Goal-directed EEG activity evoked by discriminative stimuli in reinforcement learning

David Luque, Joaquín Morís, Jacqueline A. Rushby, Mike E. Le Pelley

Luque, D., Morís, J., Rushby, J. A., & Le Pelley, M. E. (2015). Goal-directed EEG activity evoked by discriminative stimuli in reinforcement learning. *Psychophysiology, 52*, 238-248, doi: 10.1111/psyp.12302.

Goal-directed EEG activity evoked by discriminative stimuli in reinforcement learning

David Luque[1,2], Joaquín Morís[1,2], Jacqueline A. Rushby[3], Mike E. Le Pelley[3]

[1] Department of Basic Psychology, University of Málaga; Málaga, 29017; Spain

[2] Institute of Biomedical Research of Málaga (IBIMA); Málaga; Spain

[3] School of Psychology, University of New South Wales; Sydney, NSW 2052; Australia

Correspondence concerning this article should be addressed to David Luque, School of Psychology, University of Málaga. Campus de Teatinos s/n CP 29017, Málaga, Spain. Email: david.luque@gmail.com. Phone number: +34 952132630

Running Title: GOAL-DIRECTED EEG ACTIVITY

Abstract

In reinforcement learning, discriminative stimuli allow agents to anticipate the value of a future outcome, and the response that will produce that outcome. We examined this processing by recording EEG locked to discriminative stimuli during reinforcement learning in humans. Incentive value of outcomes and predictive value of stimuli were manipulated, allowing us to discriminate between outcome-related and response-related activity. Stimuli predicting the correct response differed from nonpredictive stimuli in the P2. Stimuli paired with high-value outcomes differed from those paired with low-value outcomes in a fronto-central positivity (that we argue is related to the stimulus-locked feedback correct-related positivity) and in the P3b. A slow negativity then distinguished between predictive and nonpredictive stimuli. These results suggest that, first, attention prioritizes detection of informative stimuli. Activation of mental representations of these informative stimuli then retrieves representations of outcomes, which in turn retrieve representations of responses that previously produced those outcomes.

Keywords: Reinforcement learning; goal-directed action; cognitive neuroscience; event-related potentials; feedback correct-related positivity; P2; P3; P3b.

Reinforcement learning describes how an agent adapts its responses so as to increase rewards and avoid punishments (Sutton & Barto, 1998). It is a critical component of human (and animal) cognition, as it allows agents to learn from their own actions. For example, we might experience that going to the refrigerator when hungry yields a reward of food, or that running down stairs results in a painful fall. Reinforcement learning will then mean that we are more likely to perform the former action, and less likely to perform the latter. Consistent with its central role in cognition, reinforcement learning is also implicated in certain aspects of mental disorders. For example, compulsive drug-seeking is thought to reflect a "hijacking" of the reinforcement learning system by the potent reward signals provided by drugs of abuse (Hyman, 2005). Similarly, influential accounts have implicated a dysfunction of reinforcement learning in schizophrenia, such that actions and thoughts that would normally be suppressed are instead reinforced (Frank, 2008; Waltz, Frank, Robinson, & Gold, 2007). Hence establishing the psychological and neural processes underlying reinforcement learning has the potential to shed light on these (and other) conditions.

The structure of reinforcement learning implicates three fundamental events. A discriminative stimulus, *S*, determines what type of outcome, *O* (punishment or reward), is produced by a response, *R*. For example, a green traffic light constitutes a discriminative stimulus, signaling that the response of moving into an intersection will produce a desired outcome (progress); a red traffic light signals that the same action will produce an undesirable outcome (a crash).

Theories of reinforcement learning draw a contrast between *goal-directed* (model-based) and *habitual* (model-free) behavior (de Wit & Dickinson, 2009). Goal-directed behavior is based on knowledge of the outcome produced by an action; an organism learns that in the presence of stimulus *S* it should respond *R* because it "knows" that

doing so produces outcome *O*. In contrast, habitual behavior is independent of the identity and current value of the outcome produced by an action. The experiment described in the present article used electroencephalography (EEG) to study the time-course of cognitive processes that support goal-directed control of behavior from the moment of onset of *S*. We studied these processes by measuring event-related potentials (ERPs) during presentation of *S* in a reinforcement learning task.

Processing during presentation of *S* is critical for goal-directed action, since it is assumed that the goal-directed system drives behavior by retrieving a representation of the expected *O*, and the *R* that will produce it, during the presentation of *S*. However, previous studies of the neural substrates of reinforcement learning have typically focused on the processing of *O* and *R* (e.g., Holroyd & Coles, 2002; McClure, York, & Montague, 2004).

In contrast, the current study used EEG to systematically investigate neural activity elicited by *S* in a reinforcement learning task. Notably, our experimental procedure allowed us to dissociate electrophysiological correlates of (i) the value of the *O* with which *S* was associated (the incentive value of *S*), and (ii) the extent to which *S* was associated with a particular *R* (the predictive value of *S*). This distinction is crucial, because retrieval of outcome and response information are the key information-processing steps involved in reinforcement learning. By experimentally dissociating the retrieval of response-related information from the retrieval of outcome-related information, we could use the high temporal resolution offered by EEG to distinguish and track the corresponding neural processes elicited by *S*. The stimulus-locked ERP components that we analyzed in this regard were selected on the basis of previous functional evidence marking them out as potential candidates for indices of the information-processing steps involved in goal-directed behavior; specifically, these

were (1) a fronto-central positivity from 250-350 ms, here labeled a stimulus-related positivity (SRP); and (2) the P3 (e.g., Holroyd, Krigolson, & Lee, 2011; Schevernels, Krebs, Santens, Woldorff, & Boehler, 2014; see Discussion for more details of prior studies of these components). Establishing the neural indices of these processes, and the temporal relations between them, provides a valuable test-bed for computational accounts of reinforcement learning.

We were also interested in the extent to which attention to *S* was influenced by learning about its relationship with *R* and *O*. Influential formal theories of associative learning anticipate that attention to a stimulus will vary as a function of learning about the consequences of that stimulus (e.g., Mackintosh, 1975). These theories find empirical support in a collection of studies that have produced results consistent with a bias in visual attention towards stimuli that predict important events in the environment, relative to those that do not (see Le Pelley, 2010, for a review). For example, people spend longer looking at predictive stimuli than nonpredictive stimuli (Le Pelley, Beesley, & Griffiths, 2011; Wills, Lavric, Croft, Hodgson, 2007); people are faster to respond to events occurring in the location of predictive stimuli (Le Pelley, Vadillo, & Luque, 2013); and people are faster to learn new information about predictive stimuli (Le Pelley & McLaren, 2003). Notably, using EEG, Wills et al. (2007; see also Wills, Lavric, Hemmings & Surrey, 2014) found that the magnitude of the anterior N1 and selection negativity (SN) ERP component— which has previously been characterized as an index of visual attentional processes (Hillyard & Anllo-Vento, 1998)—differed between predictive and nonpredictive stimuli. However, in all of these prior studies the relationship between *S* and *R* was confounded with the relationship between *S* and *O*. That is, 'predictive' stimuli predicted both the correct response to make, and the outcome that would be produced by that response, while 'nonpredictive' stimuli

predicted neither. As such these findings do not allow us to determine the extent to which attention to *S* is determined by response-related information (predictive value), or outcome-related information (incentive value), or both. Similarly, recent studies of the specific influence of learning about incentive value on visual attention to *S* (e.g., Anderson, Laurent, &Yantis, 2011) did not manipulate predictive value, and hence were unable to compare the influence of these two factors.

In contrast, by dissociating retrieval of response-related and outcome-related information by discriminative stimuli in a reinforcement learning task, the design of the current experiment allowed us to examine the influence of each of these properties on ERP indices of attention to these stimuli. Specifically, we examined the influence of these factors on two ERP markers of visual attention; the SN as studied by Wills et al. (2007), and the P2 (Carretié, Mercado, Tapia, & Hinojosa, 2001).

Materials & Methods

Fifteen physically and psychiatrically healthy volunteers (aged 18-35; 7 males) participated in the study. All volunteers provided informed consent according to procedures approved by the University of New South Wales Human Research Ethics Committee, and received $30 AUD for participating. They were also told that they would receive a performance-related bonus at the end of the experiment ($1 AUD per 1000 points earned; mean bonus was $14.50 AUD).

Participants completed a reinforcement learning task while EEG was recorded. Table 1 shows the task design; Figure 1A shows the structure of each trial. Our task involved three types of stimulus: 1) Non-predictive Stimuli (NPS) provided no information regarding the correct response, or the likely outcome (so these stimuli had low predictive value). 2) Predictive Stimuli-Low Value Outcome (PS-L) reliably indicated the correct response, and that the outcome would have low value (so these

stimuli had high predictive value, and low incentive value). 3) Predictive Stimuli-High Value Outcome (PS-H) reliably indicated the correct response, and that the outcome would have high value (so these stimuli had high predictive value, and high incentive value).

Two stimuli (abstract red-and-yellow shapes, taken from Wills et al. 2007) were presented sequentially on each trial; one was predictive and the other non-predictive. Whether the predictive stimulus was presented before or after the non-predictive stimulus was counterbalanced across trials. After stimulus presentation, an asterisk and a circle appeared on either side of the screen, with left/right presentation order pseudo-randomly determined on each trial. Participants had previously been informed that selecting one of these response symbols would win 100 points if it was the 'correct' response, and lose 100 points if it was 'incorrect'; selecting the other symbol gained 1 point for a correct response and lost 1 point for an incorrect response. For half of participants, the asterisk was the response symbol giving a gain/loss of 100 points, and the circle gave a gain/loss of 1 point; for remaining participants this was reversed.

Participants responded by selecting either the left- or right-hand response symbol using the 'q' or 'p' key of a standard keyboard. Subsequent feedback revealed how many points they had gained or lost on the trial; a gain of points was accompanied by the word "Correct", a loss by "Incorrect". The final screen of each trial displayed a running total of points earned.

Each of the four stimulus pairs shown in Table 1 was presented 20 times in each of 8 training blocks, giving 640 trials in total.

EEG Acquisition

Scalp EEG was recorded with tin electrodes mounted in an electrocap at 29 standard positions (electrode positions: FP1/2, F3/4, FC3/4, C3/4, CP3/4, P3/4, O1/2,

F7/8, T7/8, TP7/8, P5/6, FPz, Fz, FCz, Cz, CPz, Pz, Oz). Data were referenced to the outer canthus of the right eye (online) and then to the average mastoid recordings (offline). Vertical eye movements were monitored by an electrode placed on the infraorbital ridge of the left eye. All electrode impedances were below 5 kΩ. Electrophysiological signals were digitized at a sampling rate of 1000 Hz. Electrooculogram artifacts were corrected using the SOBI algorithm (Belouchrani, Abed-Meraim, Cardoso, & Moulines, 1997). Bad channels were interpolated using EEGlab's (Delorme & Makeig, 2004) invdist-interpolation method (mean number of bad channels per participant used in the statistical analysis = 0.03; mean number of bad channels per participant from all recorded channels = 0.6). For each participant, raw data were band-pass filtered using a Butterworth filter (roll-off 12 db/oct) in a band from 0.1-40 Hz. EEG epochs of 800ms, locked to the cue onset, were extracted. Then, the epochs of each trial type were averaged. Data were baseline-corrected by subtracting average activity during the 100ms preceding stimulus onset. Remaining trials with base-to-peak electrooculogram amplitude greater than 100 μV were excluded from analyses.

EEG Analysis

Our aim was to study the cognitive processes underlying 'mature' goal-directed behavior, i.e., when the corresponding *S*-*R*-*O* associations had been well-learned by participants. We therefore restricted ERP analyses to data from late in training, when participants' high level of performance allowed us to be confident that these instrumental relationships had been well-learned. In selecting the specific data to be analyzed we took account of the following considerations: (1) we needed trials in which performance was asymptotic in all experimental conditions; and (2) given that we did not know the signal-to-noise ratio for our experimental setting, we needed at least 50 trials per experimental condition (following the parametric study of the FRN by

Marco-Pallares, Cucurell, Münte, Strien, & Rodriguez-Fornells, 2011). On this basis, ERP analysis made use of data from blocks 7 and 8 (Figure 1B shows accuracy across all training blocks). These data comprise 80 trials featuring each of PS-H, PS-L, NPS_1, and NPS_2. ERP data were averaged for the two non-predictive stimuli, NPS_1 and NPS_2, since these stimuli were equivalent in the design, to yield an NPS condition based on 160 trials.

The selection of time-windows for ERP analysis followed two steps. First we identified the components of interest from previous literature analyzing the ERPs elicited by predictive visual stimuli presented in a non-lateralized fashion: N1, selection negativity (SN), P2, SRP and P3 (Baker & Holroyd, 2009; Holroyd et al. 2011; Dunning & Hajcak, 2007; Liao, Gramann, Feng, Deak, & Li, 2011; Schevernels et al., 2014; Wills et al., 2007, 2014). Then, we identified these components from inspection of ERP waveforms and topographical maps. Interestingly, this inspection also revealed a slow negative-going wave occurring in the final section of the *S* presentation (450-800 ms) that evolved differently across conditions. Considering some recent reports relating goal-directed action and slow wave activity (Fuentemilla et al., 2013; Morís, Luque, & Rodríguez-Fornells, 2013; Schevernels et al., 2014), we included this late slow wave in our analyses. Preliminary analyses revealed that the very earliest components (N1 and SN) were virtually identical for all types of stimuli (see Figure 1C); for the sake of brevity, we have therefore omitted further details of these analyses. The analyses that we report below are thus restricted to four time-windows, defined relative to onset of the 800ms stimulus: (1) From 150-190ms, for the P2 component; (2) From 250-350ms, for the SRP component; (3) From 350-450ms for the P3 component; (4) From 450-800ms (stimulus offset), to study the slow late negativity (cf. Brunia et al., 2011).

Stimulus-locked ERPs were measured as the mean voltage amplitude within each

time-window and electrode, with the exception of the 450-800ms window. In this latter window, the slow negativity was measured by a peak-to-peak calculation in order to avoid, as far as possible, overlapping with the P3b component. For this slow negativity, we first selected the time point in the 450-550ms time-window with the most positive voltage, and then calculated the difference between this value and the most negative voltage from the 550-800ms interval.

For each time-window, we selected for analysis the electrode in which the absolute value of the difference evoked by PS-H and NPS was maximal (see topographic maps in Figure 1C). These stimuli differ in both incentive value and predictive value and hence should produce a difference in any component relating to either of these properties.

As the P2, SRP and P3 components may be evoked very close together in time, the activities relating to each component could be to some extent superimposed. Thus, in order to better disentangle the contribution made by each specific component we used a temporal principal component analysis (PCA). This temporal PCA was performed on the average waveform of each participant for each condition at each electrode. This analysis used Promax rotation, with a kappa value of 3 (Dien, 2012).

Results

Behavioral data

Figure 1B shows the proportion of correct responses across the course of training for each of the stimulus pairs encountered by participants (see Table 1). Accuracy increased for all stimulus pairs at a similar rate as training proceeded. ANOVA with factors of stimulus pair (PS-H & NPS_1; PS-H & NPS_2; PS-L & NPS_1; PS-L & NPS_2) and training block (1 to 8) showed only a main effect of training block [$F(2.62, 36.72) =$

44.4; $p < 0.001$; $\eta_p^2 = 0.76$]; here and elsewhere, the Greenhouse-Geisser correction for degrees of freedom was performed when sphericity was violated in repeated measures ANOVA. This main effect was best fitted by a linear trend [$F(1,14) = 76.5$; $p < 0.001$; $\eta_p^2 = 0.85$]. The main effect of stimulus pair was not significant [$F(1.87, 26.24) = 2.10$; $p = 0.145$; $\eta_p^2 = 0.13$], and neither was the stimulus pair × training block interaction [$F < 1$]. These results indicate that participants learned appropriately and equally about all of the instrumental *S*-*R*-*O* relationships, and that their responses adapted appropriately to these relationships.

Figure 1B (lower panel) shows participants' response times for each trial type across the course of training. ANOVA conducted as in the previous paragraph revealed a main effect of training block [$F(1.64, 22.92) = 10.9$; $p = 0.001$; $\eta_p^2 = 0.44$], reflecting a reduction in response times across training. This main effect of training block was best fitted by a linear trend [$F(1,14) = 15$; $p = 0.02$; $\eta_p^2 = 0.52$]. The main effect of stimulus pair was not significant [$F(3, 42) = 1.36$; $p = 0.269$; $\eta_p^2 = 0.09$], and neither was the stimulus pair × training block interaction [$F(21, 294) = 1.23$; $p = 0.220$; $\eta_p^2 = 0.08$], suggesting that response times fell at a similar rate regardless of the type of trial. It should be emphasized that participants could respond only when the response symbols (the circle and asterisk) had appeared, and this occurred after the offset of both stimuli (see Figure 1A). Consequently, all responses were made *after* the stimulus-locked ERP components reported below had occurred.

EEG data

Figure 2 shows EEG waveforms across the scalp for all conditions in this study. Figure 1C shows topographical maps of differences between the conditions in each of the time-windows of interest, along with waveforms for the electrode that showed the largest difference in activity between PS-H and NPS stimuli in each time-window.

Below we report analyses of each time-window in turn.

150-190ms. The largest differences between PS-H and NPS stimuli in the P2 time-window were obtained in fronto-central electrodes, which is consistent with previous studies of the P2 (Evans & Federmeier, 2007). The maximum difference was found in electrode FCz. The main effect of stimulus type was significant, [$F(1.8, 24.8) = 4.75$; $p = .021$; $\eta_p^2 = 0.253$]. Figure 1C shows that the P2 component was greater for both types of predictive stimulus than for NPS [NPS vs. PS-H, $t(1,14) = 2.21$, $p = 0.022$, $d = 0.548$; NPS vs. PS-L, $t(1,14) = 2.2$, $p = 0.044$, $d = 0.464$]. There was no significant difference in the activity elicited by PS-H and PS-L [$t(1,14) = 0.61$, $p = 0.55$, $d = 0.088$].

250-350ms. The largest differences in the time-window of the SRP were obtained in fronto-central electrodes, which is consistent with prior research (Dunning & Hajcak, 2007). The maximum difference was found in electrode Cz. There was a main effect of stimulus type [$F(1.54, 21.5) = 4.52$, $p = 0.031$, $\eta_p^2 = 0.637$]. Pairwise comparisons indicated that PS-H elicited a larger SRP (more positive values) than PS-L [$t(1,14) = 4.57$, $p < 0.001$, $d = 0.806$] and NPS [$t(1,14) = 4.76$, $p < 0.001$, $d = 0.824$]. There was no significant difference in the activity evoked by PS-L and NPS [$t(1,14) = 0.56$, $p = 0.58$, $d = 0.062$] during this period.

350-450ms. The largest differences in the P3 time-window were obtained in parietal electrodes, with the maximum difference found in electrode Pz. This distribution is characteristic of the P3b component (Polich, 2007). There was a significant effect of stimulus type, [$F(1.82, 25.54) = 17.81$, $p < 0.001$, $\eta_p^2 = 0.974$], with significant differences between PS-H and both PS-L [$t(1,14) = 4.5$, $p < 0.001$, $d = 1.192$] and NPS [$t(1,14) = 5.58$, $p < 0.001$, $d = 1.262$], but not between PS-L and NPS

[$t(1,14) = 0.46$, $p = 0.655$, $d = 0.094$].

450-800ms. We also obtained a slow late negativity that began around 450ms from stimulus onset and lasted until stimulus offset. In this time-window the largest differences between PS-H and NPS stimuli were in parietal electrodes, with the maximum difference found in electrode P3 (we had no *a priori* distribution for this late slow component, since members of this family of evoked-potentials have quite different scalp distributions: see Brunia et al., 2011). The effect of stimulus type was significant, [F (1.83, 25.61) = 17.81, $p < 0.001$, $\eta_p^2 = 0.999$], with significant differences between all types of stimuli in the order PS-H > PS-L > NPS [PS- H vs. PS-L, $t(1,14) = 4.59$, $p < 0.001$, $d = 0.819$; PS-H vs. NPS, $t(1,14) = 7.27$, $p < 0.001$, $d = 1.161$; PS-L vs. NPS, $t(1,14) = 2.85$, $p = 0.013$, $d = 0.45$].

Principal Components Analysis

As mentioned earlier, the first three components sensitive to our manipulation (P2, SRP and P3b) occur close together in time. We used PCA to disentangle the specific contribution of each component to the overall EEG signal observed. This PCA analysis yielded three factors that showed similar distributions and time latencies to the ERP components described earlier (see Figure 3). Factor 1 had a parietal distribution, as for P3b; Factor 2 had a more central distribution and earlier latency, as for the SRP; and Factor 3 was frontal with a short latency, as for the P2. Below we report analyses of the differences between experimental conditions on the factor loadings of each of these three factors.

Factor 1: The largest differences for Factor 1 were found at electrode Pz, as for our previous analysis of the P3b component. For Factor 1, there was a significant effect of stimulus type [F (2, 28) = 17.54, $p < 0.001$, $\eta_p^2 = 0.556$], with significant differences

between all types of stimuli in the order PS-H > PS-L > NPS [PS- H vs. PS-L, $t(1,14) = 3.27$, $p = 0.006$, $d = 0.85$; PS-H vs. NPS, $t(1,14) = 6.32$, $p < 0.001$, $d = 1.65$; PS-L vs. NPS, $t(1,14) = 2.28$, $p = 0.028$, $d = 0.58$]. These results are similar to those for the time-window of the P3b described above (350-450ms). Note, however, that there was a slight difference in latency: the P3b described in the averaged EEG waveforms occurred in the 350-450 ms window, whereas Factor 1 from the PCA peaked at around 500 ms. It is also interesting that the comparison of PS-L and NPS was significant in the analysis of Factor 1, whereas this difference did not reach significance in the P3b taken from the averaged ERP analysis. These divergences could reflect some influence of the previous (SRP) and later (slow late negativity) components on the P3b component observed in the main average. In particular, previous studies investigating cognitive processing of emotional stimuli have also found that the PCA component relating to P3b can be slightly slower than the P3b observed in the averaged ERP waveform (Delplanque, Lavoie, Hot, Silvert, & Sequeira, 2004).

Factor 2: The largest differences for Factor 2 were found at electrode Cz, as for our previous analysis of the SRP component. For Factor 2, there was a significant effect of stimulus type [$F(2, 28) = 21.6$, $p < 0.001$, $\eta_p^2 = 0.607$], with significant differences between PS-H and both PS-L and NPS [PS- H vs. PS-L, $t(1,14) = 4.92$, $p < 0.001$, $d = 1.24$; PS-H vs. NPS, $t(1,14) = 5.58$, $p < 0.001$, $d = 1.44$], but not between PS-L and NPS [$t(1,14) = 0.1$, $p = 0.926$, $d = 0.03$]. These results mirror those for the time-window of the SRP described above (250-350ms).

Factor 3: The largest differences for Factor 3 were found at electrode FCz, as for our previous analysis of the P2 component. For Factor 3, there was a significant effect of stimulus type [$F(2, 28) = 4.74$, $p = 0.017$, $\eta_p^2 = 0.253$], with significant differences between NPS and both PS-H and PS-L [PS-H vs. NPS, $t(1,14) = 2.64$, $p = 0.019$, $d =$

0.68; PS-L vs. NPS, $t(1,14) = 2.16$, $p = 0.048$, $d = 0.56$], but not between PS-H and PS-L [$t(1,14) = 0.81$, $p = 0.434$, $d = 0.21$. These results mirror those for the time-window of the P2 described above (250-350ms).

In sum, PCA analyses provide strong support for the suggestion that the components that we have identified as P2, SRP and P3b are indeed these three components. In the case of P3b analysis, the PCA results differed slightly from those based on the averaged ERP waveform, which could imply that the analysis of P3b in the averaged ERP was influenced (to some extent) by overlapping activations. Notably, these differences do not compromise functional interpretations of this component, since these are similar regardless of whether we use the averaged ERP or the PCA factor (see Discussion below).

Discussion

Models of reinforcement learning propose that presentation of a discriminative stimulus can retrieve mental representations of an outcome previously associated with that stimulus, and the response that produces that outcome. Our study used ERPs to systematically examine neural activity elicited by discriminative stimuli that relates to retrieval of outcomes (incentive value) and responses (predictive value).

The first difference between conditions observed in the ERP analysis (150-190 ms after stimulus onset) related to the manipulation of predictive value. Predictive stimuli (PS-H and PS-L) elicited a larger P2 than did NPS. The P2 evoked potential is thought to be related to early visual processing and focused attention (e.g., Carretié et al. 2001). Thus, a plausible interpretation of this result is that it reflects a modulation of the attentional system as a function of the predictive value of each stimulus. This interpretation is supported by previous evidence showing that manipulations of

predictive value modulate focused visual attention (Le Pelley et al. 2011; Le Pelley et al. 2013; Wills et al., 2007).

Interestingly, Schevernels et al. (2014) showed an increased P2 magnitude when participants perceived cues that signaled the availability of reward on a particular trial. In their procedure, the reward was provided (or not) depending on the participant's performance in a subsequent visual discrimination task. Hence this task shared some similarities with an *S-R-O* reinforcement learning procedure. Notably, other cues in Schevernels et al.'s task predicted the difficulty of the visual discrimination, but this variable did not affect the P2 component. These authors therefore suggested that P2 reflected a *tuning* of the attention system to reward-related stimuli, and that P2 did not index any process related to motor preparation (in line with Anderson et al. 2011; Hughes, Mathan, & Yeung, 2012).

In our experiment P2 magnitude was modulated by the predictive value of *S*. Following the work of Schevernels et al., we interpret this finding as reflecting a modulation of attention in order to rapidly identify stimuli that predict a response—an index of an *attention for action* system (Treisman 1996). More specifically, our data suggest that tuning of this attention for action system is more strongly influenced by the extent to which a stimulus predicts a particular response than by the incentive value of the reward that is obtained by making this response.

That said, previous ERP studies have found an effect of incentive value in evoked potentials that have been implicated as markers of visual attention. For instance, Kiss, Driver, and Eimer (2009) used a visual search task in which one target stimulus signaled the availability of high reward, while a different target signaled low reward[2]. Target stimuli associated with high reward elicited a greater N2pc lateralized component than stimuli paired with low reward. As noted above, however, we found no effect of

incentive value on any ERP component related to visual attention. One plausible reason for this disparity is that the components detected in visual search experiments such as that of Kiss et al. are lateralized. Thus, it is possible that our learning task (with only one stimulus presented at a time in the center of the screen) is not sensitive to detecting modulations in lateralized components. This possibility could be assessed in future experiments by presenting the two visual stimuli simultaneously as lateralized pairs instead of sequentially.

It is also notable that we did not observe any influence of learning about stimulus properties on N1 or SN components. This contrasts with the findings of Wills et al (2007), who showed a difference in both of these components between predictive and nonpredictive stimuli (here 'predictive' stimuli had higher predictive *and* incentive value than 'nonpredictive' stimuli). Wills et al. used the same stimuli that we did, and presented these stimuli in the center of the screen. However, there remain several differences in procedure that could cause the different results. For instance, participants in Wills et al.'s experiment responded under time pressure during the presentation of the stimuli. Thus, the motor preparation and its execution might induce and require faster visual processing. In contrast, in our procedure participants responded during presentation of a separate screen after stimulus presentation. In any case, notwithstanding these differences, it is noteworthy that both our study and those of Wills et al. (2007, 2014) found modulation of attentional ERP components as a function of learning about the *S.*

The pattern of activity across conditions in the current study changed rapidly over the course of the 800ms stimulus presentation. We found that PS-H stimuli elicited greater activity than PS-L during time-windows from 250-350ms and 350-450ms after stimulus onset. Given that PS-H and PS-L differ only in the value of the predicted

reward, differences between them reflect brain activity relating to representation of outcome value. Consequently, these components provide an unambiguous index of goal-directed processes, since these processes are defined as being sensitive to reward value.

That PS-H stimuli should elicit greater neural activity than PS-L is not a foregone conclusion. While correct responses to PS-H yielded larger rewards than to PS-L, incorrect responses to PS-L produced greater losses than to PS-H[3]. Recall, however, that ERP data came from late in training (blocks 7 and 8), when participants' responses were well-adapted to the prevailing instrumental relationships such that correct responses were more common than incorrect responses (see Figure 1B). Hence the value of stimuli with regard to correct responses (favoring PS-H) dominated.

The greater goal-directed brain activity evoked by PS-H stimuli first appeared during the 250-350ms time-window, and was maximal in frontocentral regions. Up to this point, we have labeled this component a stimulus-related positivity (SRP). However, the latency and topography of this ERP is very similar to that of the feedback correct-related positivity (FCRP) that is normally evoked by positive feedback or rewards in reinforcement learning and gambling tasks (see Holroyd, et al., 2008). In other words, this component has traditionally been related to processing elicited by the outcome *O*, rather than the stimulus *S*, in reinforcement learning procedures. We believe it is likely that the SRP revealed by our study and the FCRP identified in prior research represent the same, or at least analogous, ERP components. If this is the case, then our investigation of brain activity elicited by *S* extends this prior research by suggesting that the FCRP can also index *anticipation of future reward*; that is, activation of an outcome representation by a stimulus (see also Holroyd et al., 2011, for supporting evidence from a Pavlovian learning procedure). This would further imply that the name of this

component in previous literature (feedback correct-related positivity) may be something of a misnomer, given that it need not be elicited by correct or rewarding feedback, but can also relate to the anticipation of that feedback.

The finding that the SRP/FCRP component is influenced by incentive value fits well with recent results relating FCRP to motivational and emotional processes (Riesel et al., 2012; Santesso et al., 2008). If FCRP reflects the emotional impact of a positive outcome, it seems plausible that this emotional impact might move from the time of the outcome to the time of a stimulus associated with that outcome, given enough training. Extensive prior research suggests that the anterior cingulate cortex (ACC) is responsible for producing the FCRP (see Walsh & Anderson, 2012). Moreover, anatomical and functional studies suggest that ACC constitutes a hub in which information about reinforcers becomes linked to motor centers responsible for expressing affect and executing goal-directed behavior (Shakman et al., 2011). The suggestion that the ACC is the neural generator of the fronto-central SRP component observed in our experiment is consistent with previous reports showing ACC involvement in reward-motivated behavior in macaque monkeys (e.g., Kennerley, Walton, Behrens, Buckley, & Rushworth, 2006). These findings have been used to argue that ACC plays a critical role in computing the anticipated reward value of alternative actions, particularly when action–outcome contingencies vary.

We should acknowledge an alternative possibility at this point, however. Rather than representing a stimulus-locked FCRP, it is possible that the SRP component was instead an example of a P3a. Like the FCRP, the P3a typically occurs earlier, and with a more frontal scalp distribution, than the P3b (Polich, 2007), and this pattern coincides with the timing and topographical distribution of the SRP component observed in the EEG waveform (and the PCA analysis). Classically, the P3a is observed when

infrequent stimuli are presented against a background of more frequent stimuli (e.g., Rushby, Barry, & Doherty, 2005), or when a novel stimulus is unexpectedly introduced in an oddball task (e.g., Courchesne, Hillyard, & Galambos, 1975). It has been argued that the P3a reflects the process of recruiting additional attentional resources when 'relevant' stimuli (for example, those that are new or infrequent) are perceived. This recruitment of attention then acts to facilitate later processing of these stimuli in working memory (with this memory-related processing producing the P3b; see below). This interpretation of the P3a component could accommodate the results obtained in the 250-350 ms time-window of the current study, on the assumption that PS-H stimuli were functionally similar to the 'relevant stimuli' that usually elicit the P3a. This does not seem an unreasonable suggestion, given that the PS-H stimuli were endowed with relevance in this task, as reliable predictors of a high-value outcome.

The question of whether the SRP component that we observed lying between P2 and P3b is actually stimulus-locked analogue of the FCRP or a P3a is an interesting issue that could be addressed in future experiments. More generally, this question raises the intriguing possibility that stimulus-locked components described as FCRPs in previous research on reinforcement learning were instead actually instances of P3a.

The P3b component (observed in the averaged ERP from 350-450 ms, with the corresponding PCA factor from 460-560ms) also showed sensitivity to incentive value, being greater for PS-H than PS-L or NPS stimuli. As noted above, P3b is considered a proxy of resource allocation in working memory; more specifically, this component is usually interpreted as reflecting the first stages of storing information in working memory (Polich, 2007). The magnitude of the P3b component depends on (among other factors) the subjective probability of the stimulus, the amount of information transmitted by the stimulus, and, perhaps most relevant for our study, the subjective

relevance of meaning of the stimuli (e.g., Johnson, 1986; Polich, 2007). When affective stimuli (i.e., stimuli with relevant meaning) are presented in an oddball paradigm, mean P3b amplitude is determined by the valence of the stimuli (see Olofsson, Nordin, Sequeira, & Polich, 2008). Thus, the pattern of differences observed in the current study could reflect a change in the valence of PS-H stimuli—by virtue of their association with a more valuable outcome. Future experiments could directly assess this possibility.

Our findings relating to the P3b component raise an interesting question regarding whether participants verbally categorized the stimuli. Notably, Rustemeier, Schwabe, & Bellebaum (2013) found that the earliest component affected by verbal strategies during reinforcement learning was the P3. Experiments analyzing feedback-locked ERPs have suggested that effects of learning on FCRP and P3 components reflect different processes. In a reversal learning study (Chase, Swainson, Durham, Benham, & Cools, 2011), participants were given explicit rules for changing their pattern of responses. This study found that P3 modulations indexed behavioral adjustments based on the explicit rules, whereas FCRP reflected an associative error measure, based on experience of the relationships between events (see also Walsh & Anderson, 2011). With regard to the current study, this would suggest that the observed differences in SRP (that we have argued provides an analogue of the FCRP) relate to non-verbal processing of stimuli, reflecting relatively automatic, dopamine-dependent, goal-directed associative learning processes. In contrast, P3b differences could index a posterior, declarative analysis of the outcome value information imparted by stimuli.

As noted above, the SRP and P3b showed sensitivity to incentive value, with activation in the order PS-H > PS-L = NPS (the P3b factor found in the PCA analysis also distinguished between PS-L and NPS). This is interesting, because in the final training blocks (used for ERP analysis) when the correct responses were well-learned,

PS-L were typically paired with small rewards, while NPS were paired equally often with small and large rewards (see Table 1). In this sense, NPS stimuli had an incentive value that was intermediate between PS-L and PS-H, and yet we did not observe any ERP component with activity following the order PS-H > NPS > PS-L. There are at least two reasons why this might be the case. One possibility is that simple co-occurrence with a reward is insufficient to endow a stimulus with incentive value; instead, people will associate a stimulus with a reward only if the stimulus is relevant to the occurrence of that reward. As an analogy, we will associate a good meal (reward) with a particular restaurant or chef (relevant stimulus), but not with the taxi that delivered us to that restaurant (irrelevant stimulus). Consequently in the current study, NPS—which were irrelevant to the occurrence of reward—would not develop high incentive value. An alternative, but related, possibility makes reference to the difference in attention for action that we earlier hypothesized to be indexed by the P2, suggesting greater attention to predictive stimuli (PS-H and PS-L) than to NPS. If people were ignoring NPS, then these stimuli would not become associated with rewards and/or they would not activate the representations of any rewards with which they had become associated. The current data do not allow us to decide between these two possible explanations for the low incentive value of NPS in our task.

During the final 350ms of the 800ms stimulus presentation, we detected a sustained slow negative-going deflection that was greater for PS-H than PS-L stimuli, and smallest for NPS stimuli. That is, this slow component was influenced by incentive value *and* by predictive value. The difference between PS-L and NPS suggests that it reflects activation of response representations, since predictive stimuli were consistently associated with particular responses while NPS stimuli were not. Thus, we suggest that this goal-directed component indexes the activation of a response representation

weighted by the incentive value of the outcome associated with this response. The long latency of this component resembles that of the family of negative slow potentials that include the contingent negative variation (CNV), Bereitschaftspotential (BP) and stimulus-preceding negativity (SPN: see Brunia et al., 2011, and van Boxtel & Böcker, 2004 for review). These components are indices of anticipation and preparation for upcoming relevant events, such as providing responses with significant consequences, or the delivery of monetary rewards, performance feedback, evocative photos, or painful or affectively salient stimulation (Brown, Seymour, Boyle, El-Deredy, & Jones, 2008; Donkers, Nieuwenhuis, & van Boxtel, 2005; Fuentemilla et al., 2013; Kotani, Hiraku, Suda, & Aihara, 2001; Masaki, Yamazaki, & Hackley, 2010; van Boxtel & Böcker, 2004).

The slow negativity in our study showed a posterior-parietal scalp distribution (Figure 2). Previous research suggests involvement of posterior-parietal cortex in abstract motor planning (e.g., Cohen & Andersen, 2002). These abstract representations could facilitate the coordination of different aspects of response-related behavior, like hand–eye coordination. This would explain why, in our study, a slow negativity generated in posterior-parietal cortex was larger for stimuli that allowed participants to plan a response.

As noted earlier, however, we did not have a strong *a priori* expectation of a particular scalp distribution for this slow negativity (unlike for the other ERP components that we analyzed, which have been well-characterized in previous research). Consequently, the risk of a Type I error was greater for this slow negativity because all electrodes were potential candidates for showing a significant difference. Notably, however, the *p*-values associated with the effects of stimulus type at the selected electrode (P3) were quite low: $p \leq 0.001$ for PS-H vs. PS-L and PS-H vs. NPS,

and $p = 0.01$ for PS-L versus NPS. At least the first two of these significant contrasts would survive correction for multiple comparisons over alternative electrode selections. Notwithstanding this, replication of the novel findings relating to the slow negativity would be valuable in order to confirm the conclusions drawn above.

In summary, the results of the current experiment suggest that a complex sequence of cognitive processes was initiated each time participants perceived the stimuli in our reinforcement learning procedure. More specifically, we suggest that this sequence includes cognitive processes with two consecutive objectives: First, to prime the attentional selection of important stimuli for further processing as rapidly as possible (indexed by P2); and second, to prepare participants for the subsequent outcome (indexed by SRP, P3b and slow negativity). Theories of goal-directed action are typically focused on this second set of operations. Looking more closely, the pattern of these latter ERP components might allow us to discriminate between theories of goal directed action. In particular, the temporal order of these ERP effects—with components relating to incentive value (SRP and P3b) appearing *before* a component integrating incentive value and predictive value (slow negativity)—is consistent with *outcome-response* theories of goal-directed action (e.g., de Wit & Dickinson, 2009). Such theories suggest that reinforcement learning establishes both $S \rightarrow O$ and $O \rightarrow R$ associations. So presentation of a stimulus activates a representation of the outcome with which it has been repeatedly paired (producing an early effect of incentive value), and this representation of the outcome then retrieves a memory of the response that produces that outcome (giving a later effect of predictive value). For example, presentation of PS-H would activate a representation of large reward [$S \rightarrow O$], and this representation of large reward would retrieve the memory of the response (R1) that produces that reward [$O \rightarrow R$]. Future research could further probe the goal-directed

nature of behavior following training on the reinforcement learning task by testing the influence of changes in the value of the outcome on participants' responding. That is, if behavior in this task is truly goal-directed (i.e., mediated by a representation of the expected outcome), then changing the value of that outcome—for example, by instructing participants that the points that they earn on each trial will now result in *loss* of money, rather than gain—should result in an immediate change in the pattern of responding (de Wit, Niry, Wariyar, Aitken & Dickinson, 2007).

To conclude, our findings are consistent with (1) A rapid influence of attention for action as a function of predictive value (in line with so-called attentional theories of associative learning: e.g., Mackintosh, 1975); followed by (2) Activation of cognitive processes driving goal-directed behavior as a function of incentive and predictive value (in line with *outcome-response* theories of goal-directed action: e.g., de Wit & Dickinson, 2009). An understanding of the neural correlates of these processes is clearly significant, given the fundamental importance of reinforcement learning in allowing organisms to adapt their behavior so as to exploit environmental regularities. Conversely, the current data also inform our understanding of the functional significance of the ERP components identified in our analyses (P2, SRP, P3b and slow negativity). And finally, as noted in the introduction, reinforcement learning is invoked by models of various types of mental disorder. For example, evidence suggests a disruption of goal-directed learning in schizophrenia (Waltz et al., 2007), and in particular that representations of both incentive value (Morris, Holroyd, Mann-Wrobel, & Gold, 2011) and predictive value (Morris, Griffiths, Le Pelley, & Weickert, 2013) are impaired. This raises the possibility that the stimulus-locked ERPs identified in the current research might provide biomarkers or risk indicators for such aspects of psychopathology.

References

Anderson, B. A., Laurent, P. A., & Yantis, S. (2011). Value-driven attentional capture. *Proceedings of the National Academy of Sciences of the USA, 108,* 10367-10371. doi: 10.1073/pnas.1104047108

Baker, T. E., & Holroyd, C. B. (2009) .Which way do I go? Neural activation in response to feedback and spatial processing in a virtual T-maze. *Cerebral Cortex, 19,*1708-1722. doi: 10.1093/cercor/bhn223

Belouchrani, A., Abed-Meraim, K., Cardoso, J. F., & Moulines, E. (1997). A blind source separation technique using second-order statistics. *IEEE Transactions on Signal Processing, 45*, 434-444. doi: 10.1109/78.554307

Brown, C. A., Seymour, B., Boyle, Y., El-Deredy, W., & Jones A. K. P. (2008). Modulation of pain ratings by expectation and uncertainty: Behavioral characteristics and anticipatory neural correlates. *Pain, 135,* 240–250. doi:10.1016/j.pain.2007.05.022

Brunia, C. H., Hackley, S. A., van Boxtel, G. J., Kotani, Y., & Ohgami, Y. (2011) Waiting to perceive: Reward or punishment?. *Clinical Neurophysiology, 122*, 858-868. doi: 10.1016/j.clinph.2010.12.039

Carretié, L., Mercado, F., Tapia, M., & Hinojosa, J. A. (2001). Emotion, attention, and the 'negativity bias', studied through event-related potentials. *International Journal of Psychophysiology, 41,* 75-85. doi: 10.1016/S0167-8760(00)00195-1

Chase, H. W., Swainson, R., Durham, L., Benham, L., & Cools, R. (2011). Feedback-related negativity codes prediction error but not behavioral adjustment during probabilistic reversal learning. *Journal of Cognitive Neuroscience, 23,* 936-946. doi: 10.1162/jocn.2010.21456

Cohen, Y. E., & Andersen, R. A. (2002). A common reference frame for movement

plans in the posterior parietal cortex. *Nature Reviews Neuroscience, 3,* 553-562. doi: 10.1038/nrn873

Courchesne, E., Hillyard, S. A., & Galambos, R. (1975). Stimulus novelty, task relevance and the visual evoked potential in man. *Electroencephalography and clinical neurophysiology*, *39*, 131-143. doi: 10.1016/0013-4694(75)90003-6

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,* 9-21. doi: 10.1016/j.jneumeth.2003.10.009

Delplanque, S., Lavoie, M. E., Hot, P., Silvert, L., & Sequeira, H. (2004). Modulation of cognitive processing by emotional valence studied through event-related potentials in humans. *Neuroscience letters, 356*, 1-4. doi: 10.1016/j.neulet.2003.10.014

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

de Wit, S., Niry, D., Wariyar, R., Aitken, M. R. F., & Dickinson, A. (2007). Stimulus-outcome interactions during instrumental discrimination learning by rats and humans. *Journal of Experimental Psychology: Animal Behavior Processes*, *33*,1-11

Dien, J. (2012). Applying principal components analysis to event-related potentials: a tutorial. *Developmental neuropsychology*, *37*, 497-517.

Donkers, F. C., Nieuwenhuis, S., & van Boxtel, G. J. (2005). Mediofrontal negativities in the absence of responding. *Cognitive Brain Research, 25,* 777–787. doi:10.1016/j.cogbrainres.2005.09.007

Dunning, J. P., & Hajcak, G. (2007). Error-related negativities elicited by monetary loss

and cues that predict loss. *Neuroreport, 18,* 1875-1878. doi: 10.1097/WNR.0b013e3282f0d50b

Evans, K. M, & Federmeier, K. D. (2007). The memory that's right and the memory that's left: event-related potentials reveal hemispheric asymmetries in the encoding and retention of verbal information. *Neuropsychologia, 45,*1777-1790. doi: 10.1016/j.neuropsychologia.2006.12.014

Fuentemilla, L., Cucurell, D., Marco-Pallarés, J., Guitart-Masip, M., Morís, J., & Rodríguez-Fornells, A. (2013). Electrophysiological correlates of anticipating improbable but desired events. *NeuroImage*, *78*, 135-144. doi: 10.1016/j.neuroimage.2013.03.062

Frank, M. J. (2008). Schizophrenia: A computational reinforcement learning perspective. *Schizophrenia Bulletin, 34*, 1008-1011. doi: 10.1093/schbul/sbn123

Hillyard, S. A., & Anllo-Vento, L. (1998). Event-related brain potentials in the study of visual selective attention. *Proceedings of the National Academy of Sciences of the USA, 95*,781-787. doi: 10.1073/pnas.95.3.781

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,* 679-709. doi: 10.1037/0033-295X.109.4.679

Holroyd, C. B., Krigolson, O. E., & Lee, S. (2011). Reward positivity elicited by predictive cues. *Neuroreport, 22,* 249-252. doi: 10.1097/WNR.0b013e328345441d

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

Hyman, S. E. (2005). Addiction: A disease of learning and memory. *American Journal of Psychiatry, 162*, 1414-1422. doi: 10.1176/appi.ajp.162.8.1414

Hughes, G., Mathan, S., & Yeung, N. (2012). EEG indices of reward motivation and target detectability in a rapid visual detection task. *Neuroimage, 64,* 590-600. doi: 10.1016/j.neuroimage.2012.09.003

Johnson, R. (1986). A triarchic model of P300 amplitude. *Psychophysiology, 23*, 367-384. doi: 10.1111/j.1469-8986.1986.tb00649.x

Kennerley, S. W., Walton, M. E., Behrens, T. E., Buckley, M. J., & Rushworth, M. F. (2006). Optimal decision making and the anterior cingulate cortex. *Nature Neuroscience, 9,* 940-947. doi: 10.1038/nn1724

Kiss, M., Driver, J., & Eimer, M. (2009). Reward priority of visual target singletons modulates event-related potential signatures of attentional selection. *Psychological Science, 20*, 245-251. doi: 10.1111/j.1467-9280.2009.02281.x

Kotani, Y., Hiraku, S., Suda, K., & Aihara, Y. (2001). Effect of positive and negative emotion on stimulus-preceding negativity prior to feedback stimuli. *Psychophysiology, 38,* 873–878. doi:10.1111/1469-8986.3860873

Le Pelley, M. E (2010). Attention and human associative learning. In C. J. Mitchell, & M. E. Le Pelley (Eds*.), Attention and Associative Learning: From Brain to Behaviour* (pp. 187-215). Oxford: Oxford University Press.

Le Pelley, M. E., Beesley, T., & Griffiths, O. (2011). Overt attention and predictiveness in human contingency learning. *Journal of Experimental Psychology: Animal Behavior Processes, 37*, 220-229. doi: 10.1037/a0021384

Le Pelley, M. E., & McLaren, I. P. L. (2003). Learned associability and associative change in human causal learning. *The Quarterly Journal of Experimental Psychology: Section B*, *56*, 68-79. doi: 10.1080/02724990244000179

Le Pelley, M. E., Vadillo, M. A., & Luque, D. (2013). Learned predictiveness influences rapid attentional capture: Evidence from the dot probe task. *Journal of Experimental Psychology: Learning, Memory, and Cognition, 39,* 1888-1900. doi: 10.1037/a0033700.

Liao, Y., Gramann, K., Feng, W., Deak, G., & Li, H. (2011). This ought to be good: Brain activity accompanying positive and negative expectations and outcomes. *Psychophysiology, 48,* 1412-1419. doi: 10.1111/j.1469-8986.2011.01205.x

Mackintosh, N. J. (1975). A theory of attention: Variations in the associability of stimuli with reinforcement. *Psychological Review, 82*, 276-298. doi: 10.1037/h0076778

Marco-Pallares, J., Cucurell, D., Münte, T. F., Strien, N., & Rodriguez-Fornells, A. (2011). On the number of trials needed for a stable feedback-related negativity. *Psychophysiology*, *48*, 852-860. doi: 10.1111/j.1469-8986.2010.01152.x

Masaki, H., Yamazaki, K., Hackley, S. H. (2010). Stimulus-preceding negativity is modulated by action-outcome contingency. *Neuroreport , 21,* 277–281. doi:10.1097/WNR.0b013e3283360bc3

McClure, S. M., York, M. K., & Montague, P. R. (2004). The neural substrates of reward processing in humans: the modern role of FMRI. *The Neuroscientist*,*10*, 260-268. doi: 10.1177/1073858404263526

Morís, J., Luque, D., & Rodríguez-Fornells, A. (2013). Learning-induced modulations of the stimulus-preceding negativity. *Psychophysiology*, *50*, 931-939. doi: 10.1111/psyp.12073

Morris, R., Griffiths, O., Le Pelley, M. E., & Weickert, T. W. (2013). Attention to irrelevant cues is related to positive symptoms in schizophrenia. *Schizophrenia bulletin*, *39*, 575-582. doi: 10.1093/schbul/sbr192

Morris, S. E., Holroyd, C. B., Mann-Wrobel, M. C., & Gold, J. M. (2011). Dissociation of response and feedback negativity in schizophrenia: electrophysiological and computational evidence for a deficit in the representation of value. *Frontiers in Human Neuroscience*, *5*, 123. doi: 10.3389/fnhum.2011.00123

Olofsson, J. K., Nordin, S., Sequeira, H., & Polich, J. (2008). Affective picture processing: an integrative review of ERP findings. *Biological psychology*, *77*, 247-265. doi: 10.1016/j.biopsycho.2007.11.006

Polich, J. (2007). Updating P300: an integrative theory of P3a and P3b. *Clinical neurophysiology*, *118*, 2128-2148. doi: 10.1016/j.clinph.2007.04.019

Riesel, A., Weinberg, A., Endrass, T., Kathmann, N., & Hajcak, G. (2012). Punishment has a lasting impact on error-related brain activity. *Psychophysiology*, *49*, 239-247. doi: 10.1111/j.1469-8986.2011.01298.x

Rushby, J. A., Barry, R. J., & Doherty, R. J. (2005). Separation of the components of the late positive complex in an ERP dishabituation paradigm. *Clinical Neurophysiology*, *116*, 2363-2380. doi: 10.1016/j.clinph.2005.06.008

Rustemeier, M., Schwabe, L., & Bellebaum, C. (2013). On the relationship between learning strategy and feedback processing in the weather prediction task—Evidence from event-relatedpotentials. *Neuropsychologia*, *51*, 695-703. doi: 10.1016/j.neuropsychologia.2013.01.009

Santesso, D. L., Steele, K. T., Bogdan, R., Holmes, A. J., Deveney, C. M., Meites, T. M., & Pizzagalli, D. A. (2008). Enhanced negative feedback responses in remitted depression. *Neuroreport, 19*, 1045-1048. doi: 10.1097/WNR.0b013e3283036e73

Schevernels, H., Krebs, R. M., Santens, P., Woldorff , M. G., & Boehler, C. N. (2014) Task preparation processes related to reward prediction precede those related to task-difficulty expectation. *NeuroImage, 84,* 639-647.

doi:10.1016/j.neuroimage.2013.09.039

Shackman, A. J., Salomons, T. V., Slagter, H. A., Fox, A. S., Winter, J. J., & Davidson, R. J. (2011). The integration of negative affect, pain and cognitive control in the cingulate cortex. *Nature Reviews Neuroscience*, *12*, 154-167. doi: 10.1038/nrn2994

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

Treisman, A. (1996). Selection for perception for selection for action: A reply to Van der Heijden. *Visual Cognition, 3,* 353-357

Van Boxtel, G. J., & Böcker, K. B. E. (2004). Cortical measures of anticipation. *Journal of Psychophysiology, 18,* 61–76. doi:10.1027/0269-8803.18.2–3.61

Walsh, M. M., & Anderson, J. R. (2011). Modulation of the feedback-related negativity by instruction and experience. *Proceedings of the National Academy of Sciences of the USA, 108,* 19048-19053. 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 & Biobehavioral Reviews*, *36*, 1870-1884. doi: 10.1016/j.neubiorev.2012.05.008

Waltz, J. A., Frank, M. J., Robinson, B. M., & Gold, J. M. (2007). Selective reinforcement learning deficits in schizophrenia support predictions from computational models of striatal-cortical dysfunction. *Biological psychiatry*, *62*, 756-764. doi: 10.1016/j.biopsych.2006.09.042

Wills, A. J., Lavric, A., Croft, G. S., & Hodgson, T. L. (2007). Predictive learning, prediction errors, and attention: Evidence from event-related potentials and eye tracking. *Journal of Cognitive Neuroscience*, *19*, 843-854. doi:

10.1162/jocn.2007.19.5.843

Wills, A. J., Lavric, A., Hemmings, Y., & Surrey, E. (2014). Attention, predictive learning, and the inverse base-rate effect: Evidence from event-related potentials. *NeuroImage*, *87*, 61-71. doi: 10.1016/j.neuroimage.2013.10.060

Autor Note

This work was supported by grants FT100100260 from the Australian Research Council, P11-SEJ-7896 from the Andalusian Government (Spain) and PSI2011 -24662 from the Spanish Ministry of Economy and Competitiveness. This work was conducted while David Luque was a Visiting Fellow at the School of Psychology of the University of New South Wales. This visit was financed by the Spanish Ministry of Education, Culture and Sports through the 'José Castillejo' program.

For reprints, please contact David Luque, School of Psychology, University of Málaga. Campus de Teatinos s/n CP 29017, Málaga, Spain.
Email: david.luque@gmail.com.

Footnotes

1. Throughout this article, we refer to this component as the feedback correct-related positivity, rather than using the label "feedback-related negativity" which appears in much of the previous literature on reinforcement learning (see Holroyd & Coles, 2002). This reflects recent suggestions that this component reflects positivity generated during positive feedback, rather than negativity generated during negative feedback; an idea that has received empirical support (Holroyd, et al., 2008).

2. In this task the feature of the target stimulus that signaled reward availability did not predict the response that was to be made to that stimulus. As such this task does not have the classic *S-R-O* structure of reinforcement learning that was implemented in the current procedure.

3. This also explains why participants' response accuracy increased at a similar rate for trials featuring PS-H and PS-L stimuli: there was a strong drive to learn about both types of stimuli in order to gain large numbers of points (by making the correct response on trials with PS-H) and to avoid losing large numbers of points (by making the incorrect response on trials PS-L).

Legends

Table 1: Experiment design. Stimuli were abstract images. Each image was assigned to one of the following roles in a counterbalanced way: PS-H: Predictive stimulus with high value outcome; PS-L: Predictive stimulus with low value outcome. NPS_1, NPS_2: Non-Predictive stimuli. R1 and R2 were different response options. The table shows that (for example) when stimulus pair PS-H & NPS_1 was presented, selecting R1 was scheduled to give an outcome of +100 points on a random 16 of the 20 presentations in each block, and –100 points on the other 4 presentations. Selecting R2 for this same stimulus pair was scheduled to give an outcome of –1 point on 16 presentations and +1 point on the remaining 4. Hence R1 was the 'correct' response for this stimulus pair (and for stimulus pair PS-H & NPS_2), since making this response gave the higher average payoff, and R2 was the 'incorrect' response. Similarly, R2 was the 'correct' response and R1 the 'incorrect' response for stimulus pairs PS-L & NPS_1, and PS-L & NPS_2.

Figure 1. (A) Trial structure. Participants were instructed to respond when the circle and asterisk were displayed (labeled 'Response' in the figure). (B) Proportion of correct responses (upper panel) and mean response times (lower panel) to each of the four stimulus pairs encountered by participants (see Table 1) across the course of training. Error bars show standard error of the mean. (C) Grand average ERPs during presentation of each type of discriminative stimulus. Topographical figures map the differences between stimuli that were predictive of high-value reward (PS-H) and stimuli that were not predictive (NPS) [first row of topographical figures], and between PS-H and stimuli that were predictive of low-value reward (PS-L) [second row of topographical figures]. Relative scale of each topographical map (minimum/maximum values): 150-190ms, −1/1.4 μV; 250-350ms, −1.4/3.6 ,350-450ms, −1.9/4.1 μV; 450-

800ms, −1/5.1 μV. Below the scalp topographies are shown the mean evoked activity from 100ms (baseline) before stimulus onset to stimulus offset (800 ms) for each relevant electrode (FCz, Cz, Pz and P3 respectively). Data shown in these graphs were filtered with a low-pass filter of 20 Hz for presentational purposes. Time-windows of interest are represented in shaded columns.

Figure 2. Grand average waveforms elicited by each type of discriminative stimulus (PS-H, PS-L and NPS; see Table 1 and text), in a series of electrodes distributed across the scalp.

Figure 3. Results of the Principal Components Analysis of the EEG data, showing the three principal components extracted (Factors 1 to 3). Topographical maps show, across the scalp, the differences in each factor between stimuli that were predictive of high-value reward (PS-H) and stimuli that were not predictive (NPS) [first row of topographical figures], and between PS-H and stimuli that were predictive of low-value reward (PS-L) [second row of topographical figures]. Below the scalp topographies are shown the grand average waveforms for each factor, for each type of stimulus (PS-H, PS-L and NPS), for relevant electrodes (FCz, Cz, Pz). Time-windows of interest are represented in shaded columns.

Table 1. Types and number of learning trials.

FREQUENT (16 of each type per block)			RARE (4 of each type per block)		
Stimulus pair	*Response*	*Outcome*	*Stimulus pair*	*Response*	*Outcome*
PS-H & NPS$_1$:	R1 →	+100 points	**PS-H & NPS$_1$:**	R1 →	–100 points
	R2 →	–1 point		R2 →	+1 point
PS-H & NPS$_2$:	R1 →	+100 points	**PS-H & NPS$_2$:**	R1 →	–100 points
	R2 →	–1 point		R2 →	+1 point
PS-L & NPS$_1$:	R1 →	–100 points	**PS-L & NPS$_1$:**	R1 →	+100 points
	R2 →	+1 point		R2 →	–1 point
PS-L & NPS$_2$:	R1 →	–100 points	**PS-L & NPS$_2$:**	R1 →	+100 points
	R2 →	+1 point		R2 →	–1 point

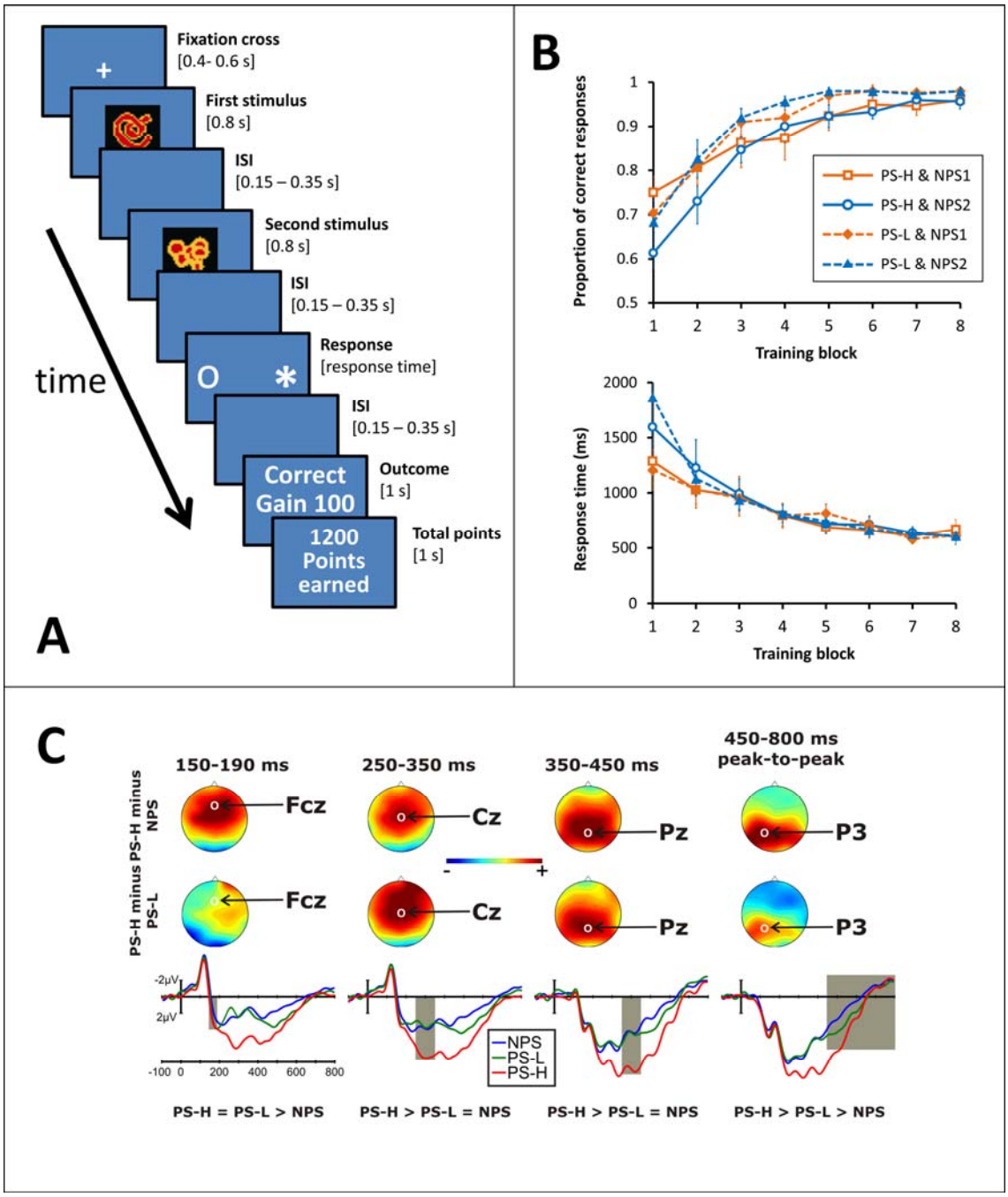
A
Fixation cross
[0.4- 0.6 s]
First stimulus
[0.8 s]
ISI
[0.15 – 0.35 s]
Second stimulus
[0.8 s]
ISI
[0.15 – 0.35 s]
Response
[response time]
ISI
[0.15 – 0.35 s]
Outcome
[1 s]
Total points
[1 s]
Correct
Gain 100
1200
Points
earned
time
B
Proportion of correct responses
Training block
PS-H & NPS1
PS-H & NPS2
PS-L & NPS1
PS-L & NPS2
Response time (ms)
Training block
C
150-190 ms
250-350 ms
350-450 ms
450-800 ms
peak-to-peak
PS-H minus NPS
PS-H minus PS-L
Fcz
Cz
Pz
P3
NPS
PS-L
PS-H
PS-H = PS-L > NPS
PS-H > PS-L = NPS
PS-H > PS-L = NPS
PS-H > PS-L > NPS

Figure 1

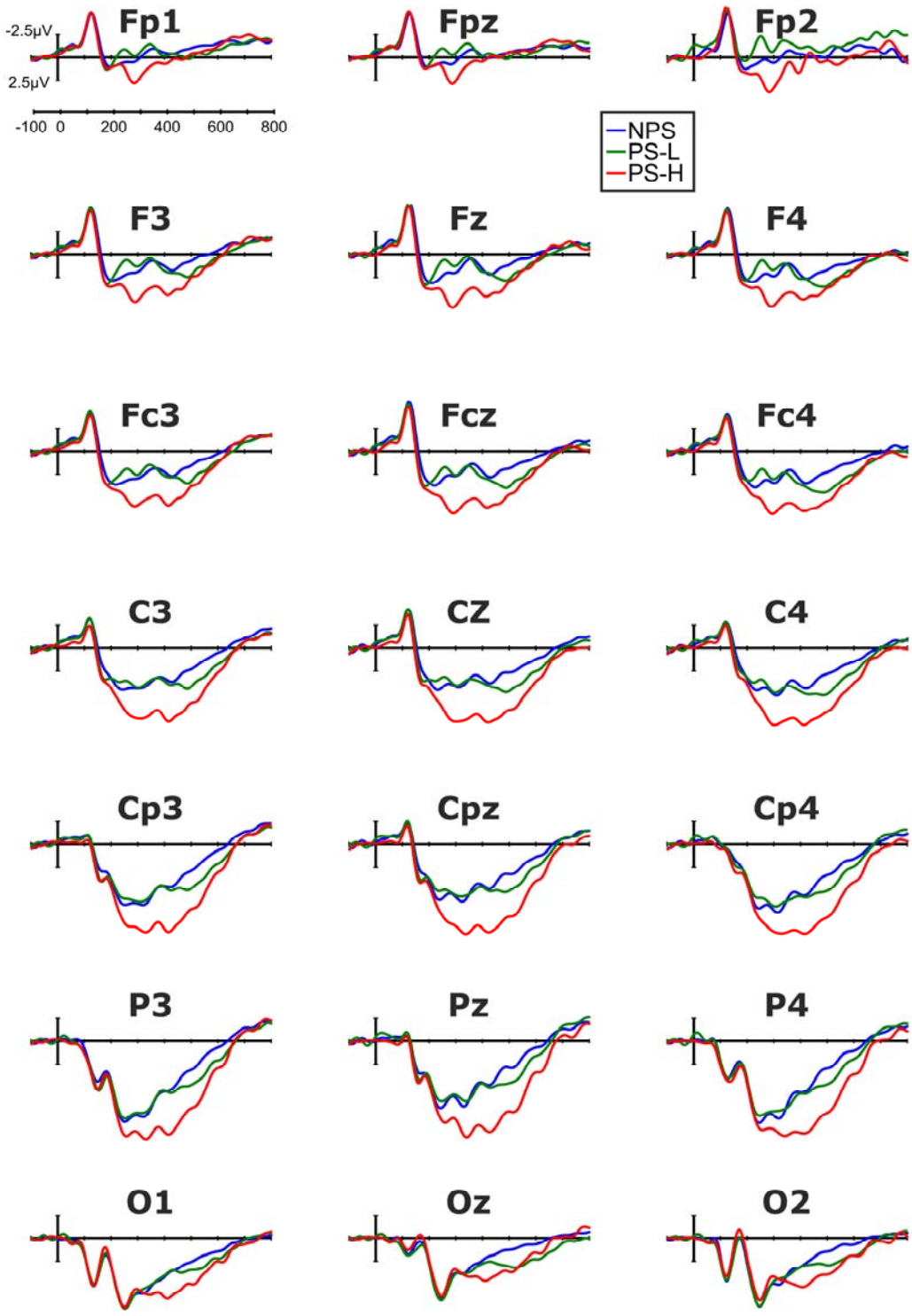
Fp1
-2.5μV
2.5μV
-100 0 200 400 600 800
Fpz
Fp2
NPS
PS-L
PS-H
F3
Fz
F4
Fc3
Fcz
Fc4
C3
CZ
C4
Cp3
Cpz
Cp4
P3
Pz
P4
O1
Oz
O2

Figure 2

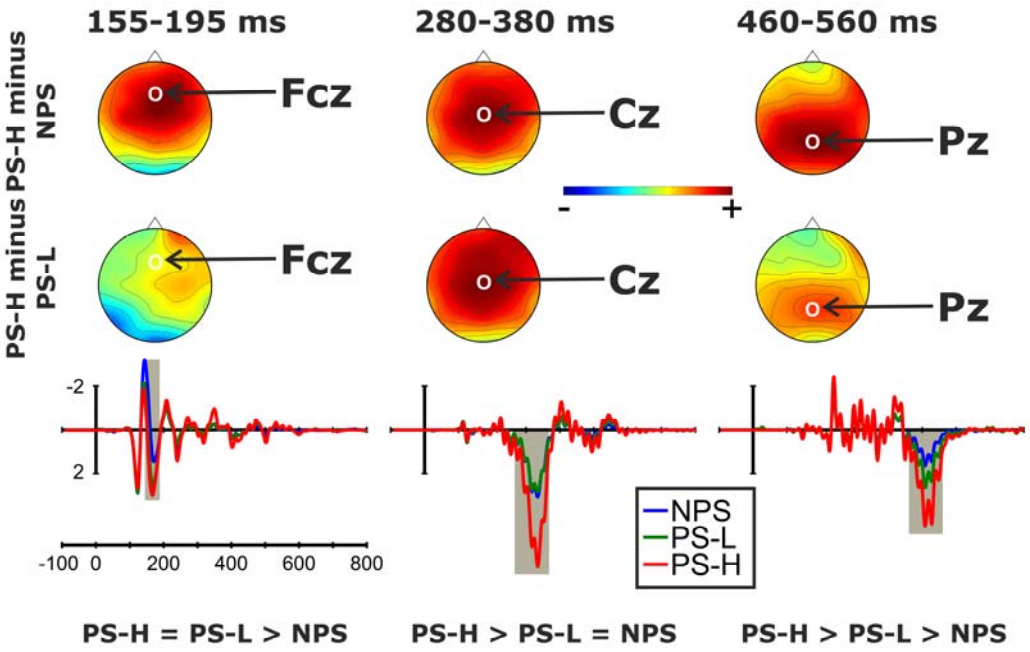
155-195 ms
280-380 ms
460-560 ms
PS-H minus NPS
PS-H minus PS-L
Fcz
Cz
Pz
-
+
-2
2
-100 0 200 400 600 800
NPS
PS-L
PS-H
PS-H = PS-L > NPS
PS-H > PS-L = NPS
PS-H > PS-L > NPS

Figure 3