Predictive learning, prediction errors and attention:

Evidence from event-related potentials and eye-tracking.

A.J. Wills, A. Lavric, G.S. Croft and T.L. Hodgson

School of Psychology, University of Exeter,

Perry Road, Exeter. EX4 4QB. England.

Tel: 44 (0) 1392 264650

Fax: 44 (0) 1392 264623

E-mail: a.j.wills@exeter.ac.uk

Accepted version.

Abstract

Prediction error ("surprise") affects the rate of learning: we learn more rapidly about cues for which we initially make incorrect predictions than cues for which our initial predictions are correct. The current studies employ electrophysiological measures to reveal early attentional differentiation of events that differ in their previous involvement in errors of predictive judgment. Error-related events attract more attention, as evidenced by features of event-related scalp potentials previously implicated in selective visual attention (selection negativity, augmented anterior N1). The earliest differences detected occurred around 120 ms after stimulus onset, and distributed source localization (LORETA) indicated that the inferior temporal regions were one source of the earliest differences. In addition, stimuli associated with the production of prediction errors show higher dwell times in an eye-tracking procedure. Our data support the view that early attentional processes play a role in human associative learning.

Determining the extent to which one event predicts another is one of the most fundamental forms of learning. Classic theorists assumed that predictive learning occurred whenever two events were contiguous (Pavlov, 1927). However, more recent analyses indicate that learning also requires that the second event be somewhat unexpected (Kamin, 1969). That is, predictive learning appears to be driven by prediction errors rather than simple contiguity, and it occurs at a rate related to the discrepancy between what is predicted on the basis of the first event, and what actually occurs.

Why does predictive learning appear to be error-driven? Associative theories assume that prediction errors affect the rate at which associations between representations of the two events form (Mackintosh, 1975; Pearce & Hall, 1980; Rescorla & Wagner, 1972; Schultz, Dayan, & Montague, 1997) whilst reasoning accounts assume that predictive learning occurs through a process of high-level reasoning (De Houwer, Beckers, & Vandorpe, 2005). Proponents of each type of account have uncovered behavioral phenomena potentially problematic for the other (De Houwer & Beckers, 2002; Le Pelley, Oakeshott, & McLaren, 2005), and the case for multi-process accounts of predictive learning is frequently made (Ashby, Alfonso-Reese, Turken, & Waldron, 1998; Erickson & Kruschke, 1998). Given this, many neuroscientific investigations have understandably sought to examine predictions of particular theories, rather than attempt to distinguish between such broad and non-exclusive classes of theory. For example, one recent investigation provided evidence that the BOLD fMRI signal in prefrontal cortex conforms to the predictions of the Rescorla-Wagner associative theory (Fletcher et al., 2001), and another (O'Doherty, Dayan, Friston, Critchley, & Dolan, 2003) demonstrated that activity in the striatum conformed to the predictions of the temporal difference model (Schultz, Dayan, & Montague, 1997).

The goal of the studies reported in the current paper was to investigate a prediction made by a number of associative theories, including Pearce-Hall theory (Pearce & Hall, 1980). Pearce-Hall theory states that predictive learning is error-driven because the learner has limited stimulus

processing capacity. In order to make maximal use of these limited resources, the extent to which a stimulus is processed is modulated by its previous involvement in prediction errors. Specifically, a stimulus whose consequence is well-predicted is processed to a lesser extent than a stimulus that has recently been followed by surprising or unexpected events. This leads to the prediction that stimuli whose consequences are uncertain receive more attention than stimuli whose consequences are well-predicted.

A different but related proposal (Kruschke, 2001; Mackintosh, 1975) is that attention is distributed amongst the features of a presented stimulus in accordance with the extent to which those features predict an outcome. Specifically, features that were previously good predictors of an outcome are assumed to attract more attention than features that were previously poor predictors of an outcome. Mackintosh-Kruschke theory is not typically framed in terms of limited processing capacity, although such an interpretation is not unreasonable. Whilst the Mackintosh-Kruschke and Pearce-Hall theories may seem to be contradictory, in that the relationships between prediction error and attention they postulate are opposite, they can in fact be considered to be complementary. Mackintosh-Kruschke theory makes predictions about the relative amounts of attention different features of the presented stimulus will receive, whilst Pearce-Hall theory makes predictions about changes in the absolute amount of attention directed to the entire stimulus.

Indirect evidence for the presence of Mackintosh-Kruschke attentional processes in human predictive learning is provided by the effects of prior predictiveness on the rate of subsequent learning. For example, Lochmann and Wills (2003) trained adults on a task where some features of the presented stimuli were predictive of an outcome whilst other features were non-predictive. In a subsequent phase, all stimulus features were fully predictive of a novel outcome; nevertheless, the previously predictive cues were learned about more rapidly than the previously non-predictive cues.

Indirect evidence for the presence of Pearce-Hall attentional processes in human predictive learning comes from the BOLD response that is observed in certain brain regions to the unexpected occurrence and unexpected omission of outcomes. Pearce-Hall theory predicts increased attention as a result of both the unexpected occurrence and the unexpected omission of an outcome, and hence the observation that the BOLD responses in the hippocampus, superior frontal gyrus and cerebellum increase to both types of event (Ploghaus et al., 2000) has been taken by some as support for this type of associative theory. In other brain regions, for example the ventral putamen, unexpected occurrence of an outcome leads to an increase in BOLD signal whilst the unexpected omission of the outcome attenuates the BOLD signal (O'Doherty et al., 2003), which is more in line with the predictions of non-attentional theories such as temporal difference theory (Schultz et al., 1997)

One limitation of Ploghaus et al. (2000), and a number of other studies (Fletcher et al., 2001; O'Doherty et al., 2003), is that the unexpected events are more novel than the expected events. For example, Ploghaus et al. (2000) compare the first trial on which a painful stimulus follows a colored light with the second trial on which this occurs. The first trial is assumed to have a higher prediction error than the second, because the painful stimulus is less expected on the first trial than on the second. However, it is also the case that both the light and the painful stimulus are less novel on the second trial than on the first. Novel events will tend, on the whole, to have larger prediction errors than familiar events, but events of equal frequency can differ in the hypothesized magnitudes of their prediction errors. Critically, it is prediction error rather than frequency per se that drives learning in most associative theories. A number of more complex experimental designs that employ multiple training phases and multi-feature stimuli allow frequency to be equated whilst maintaining differences in prediction error (e.g. Turner et al., 2004). Using such a design, Turner et al. (2004) confirmed that both the unexpected omission and unexpected occurrence of an outcome were associated with increased BOLD activity (in lateral frontal cortex).

In summary, behavioral and neuroimaging studies have thus far provided some indirect evidence of the involvement of attentional processes in human predictive learning. In Experiment 1, we sought to extend and strengthen this evidence by exploiting the temporal resolution of electrophysiological measures to determine whether stimuli differing in their prediction error also differ in the amount of early attentional resources they are allocated. Electrophysiological measures have previously been used successfully in the study of predictive learning (e.g. Holroyd, Nieuwenhuis, Yeung, & Cohen, 2003).

There is an extensive pre-existing literature on the Event-Related Potential (ERP) correlates of selective attention. Two sets of ERP components have been implicated in visual selective attention (Hillyard & Anllo-Vento, 1998). When spatial position determines the amount of attention allocated to a stimulus, attended and non-attended stimuli differ in the magnitude of the ERP components P1, posterior N1 and anterior N1, all three having a larger amplitude for attended stimuli (Clark & Hillyard, 1996). The magnitude of components from this set, often referred to as "exogenous components", can also be modulated by increasing the demand on visual discrimination of the stimulus, even when spatial position is held constant (Vogel & Luck, 2000). These spatial and non-spatial modulations of exogenous components are consistent with their interpretation in terms of a sensory enhancement mechanism (Hillyard & Anllo-Vento, 1998) that is relatively non-specific with regard to individual features of stimuli, such as color, orientation, etc. Selective attention to individual features is associated with another set of ERP components: a selection negativity (SN), with a posterior scalp distribution, often accompanied by a selection positivity (SP) at anterior scalp sites (Hillyard & Anllo-Vento, 1998). This set of components is particularly relevant in the context of the current studies, in which shape distinguishes the stimuli to be contrasted. More specifically, support for the involvement of an early attentional process in human associative learning would be provided if the magnitude of the SN and/or SP components to a stimulus previously involved in many

prediction errors was larger than the magnitude of the component to a stimulus involved in relatively few prediction errors (but had occurred with equal frequency).

It would also seem reasonable to expect prediction error to modulate the so-called "exogenous" attentional components (P1, N1): early differentiation of the stimulus associated with many prediction errors from the stimulus associated with few prediction errors may lead to enhanced subsequent perceptual processing of the former and/or a suppressed processing of the latter. As discussed, such sensory enhancement/suppression is reflected in the amplitude of the P1, posterior N1 and anterior N1 components. Experiment One tested these predictions by using multi-electrode electrophysiological recordings, and ERP component and distributed source localization analyses to examine the expected ERP effects and establish whether stimuli associated with many prediction errors result in higher activation of the cortical circuitry known to be involved in visual attention than stimuli associated with few prediction errors.

Experiment One

Experiment One employed a forward cue competition design. Forward cue competition is a design commonly employed in the study of prediction errors in learning and the direction of any reliable effect is well known. The design employed is summarized in Table 1; the letters represent the abstract stimuli employed. Hence, in the first part of the experiment, some stimuli predict an outcome (a fictitious fever) whilst others predict the absence of that outcome. In the second part of the experiment, these stimuli are paired with novel stimuli. On AX trials, participants tend to expect an outcome from the outset, and hence X is involved in few prediction errors. On BY trials, participants tend not to expect an outcome initially, and hence Y is involved in rather more prediction errors. As a consequence, participants are predicted to learn more about Y than X, and this is assessed in the final part of the experiment. Stimuli X and Y are presented isolation, and participants' propensity to respond "fever" to X and Y is assessed.

The prediction of error-driven associative learning (and some reasoning accounts) is that participants are more likely to respond "fever" to Y than to X, as a result of Y having previously been involved in more prediction errors. In this experiment, participants receive "data missing" feedback on the X and Y trial types – in other words, they are told that it is not known whether the outcome in fact occurred. The other trial types in phases 1 and 2 are fillers, and the other trial types in phase 3 maintain the learning established in phases 1 and 2.

The forward cue competition design employed has an advantage over simpler designs in that the target stimuli which differ in their previous involvement in prediction errors (X and Y in phase 3) occur with equal frequency. Nevertheless, our design is still relatively simple compared to some behavioral-only studies of forward cue competition and, as such, does not provide as much information about behavioral performance as these more complex designs. In particular, the forward cue competition design can potentially be broken down into two sub-designs that are described as *forward blocking* (A+ followed by AX+) and *reduced overshadowing* (B- followed by BY+). The relative contribution of these two components to forward cue competition can potentially be assessed by the introduction of further trial types in phases 2 and 3. We decided not to include such trial types in order to keep the difficulty and length of the task within acceptable limits for our participants. The constraints of an ERP methodology meant we had to employ large numbers of small, abstract stimuli in order to maximize our ability to detect reliable, artifact-free ERP components, and this limited the complexity of the behavioral design we could employ.

In a forward cue competition design (Table 1), attentional theories of associative learning predict that Y will attract more attention than X in phase 3. In Mackintosh-Kruschke theory attention will be directed away from X in the AX trials of phase 2 because it is being presented in the presence of a stimulus that already predicts the outcome well (A). This will not happen to Y in BY trials in phase 2 because B does not predict the outcome of BY trials in phase 2. Hence, Y will attract more attention than X in phase 2. It is a prediction of Mackintosh-Kruschke theory

that these attentional differences will persist, at least initially, when X and Y are subsequently presented in isolation. The behavioral literature suggests that attentional differences do indeed persist in this way (e.g. Lawrence, 1952; Lochmann & Wills, 2003).

Pearce-Hall theory also predicts that Y will attract more attention than X in phase 3. In phase 2, the outcome of AX trials is well-predicted from the outset, so the amount of attention attracted by X will decline substantially across phase 2. In contrast, the outcome of BY trials in phase 2 is not well-predicted initially and hence the decline in attention to Y will be slower. Although Pearce-Hall predicts that attention to both X and Y will eventually decline to zero when learning is complete, this is a limiting case that arguably may never be reached in practice. As in Mackintosh-Kruschke theory, attentional differences are predicted to persist, at least initially, when X and Y are presented in isolation.

Whilst the predictions of attentional theories of predictive learning concerning X and Y in phase 3 are unambiguous, one might reasonably argue that a more direct test of these theories would be to measure the amount of attention X and Y attract during phase 2, when learning is occurring, rather than in phase 3 after learning has occurred. Such a test is precluded due to the limitations of ERP methodology. X and Y appear in compound with other stimuli (A and B) in phase 2, and it is extremely difficult to isolate the neurophysiological response elicited by individual stimuli that are presented simultaneously. Attentional differences between the AX and BY stimulus compounds would be relatively uninformative because such differences could be due to a number of different mechanisms. For example, B may attract more attention than A in phase 2 because B is novel in the context of an outcome (having previously only appeared in the context of no outcome). Attentional differences between X and Y in phase two are assessed in Experiment 2 with eye-tracking.

Method

Participants

Twenty-one students were paid 12 GBP for a 2 hour session. All participants were right-handed. One participant's data were discarded due to excessive EOG artifacts. The remaining twenty participants (11 females; age: mean 20.85, s.d. 3.96, range 18-36) were the subject of all subsequent analyses.

Apparatus

Stimulus presentation and response collection was via a PC-compatible computer and the E-prime package (Version 1.1, Psychology Software Tools, Pittsburgh, USA). The EEG was recorded from 64 Ag/AgCl electrodes mounted in an elastic cap (ElectroCap International Inc., Eaton, Ohio, USA), with a forehead (AFz) ground and a vertex (Cz) reference. Two of the available 64 channels were used for recording the horizontal EOG (at the outer canthi of both eyes); two for recording the vertical EOG (supra- and sub-orbitally at the right eye) and two were placed on the earlobes for off-line re-referencing in component amplitude analyses. Scalp channels (58) were placed in accordance with the extended 10-20 (10%) convention. The EEG was sampled at 500 Hz, 0.016Hz –100Hz bandpass filtered, and amplified using BrainAmp amplifiers (BrainProducts Ltd, Munich, Germany).

Stimuli

Twenty-four abstract pictures were selected from the thirty-six used in a previous study (Wills & McLaren, 1997) , re-colored red with a yellow outline, and presented against a black background. The pictures were 0.64^{o} of visual angle in diameter, presented inside a white outline square 2.5^{o} in visual angle. On trials where one picture was presented, it was positioned in the center of the square. In accordance with attentional theories of predictive learning, forward cue competition appears to be facilitated by the spatial separation of the features in compound stimuli (Glautier, 2002) so, where two pictures were presented in this experiment they were

spatially separated. Specifically, they were vertically aligned, one appearing 0.36^{o} of visual angle above the mid-point, and the other an equivalent distance below.

Procedure

Participants were asked to imagine that they worked for a medical referral service, and that their job was to predict a fictitious disease ("Jominy fever") on the basis of "cell bodies" in patients' blood samples (represented by abstract pictures). The twenty-four pictures of cell bodies were, separately for each participant, randomly divided into six cell types (four cell bodies each) corresponding to stimulus types A, B, I, J, X and Y in Table 1.

The structure of each trial is illustrated in Figure 1. Trials began with the presentation of an outline square. After one second, one or two "cell bodies" appeared inside the square. Participants were expected to make either a "fever" or a "no fever" response by pressing one of two keys on a standard PC keyboard. Allocation of "fever" and "no fever" responses to these two keys was counter-balanced across participants. Once the participant had responded, the abstract pictures and outline square were replaced with a feedback message that indicated whether the participant's response was correct or incorrect, and also indicated the correct response. If no response was made within two seconds of the onset of the "cell bodies", the screen cleared and the message "Time out!" was presented for 1.5 seconds. The next trial followed immediately after this message. In the final phase of the experiment, X and Y trials were followed by the uninformative feedback message "????? - DATA MISSING"

The experiment had three phases, as shown in Table 1. Trial order within each phase was randomized within each of several sequential blocks; starts of blocks were not signaled to participants in any way. Block length was 12 trials for phase one, with each of the three trial types (A+, B- and I-) occurring once for each of the four stimuli that comprised each stimulus type (i.e. 4 A stimuli, 4 B stimuli, and 4 I stimuli). Block length in phase two was 24 trials (3 trial types x 4 pictures x 2 screen positions [e.g. A upper, X lower & X upper, A lower]). For

phase three, block length was 48 trials (2 two-picture trial types x 4 pictures x 2 screen positions + 4 one-picture trial types x 4 pictures x 2 presentations). There were 16 blocks in phase one, 6 blocks in phase two, and 6 blocks in phase three.

Electrophysiological analysis

A 40 Hz low-pass (FIR) filter (12dB/octave) was applied to the EEG data off-line. Off-line re-referencing was performed with linked earlobes serving as the new reference in amplitude analyses, and average reference serving as the new reference in source localization. ERP segments (500 ms plus 100 ms pre-stimulus baseline) were time-locked to the presentation of X and Y stimuli in Phase 3, resulting in 48 ERP epochs for each condition. All epochs were visually inspected and those containing EOG, muscle, amplifier and other artifacts were removed. Individual data-sets containing less than 30 artifact-free epochs in either of the conditions were excluded from the analyses (1 participant was excluded in this way). The two conditions did not differ in the number of artifact-free epochs (X: mean 43.9, sd. 3.0; Y: mean 44.4, sd. 4.1; $t(19) = 0.87$, $p = 0.4$).

Since the early ERP components under scrutiny occur in immediate temporal vicinity of each other and are likely to show substantial overlap, temporal principal components analysis (PCA) was conducted on the ERPs in order to disentangle temporally overlapping ERP effects. The 250 time points were the variables in this analysis, and there were 2320 cases (20 participants by 2 conditions by 58 electrodes). Varimax rotation was employed and eigenvalue ≥ 1 was used as the PCA component identification criterion. This analysis results in a number of loading-by-time functions, which are statistically orthogonal components of the ERP amplitude-by-time functions from which they are derived. From the identified PCA components, we selected those whose loading-by-time function unambiguously corresponded to the amplitude-by-time function of the ERP components under investigation (e.g. AN1, N1 and SN). Linear regression was used to obtain factor scores for each PCA component (Donchin & Heffley,

1978). The scores of a given PCA component express its magnitude at each electrode/condition/subject. These PCA component scores were then averaged within 5 scalp regions in each hemisphere: frontal left (FP1, AF3, F1, F3, F5, F7), central left (FC1, FC3, FC5, C1, C3, C5), temporal left (T7, CP5, TP7, P7), parietal left (CP1, CP3, P1, P3, P5), parieto-occipital left (PO1, PO3, O1, PO7) and the corresponding symmetric regions/electrodes in the right hemisphere. Condition x Region x Hemisphere ANOVAs were run separately on the scores of PCA components corresponding to the ERP components of interest. Region-wise t-tests were performed subject to reliable ANOVA effects involving the Condition factor.

Cortical localization

Low-Resolution Electromagnetic Tomography (LORETA - Pascual-Marqui, 1999) was used for computing the 3-D intracerebral distribution of current density underlying observed scalp ERP effects. LORETA solves the inverse problem by assuming related strengths and orientations of sources (no assumption is made about their number). Mathematically, this is implemented by finding the smoothest of all possible activity distributions. The method has been extensively validated (e.g. Mulert et al., 2004) and is currently one of the most widely used source localization techniques in EEG. It has been used previously in cognitive ERP investigations (e.g. Lavric, Pizzagalli, & Forstmeier, 2004).

LORETA computes, at each voxel, current density as the linear weighted sum of the scalp electric potentials. The LORETA version used in the present study (Pascual-Marqui, 1999) was registered to the MNI305 brain atlas. The computations are restricted to cortical gray matter and hippocampi. The spatial resolution of the method is 7 mm and the solution space consists of 2394 voxels. Brodmann's area (BA) and region labels are provided by the LORETA software. For a specific MNI coordinate, LORETA first determines the nearest gray matter voxel using a lookup table created via the Talairach Daemon (Lancaster et al., 2000) and then estimates a conversion from MNI space to Talairach space using the transform method

suggested by Brett, Johnsrude and Owen (2002). LORETA solutions were first obtained for each participant in each condition and for each time point in the 25ms time windows defined on the basis of the observed waveform differences. Subsequently, time-points were averaged to obtain a solution for each condition and participant. These averaged solutions were then submitted to voxel-wise t-tests.

Results

Unless otherwise stated, all tests of statistical significance are assessed against an α of 0.05.

Behavioral results

In the final block of phase one, mean proportion of "fever" responses was 0.90 to A trials, 0.03 to B trials and 0.03 to I trials. The difference between A and B trials was significant, $t(19) = 18.92$, as was the difference between A and I trials, $t(19) = 18.92$. The difference between B and I did not approach significance, $t(19) = 0$.

In phase two (see Fig. 2), a two-factor repeated-measures ANOVA revealed that proportion of "fever" responses was significantly affected by trial type, $F(2,38) = 158.24$, and by trial block, $F(5, 95) = 25.19$. There was also a significant interaction between these two factors, $F(10, 190) = 40.15$. A Greenhouse-Geisser correction for non-sphericity was applied in this analysis, and in all subsequent analyses where it was appropriate to do so (uncorrected degrees of freedom are reported).

In phase three, mean proportion of "fever" responses was 0.45 to trial type X, and 0.72 to trial type Y. This difference was significant, $t(19) = 3.78$. Fifteen out of twenty participants made more "fever" responses to trial type Y than to trial type X. Mean error rates for the other trial types in this phase were, A: 4%, AX: 2%, BY: 9%, B: 18%. Mean reaction times were also slightly slower for trial type X (807ms) than for trial type Y (767 ms). This difference was

significant, t(19) = 2.34. Mean reaction times for the other trial types in this phase were, A: 705ms, AX: 835ms, BY: 889ms, B: 813ms. Across the experiment, 0.3% of trials were lost due to time-outs.

ERPs

Attentional associative theories of predictive learning predict that Y will attract more attention than X in phase three. These theories also make predictions about the amount of attention attracted by X and Y in phase two, but it would be extremely difficult to assess these differences through ERPs, for the reasons outlined earlier. The analyses reported below are therefore based around time-locked ERPs to the pictures of cell types Y and X presented in phase three. These stimuli were associated with differences in prediction error in phase two and the ensuing attentional differences were predicted to persist sufficiently into phase three to be detectable.

Inspection of the middle panel of Figure 3A, reveals that cue type Y was associated with a larger ERP amplitude than cue type X in the temporal range of the N1 component (155-180ms). However, the difference between the two cue types persists for about another 100 ms beyond the N1 peak, suggesting the presence of a Selection Negativity (SN). SN is most clearly visualized via a difference waveform, which is also shown in Figures 3A, middle panel. This difference waveform illustrates the presence of a SN between about 140ms and 290s post-stimulus onset. The two cue types also diverged in the anterior N1 component, with larger amplitude in response to cue type Y than cue type X (see Fig. 3A, top).

The temporal PCA, performed on ERPs to distinguish between overlapping ERP effects, found three PCA components whose time-courses corresponded well to AN1, N1 and SN, and which accounted for 5.8%, 3.8% and 9.6% of the variance, respectively (see Fig. 3A, bottom). The scores of these PCA components were analyzed via three separate ANOVAs, one for each of the three components (i.e. AN1, N1 and SN), with factors trial type (X vs. Y), region and hemisphere. Given the pre-existing knowledge of the circumscribed scalp distributions of AN1,

N1 and SN, each ANOVA involved just the appropriate subset of scalp regions – anterior regions (frontal and central) for AN1, and posterior regions (parietal and parietal-occipital) for N1 and SN.

For the PCA component corresponding to AN1 a reliable main effect of trial type was found, $F(1,19) = 5.89$; no other main effects or interactions were significant. The PCA component corresponding to N1 was analyzed in a corresponding manner, but no significant main effects or interactions were found. The PCA component corresponding to SN was also analyzed in a corresponding manner, revealing a significant interaction between trial type and hemisphere, $F(1, 19) = 7.03$, and a significant three-way interaction between trial type, region, and hemisphere, $F(1,19) = 8.66$ (no other main effects or interactions were significant). These interactions were explored by comparing trial type (X vs. Y) in each of the two regions (parietal and parietal-occipital) over each hemisphere, separately. The X vs. Y differences in the PCA component scores in the four regions were: 0.21 (parietal left), 0.33 (parietal right), 0.15 (parietal-occipital left), 0.41 (parietal-occipital right). The reliability of these four differences was assessed by t-tests; a reliable effect of trial type was found in the right parietal-occipital region, $t(19) = 2.19$.

Cortical localization (LORETA)

Topographic maps of the difference between ERPs to cue types X and Y across four different time windows are shown in Figure 3B. As can be seen, ERP differences appear to be at anterior scalp regions during the time window of the AN1 component, and at posterior scalp regions during the N1-SN time window. LORETA analysis provides a method of estimating the cortical locations of these differences. For this analysis (see Figure 3C), two 25 ms time-windows were set where the ERP differences were the largest (AN1 range, 110-135ms; N1-SN range, 155-180ms). In both time windows of interest, greater current density was found only for stimulus type Y. Applying a significance level of $p < 0.01$ (uncorrected) to voxel-by-voxel t-tests

revealed greater current density to Y than to X in the left inferior temporal region in the earlier time window, and in the left superior parietal region in the later time window.

Discussion

Consistent with our predictions, early ERP components, previously associated with selective attention, distinguished between cue types X and Y in phase three. Since cue types in our experiment can only be differentiated by shape, one would expect the ERP differences observed to be those that have previously been associated with selective attention to individual features of stimuli. The selection negativity (SN) is one such difference, and we found a significant SN for cue type Y relative to cue type X in this experiment. The SN we observed extended between 140ms and 280ms post-stimulus onset. Given the partial overlap of the observed SN with the posterior N1 peak, we examined and confirmed the reliability of the SN as a statistically independent component using temporal PCA. The presence of an independent posterior N1 difference in addition to the posterior selection negativity cannot be ruled out on the basis of these analyses, but no evidence of a reliable difference between the cue types in the posterior N1 was obtained, when its unique contribution was assessed via PCA.

In addition to the selection negativity observed at 140-280ms after stimulus onset, we observed an even earlier ERP effect – an AN1 (anterior N1) component with a higher amplitude for cue type Y than for cue type effect. This effect was observed at around 100-150ms post stimulus onset, and its reliability as an independent effect was confirmed via temporal PCA and ANOVA. It has previously been demonstrated that so-called "exogenous" components such as posterior and anterior N1 can be modulated by demand on visual discrimination, even where the stimuli being discriminated appear in the same spatial locations (Vogel & Luck, 2000). The difference observed in our experiment may therefore reflect enhanced visual discrimination of

cue type Y, possibly by sensory amplification of all or some of the features of this cue type, as soon as it begins to be differentiated from X by the perceptual system.

We believe the difference between the ERPs to images in sets X and Y is a consequence of the participant learning about the images' differing relationship to the prediction errors made in phase two. In phase two, participants predicted the outcome of AX trials throughout, whilst accurate prediction of the outcome of BY trials was acquired more slowly. The greater prediction error in BY trials compared to AX trials is assumed by attentional theories of associative learning to result in greater attention to Y images than to X images, which is also what the AN1 and SN differences between these two trial types indicate. The X and Y trials occurred with equal frequency, so relative novelty is not a confounding factor in this experiment. The specific "cell bodies" used in the X and Y stimulus sets were randomized across participants, so this difference in attention is unlikely to be due to differences in the basic perceptual properties of the X and Y sets.

On the basis of current knowledge of functional brain anatomy, it seems reasonable to suggest that the early ERP difference (AN1) reflects early attentional differentiation in perceptual identification areas and the associated sensory amplification/suppression, whereas the SN difference reflects the involvement of selective attention circuitry. The LORETA solutions, shown in Figure 3C, are consistent with this view. Although the results from LORETA contrasts only survive the uncorrected significance threshold and as such should be seen as exploratory, the foci they reveal have been previously documented in studies of the functional anatomy of visual selective attention.

In the earlier time window (110-135ms) we found more activity to Y images than to X images in the inferior temporal region, whilst in the later time window (155-180ms) we found more activity to Y images than to X images in posterior parietal cortex. Thus, the functional anatomy reveals the expected shift from differences in early object-identification regions to later differences in regions well-known to be implicated in selective visual attention (Kim et al., 1999;

Nobre, Gitelman, Dias, & Mseulam, 2000). Our cortical localizations of attentional differences also have precedents in the study of human predictive learning. Whilst Turner et al.'s (2004) analysis concentrated on a region of interest in prefrontal cortex defined by a previous study (Fletcher et al., 2001) they also presented certain differences in other regions, including posterior parietal regions. The current study supports these findings. In addition, it provides a detailed time-course of activity in these regions, and links them to scalp waveform components (SN and AN1) whose functional significance has been extensively investigated.

There are also some notable differences between our cortical localizations, and those reported in previous studies of predictive learning. For example, previous neuroimaging work on predictive learning emphasizes the role of the striatum (e.g. O'Doherty et al. , 2003). However, the absence of localizations in the striatum in our study is unsurprising. The sensitivity of EEG to deep brain regions such as the striatum is very limited (indeed, the striatum is not even included in the LORETA solution space). Of more interest is the fact that some other human studies (Fletcher et al., 2001; Turner et al., 2004) converge on the observation that lateral frontal cortex is involved in predictive learning. Although EEG and LORETA do detect current density changes originating in the lateral frontal cortex (e.g. Lavric, Pizzagalli, & Forstmeier, 2004), we did not find differences in this region in the analyzed 500 ms following the presentation of the stimulus. It is possible that the activation detected with fMRI reflects a modulation by prefrontal regions of visual attention circuitry. Such modulation is likely to have a more continuous (across trials) and slow character and is hence perhaps unlikely to be reflected in rapid event-locked potential changes such as the ones reported here.

Experiment Two

One limitation of Experiment one is that its demonstration of attentional differences is confined to phase three, by which time learning has been completed. The persistence of

attentional differences is predicted by certain attentional associative accounts (e.g. Mackintosh, 1975) and there is behavioral evidence that attentional persistence can indeed occur in human associative learning tasks (e.g. Lochmann & Wills, 2003). Nevertheless, in order to bridge the gap between prediction error differences and the learning in phase two and the observed attentional ERP effects, it is important to demonstrate the presence of attentional differences in phase two. Crucially, in order to provide such evidence one needs to disentangle the attention to the individual cues X and Y from the cues which appeared on the screen at the same time (A and B). This is critical because attentional differences between compounds AX and BY in phase two may be due to a simple novelty detection mechanism related to cues A and B: cue B changes its outcome relative to phase one while cue A does not. It would be extremely difficult to use ERPs to measure the correlates of attention to individual cues that appear in compounds. Therefore, we turned to a technique that can accomplish this relatively easily: eye-tracking. Previous evidence indicates that eye gaze can be used as a overt measure of attention in tasks of this type (e.g. Kruschke, Kappenman, & Hetrick, 2005; Rehder & Hoffman, 2005).

Experiment two employed very similar behavioral procedures to Experiment One - the only difference being that the stimuli were enlarged and positioned further apart to facilitate effective eye-tracking (in Experiment One stimuli were small and tightly positioned to minimize eye-movement artifacts in the EEG).

Method

Participants

Sixteen students and staff (10 females; age: mean 25.88, s.d. 8.82, range 21-56) took part in Experiment Two. Each participant was paid 4 GBP for a 50 minute experimental session. None of the participants had taken part in Experiment One.

Apparatus

Stimulus presentation and response collection was via a PC-compatible computer and the E-prime package (Version 1.1, Psychology Software Tools, Pittsburgh, USA). Eye movements were recorded using an EyeLink II system (SR Research Ltd., Osgoode**,** Canada), a video based eye-tracker with head movement compensation system. The sampling rate was 500Hz. Pupil position was monitored (right eye only) via a miniature infrared CCD video camera mounted on an adjustable headband. Participants were instructed to keep head movements to a minimum and no active restraint of head movements was required to obtain sufficiently accurate gaze position recordings. The stimulus presentation PC initiated and terminated eye tracking recording blocks on each experimental trial via a TTL interface box connected to the eye-tracker PC.

Stimuli

The stimuli were enlarged relative to Experiment One, but were otherwise identical to those used in that experiment. Pictures subtended 6.2 degrees of visual angle and, where two pictures were presented at the same time, one was presented 7.7 degrees of visual angle above the center and the other at an equal distance below the center. There was no white outline square in Experiment Two.

Procedure

Eye movements were recorded during phase one and two of Experiment Two[2]. The experimental procedure was identical to that employed in Experiment One, with the following exceptions. In order to correct for drift in eye movement position accuracy, experimenter-controlled drift corrections were performed at 24-trial intervals. These drift corrections comprised of a brief 1.5 second message telling participants to “focus on the cross in the center of the screen”, followed by a fixation cross. The offset of this cross was controlled by the experimenter, who manually initiated the next trial once the drift correction had been performed by the eye-tracking software.

The time taken to perform this offset correction procedure varied across trials and participants between approximately 3 and 5 seconds.

Eye movement analysis

Eye movements were viewed and analyzed offline using the EyeLink DataViewer software. This software automatically detects saccadic eye movements and parses the eye movement data into individual fixations using a combined position / velocity / acceleration criterion (a saccade was defined as a period where eye velocity was greater than 30 degrees/sec, eye acceleration was greater than 8000 degrees / sec and the eye had deviated at least 0.1 degree from its starting position. Fixations were defined as periods between saccades. Blink artifacts were automatically removed from the data by the DataViewer software. The mean position, duration and number of fixations in each stimulus "region of interest" on each trial were outputted from the software for further statistical analysis. Regions of interest in this experiment were predefined as 210 x 210 pixel squares corresponding to the size and positions occupied by the picture stimuli. The total viewing or "dwell" time for each of these regions of interest was calculated for this data for each trial and stimulus of interest (i.e. sum of all individual fixation durations for each stimulus on each trial).

Results

All tests of statistical significance were assessed against an α of 0.05.

Behavioral results

The behavioral results were basically equivalent to those found in Experiment one. In the final block of phase one, the mean proportion of "fever" responses was 0.91 to A trials, 0.05 to B trials and 0.11 to I trials. The difference between A and B trials was significant, $t(15) = 7.582$, as

was the difference between A and I trials, t(15) = 7.931. The difference between B and I did not approach significance, t(15) = 0.103. In phase two, a two-factor repeated-measures ANOVA revealed that proportion of "fever" responses was significantly affected by trial type, F(2, 30) = 87.087, and by trial block, F(5, 75) = 29.922. There was also a significant interaction between these two factors, F(10, 150) = 14.150.

The mean response latencies were also analyzed for each trial type in phase two. AX+ trials had a mean latency of 1408 ms and 973 ms, BY+ trials were 1525 ms and 1019 ms and IJ- trials were 1480 ms and 1153 ms for the first and last blocks respectively. A two-factor (trial type, 3 levels; trial block, 6 levels) repeated measures ANOVA was applied to these data and revealed two main effects. There was significant effect of trial type, F(2, 30) = 12.450, and a significant effect of trial block, F(5, 75) = 22.360. The interaction between these two factors was not significant, F(10, 150)=1.300.

In phase three, the mean proportion of "fever" responses was 0.50 to trial type X and 0.77 to trial type Y. Thirteen out of sixteen participants made more "fever" responses to trial type Y than to trial type X. Mean error rates for the other trial types in phase three were, A: 11%, AX: 5%, BY: 15%, B: 22%. A two factor (trial type, 2 levels; trial block, 6 levels) repeated measures ANOVA was applied to the proportion of "fever" responses to X and Y stimuli in phase three and yielded a significant different between these two trial types, F(1, 15) = 11.719. ANOVA found no significant effect of trial block, F(5, 75)=2.437,and the trial block by trial type interaction was not significant either, F(5, 75)=1.576. A further two factor (trial type, 2 levels; trial block, 6 levels) repeated measures ANOVA applied to the response latencies for X and Y in phase three revealed a significant effect of trial block, with response becoming slightly faster for both trial types as phase three progressed, F(5,75) = 5.570. There was no significant difference between the trial types, F(1, 15)=0.030; mean response latency was 832ms to X and 829ms to Y. The interaction between trial type and trial block was not significant, F(1, 15)=2.194.

Eye-tracking results

Mean total dwell times (see *Method*) for the critical stimuli in phase two were calculated. The mean dwell times for X and Y across the phase were 337 ms and 435 ms per trial respectively. A two factor (trial type, 2 levels; trial block, 6 levels) repeated measures ANOVA revealed a significant difference between X and Y, $F(1, 15) = 9.346$. There was also a significant effect of trial block, $F(5, 75) = 13.528$, with dwell times decreasing as phase two progressed. The block by trial type interaction was not significant, $F(5, 75)=1.088$.

As manual response times were found to vary as a function of block in the experiment (see above), dwell times on regions of interest were also calculated as a percentage of total viewing time for each trial. For each AX trial, percentage dwell time for X was calculated as $100x / (a+x)$, where x was the dwell time in region of interest for stimulus X, and a was the dwell time in the region of interest for stimulus A. For each BY trial, percentage dwell time for Y was calculated in the corresponding manner: $100y / (y + b)$. A two factor (trial type, 2 levels; trial block, 6 levels) repeated measures ANOVA revealed a significant difference between percentage dwell time on X and Y, $F(1,15)=9.69$. Mean percentage dwell time was 37% for X, and 46% for Y. The effect of block did not approach significance, $F(5, 75) = 0.44$. The interaction between trial type and block was not significant, $F(5,75)=1.98$, $0.10 > p > 0.05$, although the trend was towards a divergence of percentage dwell times as phase two progressed.

A further analysis was carried out to test the hypothesis that what is learned about X and Y (as indexed by the proportion of "fever" responses to each in phase three) should be linked to the amount of attention directed to each during learning (as indexed by the mean dwell times for X and Y in phase two). If this hypothesis is correct a relationship should be apparent between the mean differences in dwell times for X and Y in phase two and the mean differences in proportion of "fever" responses in phase three for X and Y. Consistent with the hypothesis this analysis revealed significant positive correlation between these two variables, $r(16) = 0.526$.

Discussion

The objective of Experiment Two was to examine the eye-movement correlates of attention to cue types X and Y in phase two, separately from cue types with which they were paired (B and Y). Mean and percentage viewing (dwell) times were used as measures of attention to cues. In accordance with attentional accounts of predictive learning, we expected that early in phase two participants would dedicate more time to viewing the cue associated with larger prediction errors (Y), compared to the cue that generated smaller prediction errors (X). The results unequivocally support this prediction: both mean and percentage dwell times were reliably higher for cue type Y than cue type X. This difference in dwell times did not change significantly across phase two, suggesting that the difference arose early on in phase two.

General Discussion

The present studies examined the role of attention in predictive learning through the measurement of brain potentials (Experiment One) and eye movements (Experiment Two). The outcomes from both procedures are consistent with the idea that the amount of attentional resources allocated to a cue is positively related to the size of the prediction error it has previously produced. First, ERP components that have been implicated in selective attention were found to have larger amplitudes in response to cue types previously associated with larger prediction errors. The cortical origins of these differences in scalp-recorded ERPs (as estimated by low-resolution electromagnetic tomography) were found to be in areas closely associated with object recognition and visual attention. Second, participants dedicated more time to viewing cues that generated larger prediction errors.

The relationship between attention and prediction errors reported in these studies was predicted on the basis of certain attentional associative theories (Kruschke, 2001; Pearce & Hall,

1980), and the presence of the effects we report is consistent with such theories. Inevitably, other types of theory can also accommodate our findings if one allows the introduction of additional assumptions into those theories. The nature of these assumptions, and their implications for future research, are discussed below.

Higher-order reasoning theories of predictive learning (De Houwer, Beckers, & Vandorpe, 2005), more or less by definition, do not invoke early attentional differences in their explanations of how phenomena such as forward cue competition occur. Nevertheless, the attentional effects we have observed could be incorporated within such accounts via the assumption that attentional differences are the top-down product of high-level reasoning. Indeed, such an argument has recently been forwarded by De Houwer, Beckers and Vandorpe (2005) and could be seen as deriving some support from the prefrontal cortex activation observed in some fMRI studies of predictive learning (e.g. Turner et al., 2004). We did not observe prefrontal involvement in our studies, but this may be due to the tightly trial-locked and time-specific nature of the ERP methodology we employed. The question of whether the attentional differences we observe are the top-down result of high-level reasoning processes or the result of the lower-level, automatic processes that are sometimes assumed to be implied by associative accounts, is an important topic for future research.

Non-attentional associative learning theories (e.g. Rescorla & Wagner, 1972) could also accommodate our results by the introduction of certain assumptions. Specifically, such accounts could postulate that cues with high associative strength attract more attention than cues of low associative strength (cue Y is more likely to produce a "fever" response than cue X in our experiments, so such associative theories would predict that cue Y has higher associative strength). Whilst the introduction of attentional processes into non-attentional theories might, at first sight, appear to render the two types of theory equivalent, the proposal is interesting in the sense that it appears to make opposite predictions to Kruschke-Mackintosh and Pearce-Hall attentional theories in certain situations. Consider, for example, a slightly modified design in

which AX and BY predict the absence of fever in phase two. In such a design, non-attentional associative theories would predict that the associative strength of Y would end up being higher than the associative strength of X (which would be negative, in order to prevent the prediction of an outcome on the basis of A). Proponents of non-attentional accounts have previously argued that activation should be higher for Y than for X in this design, and such an effect has been observed for dopamine neurons in an animal model (Tobler, Dickinson, & Schultz, 2003). In contrast, attentional accounts would predict the opposite result – X should attract more attention than Y because X will have been involved in more prediction errors in phase two than Y (because participants will initially and incorrectly predict an outcome on AX trials).

In summary, the current study provides detailed insights into the electrophysiological correlates (temporal and anatomical), and the oculomotor correlates, of human associative learning. The more prediction errors an event has been involved in, the greater the early attentional resources that are directed towards it. The production and function of this attentional differentiation is a matter for further research.

References

Ashby, F. G., Alfonso-Reese, L. A., Turken, A. U., & Waldron, E. M. (1998). A neuropsychological theory of multiple systems in category learning. *Psychological Review, 105*(3), 442-481.

Brett, M., Johnsrude, I. S., & Owen, A. M. (2002). The problem of functional localization in the human brain. *Nature Reviews Neuroscience, 3*, 243-249.

Clark, V. P., & Hillyard, S. A. (1996). Spatial selective attention affects early extrastriate but not striate components of the visual evoked potential. *Journal of Cognitive Neuroscience, 8*, 387-402.

De Houwer, J., & Beckers, T. (2002). Higher-order retrospective revaluation in human causal learning. *Quarterly Journal of Experimental Psychology, 55B*(2), 137-151.

De Houwer, J., Beckers, T., & Vandorpe, S. N. (2005). Evidence for the role of higher order reasoning processes in cue competition and other learning phenomena. *Learning & Behavior, 33*(2), 239-249.

Donchin, E., & Heffley, E. F. (1978). Multivariate analysis of event-related potential data: a tutorial review. In D. Otto (Ed.), *Multidisciplinary Perspectives in Event-Related Brain Potential Research* (pp. 555-572). Washington, DC: Government Printing Office.

Erickson, M. A., & Kruschke, J. K. (1998). Rules and exemplars in category learning. *Journal Of Experimental Psychology: General, 127*(2), 107-140.

Fletcher, P. C., Anderson, J. M., Shanks, D. R., Honey, R., Carpenter, T. A., Donovan, T., et al. (2001). Responses of human frontal cortex to surprising events are predicted by formal associative learning theory. *Nature neuroscience, 4*(10), 1043-1048.

Glautier, S. (2002). Spatial separation of target and competitor cues enhances blocking of human causality judgements. *Quarterly Journal of Experimental Psychology, 55B*(2), 121-135.

Hillyard, S. A., & Anllo-Vento, L. (1998). Event-related brain potentials in the study of visual selective attention. *Proc. Natl. Acad. Sci, 95*, 781-787.

Holroyd, C. B., Nieuwenhuis, S., Yeung, N., & Cohen, J. D. (2003). Errors in reward prediction are reflected in the event-related brain potential. *NeuroReport, 14*(18), 2481-2484.

Kamin, L. J. (1969). 'Attention-like' processes in classical conditioning. In M. R. Jones (Ed.), *Miami symposium on the prediction of behavior: aversive stimulation*: University of Miami Press.

Kim, Y.-H., Gitelman, D. R., Nobre, A. C., Parris, T. B., LaBar, K. S., & Mseulam, M. M. (1999). The large-scale neural network for spatial attention displays multifunctional overlap but differential asymmetry. *Neuroimage, 9*, 269-277.

Kruschke, J. K. (2001). Toward a unified model of attention in associative learning. *Journal of Mathematical Psychology, 45*, 812-863.

Kruschke, J. K., Kappenman, E. S., & Hetrick, W. P. (2005). Eye gaze and individual differences consistent with learned attention in associative blocking and highlighting. *Journal of Experimental Psychology-Learning Memory and Cognition, 31*(5), 830-845.

Lancaster, J. L., Woldroff, M. G., Parsons, L. M., Liotti, M., Freitas, C. S., Rainey, L., et al. (2000). Automated Talairach atlas labels for functional brain mapping. *Human Brain Mapping, 10*, 120-131.

Lavric, A., Pizzagalli, D., & Forstmeier, S. (2004). When 'go' and 'nogo' are equally frequent: ERP components and cortical tomography. *European Journal of Neuroscience, 20*, 2483-2488.

Lawrence, D. H. (1952). The transfer of a discrimination along a continuum. *Journal of Comparative and Physiological Psychology, 45*, 511-516.

Le Pelley, M. E., Oakeshott, S. M., & McLaren, I. P. L. (2005). Blocking and unblocking in human causal learning. *Journal of Experimental Psychology: Animal Behavior Processes, 31*(1), 56-70.

Lochmann, T., & Wills, A. J. (2003). Predictive history in an allergy prediction task. In F. Schmalhofer, R. M. Young & G. Katz (Eds.), *Proceedings of EuroCogSci 03: The European Conference of the Cognitive Science Society* (pp. 217-222). Mahwah, NJ: Lawrence Erlbaum Associates.

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

Mulert, C., Jäger, L., Schmitt, R., Bussfeld, P., Pogarell, O., Möller, H.-J., et al. (2004). Integration of fMRI and simultaneous EEG: towards a comprehensive understanding of localization and time-course of brain activity in target detection. *Neuroimage, 22*, 83-94.

Nobre, A. C., Gitelman, D. R., Dias, E. C., & Mseulam, M. M. (2000). Covert visual spatial orienting and saccades: Overlapping neural systems. *Neuroimage, 11*, 210-216.

O'Doherty, J., Dayan, P., Friston, K., Critchley, H., & Dolan, R. (2003). Temporal difference models and reward-related learning in the human brain. *Neuron, 28*, 329-337.

Pascual-Marqui, R. D., Michel, C. M., & Lehmann, D. (1995). Segmentation of brain electrical activity into microstates: model estimation and validation. *IEEE Transactions on Biomedical Engineering, 42*, 658-665.

Pascual-Marqui, R. D. (1999). Review of methods for solving the EEG inverse problem. *International journal of bioelectromagnetism, 1*, 75-86.

Pavlov, I. P. (1927). *Conditioned reflexes*. New York: Dover.

Pearce, J. M., & Hall, G. (1980). A model for Pavlovian learning: Variations in the effectiveness of conditioned but not of unconditioned stimuli. *Psychological Review, 87*, 532-552.

Ploghaus, A., Tracey, I., Clare, S., Gati, J., Rawlins, J. N. P., & Matthews, P. M. (2000). Learning about pain: The neural substrate of the prediction error for aversive events. *Proceedings of the National Academy of Sciences, 97*(16), 9281-9286.

Rehder, B., & Hoffman, A. B. (2005). Eyetracking and selective attention in category learning. *Cognitive Psychology, 51*(1), 1-41.

Rescorla, R. A., & Wagner, A. R. (1972). A theory of Pavlovian conditioning: Variations in the effectiveness of reinforcement and nonreinforcement. In A. H. Black & W. F. Prokasy (Eds.), *Classical conditioning II: Current research* (pp. 64 - 99). New York: Appleton-Century-Crofts.

Schultz, W., Dayan, P., & Montague, P. R. (1997). A neural substrate of prediction and reward. *Science, 275*, 1593-1599.

Tobler, P. N., Dickinson, A., & Schultz, W. (2003). Coding of predicted reward omission by dopamine neurons in a conditioned inhibition paradigm. *The Journal of Neuroscience, 23*(32), 10402-10410.

Turner, D. C., Aitken, M. R. F., Shanks, D. R., Sahakian, B. J., Robbins, T. W., Schwarzbauer, C., et al. (2004). The role of lateral frontal cortex in causal associative learning: Exploring preventative and super-learning. *Cerebral Cortex, 14*(8), 872-880.

Vogel, E. K., & Luck, S. J. (2000). The visual N1 component as an index of a discrimination process. *Psychophysiology, 37*, 190-203.

Wills, A. J., & McLaren, I. P. L. (1997). Generalization in human category learning: A connectionist explanation of differences in gradient after discriminative and non-discriminative training. *The Quarterly Journal of Experimental Psychology, 50A*(3), 607-630.

Notes

[1] In Figure 3A, there is a difference between conditions Y and X in electrode FCz in a limited sub-stretch of the baseline just preceding the stimulus onset. Every presentation of pictures of fictitious cells X or Y was preceded by a 1000 ms presentation of the empty square, in which cell X or Y was subsequently placed. Thus, the preceding stimulus was constant. To assess whether there were any reliable differences anywhere in the baseline (including the stretch in question), we performed statistical comparisons of the baselines for X and Y (following baseline correction). We used a robust procedure, ideally suited for the identification of global differences across ranges of time-points: TANOVA (Pasqual-Marqui et al., 1995; see also http://www.unizh.ch/keyinst/NewLORETA/LORETA01.htm). TANOVA compares the ERPs time-point-by-time-point and identifies time-points showing significant differences, while controlling for alpha inflation in multiple tests by permutations. No time-points in the baseline showed reliable differences across conditions, including the time-points in the range under scrutiny (in this range all p values were >.4). Incidentally, when run on the post-stimulus onset

ERPs, TANOVA did find series of time-points showing significantly differences in the N1-SN as well as AN1 ranges.

[2] Eye movements were not recorded during phase three. Eye-tracking data from the target trials in phase 3 (X and Y) would have been of limited use, as each stimulus had but a single component. Measures such as dwell time to that single component add relatively little to the information already available from the participants' behavioral reaction times. Participant comfort was also an issue – our apparatus was too uncomfortable to be worn for the whole length of this fairly long (45 minute) experiment.

Author Note

This research was supported by a BBSRC grant 9/S17109, and EC Framework 6 project grant 516542 (NEST) to the first author. The authors would like to thank Jan De Houwer, and two anonymous reviewers, for their helpful comments. Related research can be found at www.willslab.co.uk

Figure Captions

Figure 1. Trial structure.

Figure 2. Proportion of "fever" responses made across phase two of the experiment, shown for AX+ (with solid circles), BY+ (with solid squares) and IJ- (with solid diamonds) trial types.

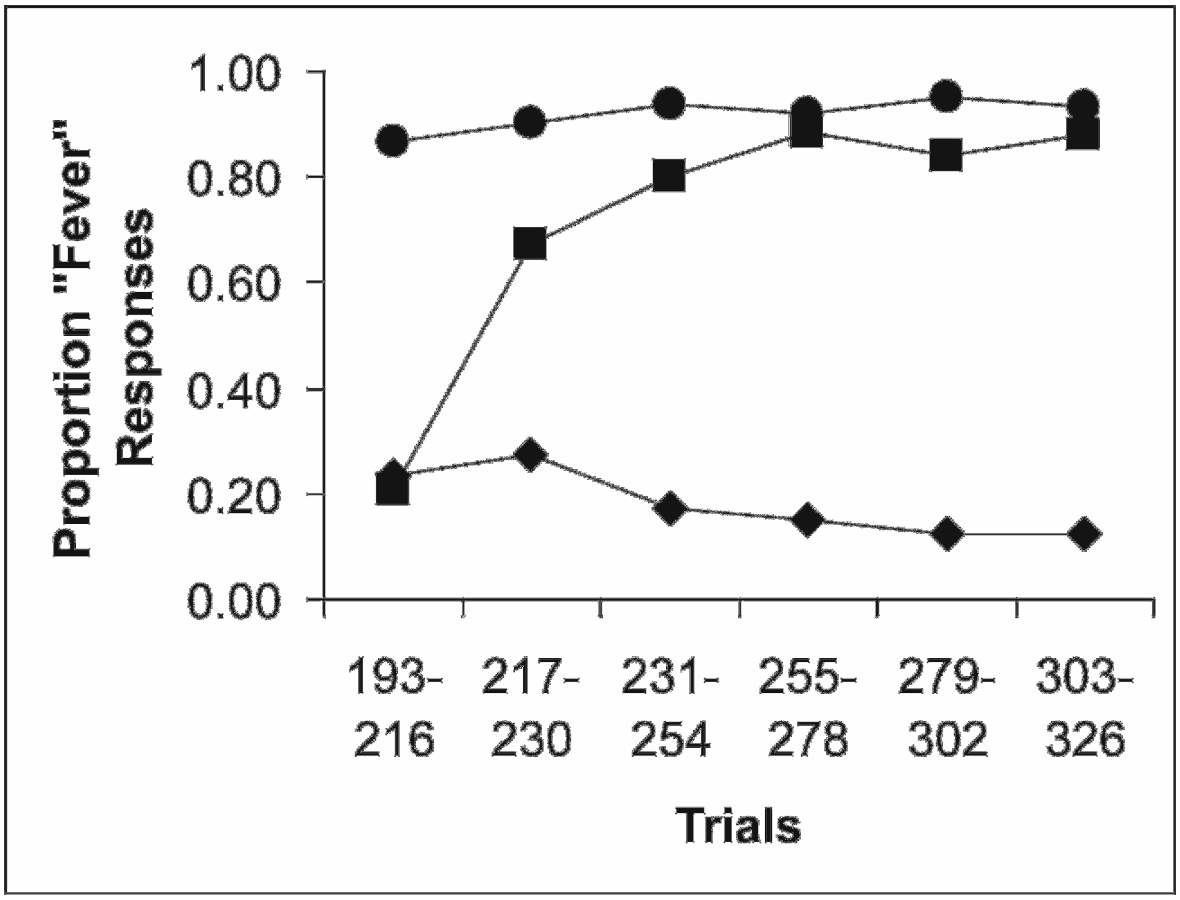
Figure 3. **(A).** The AN1 component at a anterior-central (FCz) electrode[1] and N1 and SN at a parieto-occipital electrode (PO4). Below, the raw loadings from the temporal PCA are displayed. The PCA components corresponding to the ERP components of interest (AN1, N1 and SN) are shown in bold ; PCA components that differentiated reliably in the statistical analysis between cue types Y and X are shown with solid lines (they correspond to ERP components AN1 and SN). **(B).** Topographic maps of the difference between ERPs to cue types Y and X. **(C).** Voxel-by-voxel LORETA t-tests comparing X and Y in the two time-windows of interest; t-values thresholded at $p<0.01$, uncorrected.

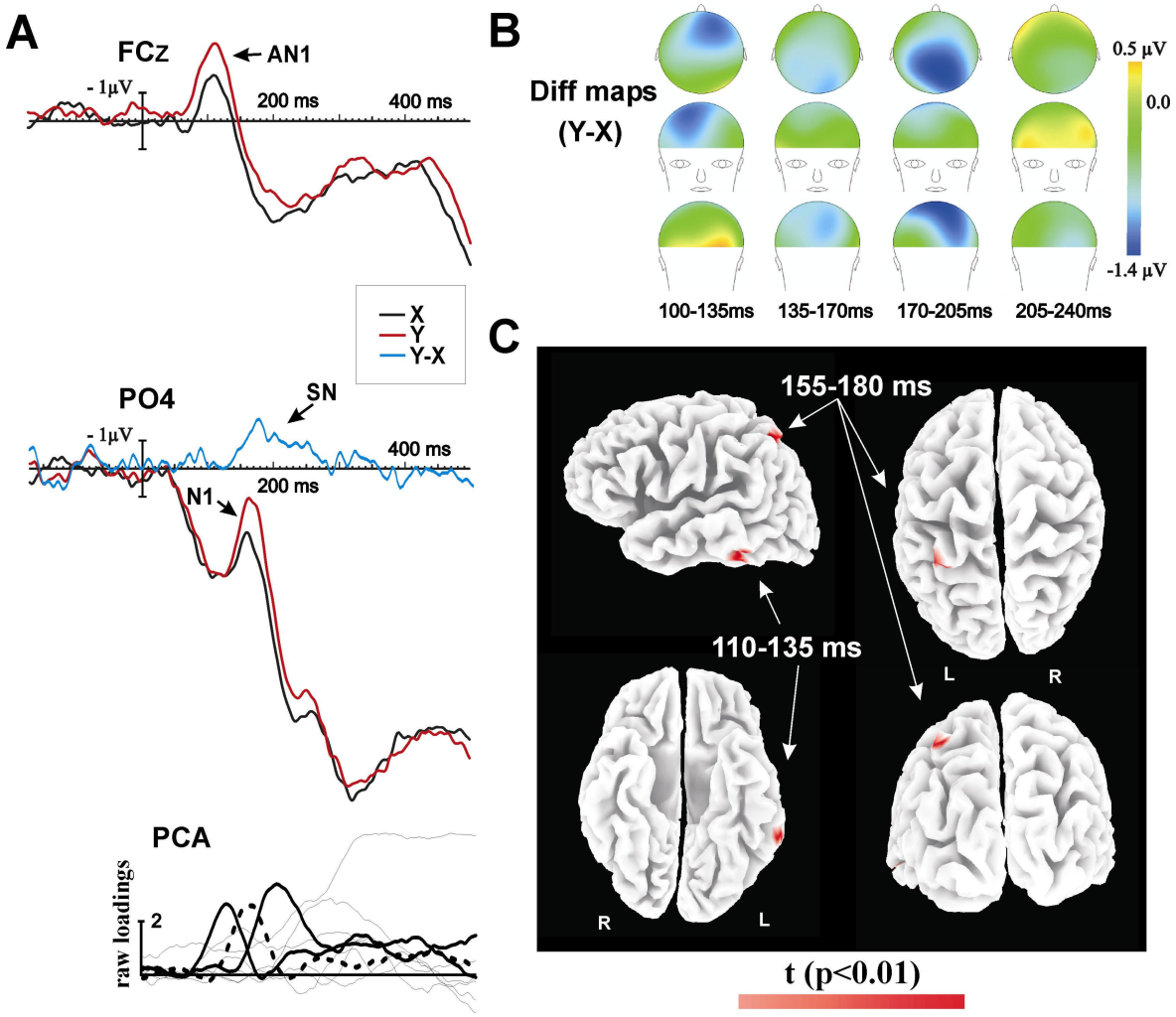
JoCN only accepts tables and figures as separate files. I've pasted these files into this PDF.

Phase 1	Phase 2	Phase 3
A → fever (A+)	AX → fever (AX+)	X → Data missing (X?)
B → no fever (B-)	BY → fever (BY+)	Y → Data missing (Y?)
I → no fever (I-)	IJ → no fever (IJ-)	A → fever (A+)
		B → no fever (B-)
		AX → fever (AX+)
		BY → fever (BY+)

Table 1. Structure of the learning task. Letters represent the abstract forms used as stimuli. Conventional learning theory notations for each trial type are presented in parentheses.


1 second
RESPONSE
Correct!
Jominy fever
1.5 seconds


Proportion "Fever" Responses
1.00
0.80
0.60
0.40
0.20
0.00
193-216
217-230
231-254
255-278
279-302
303-326
Trials


A
FCz
AN1
- 1μV
200 ms
400 ms
X
Y
Y-X
PO4
SN
- 1μV
400 ms
200 ms
N1
PCA
raw loadings
2
B
Diff maps
(Y-X)
0.5 μV
0.0
-1.4 μV
100-135ms
135-170ms
170-205ms
205-240ms
C
155-180 ms
110-135 ms
L
R
R
L
t (p<0.01)