIATE PATTERNS OF DELTA BAND ACTIVITY SHOW THE EMERGENCE OF LOSS AVERSION AFTER INTEGRATION OF VALUE

Introduction

Among the most common types of decisions made by humans are value-based decisions (ex. where to go for dinner, whether to make a risky gamble or play it safe, etc.). However, the efficiency in which we engage in these decisions is far from optimal; a large body of literature has demonstrated that humans are prone to distortions in valuation that manifest themselves in systematic biases during decision making (Kahneman & Tversky, 1979). Among the most well-known of these biases is loss aversion, in which individuals tend to prefer making decisions that avoid losses rather than those that result in equivalent gains. This tendency is believed to be the result of a valuation distortion best described by asymmetric differences in the calculation of subjective value (i.e. utility) for gains and losses. More specifically, Prospect Theory proposed that the utility function for losses is steeper than that for gains, suggesting that the psychological impact of losing is about twice as large as that of the equivalent gain (Fox & Poldrack, 2008; Kahneman & Tversky, 1979; Tversky & Kahneman, 1992). This has been succinctly captured by the adage, "losses loom larger than gains" (Tversky & Kahneman, 1991).

To date, it is not clear at what stage of processing these valuation distortions occur. In many instances, decision making takes place in two discrete stages; an initial valuation of multiple pieces of evidence, followed by an integration of that evidence into a *decision variable* (DV) upon which some sort of criterion can be imposed to guide

behavior (Shadlen & Kiani, 2013; Wyart et al., 2012). This two-stage approach provides a framework to directly assess when distortions such as loss aversion manifest themselves. For example, loss aversion could be due to an overweighting of losses during an initial valuation stage for each individual piece of evidence being considered for a final decision. Alternatively, information might be represented adequately at the initial valuation stage, but then integrated in a biased manner (i.e. during the calculation of expected value). Moreover, within each stage, such a loss-aversion bias during integration could occur in different manners. For example, it may represent a shift in the criterion with regard to where exactly the boundary between gains and losses is set (i.e. an intermediate categorical boundary vs. a shifted categorical boundary). Another possibility is that there is an asymmetry with regard to the precision with which gains versus losses are represented. Specifically, loss aversion would imply that losses are represented in an all-or-none manner (consistent with a very steep subjective value function in the loss domain), whereas gains may be represented in a more fine-grained manner, allowing a distinction between low and high gains (i.e. a selective continuous). See Figure 4.1 for a schematic of when biases such as loss aversion might occur.

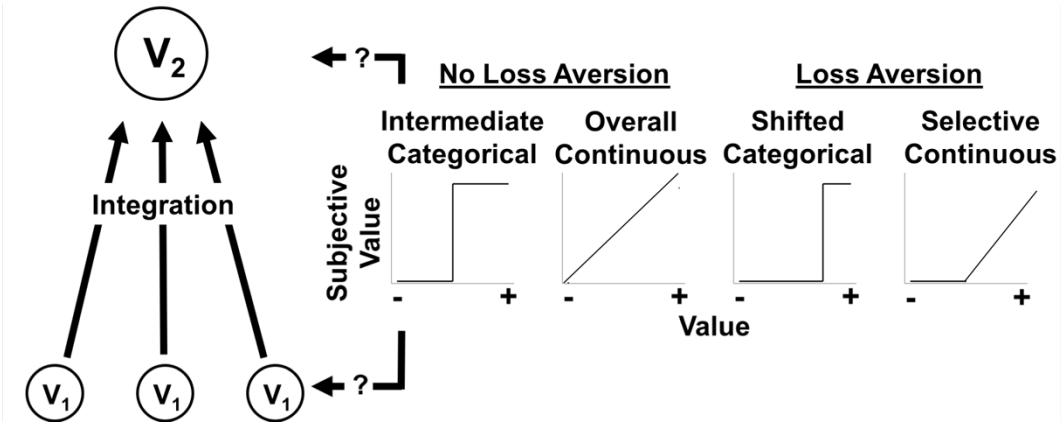

Figure 4.1. Schematic representation of stages of processing for value information. Right side illustrates different possible representations of value, which could occur either during initial valuation or later integration.

Investigating what stage of processing may produce distortions of valuation necessitates a task where optimal performance requires both valuation and integration stages. A model task that meets these requirements is the multi-sampling perceptual decision making task employed by Wyart, de Gardelle, Scholl, and Summerfield (2012). In their task, subjects were presented with a stream of eight Gabor patches at a rate of 4 Hz, and then had to decide whether the overall orientation of the stream of patches fell more along cardinal or diagonal axes. In order to optimally perform this task, the orientation of each Gabor patch should be transformed into a DV that categorizes each patch as either cardinal or diagonal (i.e. valuation). Critically, this DV would need to be updated throughout the trial with the presentation of each subsequent Gabor patch (i.e. integration) in order for the correct stimulus category (cardinal or diagonal) to be estimated more precisely. This basic multi-sampling design could be adapted to feature value based, rather than perceptual, decision making to better address what stage of processing loss aversion may occur.

Previous work has examined how loss aversion might be represented in humans through activity in putative valuation circuits such as the ventromedial prefrontal cortex (VMPFC) (Canessa et al., 2013; Tom et al., 2007) as well as through event-related potentials that respond to value such as the feedback-related negativity (FRN) (Kokmotou et al., 2017; San Martín, René et al., 2010). However, neither of these are appropriate measures for a multisampling context where information is presented at a rapid rate. A candidate neural signal that been demonstrated to be informative in such a context are EEG oscillations in the delta band (~1-3 Hz) (Wyart et al., 2012). In their task, Wyart et al. (2012) found that the DV associated each Gabor patch could be

predicted along various points in the phase of oscillations in the delta band (~1-3 Hz), suggesting that these low frequency EEG oscillations may carry information about DVs that could be subsequently integrated to inform behavior. In addition to its utility in a multisampling context, numerous other studies have suggested that the power of oscillations in the delta band may contain information about rewards (Bernat, Nelson, & Baskin-Sommers, 2015; Foti, Weinberg, Bernat, & Proudfit, 2015; Knyazev, 2007; 2012). Taken together, this suggests that delta power may contain information that could support the representations of value necessary to guide choice during value based decision making.

Rather than relying on traditional, univariate analyses, it could also be possible to map the representational space of value based information in delta power using a series of newer computational methods have been previously used "decode" complex, multivariate patterns of neural activity (Haxby et al., 2001; Kriegeskorte, 2008; Kriegeskorte et al., 2008). These methods, along with others such as representational similarity analysis (RSA; (Kriegeskorte, 2008)), and those utilized by Wyart et al. (2012) provide a basis upon which to design studies to better probe the informational content carried by various types of neural data. For example, these multivariate decoding methods could be leveraged to create confusion matrices to see whether or not information about value is arranged in a meaningful manner. In the case of the current work, confusion matrices could be used to assess whether delta power contains information that supports the presence and position of a categorical gain/loss boundary, as well as whether or not there is sensitivity to the magnitude of gains or losses. Such representations would directly inform how loss aversion might manifest in representations of value.

Taken together, the current work employed a novel approach to investigate what stage of processing of value information produces distortions such as loss aversion. We recorded EEG from human subjects as they performed a novel multi-sampling gambling task based on the multi-sampling perceptual decision making task of Wyart et al. (2012). During each trial, subjects were presented with a stream of eight pieces of value information (hereby referred to as elements) that they had to integrate in order to decide whether to select either a safe bet or a risky gamble. Critically, analyses took place separately at the element (valuation) level and the entire stream (integration) level. We began by decoding ranges of element-level or stream-level values subjects experienced over the course of the task from multivariate patterns of delta power across the scalp. We then mapped the representational space of value by creating 4x4 confusion matrices that reflected the distribution of decoding accuracies for both high and low gains and losses. If the value information carried in delta power is susceptible to distortions such loss aversion, we would expect one of the following: 1) a shift in a categorical gain/loss boundary that resulted in only high gains to be categorized as gains, or 2) a sensitivity to the magnitude of gains, but not losses, consistent with a very steep subjective value function in the loss domain. Due to the absence of previous work examining when loss aversion may occur during the processing of value information, the current work did not have strong predictions as to whether loss aversion would occur at the element or stream level. Overall, this study provides a novel way to investigate the possible underlying structure of and temporal manifestation of loss aversion by using a task and analytical approach that decouples two fundamentally different levels of processing.

Methods

Participants

A total of 24 subjects **(**3 male**)** were recruited in return for monetary compensation ranging between $20 and $40**.** All subjects had normal or corrected-to-normal vision and gave written informed consent according to procedures approved by the University of Oregon Institutional Review Board.

Stimulus Displays

Stimuli for the task were generated in Matlab using the Psychophysics toolbox extension (Brainard, 1997). They were presented on a 17-inch flat cathode ray tube computer screen. Subjects viewed the screen from a distance of approximately 100 cm.

Experimental Task

We recorded EEG from subjects as they performed 875 trials of a novel multisampling gambling task based on the perceptual decision making paradigm of Wyart et al. (2012). Each trial was initiated with the spacebar, after which subjects were presented with a ring that had 4 colored dots (orange or blue) on its outside border, each 90 degrees apart. The dots 180 degrees opposite of each other were always the same color. The color of the dots served as a visual landmark for a continuous range of monetary values represented by the ring. One color represented the maximum amount of money ($20), and the other color represented the minimum amount of money ($0). After the ring was presented for a variable interval between 875-1025 milliseconds, eight black dots (hereby referred to as elements) were presented serially at a rate of 4 Hz at various positions inside the border of the ring. Subjects were instructed that a given element's position in the ring indicated how much money that element was worth (a value that

could range between $0-$20). 500 milliseconds after the offset of the eighth element, choice options appeared on either side of the fixation cross: a gamble, or a safe bet of $10 (these appeared as "BET" and $10, respectively). If subjects chose to gamble, they could receive the monetary value of *one* of the eight elements that appeared on that trial, selected at random (there was no feedback indicating how much they received if they gambled). If they instead chose the safe bet, they could receive $10. Choice was indicated using the "z" key to select the left option, and "?/" key to select the right option. Once a choice was made, a border appeared around their selection for 250 milliseconds, after which would mark the conclusion of the trial (see Figure 4.2 for an example of a trial). Critically, no feedback was presented following gambles, as it has been suggested that providing feedback for gambles can distort subjects' overall risk preferences (Stanton et al., 2011). At the end of the experiment, one trial was drawn at random, and the monetary outcome of that trial (the value of a gamble or a $10 safe bet) was applied to a $20 endowment given to the subject prior to the study.

The orientation of the ring shifted randomly from trial to-trial, decoupling the spatial position of the elements from the monetary value they were associated with. The value of each elements was randomly drawn from a uniform distribution ranging from $0-$20. The average expected value of the entire stream of elements for a given trial was normally distributed around $10. The color of the landmarks demarcating the minimum and maximum monetary values on the ring were counterbalanced across subjects. Finally, choice options were randomly presented on either the left or the right so that subjects could prepare a motor response prior to the presentation of all elements.

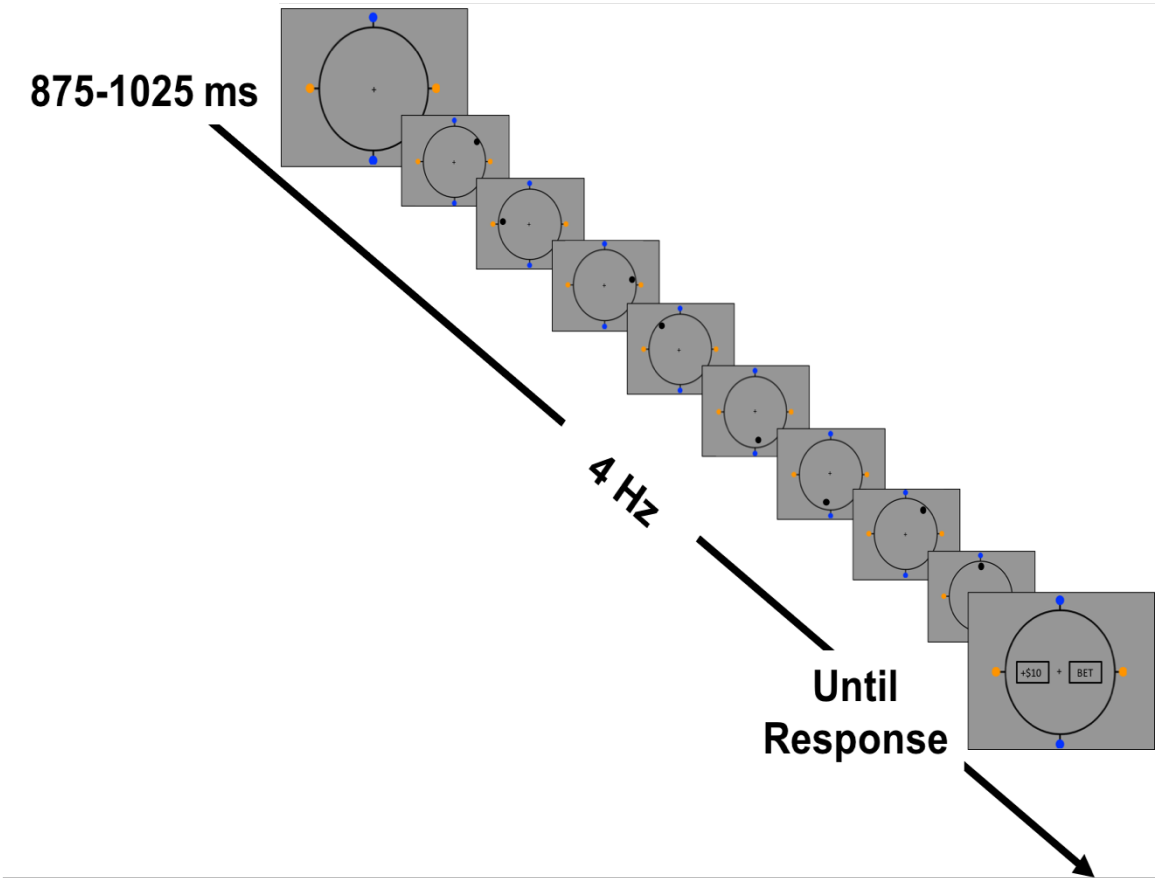

Figure 4.2. Timeline of a trial for the multi-sampling gambling task.

Electrophysiological Recording, Preprocessing, and Analysis

EEG was recorded from a cap with 22 embedded Ag/AgCL electrodes specifically designed for EEG recording (Electrocap International) using the 10/20 system of electrode placement. Electrode sites included F3, FZ, F4, T3, C3, CZ, C4, T4, P3, PZ, P4, P03, P04, P0Z, T5, T6, O1, O2, OL, and OR. All electrode sites were referenced to the left mastoid, and all data was re-referenced off-line to the algebraic average of the left and right mastoids. Horizontal electrooculogram (EOG) were recorded from electrodes placed approximately 1 cm to the left and right of the external canthi of the eyes of each subject to measure horizontal eye movements. Vertical EOG were recorded to detect subject eye blinks from an electrode placed beneath the left eye

and referenced to the left mastoid. Impedances of all channels were kept below 5 kΩ. EEG and EOG signals were amplified with an SA Instrumentation amplifier with a bandpass filter of 0.01-80 Hz and were digitized at 250 Hz in LabView 6.1 running on a PC.

EEG data were preprocessed offline in MATLAB (MathWorks) using the EEGLAB toolbox (Delorme & Makeig, 2004) and custom scripts. For each trial, raw EEG was segmented into 4700 millisecond epochs time-locked to the onset of the first element (700 milliseconds pre-stimulus, 4000 milliseconds post-stimulus). This large window was used to account for edge artifacts (M. X. Cohen & Cavanagh, 2011). A smaller -500 milliseconds pre-stimulus to 2350 milliseconds post-stimulus window was used for subsequent analyses. Trials with vertical eye movements in excess of 300 μV across a 200 millisecond sliding window, or horizontal eye movements in excess of 35μV across a 250 millisecond were excluded from analyses. Estimates of EEG power were obtained using the following methods described by Cohen and Cavanagh (2011). First, EEG data for each epoch were decomposed into time-frequency representations using a fast-Fourier transform. The power spectrum of this time-frequency representation was then multiplied with the power spectrum of complex Morelet wavelets with a frequency range of 1-25 Hz in 25 logarithmically spaced steps before taking the inverse fast-Fourier transform. Power was defined as the average squared amplitude of the resulting complex signal, and delta band power was defined as the average power between 1-3 Hz. Reported results are based on EEG data averaged across all electrode sites (Wyart et al., 2012).

Multivariate Pattern Analysis

General method.

In order to investigate whether patterns of delta power carries information about the manifestation of loss aversion on value representations during the element or stream level, we performed a series of multivariate pattern analyses using a naïve Bayes algorithm. First, all artifact-free trials were randomly partitioned into three blocks. To remove collinear processes (e.g., sensory adaptations), all data points were transformed into z-scores to de-mean power differences among electrodes, removing any univariate signals. Each block was assigned as a training set, which was used to develop a naïve Bayes function to decode the condition label of an independent test set (described in next section). We used a hold-one-block-out cross-validation procedure, in which power estimates from two of three blocks served as the training set and those from the remaining block served as the test set. This process was repeated until each block had served as the test set. To improve the signal-to-noise ratio of the analysis, we introduced an iterative procedure in which the entire process described above was repeated 35 times to obtain the average of results across iterations. Decoding results were averaged across both iterations and subjects.

Element vs. stream analyses.

The previously described method was applied slightly differently for element vs. stream level analyses. For element level analyses, each epoch of delta power was further subdivided into 600 millisecond (150 time samples at 250 Hz) segments time-locked to the onset of each of the eight elements. Thus, for this level of analysis, each trial yielded eight unique epochs of delta power to use for the decoding analyses (and thus, increased

the size of the dataset eightfold). Each element was assigned one of four condition labels to be used for decoding. These labels were based on the monetary value of the element, and were defined as follows: High loss ($0-$5), low loss ($5-$10), low gain ($10 – $15), and high gain ($15 - $20). Critically, gains and losses were defined as being *relative* to the $10 safe bet. Prior to averaging across iterations, the decoding analysis resulted in an c × c × t × i matrix of decoding accuracies for each subject. Here *c* is the number of condition labels, *t* is the number of time samples, *i* is the number of iterations. After averaging across iterations, the result was 150 4x4 confusion matrices, where each cell represented the average decoding accuracy (as a percentage) between the expected condition label and the label reported by the classifier. For element-level analyses, we averaged the time period just before the offset of the element until the end of the element epoch (248 – 600 milliseconds). It is important to note that despite using 600 millisecond epochs at the element level, where each epoch contains more than one element, it has been established using this method that the onset of an additional element does not disrupt the encoding of the previous element (Wyart et al., 2012).

For stream level analyses, each trial's entire 500 milliseconds pre-stimulus to 2350 milliseconds post-stimulus epoch (713 time samples at 250 Hz) of delta power was assigned one of four condition labels based on the average expected value of the entire stream of eight elements. Since the expected value of each stream was normally distributed around $10, the tails of this expected value distribution would contain fewer observations. To account for this, four bins were created based on average expected value, each with an approximately equal number of observations. These bins served as the condition labels, and were defined as follows: High loss *(M* = $5.59), low loss (*M* =

$8.63), low gain (*M* = $11.47), and high gain (*M* = $14.41). Again, gains and losses were defined as relative to the $10 safe bet. As a first step to determine whether multivariate patterns of delta band activity contained any information about expected value, we ran a modified version of the decoding analysis described above to produce a t x i matrix where *t* is the number of time samples and *i* is the number of iterations, that described overall decoding accuracy of the four expected value categories. Averaging across iterations resulted in a vector of decoding accuracies over time for the entire stream of elements. Next, we computed confusion matrices in the same manner described for the element level, but instead used the expected value condition labels. For stream-level analyses, we averaged the time period just after the offset of the entire stream of elements until the end of the stream epoch (2004 – 2350 milliseconds).

Assessment of neural representations of loss aversion.

In order to test for neural representations of loss aversion, all confusion matrices were first tested for two specific effects: 1) an intermediate categorical effect determining whether there is a categorical boundary that separates cells in the matrix defined by gain-related and loss-related condition labels (relative to the intermediate reference $10 point), and 2) an overall continuous effect determining whether there is overall sensitivity to the magnitude of the outcome (low or high). The presence or absence of these two effects allows us to infer the presence of representations that would be consistent with loss aversion. As mentioned before, loss aversion could potentially manifest itself in two ways; as a shift in the criterion upon which the boundary for gains and losses is defined (shifted categorical), or a selective sensitivity to the magnitude of gains, but not losses (selective continuous). Evidence for an intermediate categorical effect rules out the

possibility of a shifted categorical effect. Likewise, evidence of an overall continuous effect rules out the possibility of a selective continuous effect. In the case of the selective continuous effect of loss aversion, in addition to the lack of an overall continuous effect, it would also need to be established that there was magnitude effect present for only gains. The intermediate categorical effect was defined as the comparison between the correctly identified loss and gain categories, and incorrectly identified loss and gain categories (i.e. the lower left and upper right quadrants of the upper left and lower right quadrants of the confusion matrix; category ON vs category OFF). The overall continuous effect was defined as the comparison between the cells along the diagonal and those immediately off the diagonal for the quadrants representing correctly identified loss and gain information (i.e. lower left and upper right). See Figure 4.3 for a visualization of these potential effects.

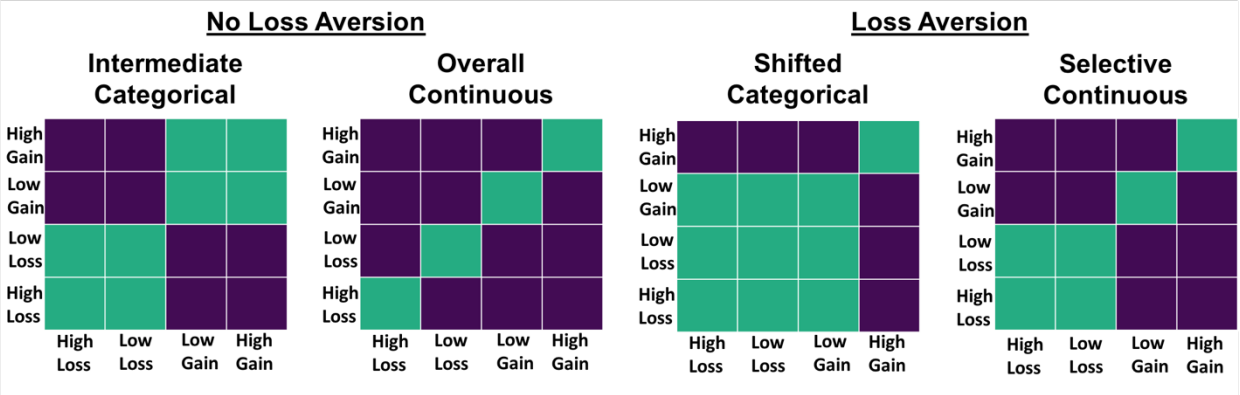

Figure 4.3. Possible representations of value for delta power.

Results

Betting Behavior and Response Times are Sensitive to Expected Value of the Stream.

As a first step, in order to determine whether subjects' betting behavior was sensitive to the expected value of the stream of elements, we predicted trialwise bets for

each of the four bins of expected value (described previously). Assuming subjects are motivated to maximize the amount of money they could earn in this task, we would expect for the proportion of bets to increase as a function of the expected value of the stream of elements (i.e., the higher the expected value, the more likely the bet). Indeed, we found that across increasing ranges of expected value, the proportion of bets increased linearly, (b = 4.92), z = 10.26, p < .001 (Figure 4.4A). In terms of overall betting behavior, this sample was relatively risk averse, with an average betting rate of 33%. In addition to the linear effect of expected value on betting behavior, there was also a quadratic effect of expected value on subject response times, (b = 74.11), z = 3.57, p < .01. Response times were overall faster for the lowest and highest ranges of expected value, and slower for the intermediate ranges, where the average expected values fell closer to the gain/loss boundary (i.e., $10, the value of the safe bet). This suggests that choices are likely easier in very low or very high expected value streams, whereas they are more difficult as the expected value of the stream approaches the value of the safe bet (Figure 4.4B). Overall, these results suggest that subjects are sensitive to the expected value of the entire stream of elements, and to some degree are calculating the expected value of the stream and making choices accordingly.

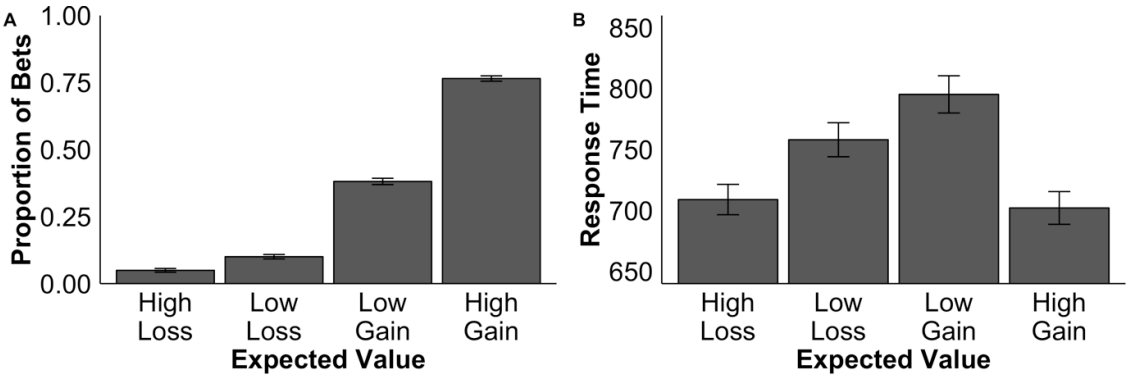

Figure 4.4. **A.** Betting behavior across different levels of expected value. **B.** Response times across different levels of expected value.

Each Element Uniquely Predicts Betting Behavior

In addition to the expected value of the stream predicting betting behavior, we also wanted to take a more granular approach and determine whether the raw monetary value of each of the eight elements predicted trialwise betting behavior. Furthermore, we wanted to determine whether there was any systematic over/under weighting of each element of value information based upon its serial position within the stream. To test this, we constructed a binary logistic hierarchical linear model where we predicted choice by the raw monetary value of each of the eight elements (entered into the model as separate predictors). Each element significantly predicted whether or not subjects chose to bet (all p values < .001), and each element was weighted approximately equally when predicting choice, regardless of serial position. This suggests that the position of a given element within the stream did not result in systematic over/under weighting while being considered for the final choice. Figure 4.5 shows the regression coefficients for each element position from this model (a coefficient value of 0 would indicate that a given element position was *not* predictive of choice).

Valuation at the Element Level Shows No Sign of Loss Aversion

Hierarchical linear models were constructed in R (R Core Team, 2016) using the lme4 package (Bates et al., 2015) to assess the presence of the categorical and continuous effects described previously at the *element level*. All analyses took place at the aggregate, subject level confusion matrices for the decoding of the condition label for each element. Separate models were created to test for both categorical and continuous representations of value present in the confusion matrices. Both models specified random intercepts and random slopes for their respective factors, and trials (level 1) were nested within subjects

(level 2). There was a large intermediate categorical effect, suggesting that delta power

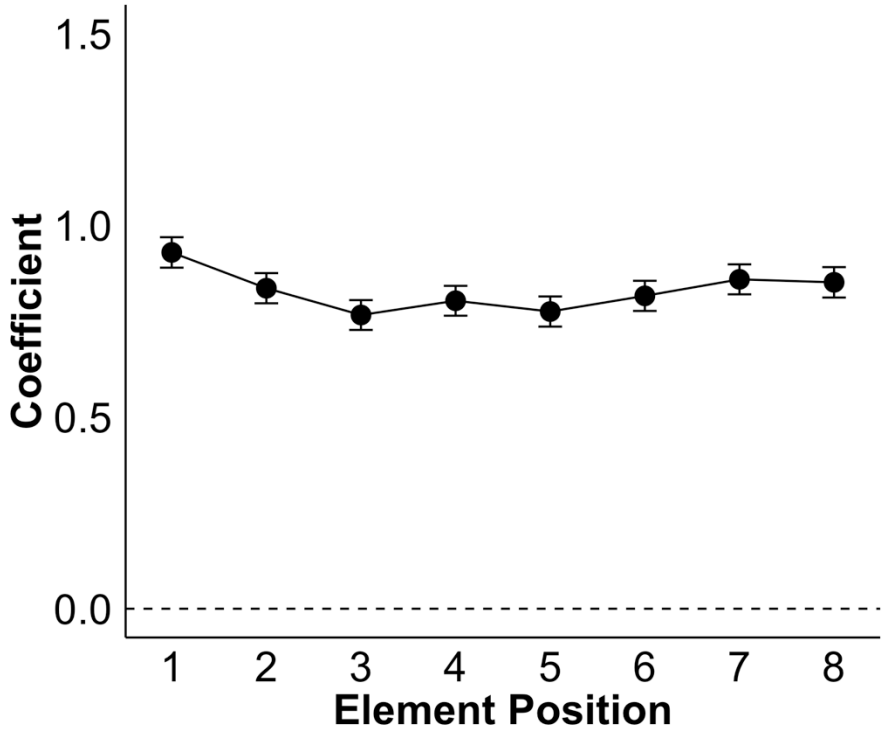

Figure 4.5. Regression coefficients predicting choice for each element position.

that contains information that differentiates between gain and loss related information for each element in the current task, and losses and gains (high and low) are more likely to be confused with category labels within the same category, (b = 0.17), t = 6.42, p < .001. This finding rules out the possibility of loss aversion manifesting as a shifted categorical effect at the element level. There was also no evidence of a difference in decodability between gains and losses, suggesting that they were equally represented in the delta band, (b = -0.01), t = -1.04, p = *ns*. Additionally, there was no evidence for an overall continuous representation of value at the element level, (b = -0.02), t = -0.78, p = *ns*. Finally, there was no evidence for loss aversion via a selective continuous effect; there was no difference in decodability between high and low outcomes between gains and losses, (b = 0.01), t = 0.49, p = *ns*. Overall, at the element level, it appears as though

gains and losses are processed in a strictly intermediate categorical manner, with no evidence for loss aversion by either a shift in the categorical boundary, or selective representation of magnitude for gains (Figure 4.6A).

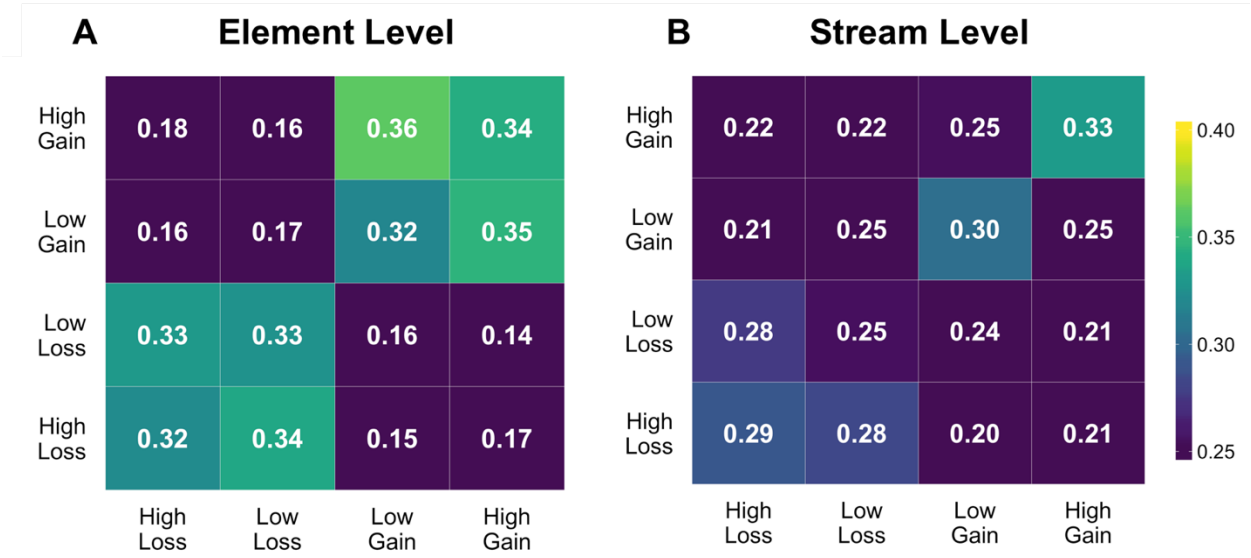

Figure 4.6. **A.** Confusion matrix of classification accuracy at the element level. **B.** Confusion matrix of classification accuracy at the stream level.

Valuation at the Stream Level Shows Some Signs of Loss Aversion

To first determine whether delta power contained any information that could reliably distinguish between average levels of expected value, we computed decoding accuracy for the levels of expected value over the course of the trial. Visualization of decoding accuracy over time showed that delta power contains information that results in above chance (25%) decodability of expected value. As would be expected by an expected value signal that would require computing the average value of the stream, rather than relying on any single element, above chance decodability began well into the trial (approximately 1000 milliseconds after the onset of the stream) and increased as the trial progressed, peaking towards the end of the epoch (Figure 4.7). This further

motivated analyzing the confusion matrices at the stream-level to assess the presence of the categorical, continuous, and loss aversion effects discussed previously.

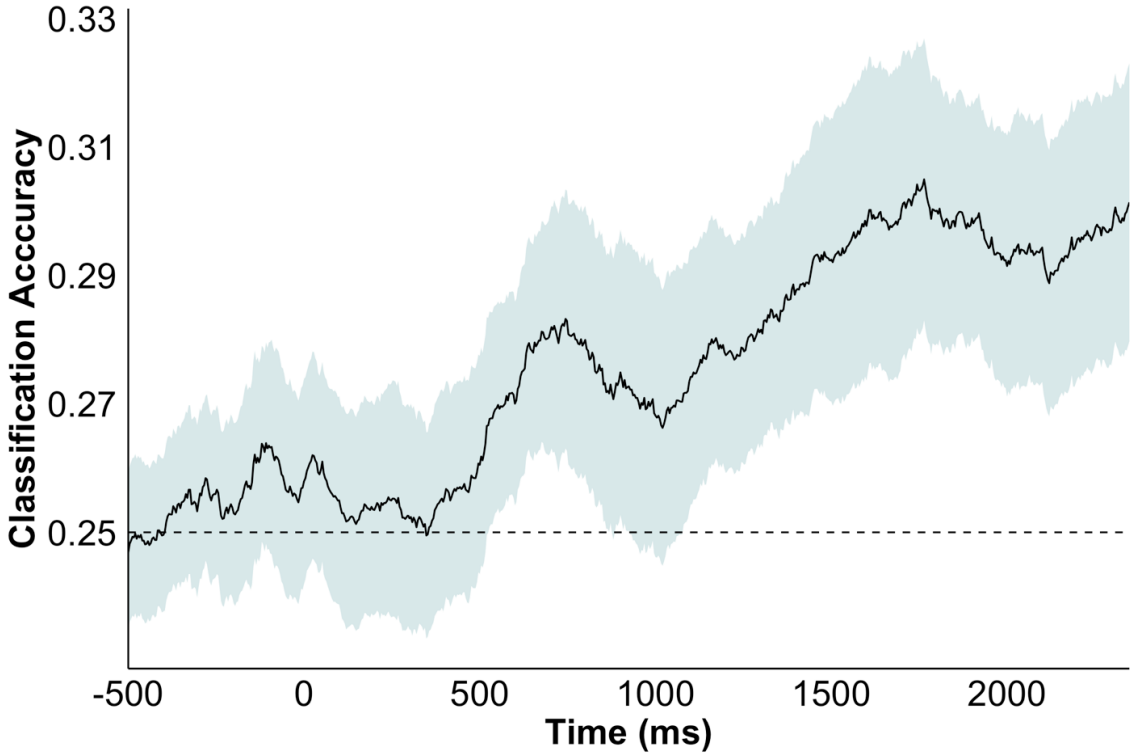

Figure 4.7. Classification accuracy of expected value across the course of a trial.

The same methods used for element-level analyses of confusion matrices were also used for the stream level confusion matrices. Like at the element-level, there was an intermediate category effect, suggesting that delta power contains information that differentiates between gain and loss related information for the entire stream of elements in the current task, and that losses and gains (high and low) are more likely to be confused with category labels within the same category, (b = 0.06), t = 2.72, p < .05. Compared to the element level, the size of this category effect was markedly smaller, but this finding rules out the possibility of loss aversion manifesting as a shifted categorical effect at the stream level. Also like at the element level, there was no evidence of a difference in decodability between gains and losses, suggesting that they were equally

represented in the delta band, (b = -0.03), t = -1.14, p = *ns*. There was also no evidence for an overall continuous representation of value at the stream level, (b = -0.01), t = -0.20, p = *ns*. However, there was evidence for loss aversion; there was a distinction between high and low magnitude outcomes specifically for gains, with no such effect for losses, (b = 0.07), t = 2.15, $p < .05$ (Figure 4.6B). This is consistent with a selective continuous representation of gains, but not losses, at the stream level. Overall, at the stream level, gains and losses appear to be processed as categorically distinct, along with some sensitivity to the magnitude of gains, but not losses, a result consistent with loss aversion. More specifically, a steeper utility function for losses than that for gains would suggest that the losses are represented in a more "all or none" manner, whereas gains may be represented with more fine detail with regards to magnitude.

Discussion

Humans engage in a wide variety of value-based decisions, but the efficiency in which we engage in them is far from optimal. Distortions in valuation such as loss aversion lead to preferences that favor making decisions to avoid losses, rather than those that result in equivalent gains (Kahneman & Tversky, 1979). Although Prospect Theory has proposed that loss aversion is the product of asymmetric utility functions between gains and losses that result in losses having about twice as much more "psychological impact" than gains, to date no studies have systematically investigated when these biases might occur during valuation. The current work sought to clarify this by implementing a novel multi-sampling gambling task that required subjects to process individual elements of value information and integrate them in order to optimize decision making. Such a task allowed us to investigate whether or not loss aversion occurs at the level of initial

valuation, or later integration. In addition, through the use of multivariate pattern classification techniques, we were not only able to investigate *when* loss aversion might occur, but also *how* it might occur. Here, we demonstrated that delta power, a signal previously shown to be associated with evidence integration (Wyart et al., 2012) and reward (Bernat et al., 2015; Knyazev, 2007; 2012), carries information about representations of value in the current task. Specifically, we showed that gains and losses are processed as binary and categorically distinct for individual elements of a stream of value information, whereas once integrated, the stream-level expected value showed magnitude sensitivity for gains, but not losses. This is entirely consistent with a steeper utility function for losses relative to that for gains, which is predicted by prospect theory (Kahneman & Tversky, 1979), and suggests that distortions in valuation like loss aversion may occur farther downstream during the integration of value information.

The use of a multi-sampling task adapted from Wyart et al. (2012) was critical in assessing what stage of valuation loss aversion might occur. Such a method is a departure from the overwhelming majority of other value based decision making tasks, where all necessary decision-relevant information is presented explicitly upfront (Canessa et al., 2013; Chib, Rangel, Shimojo, & O'Doherty, 2009; Gehring & Willoughby, 2002; Knutson, Rick, Wimmer, Prelec, & Loewenstein, 2007; Krajbich, Lu, Camerer, & Rangel, 2012; Plassmann, O'Doherty, & Rangel, 2010; Rangel et al., 2008; Tom et al., 2007). Such a design makes it difficult to differentiate between any initial valuation period and a later integration period. For example, in the case of the presentation of a mixed gamble (a commonly used method in studies examining loss aversion; (Canessa et al., 2013; Stanton et al., 2011; Tom et al., 2007)) it is unclear not *if*, but *when* they are

processing the different aspects of information (i.e. probability, magnitude, gain/loss, etc.). The multisampling approach is ideal for decoupling initial processing vs. later integration since it tightly controls the timing and type of information being presented to the subject.

Choice Behavior Sensitive to Both Element Values and Stream Expected Value

Overall, the task used in the current work proved to be effective as indexed by subject choice behavior and response times. At the element level, binary logistic regression for the raw monetary value of each element showed that they were all uniquely predictive of the subjects' choices to gamble or not, regardless of their serial position within the stream. Furthermore, each element was approximately equally predictive at each serial position, a result that mirrors those of Wyart et al (2012), whose task the current work was adapted from. In additional to the contribution of individual elements on behavior, subjects' decisions to gamble were sensitive to the level of expected value of the stream. As the expected value grew larger, betting behavior increased linearly. This strongly suggests that subjects could successfully integrate the values of the elements into an overall expected value for the stream. Particularly telling is the response time data; rather than the different levels of expected value having equivalent response times, subjects made choices more quickly when the expected value was either very low or very high, when compared to intermediate levels above and below the value of the safe bet. This quadratic effect in response times for levels of expected value suggests that very high and low levels of expected value were easy to recognize, and produced faster responses, whereas intermediate values required more deliberate consideration before making choice. This further reinforces that subjects were sensitive to the expected value

of the stream, and provides additional justification for the use of multi-sampling tasks when wanting to examine the integration of information.

Delta Power is Informative of Value Representations at Both the Element and Stream Level

As with previous work, the current work showed that delta power likely carries information about underlying representations of value or reward (Bernat et al., 2015; Knyazev, 2007; 2012). At the element level, delta power suggests that value is represented in a purely categorical gain/loss manner (relative to the value of a safe bet). We found no evidence for loss aversion either through shifting the boundary by which gains and losses are defined, or by a selective representation of the magnitude of gains, but not losses. Similarly, at the stream level, we found evidence for a categorical boundary that separated gain and loss-related levels of expected value. However, there was evidence of loss aversion at the stream level; delta power distinguished between high and low levels of expected value for gains, but not losses, with no evidence for a categorical shift in the gain/loss boundary. This is consistent with asymmetric utility functions for gains and losses, whereby the function for losses is markedly steeper than that for gains (Kahneman & Tversky, 1979). In principle, this would result in selectively poorer "resolution" with regards to being able to differentiate between magnitudes of losses, and instead treating all losses approximately equally. Meanwhile, the comparatively less steep utility function for gains would more likely accommodate for differentiation between gains of different magnitudes. Additionally, this result is similar to what some have found with the feedback-related negativity (FRN); San Martin et al. (2010) found that in addition to categorically distinguishing between gains and losses

following feedback during a gambling task, the FRN was also selectively sensitive to the magnitude of gains, but not losses.

Neural Representations of Loss Aversion Occur Following Value Integration

The current work helps address two fundamental questions about the nature of valuation distortions such as loss aversion: 1) At what stage of processing do they occur? And, 2) How are they represented through underlying neural signals? The evidence provided here suggests that value distortions like loss aversion occur during later stages of processing following integration of value evidence, rather than during initial valuation. Furthermore, through the use of multivariate decoding techniques (Haxby et al., 2001; Kriegeskorte, 2008; Kriegeskorte et al., 2008; Norman, Polyn, Detre, & Haxby, 2006), that can map the representational space of underlying neural signals, the current work suggests that loss aversion is the result of a shift from a categorical representation of gains and losses during the initial stage of processing to a selectively continuous representation of the magnitude of gains, but not losses, following integration of information. Although the current work helps clarify when and how loss aversion may occur, it does not address *why* this might happen. For example, it may be inefficient for a system that monitors the value of items in the environment to place heavy biases on the representations of value for those items during initial valuation. Compared to a more unbiased assessment of the true value of items (or category of items, as observed for value representations at the element levels), biasing their representations of value during early stages of processing might result in a decision variable that is too skewed to facilitate optimal choice. This however, would need to be investigated more directly.

Limitations and Future Directions

Although the design of the current work offered a novel opportunity to investigate when and how loss aversion might manifest, it is not without its limitations. For example, average betting rates for this sample were approximately 33%. This suggests, to some degree, that individuals were loss averse overall and were biased to play it safe unless the expected value was very high. Even when the expected value of the stream was higher than the value of the safe bet, but not drastically so (low gain condition), betting rates were fairly low (~38%). This same range of expected value also produced the longest response times. Although outwardly these behaviors seem consistent with a loss verse attitude, it is difficult to disentangle an effect due to true loss aversion and one simply due to task difficulty. The current task presented stimuli at a rapid rate and thus demanded a great deal of attention to optimally perform. It is possible that the overall low betting rates were also due to the task being difficult, and subjects simply relying on a more conservative strategy rather than truly experiencing loss aversion. However, the choice and response time patterns indicate that they were able to track the expected value of the stream. It is indeed possible to make this task less demanding (i.e. fewer elements and slower presentation rate), but that would to be carefully balanced because the current experimental session was already approximately three hours.

With regards to the use of a multi-sampling gambling task, an open question that remains is exactly how much evidence needs to be integrated before loss aversion begins to manifest itself at the neural level. This question is motivated by the fact that reliable decodability of the expected value of the stream started approximately 1000 milliseconds into a trial (approximately halfway through). This suggests that it may not be necessary to

integrate all information before a reliable decision variable is computed, and leaves open the question as to whether or not differing levels of integrated evidence produce differing levels of valuation distortions. Are underlying value representations differentially susceptible to valuation distortions based upon how much evidence has been integrated?

Conclusions

The current work used a novel multi-sampling gambling task in conjunction with multivariate pattern classification methods to investigate when and how distortions in valuation such as loss aversion might manifest. Delta power seemed to carry information about representations of value at both the individual element and aggregate stream level. Patterns of delta power visualized through confusion matrices revealed that losses and gains are represented in a purely categorical manner with no sensitivity to their magnitude at the individual element level. In contrast, at the aggregate stream level, delta power showed sensitivity for the magnitude of gains, but not losses, consistent with asymmetric utility functions predicted by Prospect Theory. This marks the first attempt to systematically classify when and how loss aversion might occur, and motivates the use of multi-sampling and pattern classification approaches to investigate a broader range of topics where distinguishing between different levels of processing is required.

CHAPTER V

WHAT ARE THE DYNAMICS OF PROCESSING INTRINSIC AND EXTRINSIC VALUE INFORMATION?

Study two demonstrated that information about losses and gains was processed independently at the element level, with a sharp categorical boundary separating the two. Subjects were able to track this information, and integrate it into an expected value signal, as evidenced by both their betting behavior and through robust representations of ranges of expected value in multivariate patterns of delta power. Evidence for the representation of such an integrated signal began to appear after approximately 1000 milliseconds into the trail. However, the information being integrated form the elements was information that was intrinsic to the elements themselves. In many cases, value can be derived from characteristics that are *extrinsic* to the item (e.g. the cost of effort associated with obtaining the item). Are intrinsic and extrinsic categories of value information integrated along a similar timecourse? Are these two types of information processed in a parallel or serial manner? These questions, among others were explored in study 3.

CHAPTER VI

FIXATIONS REVEAL SERIAL PROCESSING OF VALUE AND EFFORT DURING VALUE GUIDED CHOICE

Introduction

With regards to making value based decisions, a large body of work has suggested that the brain assigns a value to all options under consideration to allow for effective comparison, and ultimately, an advantageous choice (Krajbich, Armel, & Rangel, 2010; Rangel et al., 2008). In most instances, in order to effectively extract value information from a set of items under consideration, one needs to sample information from the items via visual fixations. Indeed, a growing body of literature has examined how overt fixation behavior contributes to information accumulation during value based decisions. This work has demonstrated that eye movements and the value of items interact reciprocally; eye movements both *affect* the assignment of value to items, and the value we assign to items is *affected by* eye movements. For example, valued items may be looked at more frequently (Krajbich et al., 2010; 2012; Krajbich & Rangel, 2011; Towal, Mormann, & Koch, 2013), but the time we spend looking at an item can also increase its perceived value (Armel, Beaumel, & Rangel, 2008; Lim, O'Doherty, & Rangel, 2011; Towal et al., 2013).

Value is often associated with the intrinsic characteristics of the item itself (i.e. taste, appearance, etc.). However, value can also be affected by characteristics that are *extrinsic* to the item itself; for example, the cost or effort associated with attaining the item (Rangel & Hare, 2010). Often, extrinsic information can be known before the object itself comes into focus and therefore can set the context of the item-specific valuation

process in a top-down manner. For example, a scenic view of Half Dome might be a highly valued experience, but comes at the cost of having to be reached via a difficult hike. Additionally, consider a situation where a hungry diner sees the prices for menu items at a well-reviewed, but expensive, restaurant. In each of these scenarios, value can be construed as the absolute difference between these intrinsic and extrinsic characteristics (Rangel & Hare, 2010).

The comparison between intrinsic and extrinsic characteristics has been examined in the case of multi-attribute decision making (ex. taste vs. health (Hare, Malmaud, & Rangel, 2011), taste vs. physical effort (Harris & Lim, 2016), monetary reward vs. physical effort (Klein-Flugge, Kennerley, Friston, & Bestmann, 2016), erotic images vs. physical and mental effort (Prévost, Pessiglione, Météreau, Cléry-Melin, & Dreher, 2010)). However, it has not been examined how intrinsic and extrinsic factors interact as we accumulate decision-relevant information during the valuation process. Specifically, we were interested in how the pre-cuing of extrinsic effort affects fixation and valuation behavior, and to what degree extrinsic and intrinsic information are integrated in a continuous and parallel manner, or whether we process these different types of information within distinct, serial stages.

To lay out our questions and competing alternative hypotheses in greater detail, it is useful to first present our paradigm. In the present study we recorded eye movements from subjects as they engaged in a decision making task where they first rated a series of snack food items, and then had to choose among pairs of items that could ultimately be chosen for them to keep. This task was largely based on previous work using similar paradigms to investigate the role that overt visual behavior has on valuation during

decision making (Armel et al., 2008; Krajbich et al., 2010; 2012; Krajbich & Rangel, 2011). However, the key difference is that we included an effort manipulation in the form of math problems that would need to be solved in order to choose particular items. In half the trials, subjects were presented with a pre-cue prior to the onset of the to-be-decided-upon items that associated one of the items with needing to solve a math problem in order to choose it (more specific detail about the task can be found in the methods section). This departure from previous designs allows us to more directly assess how value and effort are processed as indexed through visual behavior.

More specifically, this design allows us to address several questions. First, how does foreknowledge of effort affect first fixations? Based on the notion that higher-valued objects draw attention (Krajbich et al., 2010; 2012; Krajbich & Rangel, 2011; Towal et al., 2013) one might expect that adding effort to an object in an a-priori manner would drive fixations away from that object. Alternatively, we also need to consider the possibility that assigning effort to one of the two potential locations with a fully predictive pre-cue makes this location "stand out" and attracts fixations to it.

Furthermore, how is information about value and effort integrated? Are these two types of information integrated in a parallel or serial manner? If these two types of information were integrated in a parallel manner, we might expect that the initial fixation is, if anything, influenced only by the effort pre-cue. However, remaining fixations should be determined by an interaction of both value and effort—specifically, where the objects with highest conflict (no effort/low value or high effort/high value) attract the largest proportion of fixations. If information about value and effort was instead processed in a serial manner, similarly, we might expect that the initial fixation is

influenced only by the effort pre-cue. In a second phase, fixations should also start to reflect value, but in a manner that is additive with effort, followed by a third phase during which the two interact. Additionally, it would be important to determine the utility of how information integration (as assessed through fixation behavior) contributes to actual choice behavior. A schematic representation of these parallel and serial information processing hypotheses can be found in Figure 6.1.

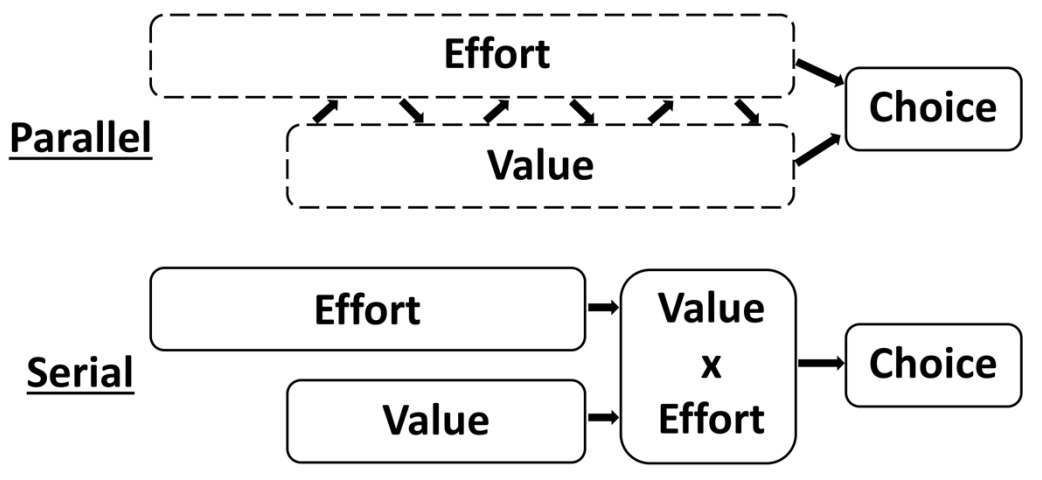

Figure 6.1. A schematic representation of parallel (top) and serial (bottom) processing of intrinsic (value) and extrinsic (effort) information.

Methods

Participants

A total of 40 subjects **(**19 male**)** were recruited in return for course credit or a monetary compensation of $30 in addition to one food item. All subjects had normal or corrected-to-normal vision and gave written informed consent according to procedures approved by the University of Oregon Institutional Review Board.

Experimental Task and Stimuli

All subjects were asked to refrain from eating in the two hours prior to the experimental session. The experiment was largely based on previous work examining both decision making and eye tracking (Armel et al., 2008; Krajbich et al., 2010; 2012; Krajbich & Rangel, 2011; Polanía, Krajbich, Grueschow, & Ruff, 2014; Polanía, Moisa, Opitz, Grueschow, & Ruff, 2015), and took place in three major phases: an initial slideshow, a ratings phase, and a decision phase. The primary stimuli for the task consisted of a set of images of 88 unique snack foods (chips, candy bars, fruits, nuts, etc.). Each image consisted of a single food item against a black background. Stimuli for the task were presented at a size of 300 x 300 pixels in Matlab using the Psychophysics toolbox extension (Brainard, 1997).

Slideshow.

Subjects first progressed through a slideshow of images of the 88 unique food items. Each item was presented for at least one second, after which subjects could advance the slideshow at their own pace by pressing the spacebar on the keyboard. The purpose of the slideshow was to familiarize subjects with the variety of foods available (they were also shown the actual inventory of these foods stored in the lab). Subjects were explicitly aware that they would be rating each of these items later in the task, but they were not informed of what criteria the items would be rated on. Items were presented in a random order for each subject.

Ratings.

Following the slideshow, subjects were required to rate each individual item based on how interested they would be in eating the item after the study. Items were rated

on a 20-point scale ranging from 1 ("I am not at all interested in eating this item") to 20 ("I am incredibly interested in eating this item"). Like the slideshow, items were presented in a random order and subjects were required to move an on-screen slider to the rating that best described how they felt about the current item on screen using the left and right arrow keys. Ratings were confirmed by pressing the spacebar, at which point subjects would be presented with the next item. The initial starting point of the slider was randomized on each trial to prevent anchoring effects (Krajbich et al., 2010), and subjects were encouraged to rate each item as specifically as possible.

Decision task.

During the decision task, subjects were required to choose among pairs of items; each item that was chosen had a chance of being selected for the subject to take home at the conclusion of the experiment via a lottery. In order to generate trials for the decision task, subjects' ratings from the ratings phase served as the input to an algorithm that found all unique pairs of the 88 items and assigned them to one of four value bins based on whether the highest rated item – lowest rated item = [1, 2, 3, 4] (Polanía et al., 2014; 2015). 88 unique pairs were randomly selected from each bin to serve the trials for the decision task. In the event that the algorithm could not generate the required number of trials, the bin widths were doubled (this only occurred with two subjects). All subjects completed 352 trials of the decision task (16 blocks of 22 trials).

At the beginning of each trial, subjects were presented with two empty square frames (one presented on the left side of the screen, one presented on the right side of the screen) for 1000 ms. The color of the frames (orange or blue) served as a pre-cue that informed the subjects whether or not the item that would appear inside that frame for that

trial would be associated with an additional math task (described in a later section). Half of the trials in each value bin were associated with math, and the colors to denote math vs. no-math items were counterbalanced across subjects. After the presentation of the empty frames, subjects were required to maintain uninterrupted fixation on a central fixation cross for 1000 ms. Following a successful 1000 ms of continuous fixation, the food items for that trial appeared simultaneously inside the frames. Once the items were presented, subjects were free to examine them and choose which item they wanted to be included in the lottery using the left and right arrow keys. There were no time constraints during the choice period, and trials were separated by a 500 millisecond intertrial interval. A basic outline of the decision portion of a trial can be seen in Figure 6.2A**.**

Math task.

At the end of the trial, if subjects had chosen an item that required them to complete the additional math task according to the pre-cue, an equation would be presented above the central fixation cross that needed to be solved in order for the experiment to progress. Equations consisted of seven unique numbers between one and nine, with addition and subtraction between them. Subjects were required to provide the solution to the addition and subtraction of these seven numbers using the numeric keypad on the right side of the keyboard, and submitting the answer using the enter key. If an incorrect answer was provided, feedback was provided in the form of the words "Invalid Response" printed in red under the equation for 1000 ms. If the correct answer was provided, the word "Correct" was printed in green under the equation. The solution to each equation always ranged between one and nine, although intermediary values while solving the equations could fall outside these bounds (subjects were explicitly aware of

this). There were no time constrains during the math trials, and subjects had as many attempts as necessary to provide the correct answer. An example of the math portion of a math trial if they chose the item associated with math can be seen in Figure 6.2B.

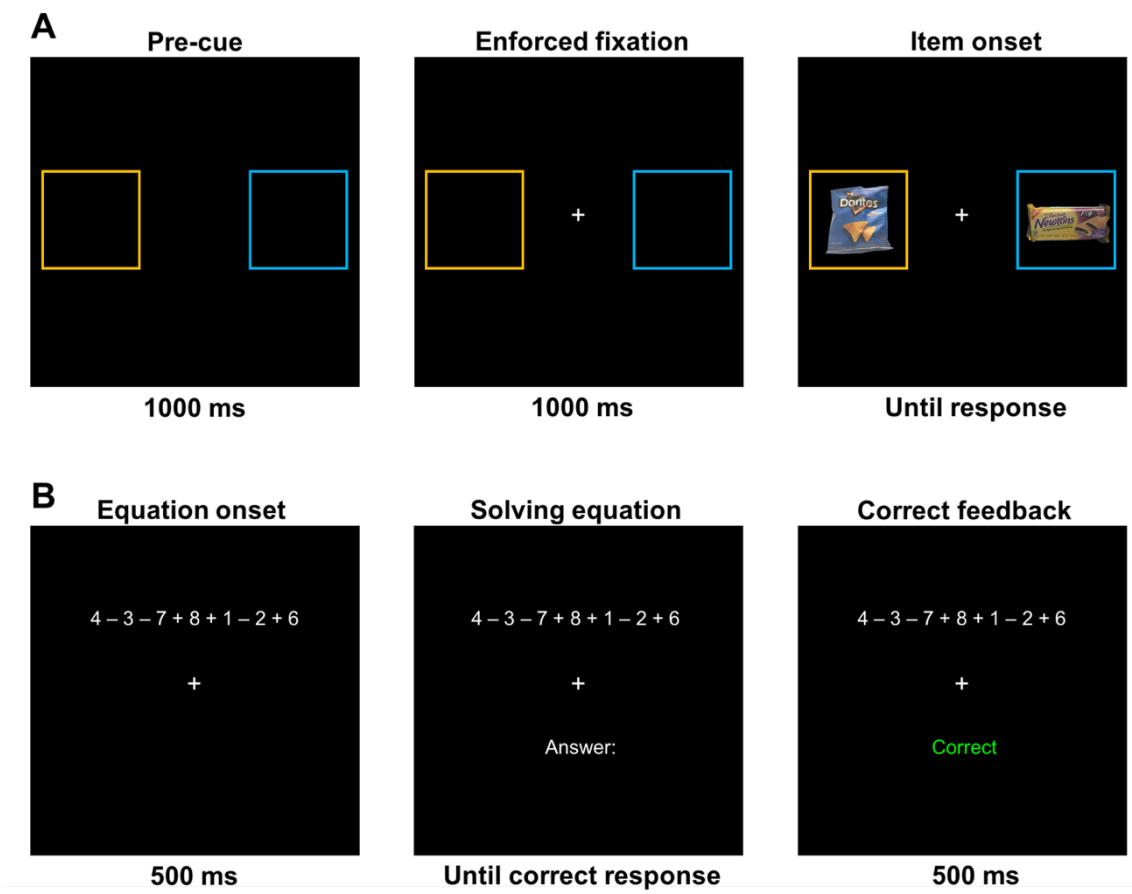

Figure 6.2. **A.** Example of a trial of the decision task (this trial contains a pre-cue for math). **B.** Example of the math task.

Eye tracking.

Subjects were seated with their chin stabilized by a chin rest with their eyes approximately 50 cm from the monitor. A 17-inch CRT monitor set to 1024 × 768 resolution was used for stimulus presentation. Eye movements were measured using the SR Research desk-mounted Eyelink 1000, controlled by the Eyelink Toolbox in MATLAB (Cornelissen, Peters, & Palmer, 2002) at a rate of 1000 Hz. Fixations were

recorded when neither a blink nor a saccade was present, and saccades were defined for each pair of successive data samples for which the velocity of eyes exceeded 30°/s or the acceleration surpassed 8,000 °/s^2. Eye movements were recorded during the decision task only.

Calculating Fixation Probabilities to Targets Across Time.

In order to investigate the temporal dynamics of fixation behavior, we computed the probability of fixations to various targets during the decision task across the duration of each trial. This was accomplished by first specifying a target during the decision task (i.e. preferred item vs. non-preferred item, item associated with math vs. item not associated with math, item on the right vs. item on the left, etc.) and extracting a 2000 millisecond epoch of data following the onset of the target for each trial. Epochs were downsampled to 40 Hz (81 timepoints) and from these we could extract whether subjects were fixating on the target separately for each timepoint. These data were later averaged across various time windows for specific analyses (described later).

Results

Response Time and Choice Consistency Effects

All reported analyses were carried out by constructing hierarchical linear models in R (R Core Team, 2016) using the lme4 package (Bates et al., 2015). All models specified a random intercept for each subject. We first constructed a model to see if the difference in value between items under consideration (i.e. choice difficulty) and the presence of math were predictive of trialwise response times. For the factor associated with the presence of math, we constructed two non-orthogonal contrasts comparing 1) no

math vs. math on the preferred item, and 2) math on the non-preferred item vs. math on the preferred item. Overall, there was a linear decrease in response times as the difference in value between the two items increased (i.e. as the choice became easier) (b = -114.87), t = -4.00, p < .001. On average, response times on trials with no math (M = 2311.90, SE = 17.42) were significantly lower when compared to those where math was associated with preferred item (M = 2510.56, SE = 32.40) (b = -190.72), t = -5.71, p < .001. Similarly, response times on trials where math was associated with the non-preferred item (M = 2304.74, SE = 27.73) were significantly lower than when math was associated with the preferred item, (b = -194.08), t = -5.02, p < .001 (Figure 6.3A).

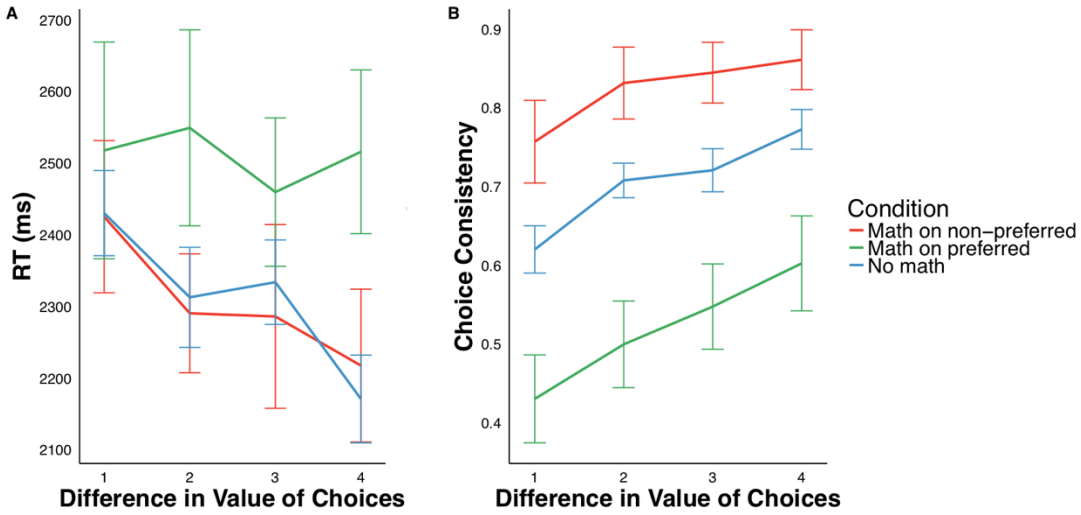

Figure 6.3. **A.** Response times as a function of the difference in value between items across each math condition. **B.** Choice consistency as a function of the difference in value between items across each math condition.

In addition to response time, we also constructed a model to see if the difference in value between items under consideration and the presence of math were predictive of trialwise selection of the preferred item. Here, the preferred item was defined as the item that was received the higher rating during the previous rating task. For the factor associated with the presence of math, we constructed two non-orthogonal contrasts

comparing 1) no math vs. math on the preferred item, and 2) no math vs. math on the non-preferred item. Overall, there was a linear increase in tendency to choose the preferred item as the difference in value between the two items increased (b = 0.57), z = 8.87, p < .001. The presence of math on the preferred item *decreased* choice of the preferred item (M =0.52, SE = 0.02) compared to the no math condition (M = 0.71, SE = 0.008), (b = 0.82), z = -5.78, p < .001. Conversely, the presence of math on the non-preferred item *increased* choice of the preferred item (M = 0.82, SE = 0.01) compared to the no math condition, (b = 0.80), z = 6.70, p < .001 (Figure 6.3B).

Basic Properties of the Visual Search

As a first step in examining the properties of the visual search, we tested to see whether first fixations were drawn to the preferred item, in addition to how first fixations were influenced by the presence of math. For the factor associated with the presence of math, we constructed two non-orthogonal contrasts comparing 1) no math vs. math on the preferred item, and 2) no math vs. math on the non-preferred item. Here, first fixations were significantly more likely to land on the preferred item when math was associated with the preferred item (M = 0.57, SE = 0.02), compared to the no math condition, (M = 0.50, SE = 0.007) (b = 0.30), z = 7.29, p < .001. Conversely, first fixations were significantly less likely to land on the preferred item when math was associated with the non-preferred item (M = 0.56, SE = 0.02), compared to the no math condition, (b = -0.27), z = -6.50, p < .001. It is important to note that the probability of fixations to preferred/non-preferred items are inversely related to one another, since subjects only have two targets to fixate on. Overall there is no evidence to indicate that first fixations are drawn by the value of the item; rather, the presence of math (which is known in

advance by the subject) seems to draw first fixations (Figure 6.4). Furthermore, there was no evidence for any systematic bias for first fixations to go to the left or right item.

Importantly, we also examined what effect item preference (preferred vs. non-preferred) and presence of math had on total dwell time to items during the decision task. For the factor associated with the presence of math, we constructed two non-orthogonal contrasts comparing 1) no math vs. math on the preferred item, and 2) math on the non-preferred item vs. math on the preferred item. Overall, subjects dwelled longer on the preferred item (M = 767.01, SE = 6.50) compared to the non-preferred item (M = 718.83, SE = 7.09), (b = 49.52), t = 5.78, p < .001. Subjects also looked longer at a given item when math was associated with the preferred item (M = 789.00, SE = 7.12), compared to the no math condition, (M = 715.16, SE = 7.12) (b = 81.54), t = 5.80, p < .001. The same pattern was true comparing dwell times between the preferred item and non-preferred item (M = 724.59, SE = 9.43), (b = 89.68), t = 5.49, p < .001 (Figure 6.5A).

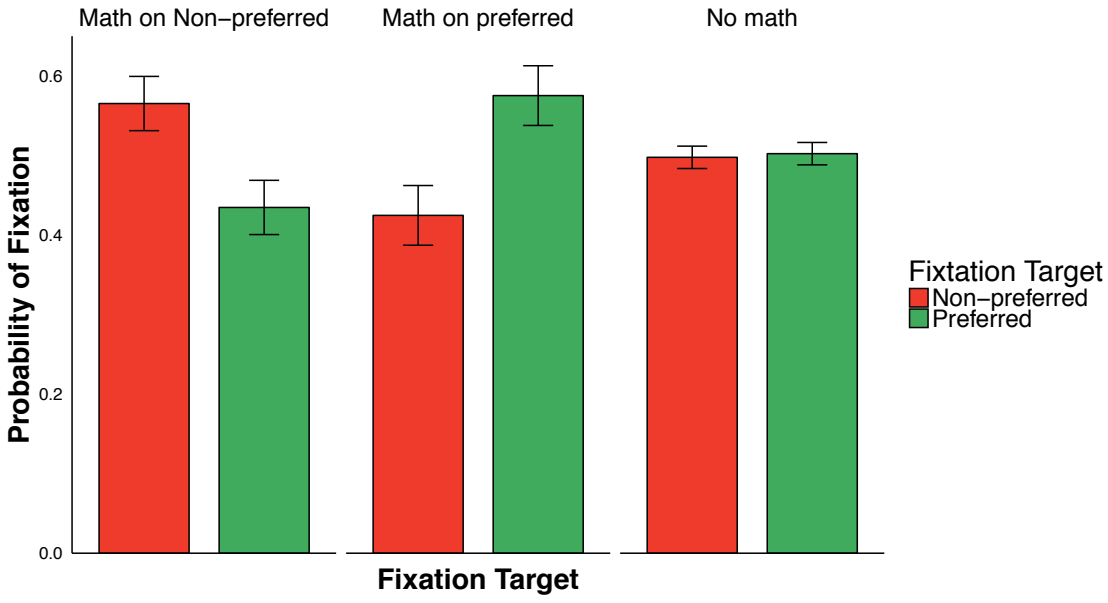

Figure 6.4. Fixation probabilities to preferred and non-preferred items for each math condition.

Finally, we examined whether or not fixating back and forth between items differed as a function of the presence of math, and the difference in value between the items. For this analysis, we computed the number of *unique* back and forth fixations between items per trial; consecutive fixations to the same item did not count towards this. We first constructed a model to see if the difference in value between items under consideration (i.e. choice difficulty) and the presence of math were predictive of trialwise back and forth fixations. For the factor associated with the presence of math, we constructed two non-orthogonal contrasts comparing 1) no math vs. math on the preferred item, and 2) math on the non-preferred item vs. math on the preferred item. Overall, there was a linear decrease in back and forth fixations as the difference in value between the two items increased (i.e. as the choice became easier) (b = -.11), t = -5.22, p < .001. On average, there were fewer back and forth fixations on trials with no math (M = 2.85*, SE* = 0.01) when compared to trials where math was associated with preferred item (M = 2.92*, SE* = 0.02) (b = -0.07), t = -2.60, p < .01. This difference in back and forth fixations between these two conditions increased as a function of the difference in value between the items under consideration, (b = -0.16), t = -3.14, p < .01. Similarly, there were fewer back and forth fixations on trials where math was associated with the non-preferred item (M = 2.75*, SE* = 0.02) compared to trials where math was associated with the preferred item, (b = -0.17), t = -5.78, p < .001. Again, this difference in back and forth fixations between these two conditions increased as a function of the difference in value between the items under consideration, (b = -0.18), t = -3.11, p < .01. These effects are illustrated in Figure 6.5B.

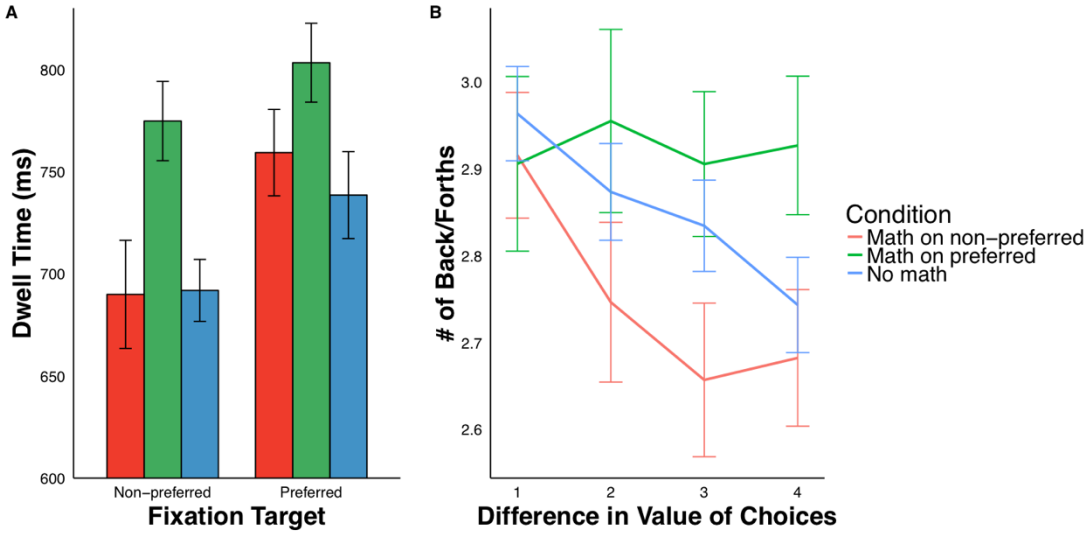

Figure 6.5. **A.** Dwell times to preferred and non-preferred targets for each math condition. **B.** Number of back and forth fixations as a function of the difference in value of choices for each math condition.

Fixation Behavior Over Time Shows Prioritization of Different Item Characteristics.

In order to more closely examine the dynamics of fixation behavior, we calculated fixation probabilities to various targets across time over the course of a trial following the onset of the items (described in methods). For simplicity, only 4 conditions were used for analyses: fixations to the preferred item when math was associated with the preferred item, fixations to the non-preferred item when math was associated with the non-preferred item, fixations to the preferred item in the no math condition, and fixations to the non-preferred item in the no math condition. This effectively reduced the factor structure to 2 (preferred/non-preferred) x 2 (math/no math), which makes it easier to interpret effects purely due to item preference or the presence of math. The excluded conditions (fixations to the preferred item when math is on the non-preferred item, and fixations to the non-preferred item when math is on the preferred item) are perfectly complimentary to the two included math conditions, and are thus redundant for further

analysis. Additionally, for ease of interpretation, we binarized fixation probabilities to convert them to a binary hit/no hit factor for a given target. Time courses of the fixation probabilities for these four conditions can be seen in Figure 6.6. Based on visual inspection of the fixation probabilities across time, we identified three time windows (in ms) of interest for subsequent analyses: an early time window (300 – 500), a middle time window (800 - 1300), and a late time window (1500 – 2000).

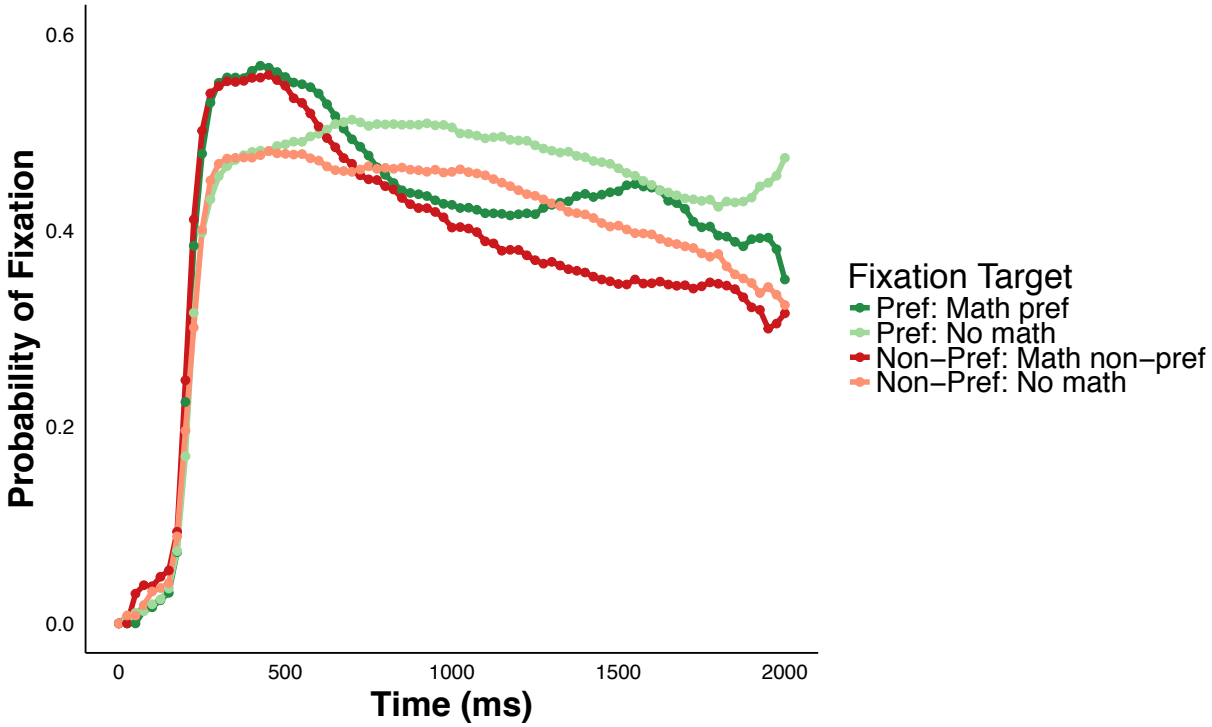

Figure 6.6. Timecourses of fixation probabilities to preferred and non-preferred targets in math and no math conditions.

For the early time window (300 – 500 ms), fixations were significantly more likely to be drawn to an item associated with math (M = 0.58, SE = 0.006), regardless of item preference, compared to an item in the no math condition (M = 0.51, SE = 0.004), (b = 0.31), z = 3.77, p < .001. In contrast, during the middle time window (800 – 1300 ms), fixations were significantly *less* likely to be drawn to an item associated with math (M = 0.63, SE = 0.005), regardless of item preference, compared to an item in the no

math condition (M = 0.71, SE = 0.003), (b = -0.29), z = -4.27, p < .0001. Additionally, fixations were significantly more likely to be drawn to a preferred item (M = 0.71, SE = 0.004), regardless of the presence of math, compared to a non-preferred item (M = 0.66, SE = 0.004), (b = 0.26), z = 5.82, p < .0001. Similar patterns were found for the late time window (1500 – 2000 ms). Again, fixations were significantly *less* likely to be drawn to an item associated with math (M = 0.46, SE = 0.006), regardless of item preference, compared to an item in the no math condition (M = 0.49, SE = 0.004), (b = -0.23), z = -4.29, p < .001. Similarly, fixations were significantly more likely to be drawn to a preferred item (M = 0.51, SE = 0.005), regardless of the presence of math, compared to a non-preferred item (M = 0.44, SE = 0.005), (b = 0.25), z = 5.79, p < .001. Finally, there was a significant interaction that indicated that overall, late period fixations only show a math-related aversion for non-preferred items; there was no difference in the likelihood of late fixations landing on a preferred item associated with math, and a preferred item in the no math condition, (b = 0.19), t = 2.94, p < .01. This suggests some degree of conflict between preference and math-related effort for later fixations. Overall, this pattern is consistent with the serial, rather than parallel processing of effort and value, with both types of information being integrated in the late stages of visual search. An overview of these patterns is displayed in Figure 6.7.

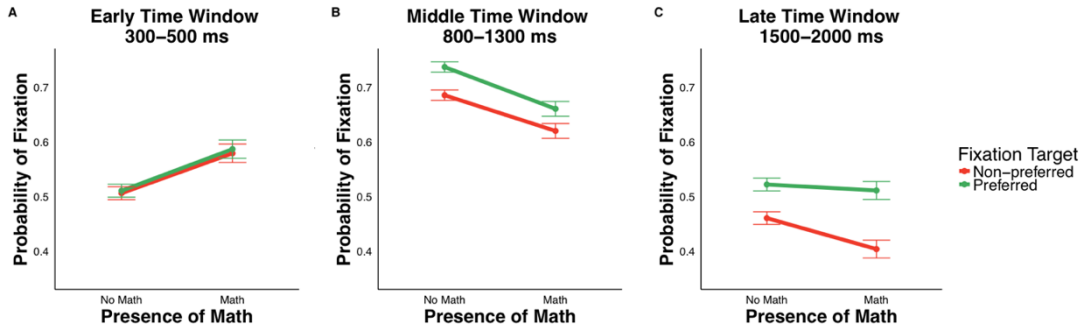

Figure 6.7. Fixation probabilities to preferred and non-preferred targets in the math and no math conditions across early, middle, and late time windows.

Late Stage Integration of Effort and Value Information Predicts Choice Behavior.

The previous patterns of fixation behavior suggest that information about effort and value is processed over time in a serial manner, and the two types of information are ultimately integrated in late stages prior to choice. If these patterns of fixation behavior actually represent underlying valuation processes, it would stand to reason that they would predict choice behavior. To examine this, we constructed a binary logistic hierarchical linear model predicting trialwise choice of the item on the right using the following predictors: the presence of math on the left (contrasted with the no math condition), the presence of math on the right (contrasted with the no math condition), the total value advantage of the item on the right, as well as the binarized occurrence of a fixation landing on the item on the right during the each of the previously described early, middle, and late time windows (these values were never interacted with each other). The specific result that would further support the previous finding that late fixations integrate information about effort and value would be a significant three-way interaction between the presence of math on the right item, the value of the right item, and late fixations landing on the right item. It should be noted that the presence of this result is all that would be required to further support the serial integration of effort and value information

during later fixations; the basic prediction is agnostic to the direction of the effect. The main goal is to demonstrate that different types of information simultaneously (i.e. in an integrated manner) contribute to choice behavior.

The full output of this binary logistic hierarchical linear model is displayed in Table 1**.** Consistent with the previously reported effects on choice consistency, the main effects for the presence of math contributed to overall choice of the right item; math on the left increased the likelihood of choosing the right item, whereas math on the right decreased it. Unsurprisingly, there was also a main effect of the value advantage of the item on the right, suggesting that the greater the value advantage of the item on the right, the greater the likelihood of choosing it. Furthermore, fixating on the right item during each of the three time periods (early, middle, late) independently predicted choice of the right item. Importantly, it should be noted that the effects for each of these time periods were still significant (and in the same direction) when included individually in a reduced model. Table 1 shows that there were largely no significant two-way interactions between any of the predictors. The only two-way interaction that reached significance was an interaction between the value advantage of the item on the right and fixating on the item on the right during the early time period. This suggests to some degree early representation of value information, which was not captured in the previous models of fixation behavior. The only significant three-way interaction was the predicted interaction between the presence of math on the right, the value advantage of the item on the right, fixating on the item on the right during the late time period. This lends further support to the idea that information about effort and value is integrated in during a later period (as

Table 1

Results of Binary Logistic Hierarchical Linear Model Predicting Choice of Item on the Right

Variable	*Estimate*	*std. Error*	*z*	*p*
(Intercept)	-0.95	0.07	-12.81	**<.001**
mathLeft	0.65	0.16	3.99	**<.001**
mathRight	-0.75	0.17	-4.51	**<.001**
rightVal_adv	0.28	0.03	10.83	**<.001**
early_fix	0.37	0.06	6.63	**<.001**
middle_fix	0.53	0.06	8.67	**<.001**
late_fix	0.80	0.05	16.15	**<.001**
mathLeft: rightVal_adv	-0.02	0.06	-0.33	.744
mathRight: rightVal_adv	0.08	0.06	1.36	.174
mathLeft: early_fix	0.03	0.13	0.23	.816
mathRight: early_fix	0.20	0.13	1.59	.112
mathLeft: middle_fix	-0.22	0.14	-1.53	.127
mathRight: middle_fix	0.09	0.14	0.68	.494
mathLeft: late_fix	0.21	0.11	1.91	.056
mathRight: late_fix	0.04	0.11	0.37	.711
rightVal_adv: early_fix	0.06	0.02	2.73	**.006**
rightVal_adv: middle_fix	0.01	0.02	0.33	.744
rightVal_adv: late_fix	-0.02	0.02	-1.12	.262
mathLeft: rightVal_adv: early_fix	-0.03	0.05	-0.60	.546
mathRight: rightVal_adv: early_fix	-0.03	0.05	-0.54	.590
mathLeft: rightVal_adv: middle_fix	0.02	0.05	0.42	.675
mathRight: rightVal_adv: middle_fix	-0.04	0.05	-0.85	.394
mathLeft: rightVal_adv: late_fix	-0.01	0.04	-0.31	.754
mathRight: rightVal_adv: late_fix	-0.09	0.04	-2.09	**.037**

Note. mathLeft/Right refers to non-orthogonal contrasts comparing the presence of math on the left/right relative to the no math condition. rightVal_adv is value if the item on the left subtracted from the value of the item on the right. early/middle/late_fix refers to whether a fixation landed on the item on the right during those time windows.

indexed by fixation behavior), and the integration of this information meaningfully contributes to actual choice behavior.

Discussion

In order to make effective value based decisions, it is often required that we extract value information from a set of items under consideration via visual fixations. Indeed, fixation behavior has been demonstrated to drive information accumulation during value based decisions across a variety of studies (Krajbich et al., 2010; 2012; Krajbich & Rangel, 2011; Lim et al., 2011). Often times, decision-relevant information can reflect either immediate, intrinsic characteristics of the items under consideration (i.e. taste, appearance, etc.), or extrinsic characteristics such as the cost or effort required to attain the items. It remains an open question whether decision-relevant information from intrinsic and extrinsic sources are processed in a similar manner. More specifically, it is unclear whether or not these types of information are processed in a continuous, parallel manner, or in distinct serial stages. The current work sought to address these questions by recording the eye movements of subjects while they performed a value based decision making task where in half of the trials, subjects were presented with a pre-cue that indicated the presence of an effortful math task that would be associated with one of the upcoming to-be-decided-upon items. This design, coupled with eye tracking, allowed us to investigate how (if at all) information about effort and value is accumulated over time, and whether the processing of this information is ultimately predictive of choice. Overall, the current work found several lines of evidence to suggest that effort and value information are initially processed in a serial manner before later integration, as assessed though fixation behavior. Furthermore, this same index of late information integration

was also predictive of actual choice behavior, suggesting that it may reflect an underlying decision-relevant signal.

Information About Effort and Value is Reflected in Basic Behavior and Visual Search

As with previous studies that have utilized similar decision making tasks, the current work found that both response times and choice consistency (i.e., the degree to which subjects consistently select items that they had previously indicated as preferred) were affected by the difference in value between the two items under consideration. Overall, larger differences in value between the two items resulted in faster response times and the more consistent choices. These results are largely in line with previous studies, and the current work overall had comparable levels of choice consistency (Krajbich et al., 2010; Krajbich & Rangel, 2011). However, unlike previous studies, the current work included an effort manipulation in half the trials in the form of a challenging math task that was associated with one of the items. The presence of math associated with either of the items had potent effects on behavior; math simultaneously inhibited and enhances choice consistency when math was associated with the preferred and non-preferred items, respectively (relative to the no math condition). Additionally, response times when math was on the preferred item did not show the same decreasing linear trend as the difference in value between the two items increased, suggesting that effort and value may interact. However, these results don't specifically speak to how and when the two types of information may interact.

In addition to response times and choice consistency, basic properties of visual search in this task seem to reflect both effort and value. In line with previous work, first

fixations were not intrinsically drawn to the more valued item (Krajbich et al., 2010; 2012); instead, they were drawn to the item associated with the math. It is likely that this was not simply due to a visual capture effect, or something akin to value-based attentional capture (B. A. Anderson, 2013), since the onset of the effort pre-cue did not occur suddenly and simultaneously with the onset of the items for a given trial. In math trials, the pre-cue was always on screen for at least two seconds (1000 ms pre-cue alone, 1000 ms enforced fixation with pre-cue still present), and thus should not have contributed to salient attentional capture effects. Instead, it is likely that first fixations were drawn to the item associated with math as part of a general strategy that prioritized determining whether the item associated with math was worth the effort, and suggests that foreknowledge of effort may be a powerful draw of attention. Importantly, in the no math condition, first fixations were random, and landed on the preferred and non-preferred items with equal frequency. Although first fixations were initially drawn to items associated with math, subjects overall looked *longer* more valuable items. Additionally, the back-and-forth fixation patterns between items decreased as a function of the difference in value between items, but only for the no math condition, or when math was associated with the non-preferred item. The overall equivalent degree of back-and-forth fixations across differences in value between the two items suggests a degree of conflict in trying to resolve effort with value by repeatedly assessing the preferred, but effortful item, and the non-preferred, but easily attainable item. Again, this suggests integration of effort and value, but these aggregate values are uninformative as to *when* integration may occur.

Fixation Behavior Supports Serial Integration of Effort and Value Information and Likely Reflects Underlying Decision-Relevant Signals

Overall, fixation behavior over the course of a trial support the hypothesis that information about effort and value is initially processed in a serial, additive manner, before being integrated to reflect the conflict the two. In the early phase of a trial, fixations are biased to the items associated with math, with no consideration for value (consistent with the reported first fixation data). Middle fixations begin to show sensitivity to value in an additive manner; a main effect for value during middle fixations indicated that more valued items are more likely to be looked at, while a main effect for the presence of math indicated that items associated with math are less likely to be looked at. Importantly, there was no interaction between the value and presence of math factors during middle fixations, which suggests that they were still being processed serially during this phase. Finally, late fixations seem to reflect the integration of information about effort and value, as reflected by a significant interaction between the presence of math and value which suggests that fixation behavior reflects the degree of conflict between two oppositional sources of information. Critically, this integration stage is predictive of subsequent choice behavior, which supports the idea that this integration reflects a decision-relevant value signal. This pattern of results indicates that the relative weighting of different types of information fluctuates over time. Here, we see evidence for an initial prioritization of extrinsic information (effort), followed by a prioritization of intrinsic information (value), and finally ending with both extrinsic and intrinsic information being considered simultaneously. Initial sampling seems to consider these sources of information independently, and the later integration of these sources of

information may be indicative of the time required to normalize them to a common scale for effective comparison, and subsequently, choice (Chib et al., 2009; Levy & Glimcher, 2012). Overall, the current work supports the hypothesis that extrinsic and intrinsic information are initially processed in a serial, rather than parallel manner.

Some previous work has attempted to investigate the temporal dynamics of extrinsic and intrinsic information during value base decision making. Harris and Lim (2016) employed a decision making task where subjects had to decide to expend physical effort to potentially receive snack foods while recording EEG. Their results are largely in line with the current work; they found that effort information was represented relatively early on following stimulus onset, whereas the integrated effort-value information was represented later on in the trial, prior to response. However, their study did not directly use that integrated signal to predict choice; rather, they traced the signals to sensorimotor areas, which are ultimately responsible for generating a response. The current work improves upon this by actually demonstrating that this later integrated signal contributes to choice behavior within the context of the task.

Limitations and Future Directions

Although the current work demonstrated that extrinsic information about effort is prioritized and processed early, the design of the current task may have artificially induced this strategy by presenting information about effort prior to the presentation of the to-be-decided upon items. However, Harris and Lim (2016) presented information about effort and value simultaneously, and found evidence for early processing of effort (as opposed to value) information. To further investigate whether the temporal ordering of different types of information influences the order in which it is processed, a future

study could systematically manipulate the temporal intervals in which effort and value information is presented. This would allow for a more detailed assessment of whether extrinsic/effort information is generally prioritized.

Conclusions

Value based decisions often require sampling between intrinsic and extrinsic sources of information, and this is often accomplished through a series of fixations. Few studies have examined whether intrinsic and extrinsic information is processed in a serial or parallel manner, and how the nature of these stages of processing contributes to choice behavior. The current work clarifies this by employing a value based decision task where subjects often had to reconcile between extrinsic (effort) and intrinsic (value) sources of information. We demonstrated through fixation behaviors that information about effort and value is processed serially, with an initial prioritization of effort information, followed by the additive processing of value information, and then finally with the integration of effort and value information into a signal that likely represents the cost-benefit difference of the items under consideration. Further, this integrated signal was predictive of subsequent choice in the task. This work extends previous work by providing a temporally precise method to characterize the dynamics of the online processing of extrinsic and intrinsic sources of value information.

CHAPTER VII

GENERAL CONCLUSIONS

The current work utilized behavioral, eye tracking, and neuroimaging methods to probe the underlying representations of decision variables (DVs) across a diverse set of tasks. The first study sought to examine what kind of information was represented in the feedback-related negativity (FRN), an event-related potential that has been shown across many studies to reflect the evaluation of ongoing events. This study was motivated by the fact that there is a degree of inconsistency regarding whether or not the FRN represents information about the valence, probability, or magnitude of outcomes. To address this, we developed a modified two-armed bandit task where we could study each of these sources of information simultaneously. Consistent with the overwhelming majority of previous work (Gehring & Willoughby, 2002; Nieuwenhuis et al., 2004; San Martín, 2012), the FRN robustly represented the valence of the outcome (i.e. losses vs. gains.). However, in contrast to two prominent theories that describe how the FRN should respond to the expectancy of an outcome (Alexander & Brown, 2011; Holroyd & Coles, 2002), we failed to observe any modulation of the FRN that could be attributed to outcome expectancy. This may have resulted from the fact that the task employed to measure the FRN explicitly provided subjects with information about the probability of outcomes, and did not require them to develop expectations by learning through trial and error, a key feature which may be required to see expectancy effects with the FRN (Holroyd et al., 2009). However, subject behavior suggested that they had fairly strong expectations about the likely outcomes of trials given that decisions to place high bets tracked the cued probability of gain on a trialwise basis. Finally, FRN amplitudes were

significantly modulated by the magnitude of outcomes, but this effect was qualified by an interaction that indicated that this was driven by the magnitude of gains rather than losses. At the very least, this study demonstrated that the FRN coarsely categorizes outcomes based on their relative "goodness" or "badness", and may also represent the magnitudes of gains and losses asymmetrically, a result consistent with the types of utility functions that are proposed to contribute to attitudes like loss aversion (Fox & Poldrack, 2008; Kahneman & Tversky, 1979).

These asymmetric utility functions are the basis for Prospect Theory, which has demonstrated that losses are approximately twice as psychologically impactful as equivalent gains. However, it is unclear when this bias might occur during the processing of value information; does loss aversion manifest itself during the processing of individual elements of value information, or is it a bias that occurs only once all relevant information has been integrated? Furthermore, is loss aversion the result of a shift in the boundary that separates losses and gains that results in only the highest gains to be perceived as gains, or instead the result of an "all or none" processing of losses, with a selective sensitivity to the magnitude of gains (a result that would be consistent with asymmetric utility functions)? To this end, the second study investigated during what stage of processing (individual element or integrated stream) distortions in valuation like loss aversion might occur, as well as what kind of representation of gains and losses might be responsible for these distortions. We accomplished this by implementing a novel multi-sampling gambling task where optimal performance requires the processing of both individual elements from a stream of value information, as well as successful integration of this stream of information in order to make informed decisions of when to

gamble or play it safe. Subject behavior suggested that they were able to successfully track the expected value of the stream of loss and gain information and placed bets accordingly. Using multivariate pattern classification methods, we probed the informational content of delta power (which has been shown to be associated with information about both value and information integration (Bernat et al., 2015; Knyazev, 2007; 2012; Wyart et al., 2012)) across the scalp in an attempt to map the representational space of losses and gains at both the individual element and stream level. At the element level, patterns of delta power suggested that losses and gains were represented in a purely categorical manner, with no indication of sensitivity to their magnitude. In contrast, at the stream level, losses were represented as "all or none" whereas gains showed magnitude sensitivity. Overall, these results suggest that biases such as loss aversion may occur farther downstream during the processing of an integrated value signal, rather than rather than during the processing of the discrete elements of information that contributed to that signal. Furthermore, this study provides further support that loss aversion is best characterized by asymmetric utility functions for losses and gains.

Although value is often associated with intrinsic properties of the item items under consideration (i.e. taste, appearance, etc.), value can also be affected by properties that are *extrinsic* to the item, such as the cost or effort associated with attaining the item (Rangel & Hare, 2010). There are many instances where there are multiple sources of information that help shape the perception of value of an item; perhaps the most intuitive way to conceptualize this is via the tradeoff of intrinsic benefits vs. extrinsic costs. A large body of work has suggested that we extract relevant value information form the

environment by continuously sampling items under consideration through visual fixations (Krajbich et al., 2010; Rangel et al., 2008). What remains unclear is whether or not information about value and effort are processed in the same manner, as well as when and how these two types of information are integrated during the decision process. To further investigate this, we utilized a task where subjects had to make decisions between pairs of food items where in some cases subjects were provided with the foreknowledge that one of the items was associated with an effortful math task. This approach allowed us to look at the dynamics of fixation behavior to determine whether effort and value information are processed serially or in parallel, as well when (if at all) the two types of information are integrated in a manner that is meaningful to choice behavior. Overall, the results suggested that fixation behavior prioritized the processing of effort information during initial stages of the decision phase, with no modulation by the value of the item; early eye fixations were rapidly drawn to items that had been pre-cued to be associated with math. This likely represented an overall tendency to determine whether the item was worth the effort of having to complete a difficult math problem. Middle fixations showed a tradeoff between effort and value information; fixations were less likely to be drawn to math, and more likely to be drawn to the valued item. Importantly, for middle fixations, the effect of value was additive, and effort and value did not have an interactive effect on fixations. Finally, during late fixations, effort and value interacted in a manner likely reflected the degree of conflict between the two. Specifically, during late fixations, the preferred item was just as likely to be looked at during both the math and no math conditions. In contrast, fixations to the non-preferred item were overall less likely in the no math condition compared to the math condition. The dynamics of fixations over time suggest

that effort and value are processed independently and serially during the early and middle stages of the decision process. However, during the later stage of the process, the two types of information are integrated and considered simultaneously. If fixation patterns during this later stage reflect a decision-relevant variable where effort and value are being considered simultaneously, it would stand to reason that fixations during this period predict choice behavior. Indeed, the interaction between effort, value, and fixations only predicted choice during the late period, but nor during the early or middle periods, suggesting that the different types of information need to be normalized to a common scale to effectively guide behavior (Chib et al., 2009; Levy & Glimcher, 2012; Rangel et al., 2008).

This body of work provides several insights regarding the nature of key decision-relevant variables. Studies one and two found that underlying representations of losses and gains indicate that they are categorically distinct from one another, and that the underlying representations of losses and gains may be derived from two independent sources of information. In study one, the FRN distinguished between gains and losses following feedback in a gambling task, an effect that has been replicated across many studies (Bellebaum et al., 2010; Gehring & Willoughby, 2002; Hajcak et al., 2005; Holroyd et al., 2004; Holroyd & Coles, 2002; Miltner et al., 1997; Weinberg, Luhmann, Bress, & Hajcak, 2012), Study two found that when probing the informational content of delta power, there was a sharp categorical boundary between losses and gains; losses were much more likely to be confused with other losses rather than gains, and vice versa. Taken together, studies two and three suggest that constructs like rewards and

punishments may actually represent two independent sources of information, rather than a single dimension with rewards and punishment at opposite ends.

Studies one and two also provide supporting evidence that losses and gains are represented asymmetrically, consistent with Prospect Theory (Fox & Poldrack, 2008; Kahneman & Tversky, 1979). Both studies showed that losses are treated in an "all or none" manner, with no sensitivity to their magnitude. In contrast, there seems to a be a selective sensitivity to the magnitude of gains (this was only present at the stream level in study two). It is important to highlight that this general pattern was observed using two very distinct neural measures: the fronto-central FRN, and the distribution of delta power across the entire scalp. Perhaps even more striking is that the observed loss/gain asymmetries not only occurred with distinct neural measures, they occurred during two distinct stages of decision making. In the case of study one, asymmetries were present during post-choice feedback period, whereas in study two, the asymmetries were present during the pre-choice decision period. Together, these findings suggest that evidence for biases such as loss aversion can be found along various points of the decision making process, but the biases only occur when all the subject has integrated all relevant loss/gain information.

Study three provided evidence that suggests that different types of decision-relevant information (i.e. effort vs. value) are processed serially. Furthermore, study three shows that they are processed independently during early stages of the decision process. This result reinforces the idea discrete sources of information are considered independently, and may exist as entirely separate dimensions. As highlighted earlier, losses and gains may exist within the context of completely independent and

unidirectional scales of punishment and reward. Likewise, value (i.e. reward) and effort may be represented in a similar fashion. Early processing of effort and reward does not immediately show integration of the two types of information; this only begins to show during the later periods of the decision process. Thus, early processing seems to reflect a more modular and independent representation of information which is later integrated into a DV. As seen in study two, it is during this later integration phase that biases in the representation of that information might occur.

Although the current work provides evidence to suggest how decision-relevant information may be processed independently (at least initially), future work could investigate the limits of this. In each study, we provided subjects with at most three types of information to consider (valence, probability, magnitude; study 1). In reality however, some decisions are much more complex. For example, deciding whether or not to uproot your family by accepting a job in another city. Does the number of different types of information to consider have any impact on the ability to process them independently? Is there a capacity limit to how much decision-relevant information you can process in a modular manner, similar to a capacity limit for working memory (Unsworth, Fukuda, Awh, & Vogel, 2014)? Is there a point where decision-relevant information immediately is chunked into coarse "good" and "bad" categories early in the decision process? These are all open questions that could be addressed with further research, and would provide valuable insight to the underlying computations necessary for decision making.

Among the most common types of decisions that humans engage in are value based decisions. The current work sought to further investigate value based decision making by probing the nature of how information is represented using behavioral,

neuroimaging, and eye tracking methods. This approach allows us to obtain a more comprehensive and complete picture of the underlying processes and computations that give rise to this fundamental aspect of our experience. Overall, this work advances the field by providing additional insight about how decision-relevant information is represented in a dynamic and flexible manner.

REFERENCES CITED

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

Anderson, B. A. (2013). A value-driven mechanism of attentional selection. *Journal of Vision*, *13*(3), 7–7. http://doi.org/10.1167/13.3.7

Armel, K. C., Beaumel, A., & Rangel, A. (2008). Biasing simple choices by manipulating relative visual attention. *... And Decision Making*.

Banis, S., & Lorist, M. M. (2012). Acute noise stress impairs feedback processing. *Biological Psychology*, *91*(2), 163–171. http://doi.org/10.1016/j.biopsycho.2012.06.009

Bates, D., Mächler, M., Bolker, B., & Walker, S. (2015). Fitting Linear Mixed-Effects Models Using lme4. *Journal of Statistical Software*, *67*(1), 1–48. http://doi.org/10.18637/jss.v067.i01

Bellebaum, C., Polezzi, D., & Daum, I. (2010). It is less than you expected: the feedback-related negativity reflects violations of reward magnitude expectations. *Neuropsychologia*, *48*(11), 3343–3350. http://doi.org/10.1016/j.neuropsychologia.2010.07.023

Bernat, E. M., Nelson, L. D., & Baskin-Sommers, A. R. (2015). Time-frequency theta and delta measures index separable components of feedback processing in a gambling task. *Psychophysiology*, n/a–n/a. http://doi.org/10.1111/psyp.12390

Botvinick, M., Weinstein, A., Solway, A., & Barto, A. (2015). Reinforcement learning, efficient coding, and the statistics of natural tasks. *Current Opinion in Behavioral Sciences*, *5*(C), 71–77. http://doi.org/10.1016/j.cobeha.2015.08.009

Brainard, D. H. (1997). The Psychophysics Toolbox. *Spatial Vision*, *10*(4), 433–436. http://doi.org/10.1163/156856897x00357

Canessa, N., Crespi, C., Motterlini, M., Baud-Bovy, G., Chierchia, G., Pantaleo, G., et al. (2013). The functional and structural neural basis of individual differences in loss aversion. *The Journal of Neuroscience : the Official Journal of the Society for Neuroscience*, *33*(36), 14307–14317. http://doi.org/10.1523/JNEUROSCI.0497-13.2013

Cavanagh, J. F., Frank, M. J., Klein, T. J., & Allen, J. J. B. (2010). Frontal theta links prediction errors to behavioral adaptation in reinforcement learning. *NeuroImage*, *49*(4), 3198–3209. http://doi.org/10.1016/j.neuroimage.2009.11.080

Chib, V. S., Rangel, A., Shimojo, S., & O'Doherty, J. P. (2009). Evidence for a common representation of decision values for dissimilar goods in human ventromedial prefrontal cortex. *The Journal of Neuroscience : the Official Journal of the Society for Neuroscience*, *29*(39), 12315–12320. http://doi.org/10.1523/JNEUROSCI.2575-09.2009

Cohen, M. X., & Cavanagh, J. F. (2011). Single-Trial Regression Elucidates the Role of Prefrontal Theta Oscillations in Response Conflict. *Frontiers in Psychology*, *2*. http://doi.org/10.3389/fpsyg.2011.00030

Cohen, M. X., Elger, C. E., & Ranganath, C. (2007). Reward expectation modulates feedback-related negativity and EEG spectra. *NeuroImage*, *35*(2), 968–978. http://doi.org/10.1016/j.neuroimage.2006.11.056

Cornelissen, F. W., Peters, E. M., & Palmer, J. (2002). The Eyelink Toolbox: Eye tracking with MATLAB and the Psychophysics Toolbox. *Behavior Research Methods, Instruments, & Computers*, *34*(4), 613–617. http://doi.org/10.3758/BF03195489

Delorme, A., & Makeig, S. (2004). EEGLAB: an open source toolbox for analysis of single-trial EEG dynamics including independent component analysis. *Journal of Neuroscience Methods*, *134*(1), 9–21. http://doi.org/10.1016/j.jneumeth.2003.10.009

Ferdinand, N. K., Mecklinger, A., Kray, J., & Gehring, W. J. (2012). The Processing of Unexpected Positive Response Outcomes in the Mediofrontal Cortex, *32*(35), 12087–12092. http://doi.org/10.1523/JNEUROSCI.1410

Foti, D., Weinberg, A., Bernat, E. M., & Proudfit, G. H. (2015). Anterior cingulate activity to monetary loss and basal ganglia activity to monetary gain uniquely contribute to the feedback negativity. *Clinical Neurophysiology : Official Journal of the International Federation of Clinical Neurophysiology*, *126*(7), 1338–1347. http://doi.org/10.1016/j.clinph.2014.08.025

Fox, C. R., & Poldrack, R. A. (2008). Prospect theory and the brain. *Handbook of Neuroeconomics*.

Gehring, W. J., & Willoughby, A. R. (2002). The medial frontal cortex and the rapid processing of monetary gains and losses. *Science (New York, N.Y.)*, *295*(5563), 2279–2282. http://doi.org/10.1126/science.1066893

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. http://doi.org/10.1111/j.1467-9280.1993.tb00586.x

Gu, R., Lei, Z., Broster, L., Wu, T., Jiang, Y., & Luo, Y.-J. (2011). Beyond valence and magnitude: a flexible evaluative coding system in the brain. *Neuropsychologia*, *49*(14), 3891–3897. http://doi.org/10.1016/j.neuropsychologia.2011.10.006

Hajcak, G., Holroyd, C. B., Moser, J. S., & Simons, R. F. (2005). Brain potentials associated with expected and unexpected good and bad outcomes. *Psychophysiology*, *42*(2), 161–170. http://doi.org/10.1111/j.1469-8986.2005.00278.x

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. http://doi.org/10.1016/j.biopsycho.2005.04.001

Hajcak, G., Moser, J. S., Holroyd, C. B., & Simons, R. F. (2007). It's worse than you thought: the feedback negativity and violations of reward prediction in gambling tasks. *Psychophysiology*, *44*(6), 905–912. http://doi.org/10.1111/j.1469-8986.2007.00567.x

Hare, T. A., Malmaud, J., & Rangel, A. (2011). Focusing Attention on the Health Aspects of Foods Changes Value Signals in vmPFC and Improves Dietary Choice. *Journal of Neuroscience*, *31*(30), 11077–11087. http://doi.org/10.1523/JNEUROSCI.6383-10.2011

Harris, A., & Lim, S. L. (2016). Temporal Dynamics of Sensorimotor Networks in Effort-Based Cost-Benefit Valuation: Early Emergence and Late Net Value Integration. *Journal of Neuroscience*, *36*(27), 7167–7183. http://doi.org/10.1523/JNEUROSCI.4016-15.2016

Harris, A., Adolphs, R., Camerer, C., & Rangel, A. (2011). Dynamic Construction of Stimulus Values in the Ventromedial Prefrontal Cortex. *PloS One*, *6*(6), e21074. http://doi.org/10.1371/journal.pone.0021074

Hauser, T. U., Iannaccone, R., Stämpfli, P., Drechsler, R., Brandeis, D., Walitza, S., & Brem, S. (2014). The feedback-related negativity (FRN) revisited: New insights into the localization, meaning and network organization. *NeuroImage*, *84*(C), 159–168. http://doi.org/10.1016/j.neuroimage.2013.08.028

Haxby, J. V., Gobbini, M. I., Furey, M. L., Ishai, A., Schouten, J. L., & Pietrini, P. (2001). Distributed and overlapping representations of faces and objects in ventral temporal cortex. *Science*, *293*(5539), 2425–2430. http://doi.org/10.1126/science.1063736

Hewig, J., Trippe, R., Hecht, H., Coles, M. G. H., Holroyd, C. B., & Miltner, W. H. R. (2007). Decision-making in Blackjack: an electrophysiological analysis. *Cerebral Cortex (New York, N.Y. : 1991)*, *17*(4), 865–877. http://doi.org/10.1093/cercor/bhk040

Holroyd, C. 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. http://doi.org/10.1037//0033-295X.109.4.679

Holroyd, C. B., Krigolson, O. E., Baker, R., Lee, S., & Gibson, J. (2009). When is an error not a prediction error? An electrophysiological investigation. *Cognitive, Affective \& Behavioral Neuroscience*, *9*(1), 59–70. http://doi.org/10.3758/CABN.9.1.59

Holroyd, C. 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. http://doi.org/10.1111/j.1469-8986.2004.00152.x

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), 2481–2484. http://doi.org/10.1097/01.wnr.0000099601.41403.a5

Holroyd, C. B., Pakzad-Vaezi, K. L., & Krigolson, O. E. (2008). The feedback correct-related positivity: sensitivity of the event-related brain potential to unexpected positive feedback. *Psychophysiology*, *45*(5), 688–697. http://doi.org/10.1111/j.1469-8986.2008.00668.x

Kahneman, D., & Tversky, A. (1979). Prospect Theory: An Analysis of Decision under Risk. *Econometrica*, *47*(2), 263. http://doi.org/10.2307/1914185

Kamarajan, C., Porjesz, B., Rangaswamy, M., Tang, Y., Chorlian, D. B., Padmanabhapillai, A., et al. (2009). Brain signatures of monetary loss and gain: outcome-related potentials in a single outcome gambling task. *Behavioural Brain Research*, *197*(1), 62–76. http://doi.org/10.1016/j.bbr.2008.08.011

Klein-Flugge, M. C., Kennerley, S. W., Friston, K., & Bestmann, S. (2016). Neural Signatures of Value Comparison in Human Cingulate Cortex during Decisions Requiring an Effort-Reward Trade-off. *Journal of Neuroscience*, *36*(39), 10002–10015. http://doi.org/10.1523/JNEUROSCI.0292-16.2016

Knutson, B., Rick, S., Wimmer, G. E., Prelec, D., & Loewenstein, G. (2007). Neural Predictors of Purchases. *Neuron*, *53*(1), 147–156. http://doi.org/10.1016/j.neuron.2006.11.010

Knyazev, G. G. (2007). Motivation, emotion, and their inhibitory control mirrored in brain oscillations. *Neuroscience and Biobehavioral Reviews*, *31*(3), 377–395. http://doi.org/10.1016/j.neubiorev.2006.10.004

Knyazev, G. G. (2012). EEG delta oscillations as a correlate of basic homeostatic and motivational processes. *Neuroscience and Biobehavioral Reviews*, *36*(1), 677–695. http://doi.org/10.1016/j.neubiorev.2011.10.002

Kokmotou, K., Cook, S., Xie, Y., Wright, H., Soto, V., Fallon, N., et al. (2017). Effects of loss aversion on neural responses to loss outcomes: An event-related potential study. *Biological Psychology*, *126*, 30–40. http://doi.org/10.1016/j.biopsycho.2017.04.005

Krajbich, I., & Rangel, A. (2011). Multialternative drift-diffusion model predicts the relationship between visual fixations and choice in value-based decisions. *Proceedings of the National Academy of Sciences of the United States of America*, *108*(33), 13852–13857. http://doi.org/10.1073/pnas.1101328108

Krajbich, I., Armel, K. C., & Rangel, A. (2010). Visual fixations and the computation and comparison of value in simple choice. *Nature Neuroscience*, *13*(10), 1292–1298. http://doi.org/10.1038/nn.2635

Krajbich, I., Lu, D., Camerer, C., & Rangel, A. (2012). The attentional drift-diffusion model extends to simple purchasing decisions. *Frontiers in Psychology*, *3*(June), 193. http://doi.org/10.3389/fpsyg.2012.00193

Kriegeskorte, N. (2008). Representational similarity analysis – connecting the branches of systems neuroscience. *Frontiers in Systems Neuroscience*, 1–28. http://doi.org/10.3389/neuro.06.004.2008

Kriegeskorte, N., Mur, M., Ruff, D. A., Kiani, R., Bodurka, J., Esteky, H., et al. (2008). Matching Categorical Object Representations in Inferior Temporal Cortex of Man and Monkey. *Neuron*, *60*(6), 1126–1141. http://doi.org/10.1016/j.neuron.2008.10.043

Levy, D. J., & Glimcher, P. W. (2012). The root of all value: a neural common currency for choice. *Current Opinion in Neurobiology*, *22*(6), 1027–1038. http://doi.org/10.1016/j.conb.2012.06.001

Lim, S. L., O'Doherty, J. P., & Rangel, A. (2011). The Decision Value Computations in the vmPFC and Striatum Use a Relative Value Code That is Guided by Visual Attention. *Journal of Neuroscience*, *31*(37), 13214–13223. http://doi.org/10.1523/JNEUROSCI.1246-11.2011

Martin, L. E., & Potts, G. F. (2011). Medial frontal event-related potentials and reward prediction: Do responses matter? *Brain and Cognition*, *77*(1), 128–134. http://doi.org/10.1016/j.bandc.2011.04.001

Miltner, W., Braun, C. H., & Coles, M. (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. http://doi.org/10.1162/jocn.1997.9.6.788

Nieuwenhuis, S., Holroyd, C. B., Mol, N., & Coles, M. G. H. (2004). Reinforcement-related brain potentials from medial frontal cortex: origins and functional significance. *Neuroscience and Biobehavioral Reviews*, *28*(4), 441–448. http://doi.org/10.1016/j.neubiorev.2004.05.003

Norman, K. A., Polyn, S. M., Detre, G. J., & Haxby, J. V. (2006). Beyond mind-reading: multi-voxel pattern analysis of fMRI data. *Trends in Cognitive Sciences*, *10*(9), 424–430. http://doi.org/10.1016/j.tics.2006.07.005

Oliveira, F. T. P., McDonald, J. J., & Goodman, D. (2007). Performance monitoring in the anterior cingulate is not all error related: expectancy deviation and the representation of action-outcome associations. *Journal of Cognitive Neuroscience*, *19*(12), 1994–2004. http://doi.org/10.1162/jocn.2007.19.12.1994

Plassmann, H., O'Doherty, J. P., & Rangel, A. (2010). Appetitive and Aversive Goal Values Are Encoded in the Medial Orbitofrontal Cortex at the Time of Decision Making. *Journal of Neuroscience*, *30*(32), 10799–10808. http://doi.org/10.1523/JNEUROSCI.0788-10.2010

Platt, M. L., & Glimcher, P. W. (1999). Neural correlates of decision variables in parietal cortex. *Nature*, *400*(6741), 233–238. http://doi.org/10.1038/22268

Polanía, R., Krajbich, I., Grueschow, M., & Ruff, C. C. (2014). Neural Oscillations and Synchronization Differentially Support Evidence Accumulation in Perceptual and Value-Based Decision Making. *Neuron*, *82*(3), 709–720. http://doi.org/10.1016/j.neuron.2014.03.014

Polanía, R., Moisa, M., Opitz, A., Grueschow, M., & Ruff, C. C. (2015). The precision of value-based choices depends causally on fronto-parietal phase coupling. *Nature Communications*, *6*, 8090. http://doi.org/10.1038/ncomms9090

Prévost, C., Pessiglione, M., Météreau, E., Cléry-Melin, M.-L., & Dreher, J.-C. (2010). Separate valuation subsystems for delay and effort decision costs. *The Journal of Neuroscience : the Official Journal of the Society for Neuroscience*, *30*(42), 14080–14090. http://doi.org/10.1523/JNEUROSCI.2752-10.2010

Proudfit, G. H. (2015). The reward positivity: From basic research on reward to a biomarker for depression. *Psychophysiology*, *52*(4), 449–459. http://doi.org/10.1111/psyp.12370

R Core Team. (2016). R: A Language and Environment for Statistical Computing.

Rangel, A., & Hare, T. (2010). Neural computations associated with goal-directed choice. *Current Opinion in Neurobiology*, *20*(2), 262–270. http://doi.org/10.1016/j.conb.2010.03.001

Rangel, A., Camerer, C., & Montague, P. R. (2008). A framework for studying the neurobiology of value-based decision making. *Nature Reviews. Neuroscience*, *9*(7), 545–556. http://doi.org/10.1038/nrn2357

Sambrook, T. D., & Goslin, J. (2014). Mediofrontal event-related potentials in response to positive, negative and unsigned prediction errors. *Neuropsychologia*, *61*(C), 1–10. http://doi.org/10.1016/j.neuropsychologia.2014.06.004

Sambrook, T. D., & Goslin, J. (2015). A neural reward prediction error revealed by a meta-analysis of ERPs using great grand averages. *Psychological Bulletin*, *141*(1), 213–235. http://doi.org/10.1037/bul0000006

San Martín, R. (2012). Event-related potential studies of outcome processing and feedback-guided learning. *Frontiers in Human Neuroscience*, *6*. http://doi.org/10.3389/fnhum.2012.00304

San Martín, R., Kwak, Y., Pearson, J. M., Woldorff, M. G., & Huettel, S. A. (2016). Altruistic traits are predicted by neural responses to monetary outcomes for self versus charity. *Social Cognitive and Affective Neuroscience*, *11*(6), nsw026–14. http://doi.org/10.1093/scan/nsw026

San Martín, René, Appelbaum, L. G., Pearson, J. M., Huettel, S. A., & Woldorff, M. G. (2013). Rapid brain responses independently predict gain maximization and loss minimization during economic decision making. *The Journal of Neuroscience : the Official Journal of the Society for Neuroscience*, *33*(16), 7011–7019. http://doi.org/10.1523/jneurosci.4242-12.2013

San Martín, René, Manes, F., Hurtado, E., Isla, P., & Ibañez, A. (2010). Size and probability of rewards modulate the feedback error-related negativity associated with wins but not losses in a monetarily rewarded gambling task. *NeuroImage*.

Santesso, D. L., Dzyundzyak, A., & Segalowitz, S. J. (2011). Age, sex and individual differences in punishment sensitivity: Factors influencing the feedback-related negativity. *Psychophysiology*, *48*(11), 1481–1489. http://doi.org/10.1111/j.1469-8986.2011.01229.x

Shadlen, M. N., & Kiani, R. (2013). Decision Making as a Window on Cognition. *Neuron*, *80*(3), 791–806. http://doi.org/10.1016/j.neuron.2013.10.047

Silvetti, M., Castellar, E. N., Roger, C., & Verguts, T. (2014). Reward expectation and prediction error in human medial frontal cortex: An EEG study. *NeuroImage*, *84*(C), 376–382. http://doi.org/10.1016/j.neuroimage.2013.08.058

Stanton, S. J., Mullette-Gillman, O. A., McLaurin, R. E., Kuhn, C. M., LaBar, K. S., Platt, M. L., & Huettel, S. A. (2011). Low- and High-Testosterone Individuals Exhibit Decreased Aversion to Economic Risk. *Psychological Science*, *22*(4), 447–453. http://doi.org/10.1177/0956797611401752

Talmi, D., Atkinson, R., & El-Deredy, W. (2013). The feedback-related negativity signals salience prediction errors, not reward prediction errors. *The Journal of Neuroscience : the Official Journal of the Society for Neuroscience*, *33*(19), 8264–8269. http://doi.org/10.1523/JNEUROSCI.5695-12.2013

Tom, S. M., Fox, C. R., Trepel, C., & Poldrack, R. A. (2007). The neural basis of loss aversion in decision-making under risk. *Science (New York, N.Y.)*, *315*(5811), 515–518. http://doi.org/10.1126/science.1134239

Towal, R. B., Mormann, M., & Koch, C. (2013). Simultaneous modeling of visual saliency and value computation improves predictions of economic choice. *Proceedings of the National Academy of Sciences*, *110*(40), E3858–E3867. http://doi.org/10.1073/pnas.1304429110

Tversky, A., & Kahneman, D. (1981). The framing of decisions and the psychology of choice. *Science*, *211*(4481), 453–458.

Tversky, A., & Kahneman, D. (1991). Loss Aversion in Riskless Choice: A Reference-Dependent Model. *The Quarterly Journal of Economics*, *106*(4), 1039–1061. http://doi.org/10.2307/2937956

Tversky, A., & Kahneman, D. (1992). Advances in Prospect Theory: Cumulative Representation of Uncertainty. *Journal of Risk and Uncertainty*, *5*(4), 297–323. http://doi.org/10.2307/41755005?ref=search-gateway:4016cc49e7a26b1170191726c2331a71

Unsworth, N., Fukuda, K., Awh, E., & Vogel, E. K. (2014). Working memory and fluid intelligence: Capacity, attention control, and secondary memory retrieval. *Cognitive Psychology*, *71*(C), 1–26. http://doi.org/10.1016/j.cogpsych.2014.01.003

Walsh, M. M., & Anderson, J. R. (2011). Modulation of the feedback-related negativity by instruction and experience, *2011*, 1–6. http://doi.org/10.1073/pnas.1117189108/-/DCSupplemental.www.pnas.org/cgi/doi/10.1073/pnas.1117189108

Walsh, M. M., & Anderson, J. R. (2012). Learning from experience: event-related potential correlates of reward processing, neural adaptation, and behavioral choice. *Neuroscience and Biobehavioral Reviews*, *36*(8), 1870–1884. http://doi.org/10.1016/j.neubiorev.2012.05.008

Weinberg, A., Luhmann, C. C., Bress, J. N., & Hajcak, G. (2012). Better late than never? The effect of feedback delay on ERP indices of reward processing. *Cognitive, Affective \& Behavioral Neuroscience*, *12*(4), 671–677.

Wyart, V., de Gardelle, V., Scholl, J., & Summerfield, C. (2012). Rhythmic Fluctuations in Evidence Accumulation during Decision Making in the Human Brain. *Neuron*, *76*(4), 847–858. http://doi.org/10.1016/j.neuron.2012.09.015

Yechiam, E., & Hochman, G. (2013). Loss-aversion or loss-attention: The impact of losses on cognitive performance. *Cognitive Psychology*, *66*(2), 212–231. http://doi.org/10.1016/j.cogpsych.2012.12.001

Yeung, N., Holroyd, C. B., & Cohen, J. D. (2005). ERP correlates of feedback and reward processing in the presence and absence of response choice. *Cerebral Cortex (New York, N.Y. : 1991)*, *15*(5), 535–544. http://doi.org/10.1093/cercor/bhh153

Zani, A., & Proverbio, A. M. (2003). Reinforcement Learning Signals Predict Future Decisions. Academic Press. http://doi.org/10.1523/JNEUROSCI.4421-06.2007