Neuroscience 228 (2013) 361–370

WHAT DO I DO NOW? AN ELECTROENCEPHALOGRAPHIC INVESTIGATION OF THE EXPLORE/EXPLOIT DILEMMA

C. D. HASSALL,* K. HOLLAND AND O. E. KRIGOLSON

Department of Psychology and Neuroscience, Dalhousie University, Halifax, Nova Scotia, Canada B3H 4R2

Abstract—To maximize reward, we are faced with the dilemma of having to balance the exploration of new response options and the exploitation of previous choices. Here, we sought to determine if the event-related brain potential (ERP) in the P300 time range is sensitive to decisions to explore or exploit within the context of a sequential risk-taking task. Specifically, the task we used required participants to continually explore their options—whether they should ''push their luck'' and keep gambling or ''take the money and run'' and collect their winnings. Our behavioral analysis yielded two distinct distributions of response times: a larger group of short-decision times and a smaller group of long-decision times. Interestingly, these data suggest that participants adopted one of two modes of control on any given trial: a mode where they quickly decided to keep gambling (i.e. exploit), and a mode where they deliberated whether to the take the money they had already won or continue gambling (i.e. explore). Importantly, we found that the amplitude of the ERP in the P300 time range was larger for explorative decisions than for exploitative decisions and, further, was correlated with decision time. Our results are consistent with a recent theoretical account that links changes in ERP amplitude in the P300 time range with phasic activity of the locus coeruleus–norepinephrine system and decisions to engage in exploratory behavior.

Key words: P300, exploration/exploitation tradeoff, decision making, reinforcement learning, ERP, learning.

INTRODUCTION

In Mill's Utilitarianism (1863/2008), he argued that humans have an inherent desire to maximize utility. As such, the decisions that we make on a day-to-day and moment-to-moment basis typically reflect a desire to maximize the reward. However, as Dennett (1986) and others have pointed out, calculating the utility of decisions in the real world can be challenging because the potential consequences of our actions are not always known. Even if utility calculations are restricted to the near future, complex or novel situations may arise that require exploring options with unknown consequences. Exploration is inherently risky but necessary in order to assess new response options or reassess old ones. The knowledge gained through exploration can later be exploited to improve subsequent decisions, and thus yield even greater increases in utility. However, one cannot always engage in exploratory behavior. Rather, one must balance exploratory behavior with exploitation—selecting the most rewarding response option as much as possible. Therefore, an optimal decision strategy for maximizing utility would entail utilizing an exploitative mode of control most of the time with occasional instances of exploratory behavior.

Experimentally, decisions to explore or exploit can be studied in tasks such as the Balloon Analog Risk Task (BART: Lejuez et al., 2002). During performance of the BART, participants must continually explore their options—either take the money they have already earned or continue gambling. The key manipulation of the BART is that, for each pump of the balloon (gamble), the amount of money earned increases along with the probability of losing all earned money. This manipulation makes each gamble increasingly risky. Thus, there is an optimal response in the BART (i.e. total number of balloon pumps) that is based on the risk and reward structure of the task (Lejuez et al., 2002), and as such, to maximize reward, participants need to explore in order to determine the optimal number of balloon pumps. Computational models of the BART suggest that people make a risk assessment prior to each pump: a decision to continue pumping or collect their accumulated reward (Wallsten et al., 2005). The Wallsten et al. (2005) model's predictions were recently corroborated by Wershbale and Pleskac (2010) who observed two distinct distributions of response times in human BART performance. Specifically, they observed that people generally made automatic, rapid responses in the BART, but occasionally paused to assess whether or not they should continue gambling. Wershbale and Pleskac (2010) hypothesized that these pauses represent the assessments predicted by earlier modeling work (Wallsten et al., 2005; Pleskac, 2008). Interestingly, the number of assessments that

*Corresponding author. Address: Department of Psychology and Neuroscience, Dalhousie University, P.O. Box 15000, Halifax, Nova Scotia, Canada B3H 4R2. Tel: +1-902-494-2923; fax: +1-902-494-6585.

E-mail address: cameron.hassall@dal.ca (C. D. Hassall).

Abbreviations: BART, Balloon Analog Risk Task; BSR, Bayesian sequential risk-taking model; EEG, electroencephalographic; ERP, event-related brain potential; fMRI, functional magnetic resonance imaging; LC–NE, locus coeruleus–norepinephrine; PFC, prefrontal cortex; sLORETA, standardized low-resolution brain electromagnetic tomography.

0306-4522/12 $36.00
http://dx.doi.org/10.1016/j.neuroscience.2012.10.040

participants made during the BART decreased over time. Importantly, this change in assessment rate is consistent with theoretical models of the exploration/exploitation dilemma. Early in learning, people need to explore more often in order to determine the reward structure of a task (e.g., the optimal number of pumps in the BART). However, once the reward structure is known, people exploit more frequently. With all of this in mind, Wershbale and Pleskac (2010) likened fast BART responses to exploitation and slower responses to exploration.

Research examining the neural basis of decisions to explore or exploit is limited (see Cohen et al., 2007 for a review). In one recent study, Cavanagh et al. (2011) suggested increased frontal theta-band oscillation as a possible neural marker of uncertainty-driven exploration. Specifically, Cavanagh and colleagues (2011) observed a correlation between medial–frontal theta power and the parameters of their reinforcement-learning model during exploration in a decision-making task. From their results, Cavanagh et al. (2011) hypothesized that midbrain regions were responsible for exploitation but that frontal brain regions took control when deciding to explore in uncertain situations. The Cavanagh et al. (2011) hypothesis is consistent with an earlier functional magnetic resonance imaging (fMRI) study that showed enhanced frontal brain activity during exploratory decisions in a four-armed bandit task (Daw et al., 2006). Cavanagh and colleagues' (2011) hypothesis is also consistent with work by Frank et al. (2009) that associated a prefrontal cortex (PFC) dopamine gene (COMT) with exploratory decisions. In particular, Frank et al. (2009) showed an effect of COMT gene dose (which they defined as the amount of methionine-encoding or *met* allele present) on uncertainty-driven exploration. The presence of the met allele is linked to increased PFC dopamine levels compared to the presence of the valine-encoding or *val* allele. Although Frank et al. (2009) were uncertain about the exact role of COMT in exploratory behavior, they suggested that the observed and known effects of the met allele implicate the PFC as the controller of uncertainty-driven exploration. Taken together, these studies suggest that switching from an exploitative to an explorative mode of control involves the intervention of frontal cognitive systems over midbrain lower-level reward-processing systems (see Mars et al., 2011, for more examples of cognitive control).

Currently, there are no definitive electroencephalographic (EEG) correlates differentiating decisions to explore or exploit. Having said that, there are good reasons to hypothesize that the event-related brain potential (ERP) in the time range of the P300 may be sensitive to this distinction. The P300 is a high-amplitude, positive ERP component with peak latency 300–500 ms following the presentation of a stimulus (Sutton et al., 1965) that has been associated with several different cognitive functions (Polich, 2007). One influential account—the context-updating hypothesis—states that the P300 reflects the updating of an internal model of the probabilistic structure of the world (Donchin, 1981; Donchin and Coles, 1988). Donchin's (1981) account arose out of earlier observations that the P300 is sensitive to stimulus frequency (Duncan-Johnson and Donchin, 1977). Consistent with the context-updating hypothesis, Nieuwenhuis et al. (2005) recently suggested that ERP changes in the P300 time range reflect the locus coeruleus–norepinephrine (LC–NE) system's response to internal decision-making processes regarding task-relevant stimuli (Aston-Jones and Cohen, 2005; Nieuwenhuis, 2011; also see Pineda et al., 1989, for early work linking the LC and the P300). The LC contains noradrenergic neurons and provides the only source of NE to the hippocampus and neocortex (Berridge and Waterhouse, 2003). Increases in LC activity, and the associated rise in NE, are linked to increased exploratory behavior in monkeys (Aston-Jones and Bloom, 1981; Usher et al., 1999; Aston-Jones and Cohen, 2005; modeled by McClure et al. (2006)). Importantly, a series of lesion, psychopharmacological, and EEG studies support the link between an ERP difference in the P300 time range and phasic changes in the activity of the LC–NE system (see Nieuwenhuis et al., 2005, for a review). Thus, given the link between the LC–NE system and exploration, and the link between the LC–NE system and the P300, it stands to reason that the amplitude of the ERP in the P300 time range may differentiate decisions to explore or exploit.

Our main purpose here was to determine whether or not ERP amplitude in the P300 time range would be sensitive to decisions to explore or exploit. To accomplish this, we had participants perform a modified version of the BART while EEG data were recorded. In terms of behavior, we expected to observe a similar distribution of response times as Wershbale and Pleskac (2010). In particular, we expected to see two distinct distributions of response times: one distribution of fast responses indicative of exploitation, and a second distribution of slow responses indicative of exploration. Importantly, we predicted that the amplitude of the ERP in the P300 time range preceding decisions to explore would be greater than the ERP amplitude in the same time range preceding decisions to exploit—a prediction derived from Nieuwenhuis and colleagues' (2005) hypothesis that ERP modulation in the P300 time range is driven by phasic changes in LC–NE activity linked to internal decision-making processes.

There is a growing body of evidence that the amplitude of the P300 is also modulated by reward magnitude (Yeung and Sanfey, 2004; Hajcak et al., 2005; Bellebaum and Daum, 2008; Wu and Zhou, 2009). The P300's sensitivity to reward magnitude is of particular importance here because the purpose of exploration is to specify or update values associated with actions, and the purpose of exploitation is to take advantage of current value assessments (Sutton and Barto, 1998). As such, we also hypothesized that the amplitude of the P300 elicited by balloon bursts would scale with the magnitude of the amount of lost reward, reflecting an update of participants' model of the probabilistic reward structure of the task.

EXPERIMENTAL PROCEDURES

Participants

Fourteen right-handed university-aged participants (2 male, mean age: 21.5 ± 1.5) with no known neurological impairments and with normal or corrected-to-normal vision took part in the experiment. All of the participants were volunteers who received monetary compensation for their participation. The participants provided informed consent approved by the Office of the Vice-President, Research, Dalhousie University, and the study was conducted in accordance with the ethical standards prescribed in the 1964 Declaration of Helsinki.

Apparatus and procedure

Participants were seated comfortably 75 cm in front of a computer monitor and used a standard USB controller to perform a computerized risk-taking task (written in MATLAB [Version 7.14, Mathworks, Natick, USA] using the Psychophysics Toolbox Extension, Brainard, 1997). To perform the task, participants pushed a button on the controller to inflate a ''balloon'' (initially a 2.8-cm diameter green circle, subtending 2.1° of visual angle) and earn money. Each trial began with the presentation of a fixation cross for one second. After one second, a green-colored balloon appeared behind the fixation cross, cuing participants to begin self-paced pumping. With each pump, the balloon either ''grew'' (increasing in size by 0.3° of visual angle) and the participant won five cents, or the balloon ''exploded'' (turned red—see below for more detail on the probability of the balloon exploding) and the participant lost all of the money he or she had won during that trial. As such, prior to each pump, participants had to decide whether or not to pump and potentially earn more money, or to stop the trial and take the money that they had already won (see Fig. 1 for timing details). After each group of 10 trials, participants were given a self-paced rest break. The experiment consisted of 300 trials in total. All trials were paid at a rate of 20:1 so that the average total payoff was $9.37 ± $0.16, with individual total payoffs ranging from $8.27 to $10.42.

Participants were informed that they would play 300 trials, but were given no prior information on the probability structure that governed the balloon exploding; rather, they were only informed ''it is up to you to decide how much to pump each balloon—some may pop after one pump, and some may not pop until the balloon fills the whole screen.'' In reality, and unbeknownst to participants, the computer program allowed a maximum of 30 pumps, and the balloon exploded randomly with a probability of $(31 - n)^{-1.4}$ on trial n.

Data collection

The experimental program recorded response time (elapsed time from the previous button press or start of trial, in ms), decision type (pump or collect), and whether or not the balloon grew or exploded. The EEG was recorded from 64 electrodes using BrainVision Recorder software (Version 1.20, Brainproducts, GmbH, Munich, Germany). The electrodes were mounted in a fitted cap with a standard 10–20 layout and were recorded with an average reference built into the amplifier (see www.neuroeconlab.com for the exact electrode configuration). Vertical and horizontal electrooculograms were recorded from electrodes placed above and below the right eye and on the outer canthi of the left and right eyes. Electrode impedances were kept below 20 kΩ at all times. The EEG data were sampled at 1000 Hz, amplified (Quick Amp, Brainproducts, GmbH, Munich, Germany), and filtered through a passband of 0.017–67.5 Hz (90 dB octave roll off).

Data analysis

For each response (balloon pump), a response time defined as the elapsed time since the previous response was recorded. Balloon pumps with a response time less than 100 ms or greater than 2000 ms were excluded from subsequent analysis. Next, we classified each balloon pump as corresponding either to a decision to explore or a decision to exploit. Based on Wershbale and Pleskac (2010), we classified response times more than three standard deviations above the mean as decisions to explore. Thus, the increase in balloon size for a successful pump prior to a long response time was marked as the time point at which participants began ''exploring'' or, in other words, considering their options. All other balloon pumps were classified as ''exploitations'', with the preceding increase in balloon size marked as the time point following which a decision was made to exploit. Thus, we were able to relabel the EEG data following data collection, and then use these revised labels to epoch the EEG data into segments containing decisions to explore or exploit.

The preprocessing of the EEG data began with the application of a 0.1–20 Hz phase shift-free Butterworth filter, following which the continuous EEG data were re-referenced to the average of the two mastoid channels. As mentioned previously, our ERP hypotheses concerned two events: decisions to explore or exploit, and balloon bursts. To test whether the amplitude of the ERP in the P300 time range was sensitive to the decision to explore or exploit, 800 ms epochs of data (from 200 ms before the increase in balloon size to 600 ms after the increase in balloon size) were extracted from the continuous EEG for each trial, channel, and participant, for each condition (explore/exploit). Following isolation of the epoched data, ocular artifacts were corrected using the algorithm described by Gratton et al. (1983). Subsequent to this, all trials were baseline corrected using a 200-ms epoch prior to stimulus onset. Finally, trials in which the change in voltage in any channel exceeded 10 μV per sampling point, or the change in voltage across the epoch was greater than 100 μV, were discarded. In total, less than 2% of the data were discarded.

Our preprocessing resulted in far more exploitation than exploration segments; as such, only exploitation segments that immediately preceded exploration segments were used in the subsequent ERP analysis. Specifically, our average ERP waveforms only included the 100 epochs corresponding to the 100 longest exploration periods and the 100 epochs (i.e. exploitation periods) immediately preceding them. Subsequent to the creation of the average ERP waveforms for each participants and condition (explore/exploit) we created difference waveforms for each participant and channel by subtracting the average exploitation waveforms from the average exploration waveforms. A visual examination of the grand average difference waveforms and a review of recent research (Polich, 2007; Duncan et al., 2009; Nieuwenhuis et al., 2010) led to a decision to quantify the magnitude of the ERP in the P300 time range as the maximum positive deflection of the difference waveform 300–450 ms following the increase in balloon size at the centro-parietal channel where the difference was maximal (channel CP2). The resulting ERP amplitudes were then statistically tested against zero using a single-sample t-test, with an assumed alpha level of .05.

To evaluate whether the amplitude of the P300 was sensitive to accumulated reward magnitude, 800 ms epochs of data (from 200 ms before balloon burst/growth onset to 600 ms after burst/growth onset) were extracted from the continuous EEG for each trial, channel, and participant for early and late balloon bursts (i.e. losses) and for the increase in balloon size immediately preceding the balloon bursts (i.e. potential gains). Early balloon bursts/growths were defined as bursts that were preceded by between 1 and 15 successful pumps. Late bursts/growths were preceded by between 16 and 30 successful

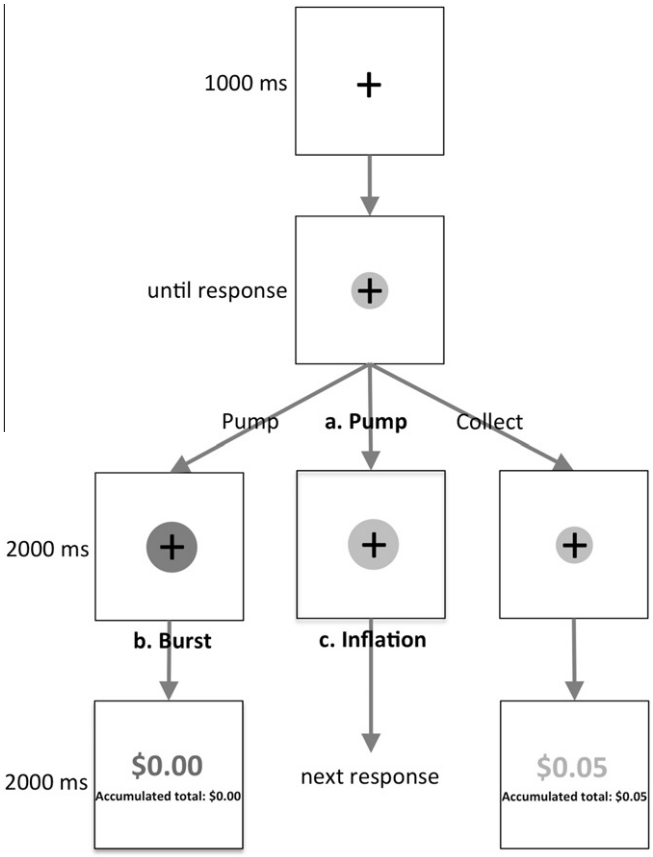

Fig. 1. Experimental design, along with timing details. Participants could respond by either pumping the balloon, or collecting the accumulated reward. Pumps could result in a successful inflation, or a balloon burst, in which case the accumulated reward for that balloon was lost. Relevant EEG data were recorded at (a) decisions to pump that were followed by a balloon inflation, (b) balloon bursts, and (c) balloon inflations.

pumps. We then preprocessed the EEG data in an identical manner as outlined above. Following preprocessing, ERPs were created by averaging the EEG data by condition for each electrode channel and participant separately for early and late gains and losses.

To quantify the P300 evoked by balloon bursts, we created a difference waveform for each participant and channel by subtracting the gain (growth) waveforms from the subsequent loss (burst) waveforms for both early and late balloons (see above). As before, the P300 was defined as the maximum positive deflection in the difference waveforms 300–450 ms following stimulus onset for each balloon burst (early/late) at electrode site Cz, where the difference was maximal. P300 amplitudes were then statistically tested against zero using a single-sample *t*-test, with an assumed alpha level of .05.

RESULTS

Time between pumps

A visual examination of the behavioral data revealed a subset of trials with longer response times—presumably, trials in which participants deliberated whether to take their accumulated money or continue playing (i.e. exploration). Long decision times (long inter-pump times) were defined as those more than three standard deviations above the mean. See Fig. 2 for a set of sample responses. Explore decision points were defined as increases in balloon size preceding long inter-pump times. All other increases in balloon size were classified as exploitations—trials in which the response time was short, suggesting an exploitative mode of control. This criterion created two separate distributions of decision times, each with a different mean ($p < .01$): shorter decision times for exploitative decisions (404 ± 31 ms), and longer decision times for exploratory decisions (798 ± 50 ms), consistent with Wershbale and Pleskac (2010). Also consistent with Wershbale and Pleskac (2010), participants explored less (3 ± 0.4% of trials for balloons numbered 51–300 compared to 15 ± 3% for balloons 1–50) as they became more familiar with the task (Fig. 3).

Exploration

Recall that we predicted exploration would lead to a larger ERP response in the P300 time range preceding longer response times, as we believed that this reflected deliberation of the decision to explore or exploit. Indeed, our analysis of the ERP waveforms in the P300 time range supported our hypothesis as we found a difference between explorative and exploitative trials that was maximal at electrode CP2. Specifically, we found a larger (more positive) ERP response in the P300 time range for exploration trials (1.79 ± 0.40 μV) relative to exploitation trials (0.47 ± 0.39 μV),

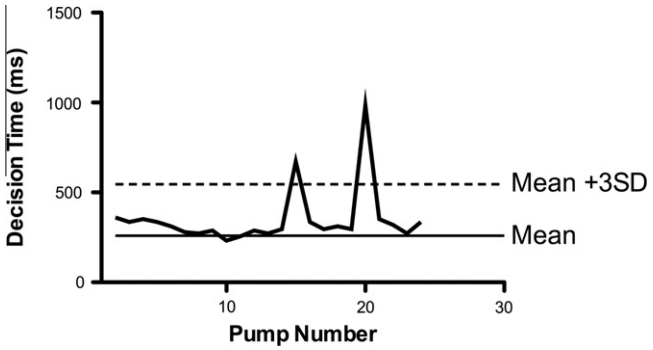

Fig. 2. Time between pumps for subject 14, balloon 10. Mean response time was characterized by short, somewhat automatic pumps (exploitations). Response times more than 3 standard deviations above the mean were classified as explorations.

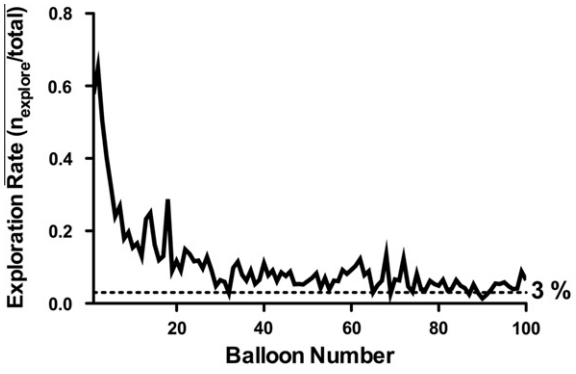

Fig. 3. Mean exploration rate. The mean number of explorations per balloon decreased over time. Only the first 100 out of 300 balloons are shown to emphasize the change in exploration rate over the first few balloons. A horizontal line is shown at 3%, the mean exploration rate for balloons 51–300.

$t(13) = 5.202$, $p < .01$ (see Fig. 4).[1] We then localized the source of the voltage difference between exploration and exploitation trials using standardized low-resolution brain electromagnetic tomography (sLORETA; Pascual-Marqui, 2002). An sLORETA analysis at 400 ms post decision (when the ERP response in the P300 time range was maximal) indicated maximal current sources in Brodmann Areas 6 and 10 within the superior frontal gyrus (Fig. 5). Finally, ERP amplitude in the P300 time range for both exploration and exploitation trials correlated positively with decision time, $r(28) = .51$, $p = .01$ (see Fig. 6).

Balloon bursts

We also wanted to see if the P300 following balloon bursts was sensitive to accumulated reward magnitude, since balloon bursts later in a trial sequence reflected a loss of more money as more money had accumulated. On average, there was an equal number of early bursts (53.0 ± 2.2) compared to late bursts (54.1 ± 3.5), $p = .8$. In line with our prediction, we found that the amplitude of the P300 scaled to reward magnitude: late high-valued pumps (defined as pumps 16–30: 35.53 ± 1.69 μV) versus early low-valued pumps (defined as pumps 1–15: 28.24 ± 2.15 μV), $t(13) = 5.00$, $p < .001$ (see Fig. 7). When all possible loss values were considered, there was a correlation between P300 peak and loss value, $r(390) = .33$, $p < .001$ (Fig. 8).

[1] We also statistically tested whether the N1 was sensitive to decisions to explore/exploit. No difference was seen between decisions to explore/exploit in the N1 time range (130–190 ms post stimulus: $t(13) = .46$, $p = .67$).

DISCUSSION

In the present study, decisions to explore in a sequential risk-taking task elicited a larger ERP response in the time range of the P300—a component sensitive to cognitive processing (Donchin, 1981; Donchin and Coles, 1988) and linked to phasic activity of the LC–NE system (Nieuwenhuis et al., 2005). Supporting our ERP result, our behavioral data mirrored previous work (Wershbale and Pleskac, 2010). We observed that response times in a sequential risk-taking task followed one of two distributions: longer response times indicative of exploration and shorter response times indicative of exploitation. Furthermore, we found that participants explored less over time as they became familiar with the probabilistic structure of the task, a result consistent with observations by Wershbale and Pleskac (2010) and reinforcement-learning theory in general (Sutton and Barto, 1998).

Computational framework

Like earlier work on exploration in humans (Daw et al., 2006; Cavanagh et al., 2011), we relied on a theoretical model (Wallsten et al., 2005; Pleskac, 2008; Wershbale and Pleskac, 2010) to identify participants' decisions to explore or exploit during task performance. Recall, decisions preceding fast responses were classified as exploitatory, while decisions preceding long responses were classified as exploratory. The validity of this criterion is critical when interpreting our findings because, while our difference wave in the P300 time range for explore/exploit decisions statistically differed from zero, it was computed by averaging over a post hoc selection of EEG segments derived from this classification system.

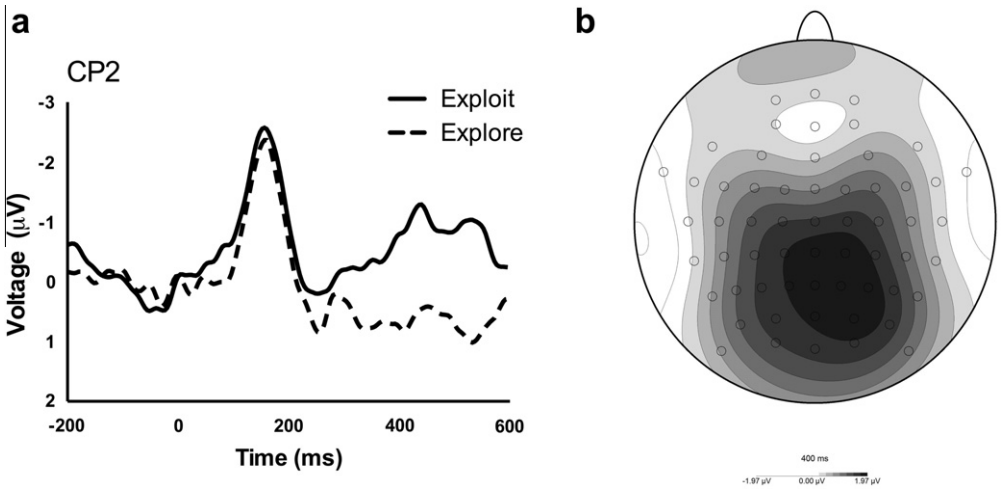

Fig. 4. Decision to explore or exploit. Note that 0 ms corresponds to the onset of the decision (balloon pump). Negative voltages are plotted up by convention. (a) Averaged ERP waveforms recorded at channel CP2 for exploration and exploitation decisions. (b) ERP topography map for the difference waveform (explore minus exploit) at 400 ms post decision.

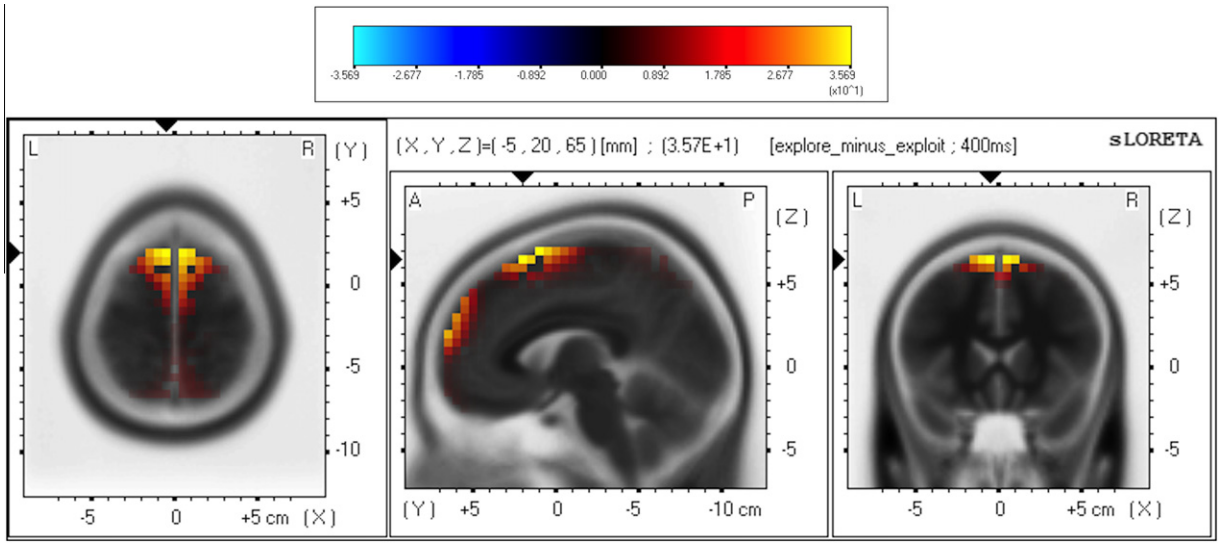

Fig. 5. sLORETA source analysis of exploration trials compared to exploitation trials at 400 ms post decision. Statistical nonparametric mapping (SnPM) at a significance level of .05 revealed differences localized in Brodmann Areas 6 (sLORETA value = 35.7) and 10 (sLORETA value = 31.8) within the superior frontal gyrus.

Previous research justifies our approach. In a seminal study, Wallsten et al. (2005) evaluated several models of BART performance by comparing their simulated outputs to human behavioral data. Wallsten et al. (2005) found some variation in exploratory behavior among individual human participants, with some participants continuing to gamble after the optimal number of pumps. To account for this, Wallsten and colleagues' model included components that decided how many pumps to make and whether to stop or keep going prior to each individual pump. In a later improvement of the Wallsten et al. (2005) model called the Bayesian sequential risk-taking model (BSR), Pleskac (2008) included an individual response bias that changed over time (see Busemeyer and Pleskac, 2009, for a review of the different components of dynamic decision-making models). Wershbale and Pleskac (2010) later amended the BSR to account for observed delays in response times so that assessments (decisions to either continue or stop) only occurred on a subset of trials. The trials associated with exploratory behavior were preceded by longer response times—explained as an increase in cognitive load linked to the decision process. Notably, and in line with human data, the model predicted that participants would tend to make fewer assessments over time, a prediction consistent with both exploratory behavior and the pattern of results we observed in our data. The most recent version of the BSR (Wershbale and Pleskac, 2010) provided a good fit for human BART data, including between-subject variation in response selection, and within-subject variation in response-time. In the present experiment, our participants' response

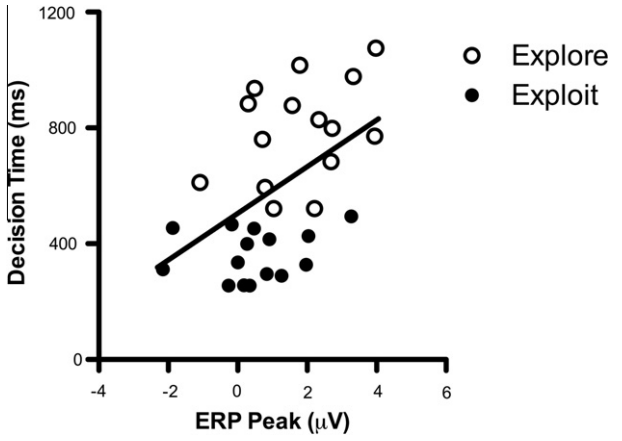

Fig. 6. Correlation between decision time (time between pumps) and magnitude of the peak of the ERP in the P300 time range, $r(28) = .51$, $p = .01$.

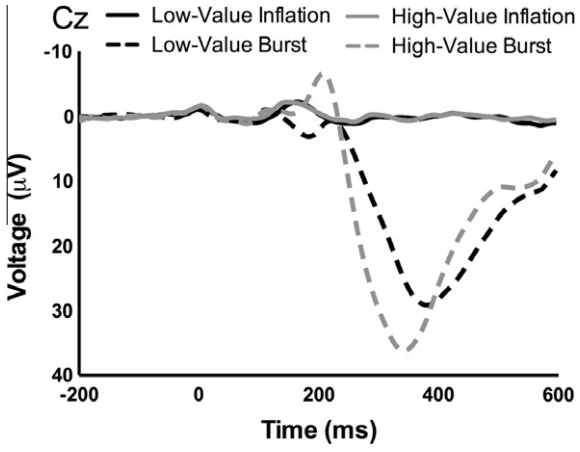

Fig. 7. Averaged ERP waveforms recorded at channel Cz for low- and high-value bursts and inflations. Note that 0 ms corresponds either to the onset of the balloon burst or the onset of the balloon inflation. Negative voltages are plotted up by convention.

time distributions mirrored Wershbale and Pleskac's (2010), thus providing strong support for the use of a response-time criterion to classify participant EEG segments as either containing decisions to explore or exploit.

The P300 and exploratory behavior

Our result that ERP amplitude in the P300 time range was larger for decisions to explore is consistent with the context-updating hypothesis of the P300 (Donchin, 1981; Donchin and Coles, 1988). Under this theoretical framework, a P300 is observed whenever new information requires an update to one's internal mental model of the world—specifically, the probabilistic framework of a particular task (Donchin and Coles, 1988). In our case, to maximize utility, participants had to learn the optimal number of pumps to undertake, a challenging task taking into account the value of a given pump and the risk associated with different balloon

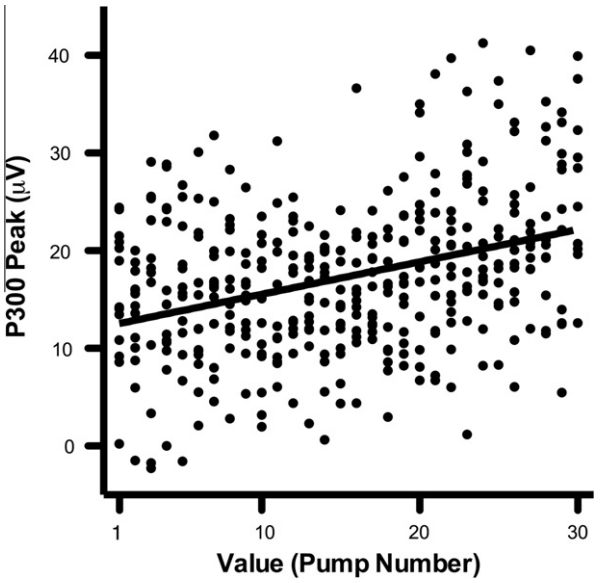

Fig. 8. Correlation between P300 peaks in response to balloon bursts and total value of the burst balloon, or total loss, $r(390) = .33$, $p < .001$. Large losses (i.e. balloon bursts preceded by many pumps) resulted in an enhanced P300.

sizes (i.e. that larger balloons entailed greater risk). Each pump, whether it resulted in a balloon burst or successful balloon inflation, thus provided information for participants. This notion is corroborated by earlier modeling work (i.e. Wershbale and Pleskac, 2010) suggesting that participants consider new information and review their potential actions at various points throughout a sequential decision-making task—points marked by longer-than-normal response times. It is at these assessment points, we claim, that participants incorporate new information into their model of the BART and then decide whether or not to continue pumping. As such, at assessment points a larger ERP in the P300 time range is observed, reflecting the incorporation of new information into the internal model and a subsequent exploratory decision. Interestingly, the length of the assessment period correlated with the amplitude of the subsequent ERP in the P300 time range (Fig. 6)—a result that further supports our hypothesis that the ERP in the P300 time range is sensitive to decisions to explore or exploit. Finally, an sLORETA source analysis (Pascual-Marqui, 2002) revealed a difference in frontal brain regions for exploration trials compared to exploitation trials, consistent with earlier research (Daw et al., 2006; Frank et al., 2009; Cavanagh et al., 2011).

An unavoidable limitation in this study arose because participants were asked to respond as quickly as they wanted to. As such, the mean response time corresponding to decisions to exploit (404 ± 31 ms) suggests that some of the EEG segments containing decisions to exploit might have overlapped with the following decision. However, that participant responses were self-paced seems an important part of the BART design, especially if a clear distinction between explorations and exploitations is to be achieved (Lejuez

et al., 2002; Wershbale and Pleskac, 2010). Although there are versions of the BART that introduce timing delays (Rao et al., 2008; Fukunaga et al., 2012), those versions do not, to our knowledge, produce the two distributions of response times necessary to classify responses as explorations or exploitations (Wershbale and Pleskac, 2010).

An alternative explanation for our findings relates to Nieuwenhuis and colleagues' (2005) hypothesis that the P300 time range is modulated by phasic activity of the LC–NE system. Interestingly, research by Usher et al. (1999) suggests that modulatory activity of the LC is responsible for regulating exploratory behavior in monkeys. Extending from this, Nieuwenhuis et al. (2005) proposed that the LC may regulate exploratory behavior in humans through the release of NE, with the change to an exploratory mode of control marked by a related increase in ERP magnitude in the P300 time range. Supporting this contention, Aston-Jones and Cohen (2005) suggested that LC phasic activity is driven by computations about value in the orbitofrontal cortex (OFC) and anterior cingulate cortex (ACC). They further suggested that the purpose of LC phasic release of NE is to break out of one behavioral routine (e.g. exploitation) to engage in a different behavior (e.g. exploration). Importantly, our data support Aston-Jones and Cohen's (2005) suggestion and the hypothesized link between the LC and the P300 (i.e. Nieuwenhuis et al., 2005) as we observed an increase in the amplitude of the ERP in the P300 time range when participants changed to an exploratory mode of control.

A second alterative explanation for our results relates to response time. Recently, Grinband et al. (2011) suggested that time on task, rather than an increase in cognitive control, might be responsible for increased frontal cortex activity. Grinband et al. (2011) asked participants to balance speed and accuracy in a Stroop task and observed that response times were slower and frontal cortex activity greater on incongruent trials compared to congruent trials. However, when slow and fast congruent trials were compared, Grinband et al. (2011) noted increased frontal activity for slower trials, even though congruency was controlled for. This somewhat controversial finding (e.g., Yeung et al., 2011) is relevant to the current study since we used response times to categorize decisions as explorations or exploitations. We observed an enhanced P300 for longer response times (classified as explorations). This is consistent with Grinband and colleagues' (2011) result, provided one is willing to extend a conflict-monitoring result to the exploration/exploitation dilemma (see Ishii et al., 2002; Khamassi et al., 2011, for some arguments supporting this comparison).

Although the body of research on the EEG correlates of the exploration/exploitation dilemma is sparse, it is growing. For example, Tzovara et al. (2012) recently used EEG to study the Daw et al. (2006) gambling paradigm and observed increased frontal brain activity prior to exploratory decisions, which they were able to define based on a computational model. Like us, Tzovara et al. (2012) compared ERPs to feedback prior to participant decisions to explore or exploit, and observed a difference. However, because Tzovara et al. (2012) only examined responses to feedback it is unclear whether their observed difference was due to the result of a decision to explore, reward evaluation, or both. Interestingly, Tzovara et al. (2012) observed that feedback ERP differences (including P300) predicted whether or not participants explored on subsequent trials. This lends further support to our second hypothesis that the P300 scales with reward magnitude, and our speculation that changing representations of value (as indexed by the P300) drive exploration (Sutton and Barto, 1998). A major strength of the present study, and one that distinguishes it from earlier work on the explore/exploit dilemma, is that we were able to examine ERP responses to the explore/exploit decisions themselves, as opposed to responses to feedback alone.

The P300 and reward magnitude

We also observed that the amplitude of the P300 was sensitive to reward magnitude. Specifically, we found a larger P300 amplitude for high-valued losses (balloon bursts) compared to low-valued losses—a result reflective of a neural representation of the magnitude of the value of taking different actions. In this case, the aforementioned representation related to the negative value associated with losses following early low-valued pumps versus later high-valued pumps. This finding is consistent with earlier work showing that the amplitude of the P300 is (a) sensitive to the magnitude of both wins and losses (Yeung and Sanfey, 2004) and (b) could be related to the motivational significance of feedback (Nieuwenhuis et al., 2005; Nieuwenhuis, 2011). Simply put, high-valued rewards and losses are more motivationally significant than low-valued rewards and losses. Of particular relevance here, Yeung and Sanfey (2004) speculated that the P300 might be impacted by the magnitude of actual and alternate outcomes (what might have been)—in other words, they speculated that the P300 reflects an objective representation of reward magnitude, regardless of whether or not the reward was actually received. In the present study, losses represented alternate outcomes: what participants might have won had they collected their money instead of gambling. Thus, our result that the P300 amplitude scaled with what might have been won supports the idea that the P300 reflects an objective representation of reward.

CONCLUSIONS

Research on the neural basis of exploration in humans has thus far lacked specific neural markers for this behavior. Here, we found that decisions to explore or exploit modulated ERP amplitude in the P300 time range in a sequential risk-taking task. Interestingly, this result is in line with a theoretical account that relates ERP amplitudes in the P300 time range to changes in phasic LC–NE activity—changes which are yoked to increased exploratory behavior (Aston-Jones and Cohen, 2005; Nieuwenhuis et al., 2005). As such, our

results (a) suggest that the amplitude of the ERP in the P300 time range is sensitive to decisions to explore or exploit and (b) relate modulation of the ERP in the P300 time range to an underlying neural system that is responsible for these changes: the LC–NE system. Of further interest, our results are in line with previous findings (e.g. Yeung and Sanfey, 2004) that demonstrate that the amplitude of the P300 scales to reward magnitude.

Acknowledgement—This research was supported by the Natural Sciences and Engineering Research Council of Canada.

REFERENCES

Aston-Jones G, Bloom FE (1981) Activity of norepinephrine-containing locus coeruleus neurons in behaving rats anticipates fluctuations in the sleep–waking cycle. J Neurosci 1(8):876–886.

Aston-Jones G, Cohen JD (2005) An integrative theory of locus coeruleus–norepinephrine function: adaptive gain and optimal performance. Annu Rev Neurosci 28(1):403–450. http://dx.doi.org/10.1146/annurev.neuro.28.061604.135709.

Bellebaum C, Daum I (2008) Learning-related changes in reward expectancy are reflected in the feedback-related negativity. Eur J Neurosci 27(7):1823–1835.

Berridge CW, Waterhouse BD (2003) The locus coeruleus–noradrenergic system: modulation of behavioral state and state-dependent cognitive processes. Brain Res Rev 42(1):33–84. http://dx.doi.org/10.1016/S0165-0173(03)00143-7.

Busemeyer JR, Pleskac TJ (2009) Theoretical tools for understanding and aiding dynamic decision making. J Math Psychol 53(3):126–138. http://dx.doi.org/10.1016/j.jmp. 2008.12.007.

Cavanagh JF, Figueroa CM, Cohen MX, Frank MJ (2011) Frontal theta reflects uncertainty and unexpectedness during exploration and exploitation. Cereb Cortex. http://dx.doi.org/10.1093/cercor/bhr332.

Cohen JD, McClure SM, Yu AJ (2007) Should I stay or should I go? How the human brain manages the trade-off between exploitation and exploration. Philos Trans R Soc Lond B Biol Sci 362(1481):933–942. http://dx.doi.org/10.1098/rstb.2007.2098.

Daw ND, O'Doherty JP, Dayan P, Seymour B, Dolan RJ (2006) Cortical substrates for exploratory decisions in humans. Nature 441(7095):876–879. http://dx.doi.org/10.1038/nature04766.

Dennett D (1986) The moral first aid manual. The Tanner Lecture on Human Values. University of Michigan.

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

Donchin E, Coles MGH (1988) Is the P300 component a manifestation of context updating? Behav Brain Sci 11(03):357–374. http://dx.doi.org/10.1017/S0140525X00058027.

Duncan CC, Barry RJ, Connolly JF, Fischer C, Michie PT, Näätänen R, Polich J, Reinvang I, Van Petten C (2009) Event-related potentials in clinical research: guidelines for eliciting, recording, and quantifying mismatch negativity, P300, and N400. Clin Neurophysiogy 120(11):1883–1908. http://dx.doi.org/10.1016/j.clinph.2009.07.045.

Duncan-Johnson CC, Donchin E (1977) On quantifying surprise: the variation of event-related potentials with subjective probability. Psychophysiology 14(5):456–467. http://dx.doi.org/10.1111/j.1469-8986.1977.tb01312.x.

Frank MJ, Doll BB, Oas-Terpstra J, Moreno F (2009) Prefrontal and striatal dopaminergic genes predict individual differences in exploration and exploitation. Nat Neurosci 12(8):1062–1068. http://dx.doi.org/10.1038/nn.2342.

Fukunaga R, Brown JW, Bogg T (2012) Decision making in the Balloon Analogue Risk Task (BART): anterior cingulate cortex signals loss aversion but not the infrequency of risky choices. Cogn Affect Behav Neurosci 12(3):479–490. http://dx.doi.org/10.3758/s13415-012-0102-1.

Gratton G, Coles MG, Donchin E (1983) A new method for off-line removal of ocular artifact. Electroencephalography Clinical Neurophysioly 55(4):468–484. http://dx.doi.org/10.1016/0013-4694(83)90135-9.

Grinband J, Savitskaya J, Wager TD, Teichert T, Ferrera VP, Hirsch J (2011) The dorsal medial frontal cortex is sensitive to time on task, not response conflict or error likelihood. Neuroimage 57(2):303–311. http://dx.doi.org/10.1016/j.neuroimage.2010.12.027.

Hajcak G, Holroyd CB, Moser JS, Simons RF (2005) Brain potentials associated with expected and unexpected good and bad outcomes. Psychophysiology 42(2):161–170. http://dx.doi.org/10.1111/j.1469-8986.2005.00278.x.

Ishii S, Yoshida W, Yoshimoto J (2002) Control of exploitation–exploration meta-parameter in reinforcement learning. Neural Netw 15(4–6):665–687.

Khamassi M, Wilson CRE, Rothé M, Quilodran R, Dominey PF, Procyk E (2011) Meta-learning, cognitive control, and physiological interactions between medial and lateral prefrontal cortex. In: Neural basis of motivational and cognitive control. MIT Press. p. 351–369.

Lejuez C, Read JP, Kahler CW, Richards JB, Ramsey SE, Stuart GL, Strong DR, Brown RA (2002) Evaluation of a behavioral measure of risk taking: the Balloon Analogue Risk Task (BART). J Exp Psychol Appl 8(2):75.

Mars RB, Sallet J, Rushworth M, Yeung N (2011) Neural basis of motivational and cognitive control. MIT Press.

McClure S, Gilzenrat M, Cohen J (2006) An exploration–exploitation model based on norepinephrine and dopamine activity. Adv Neural Inf Process Syst 18:867–874.

Mill JS (1863) Utilitarianism. In: Warnock M, editor. Utilitarianism and on liberty. Blackwell Publishing Ltd.. p. 181–235.

Nieuwenhuis S (2011) Learning, the P3, and the locus coeruleus–norepinephrine system. In: Neural Basis of motivational and cognitive control. The MIT Press. p. 209–222.

Nieuwenhuis S, Aston-Jones G, Cohen JD (2005) Decision making, the P3, and the locus coeruleus–norepinephrine system. Psychol Bull 131(4):510–532. http://dx.doi.org/10.1037/0033-2909.131.4.510.

Nieuwenhuis S, de Geus EJ, Aston-Jones G (2010) The anatomical and functional relationship between the P3 and autonomic components of the orienting response. Psychophysiology 48(2):162–175. http://dx.doi.org/10.1111/j.1469-8986.2010.01057.x.

Pascual-Marqui RD (2002) Standardized low-resolution brain electromagnetic tomography (sLORETA): technical details. Methods Find Exp Clin Pharmacol 24(Suppl. D):5–12.

Pineda JA, Foote SL, Neville HJ (1989) Effects of locus coeruleus lesions on auditory, long-latency, event-related potentials in monkey. J Neurosci 9(1):81–93.

Pleskac TJ (2008) Decision making and learning while taking sequential risks. J Exp Psychol Learn Mem Cogn 34(1):167–185.

Polich J (2007) Updating P300: an integrative theory of P3a and P3b. Clin Neurophysiol 118(10):2128–2148. http://dx.doi.org/10.1016/j.clinph.2007.04.019.

Rao H, Korczykowski M, Pluta J, Hoang A, Detre JA (2008) Neural correlates of voluntary and involuntary risk taking in the human brain: an fMRI study of the Balloon Analog Risk Task (BART). Neuroimage 42(2):902–910. http://dx.doi.org/10.1016/j.neuroimage.2008.05.046.

Sutton RS, Barto AG (1998) Reinforcement learning: an introduction. Cambridge, MA: MIT Press.

Sutton S, Braren M, Zubin J, John ER (1965) Evoked-potential correlates of stimulus uncertainty. Science 150(3700):1187–1188. http://dx.doi.org/10.1126/science.150.3700.1187.

Tzovara A, Murray MM, Bourdaud N, Chavarriaga R, del Millán JR, De Lucia M (2012) The timing of exploratory decision-making revealed by single-trial topographic EEG analyses. Neuroimage 60(4):1959–1969. http://dx.doi.org/10.1016/j.neuroimage.2012.01.136.

Usher M, Cohen JD, Servan-Schreiber D, Rajkowski J, Aston-Jones G (1999) The role of locus coeruleus in the regulation of cognitive performance. Science 283(5401):549–554. http://dx.doi.org/10.1126/science.283.5401.549.

Wallsten TS, Pleskac TJ, Lejuez C (2005) Modeling behavior in a clinically diagnostic sequential risk-taking task. Psychol Rev 112(4):862–880.

Wershbale A, Pleskac TJ (2010) Making assessments while taking sequential risks. Presented at the Cognitive Science Society, Portland, Oregon.

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

Yeung N, Cohen JD, Botvinick MM (2011) Errors of interpretation and modeling: a reply to Grinband et al. Neuroimage 57(2):316–319. http://dx.doi.org/10.1016/j.neuroimage.2011.04.029.

Yeung N, Sanfey AG (2004) Independent coding of reward magnitude and valence in the human brain. J Neurosci 24(28):6258–6264. http://dx.doi.org/10.1523/JNEUROSCI.4537-03.2004.

(Accepted 19 October 2012)
(Available online 26 October 2012)