ACCEPTED —JOURNAL OF COGNITIVE NEUROSCIENCE— September 2011

Feedback-related Brain Potential Activity Complies with Basic Assumptions of Associative Learning Theory

David Luque[1], Francisco J. López[1], Josep Marco-Pallares[2], Estela Càmara[2,3] and Antoni Rodríguez-Fornells[2,4]

[1]Department of Basic Psychology, University of Málaga. [2]Department of Physiology, University of Barcelona. [3]Department of Neuropsychology, Otto-von Guericke University. [4]Institució Catalana de Recerca i Estudis Avançats (ICREA).

Author Note

Acknowledgments. The present study was supported by research grants from Junta de Andalucía (SEJ-406 and P08-SEJ-03586) and the Ministerio de Educación y Ciencia (SEJ2007-63691/PSIC to FJ. Lopez and PSI2008-03901/PSIC to ARF). The authors would like to thank Pedro L. Cobos, Toni Cunillera, Lluis Fuentemilla, Joaquín Morís and David Cucurell for their advice and support in the completion of this research.

Correspondence concerning this article should be addressed to David Luque, Departamento de Psicología Básica (Facultad de Psicología), Universidad de Málaga, Campus de Teatinos, s/n. 29071, Málaga (Spain). E-mail: david.luque@gmail.com. Phone number: +34 952132630/ Fax: +34 952132631

Abstract

Feedback-related negativity (FRN) is an event-related potential (ERP) component that distinguishes positive from negative feedback. FRN has been hypothesized to be the product of an error signal that may be used to adjust future behavior. In addition, associative learning models assume that the trial-to-trial learning of cue-outcome mappings involves the minimization of an error term. The present study evaluated whether FRN is a possible electrophysiological correlate of this error term in a predictive learning task where human subjects were asked to learn different cue-outcome relationships. Specifically, we evaluated the sensitivity of the FRN to the course of learning when different stimuli interact or compete to become a predictor of certain outcomes. Importantly, some of these cues were *blocked* by more informative or *predictive* cues (i.e., the blocking effect). Interestingly, the present results show that both learning and blocking affect the amplitude of the FRN component. Furthermore, independent analyses of positive and negative feedback event-related signals showed that the learning effect was restricted to the ERP component elicited by positive feedback. The blocking test showed differences in the FRN magnitude between a predictive and a blocked cue. Overall, the present results show that ERPs that are related to feedback processing correspond to the main predictions of associative learning models.

Keywords: anterior cingulate cortex, blocking, dopamine, feedback-related negativity, learning

Feedback-related Brain Potential Activity Complies with Basic Assumptions of Associative Learning Theory

Learning to predict future events or outcomes from current cues represents an extremely important behavioral adaptation in both human and animals. The specific mechanisms underlying predictive learning have been the subject of considerable research. There are disputes in the literature, however, and associative learning mechanisms have been compared with more cognitive reasoning mechanisms (Shanks, 2010). The present study was designed to investigate the neurophysiological correlates and advance our understanding of the mechanisms underlying human predictive learning. Using event-related brain potentials (ERPs), the feedback-related negativity (FRN) component was recorded in a 'cue-interaction' situation where different stimuli or cues interact or compete to become a predictor of the outcome. This component has been associated with the processing of negative feedback (Gehring & Willoughby, 2002) and reinforcement learning (Holroyd & Coles, 2002). Thus, the modulation of the FRN observed during learning was directly evaluated in terms of the predictions derived from associative learning mechanisms.

Cue-interaction effects are phenomena that occur when the successful learning of cue-outcome relationships depends not only on the contingency or statistical relationship between the cue and the outcome but also on the contingency between alternative cues that are present in the situation and the outcome (see Kamin, 1968 and Dickinson, Shanks, & Evenden, 1984 for seminal demonstrations of these effects in animal and human learning, respectively). For example, in the 'blocking effect', if there are other cues in the situation that already predict an outcome, a cue will be considered to be weakly related to this outcome regardless of the number of cue-outcome pairings. Imagine that an allergist is examining the case stories of fictitious patients, which is a common scenario for participants in predictive learning studies, and learns that eating grapes consistently predicts an allergic reaction in one of the patients. If this highly predictive cue is presented on other occasions together with a new cue (e.g., eating

bananas) and this ‘compound cue’ (i.e., eating grapes and bananas) also predicts the outcome, it is unlikely that the new cue will be considered to be a true predictor of the allergic reaction regardless of the number of parings between the added cue and this allergic reaction. These cue-interaction effects originally prompted the development of associative theories, such as the Rescorla and Wagner (RW) model (1972).

The core of associative theories of predictive learning, including RW and its real-time extensions (temporal difference models, Sutton & Barto, 1981), is the assumption that the learning mechanism operates on the basis of the computation of an error signal. This error signal is conceived as the difference between the predicted and the actual outcome. The operation of the model across trials ensures that any deviation between the predicted and actual outcome (i.e., the error term) is used to adjust the strength of the associative link between the cue and the outcome. This process will continue until the error signal converges with zero (i.e., the error signal gradually diminishes until the expectation matches the outcome). In cue-interaction situations, there is already a cue that reliably predicts the outcome, and no further error signal is available to promote new learning. Thus, no other cue will become strongly associated with a particular outcome in the presence of a predictive cue. In addition, the strength of the associative link between the cue and the outcome reflects the predictive validity of the cue regarding that particular outcome relative to other cues that are present.

If associative mechanisms are a tenable explanation of predictive learning-including cue interaction phenomena- then, there should be neurophysiological mechanisms similar to those envisioned by associated models (e.g., RW) that compute error signals. Numerous studies exist that support this theory. The most direct evidence comes from studies that used single-cell electrophysiological recording techniques. Schultz and colleagues (e.g., Schultz, 2002; Waelti, Dickinson, & Schultz, 2001) showed that midbrain dopaminergic neurons display a short-latency phasic signal that can be interpreted to act as an error signal in nonhuman primates. The results of Waelti et al. (2001)’s study showed that the activity of

dopaminergic neurons diminished across trials as the outcome begins to be predicted (i.e., just as the error term predicted by associative models converges towards zero). Similar to a situation where the outcome is unpredicted, the activity of dopaminergic neurons in a cue-interaction design was significantly increased in the presence of an outcome that had been presented after a 'blocked' cue (i.e., the 'blocking effect' predicted by associative models).

Another relevant source of evidence regarding the neural basis of error-driven predictive learning in humans comes from functional magnetic resonance imaging (fMRI) studies. The most relevant evidence comes from studies where participants had the opportunity to learn cue-outcome predictive relationships from trial-by-trial feedback. Again, the results from fMRI studies (e.g., Fletcher et al., 2001; O'Doherty, Dayan, Friston, Critchley, & Dolan, 2003; Tobler, O'Doherty, Dolan, & Schultz, 2006; Turner et al., 2004) have revealed specific activities that may be related to the processing of an error signal. Specifically, Tobler et al. (2006) and Turner et al. (2004) have shown a greater activation of the orbitofrontal cortex area when participants experienced unpredicted rather than well-predicted outcomes (i.e., outcomes following a blocked rather than a predictive cue). Anatomically, this error-learning signal might be sustained in the human midbrain and conveyed through corticostriatal-midbrain circuits [see recent revisions in Càmara, Rodríguez-Fornells, Ye, & Münte (2009) and Haber & Knutsen (2010)].

Noninvasive electrophysiological recordings in humans (using ERPs and time-frequency analysis) have also provided information about the neural basis of error detection and feedback processing. For example, Wills, Lavric, Croft, and Hodgson (2007) used a cue-interaction design and showed that cues leading to prediction errors attract more attention than cues for which initial predictions are correct. Early attentional differentiation of cues that differ in their previous involvement in errors has been observed in ERPs implicated in selective visual attention, such as selection negativity and augmented anterior N1 (Clark & Hillyard,

1996; Vogel & Luck, 2000). These results suggest that error processing in human predictive learning involves early attentional processes.

FRN is a specific ERP component that is a likely neural correlate of the error term proposed in RW and similar learning models. FRN is a sharp negative deflection with a frontocentral distribution that peaks approximately 250 ms after negative feedback (i.e., a stimulus signaling that an incorrect response has been made; Gehring & Willoughby, 2002). In their *reinforcement learning theory*, Holroyd and Coles (2002) originally claimed that FRN reflected a mismatch between the expected and actual outcome, which was highly correlated with the error term described in associative theories. Thus, FRN could be a cortical electrophysiological correlate that indexed information similar to the information recorded in fMRI studies. FRN activity is greater when an unexpected outcome occurs and progressively weakens as an associated cue makes the outcome predictable. Thus, FRN decreases in magnitude as learning progresses (Cohen, Elger, & Ranganath, 2007; Eppinger, Kray, Mock, & Mecklinger, 2008; Holroyd & Coles, 2002; Müller, Möller, Rodríguez-Fornells, & Münte, 2005). *Reinforcement learning theory* suggests that the anterior cingulate cortex (ACC) receives error signal input from the midbrain dopaminergic reward system to adjust future actions that are compatible with the proposed neural circuit implementing this error signal. Indeed, numerous studies involving both human and animals (see Holroyd & Coles, 2002 for a review) have identified the ACC as a relevant locus for action planning and execution.

The main objective of the present study was to evaluate whether FRN is a possible electrophysiological correlate of the associative error term in a predictive learning task (i.e., a task in which different cue-outcome relationships had to be learned). To accomplish this, we evaluated the course of learning throughout a predictive learning task and determined whether the FRN was sensitive to cue-interaction manipulations.

The experimental procedure that we used allowed us to evaluate whether FRN computed an error term that complied with the basic assumptions of associative learning theories in two independent ways. First, we analyzed how FRN evolved over the course of learning. Because associative learning predicts that the error term will decrease as the subject learns, we expected FRN to decrease in the same way. Previous studies have also analyzed changes in FRN during learning and found that FRN decreases with learning (e.g., Holroyd & Coles, 2002 or Eppinger et al., 2008). It is important to note that the learning task in the present study differed from previous studies in which participants were asked to evaluate arbitrary cue-outcome relationships. In the present study, we used an experimental paradigm (i.e., the allergy task) that has been widely used in the human predictive learning literature (e.g., Dickinson & Burke, 1996; Matute, Arcediano, & Miller, 1996; Shanks & López, 1996; Vadillo, Castro, Matute, & Wasserman, 2008; Van Hamme & Wasserman, 1994).

The present study was the first to evaluate whether FRN is sensitive to cue-interaction manipulations. Following the rationale of Waelti et al. (2001), the present study employed a cue-interaction phenomenon to determine whether a pattern of neuronal responses complies with the basic assumptions of associative learning theory. Associative models predict that 'predictive cues' should elicit a stronger expectation of the outcome than 'blocked cues' (see Table 1). Thus, if the expected outcome is contradicted by the provided feedback, a greater FRN should be found for cues that are known to be predictive compared with cues that are known to be unpredictive or blocked (i.e., we expected to see a larger FRN in test trials where subjects were presented with a predictive cue compared with trials that present a blocked cue).

Method

Participants

A total of 25 right-handed adults participated in the present study. The experiment was divided into two experimental sessions and lasted for about five hours. The data from one participant was excluded because he/she did not complete the second session. The effective

sample size was 24 participants (mean age = 22, S.D. = 4.1, range = 19 - 38, 7 males). According to self-reports, all participants were healthy and had no history of neurological or psychiatric illness. Participants received 40 Euros for completing both sessions.

Stimuli and task

Colored stimuli were presented on a 19-inch computer screen in the center of a white rectangle (15 cm wide x 5 cm high) surrounded by a dark blue background. A total of 180 pictures of objects were used as stimuli. These objects belonged to one of the following six categories: fruits, foods (excluding fruits), clothes, office supplies, animals and toys.

Participants were asked to predict whether the presented object(s) (i.e., cue(s)) would cause an allergic reaction (i.e., outcome) in an imaginary patient on a trial-by-trial basis. At the beginning of each trial, participants were presented with a fixation cross placed in the center of the screen for 1 s prior to the presentation of the cue(s). In single-cue trials (see Table 1 for the design), the cue appeared in the center of the screen. In compound-cue trials, the two cues appeared side by side in the center of the screen (they were separated from each other by 2 cm). In these trials, the position of each cue on the screen (right or left) was randomized, and the cue was shown for 1.2 s. After the presentation of the cue, participants were asked to make a two-choice decision upon the presentation of the cue(s) by pressing either the allergy or no allergy response button. Participants were instructed to respond using their index finger, and the assignment of allergy and no allergy responses to the left or right hand was counterbalanced across the different blocks of the task. After the participants submitted their response, a blank screen was displayed for 1 s, which was followed by a feedback stimulus for 1 s. Feedback stimuli consisted of a smiling face to indicate a correct response or a sad face to indicate an incorrect response (Figure 1). The feedback stimuli informed participants whether the patient experienced an allergic reaction from the cue object.

****

Please insert Figure 1 about here

****

Experimental design

The experiment was divided into thirty blocks (5 blocks per stimulus category). The experimental design (Table 1) was repeated across the different blocks, and a different set of stimuli was employed in each of these blocks. The roles of the cues displayed in Table 1 were pseudo-randomly assigned to the different stimuli with the constraints that all objects must be from the same category in each block and none of the stimuli could be repeated. The order of block presentation was counterbalanced across all participants.

****

Please insert Table 1 about here

****

In each block, participants had to learn the relationships between the cues and the outcome (i.e., allergy) across trials using error-guided learning. Each block included 7 different trial types: 4 'learning' trial types, including the training trials to allow an evaluation of the cue-interaction effect, 2 'test' trial types in which the cue-interaction effect was measured, and 1 'filler' trial type, which was used to equate the number of allergy and no allergy outcomes in each block. Learning trials included A → Allergy trials, AB → Allergy trials (i.e., the blocked condition), and C → No allergy trials and CP → Allergy trials (i.e., the predictive condition). Ten trials of each association were presented per block. Test trials consisted of eight B and eight P trials that were equally associated with allergy and no allergy (50%) to prevent contingency learning outside of the specific relationships involved in the cue-interaction situation. The filler consisted of twenty EF → No allergy trials. For each of the learning and

filler trial types, additional 2 and 4 trials, respectively, were included. These were inconsistent trials, which meant that they were associated with the alternative outcome to their reference trial type (Table 1). These extra noisy trials were included to make the learning and the test trials look more similar (i.e., so that trials other than the test trials were of a probabilistic nature) and make generalization more likely. The inclusion of a single-phase, cue-interaction design had a similar objective. The traditional two-phase design, which has been used in classical animal conditioning experiments, would have entailed single-trial types (e.g., A → Allergy) in a first phase and compound-trial types (e.g., AB → Allergy) in a second phase. Test trials would have been presented in a different phase. The use of multiphase designs has been suggested to encourage human subjects to see the different phases as independent or unrelated, which diminishes the chance of a cue-interaction effect (Hinchy, Lovibond, & Ter-Horst, 1995; Shanks & López, 1996). Thus, including all trial types within a single phase should facilitate generalization between what is being learned across the task.

Additionaly, Eppinger et al. (2008) also used a single-phase design, including a large number of target trials as required in ERP experiments and repeating trial blocks that were functionally identical. Instead, Wills et al. (2007) programmed simultaneous replications of functionally identical blocks of a classical two-phase blocking design, rather than a sequential repetition of blocks. As there are no strong reasons to prefer one strategy over the other, we preferred to follow Eppinger et al.'s strategy as this study had already shown clear learning effects in the FRN.

The order of trial presentation within each block was randomized to prevent more than two consecutive presentations of the same trial. To allow testing of cue-interaction effect, however, test trials were only presented after learning had already taken place (i.e., test trials were only presented after the first ten trials of each trial type had already been presented). After the first ten trials of each type, test trials were randomly intermixed with the other trial types.

Procedure

Participants started by reading the instructions on a computer screen, which contained a detailed description of the experimental task. Participants were instructed to determine which stimuli caused an allergic reaction in fictitious patients. In each learning trial, participants had to predict whether the patient would suffer from an allergy. The task was divided into 30 blocks, and each block represented the case story of a different patient. Participants were instructed to consider each patient independently (i.e., no relationship existed between them). These instructions were repeated after each experimental block, which indicated the stimuli category [fruits, foods (excluding fruits), clothes, office supplies, animals or toys] that caused the allergy to the next patient. The instructions stated that all allergies were imaginary and that participants should not utilize any previous knowledge about allergies to complete the task.

Data recording

Behavioral and electroencephalogram data recording

An IBM compatible computer was used to collect behavioral data and present stimuli. Responses were registered by depressing the left buttons of two PC mice located approximately 1 m apart on either side of the computer monitor.

Event-related potentials were recorded from the scalp with tin electrodes mounted in an electrocap at 29 standard positions. The data were referenced to the outer canthus of the right eye (online) and the average mastoid recording (offline). Vertical eye movements were monitored by an electrode placed on the infraorbital ridge of the right eye. Electrode impedances were below 3 kΩ. The electrophysiological signals were filtered online with a bandpass of 0.01–50 Hz (half-amplitude cutoffs) and digitized at a rate of 250 Hz. Trials with a base-to-peak electrooculogram (EOG) amplitude greater than 50 μV, amplifier saturation, or a baseline shift exceeding 200 μV/s were excluded from analyses. Remaining EOG artifacts

were corrected using the SOBI algorithm (Belouchrani, Abed-Meraim, Cardoso, & Moulines, 1993, 1997).

Data analysis

Behavioral data

We combined the thirty learning blocks for each participant and used the proportion of allergy responses in each trial to analyze the behavioral effects. For repeated measure analysis of variance (ANOVA), the Greenhouse-Geisser correction for degrees of freedom was performed when sphericity was violated. The cases in which this correction was applied are indicated in the text.

Event-related potential data

Electroencephalography epochs were averaged with reference to the feedback onset. Data were baseline-corrected by subtracting the average activity that occurred during the 200 ms preceding the feedback onset. To analyze learning-related changes, the learning trial ERPs were averaged into four bins, and each bin corresponded to one quarter of the learning block trials (bin 1 was trials 1-3; bin 2 was trials 4-6; bin 3 was trials 7-9; and bin 4 was trials 10-12). Because this study was focused on elucidating learning effects, the data from the 4 learning trial types were combined for analysis. To analyze the cue-interaction effect, the ERP activity for target cues B and P were independently combined across the 8 test trials. Note that cues B and P were paired with the outcome during test trials at a frequency of 50%.

Statistical analyses were performed for the Fz electrode, which is a standard electrode location for FRN analysis (e.g., Ghering & Willoughby, 2002; Eppinger et al., 2008) (topographic maps Figures 4 and 6). The feedback-locked components were measured as the mean amplitude within a 100-ms time window centered around the FRN at the Fz electrode (for learning-related changes: 300 ms in the first bin, 290 ms in the second, 295 ms in the third

and 290 ms in the last bin. For the cue interaction analysis, all conditions that were analyzed had a peak at 300 ms)[1].

Results

Behavioral results

Training: proportion of allergy responses

A 4 (Cues: A, AB, C and CP) x 12 (Trials: Training trials 1-12) ANOVA was performed to analyze the effect of learning on the proportion of allergy responses. This analysis showed a main effect of Trials, $F(4.04, 92.88) = 15.84$, $MSE = 0.4$, $p < 0.001$ (Greenhouse-Geisser), a main effect of Cues, $F(1.73, 39.75) = 138.9$, $MSE = 19.79$, $p < 0.001$ (Greenhouse-Geisser) and a significant Cues x Trials interaction, $F(11, 252.94) = 23.82$, $MSE = 0.14$, $p < 0.001$. These effects can be explained by a progressive adjustment of participants' responses to the programmed contingencies (Figure 2). A visual inspection of Figure 2 suggests that CP trials were the hardest trial type to learn. The most significant difference in performance across trial types appeared to be between AB and CP trials, which both included the target cues B and P, were compound-cue trial types, and received an identical number of training trials. To confirm this finding, we compared the number of allergy responses in AB and CP trials in a 2 (Cues: AB vs. CP) x 12 (Trials: Training trials 1-12) ANOVA. This analysis showed main effects of Trials, $F(4.81, 110.62) = 30.57$, $MSE = 0.63$, $p < 0.001$ (Greenhouse-Geisser) and Cues $F(1, 23) = 33.87$, $MSE = 1.34$, $p < 0.001$. The Cues x Trials interaction was not significant $F(11, 253) < 1$. This analysis indicates that CP trials were harder to learn than AB trials, which may be explained by the interference produced by C→No allergy trials while learning the CP→Allergy association.

Each participant's performance was analyzed to check whether the overall number of correct responses increased throughout the 30 task blocks programmed. All cue types and trials were collapsed for this analysis, and a one-way ANOVA was performed over blocks 1 through

30. The main effect of blocks was significant [$F(29, 638) = 5.05$, $MSE = 0.1$, $p < 0.001$], which indicates that the sample over which ERPs were collected was not static (i.e., the performance was significantly better across the final blocks). Unfortunately, the large number of trials needed for an adequate analysis of the learning course and the cue-interaction effects prevented us from including the Block factor in the ERP analyses. Thus, similar to previous studies, all blocks were combined for this analysis (e.g., Eppinger, et al., 2008). A 2 (Cues: B vs. P) x 30 (Blocks: 1-30) ANOVA was performed, and the interaction between both factors was not significant [$F(29, 667) < 1$], which indicated that the overall improvements in performance did not interact with the cue-interaction effect.

****

Please insert Figure 2 about here

****

Test: proportion of allergy responses

The cue-interaction effect on the proportion of allergy responses was evaluated with a 2 (Cues: B vs. P) x 8 (Trials: Test trials 1-8) ANCOVA using the differences in performance between AB and CP in the last training trial as a covariable, which indexed the difference between the proportion of allergy responses in the last AB and CP trials. This allowed for a stricter analysis of the cue-interaction because it controlled for possible effects of different acquisition levels in B and P test trials. The ANCOVA showed a significant main effect of Cues, $F(1, 22) = 5.19$, $MSE = 0.51$, $p = 0.033$, a marginal main effect of Trials $F(4.2, 92.53) = 2.3$, $MSE = 0.02$, $p = 0.061$ (Greenhouse-Geisser) and a marginal Cues x Trials interaction, $F(7, 154) = 2.01$, $MSE = 0.01$, $p = 0.057$. The means showed that the effect of the Cues factor was due to a greater proportion of allergy responses in P trials compared with B trials (Mean and SD: P = 0.64, SD = 0.03; B= 0.59, SD = 0.02). We observed a cue-interaction effect, and the effect of the cue interaction was larger in the first trial than in the subsequent test trials (Figure 3). The Cues x Trial interaction, however, was not significant.

Two main behavioral results were obtained. First, participants' responses showed a progressive adjustment to the contingencies programmed across trials. Secondly, a cue-interaction effect was found in the test trials. Specifically, more allergy responses were found in P trials compared with B trials. These results replicate two basic findings of the human predictive learning literature and demonstrate a successful adaptation of a standard human predictive learning paradigm to obtain neurophysiological data. Interestingly, we had to significantly increase the duration of the task or embed the test for the cue-interaction effect within a probabilistic task, which distinguished the present study design from previous studies.

****

Please insert Figure 3 about here

****

ERP Results

Training phase

All training trials were combined into four training bins. The following section shows the feedback-locked ERPs that were recorded during the training phase.

FRN

A 4 (Cues: A, AB, C and CP) x 4 (Trials: bins 1-4) x 2 (Feedback: positive vs. negative) ANOVA was performed to analyze the effect of learning on FRN. The FRN was defined as the mean amplitude within a 100-ms time window centered around the maximum difference between positive and negative feedback at the Fz electrode. The only significant effect was the Trials x Feedback interaction, $F(3, 66) = 7.05$, $MSE = 11.7$ $p < 0.001$, which appears to be due to a progressive reduction in FRN magnitude across trial bins (Figure 4). Because the Cues factor was not significant, data from all cue types were combined for the remaining analyses. As it is shown in Figure 4, the Trials x Feedback interaction appears to be due to a progressive reduction in the FRN magnitude across trial bins. To evaluate the accuracy

of this interpretation, additional *t* tests of the effect of FRN in each bin were performed. These analyses showed a significant FRN effect in the first bin, $t(23) = 5.24$, $p < 0.001$, nonsignificant trends in the second and third bins, [$t(23) = 1.46$, $p = 0.158$ and $t(23) = 1.64$, $p = 0.114$, respectively] and the absence of FRN in the last bin, $t(23) = -0.33$, $p = 0.746$.

Learning effects on positive and negative feedback ERP components were independently evaluated. An ANOVA with Trials (i.e., bins 1-4) as a single factor was performed on both ERPs. Positive and negative feedback were analyzed independently, and the mean amplitude within the time window was used as the dependent variable for FRN analysis. For positive feedback, the effect of Trials was significant, $F(1.73, 39.73) = 8.49$, MSE = 23.16, $p = 0.001$ (Greenhouse-Geisser), and there was also a significant linear trend, $F(1, 23) = 12.69$, MSE = 39.6, $p = 0.002$. For negative feedback, the effect of Trials was not significant, $F(3, 69) = 1.67$, MSE = 3, $p = 0.181$. In addition, the linear trend was far from significant, $F(1, 23) = 1.45$, MSE = 2.02, $p = 0.241$.

****

Please insert Figure 4 about here

****

Test phase

The FRN component was evaluated for a cue-interaction effect in the test phase. In contrast with the training analysis, during testing, both target cues (B and P) were paired with the outcome in 50% of the trials. Thus, corrective feedback (i.e., either positive or negative feedback) could not be clearly established. Because the target associations needed to evaluate a cue-interaction effect are those that involve the blocked cue B (i.e., B → Allergy association) and the predictive cue P (i.e., P → Allergy association), the effectiveness of the cue-interaction manipulation on FRN measures should be elucidated by an analysis of allergy responses.

Remember that P should be strongly associated with allergy whereas B should only be weakly associated with allergy. Thus, the FRN should index the magnitude of both target associations *in allergy response trials*. Specifically, a larger FRN amplitude for allergy responses to P compared with B should be interpreted as evidence of a cue-interaction effect. For the no allergy responses, it is difficult to conceive how FRN can convey information about cue interactions. In addition, according to associative learning models, the situation is not as symmetrical as it may seem for allergy and no allergy responses. Indeed, even the blocked cue (B) was expected to be partially associated with allergic reaction (i.e., blocking is only partial with the actual contingencies programmed). In fact, more allergy responses were registered for both target cues throughout the task (Figure 3). In sum, FRN sensitivity to cue interaction may be more confidently evaluated through an analysis of allergy than of no allergy responses. Although both response-types were included in the same analysis, we also analyzed allergy and no allergy responses separately.

FRN

A 2 (Cues: B vs. P) x 2 (Responses: Allergy vs. no allergy) x 2 (Feedback: Positive vs. negative) ANOVA was performed to evaluate the sensitivity of FRN to cue interaction. This analysis showed main effects of Responses, $F(1, 23) = 9.24$, $MSE = 89.74$, $p = 0.006$ and Feedback, $F(1, 23) = 29.83$, $MSE = 271.25$, $p < 0.001$, a marginal Cues x Responses interaction $F(1, 23) = 3.7$, $MSE = 20.03$, $p = 0.067$ and, importantly, a significant Cues x Responses x Feedback interaction, $F(1, 23) = 8.33$, $MSE = 15.13$, $p = 0.008$. All remaining effects were not significant (F values < 1). The effect of Feedback indicates a main effect of FRN (i.e., a more negative activity following negative feedback compared with positive feedback). Interestingly, the Cues x Responses x Feedback interaction can be explained by a greater FRN for P than for B when participants responded allergy (Figure 5). Conversely, more similar FRNs were found for both target cues when participants responded no allergy. This

effect could also explain the main effect of Responses. Two independent ANOVAs for allergy and no allergy responses were used to analyze the origin of this second order interaction.

For allergy responses, a 2 (Cues: B vs. P) x 2 (Feedback: Positive vs. negative) ANOVA showed a main effect of Feedback (i.e., FRN), $F(1, 23) = 23.12$, $MSE = 133.7$, $p < 0.001$ and a significant Cues x Feedback interaction, $F(1, 23) = 5.25$, $MSE = 5.34$, $p = 0.031$. This interaction was produced by a stronger FRN effect from P compared with B, $F(1, 23) = 33.14$, $MSE = 2.9$, $p < 0.001$, and $F(1, 23) = 10.98$, $MSE = 3.89$, $p = 0.003$ for cues P and B, respectively. Although significant FRNs were found for both P and B, the effect was larger for P trial types, which was expected based on associative models of learning.

For the no allergy responses, a 2 (Cues: B vs. P) x 2 (Feedback type: positive vs. negative) ANOVA showed a main effect of Feedback (i.e., FRN), $F(1, 23) = 19.34$, $MSE = 137.58$, $p < 0.001$ and a significant main effect of Cues $F(1, 23) = 4.43$, $MSE = 21.31$, $p = 0.047$. In this instance, the Cues x Feedback interaction was marginal [$F(1, 23) = 3.48$, $MSE = 10.17$, $p = 0.075$]. Together with the trend for a larger FRN in B trials compared with P trials, the main effect of Cues (P = 3.45 μV vs. B = 2.51 μV) may be interpreted as going in the same direction as the cue-interaction effect reported in the allergy responses. Participants tended to expect the corrective feedback (allergy) to a lesser degree in B trials compared with P trials. Because the Cues x Feedback interaction was not significant (and for the reasons described above), the results concerning allergy responses will be centered on the discussion of the cue-interaction effect.

****

Please insert Figure 5 about here

****

When participants responded that a cue would cause an allergy, the FRN was larger for P compared with B. This result was consistent with a cue-interaction effect (i.e., due to a

stronger P → Allergy association, the unpredicted negative feedback had a larger impact in the EEG for predicted cue P compared with the blocked cue B).

Discussion

In the last 30 years, a considerable body of scientific research has been directed toward the study of mechanisms underlying predictive and causal learning in humans (e.g., López & Shanks, 2008; Shanks, 2010). Associative models of predictive learning, which are arguably the most influential models of predictive learning, establish error computation as the core of the learning process. Thus, to the extent that the mechanisms proposed by associative models are viable, it seems likely that there would be neural networks to calculate an error term in a way that is computationally equivalent to the proposed associative learning models. In the present study, we explored whether the FRN ERP component could be used as a functional correlate for this error term. We assessed whether the FRN was consistent with the gradual learning of cue-outcome associations and cue-interaction predictions of associative learning models. We hypothesized that the FRN would be sensitive to both cue-outcome associations and cue-interaction predictions.

In agreement with our hypothesis, participants' behavioral performances gradually adjusted to the programmed cue-outcome contingencies (see Figure 2). Moreover, a cue-interaction effect was shown in the test trials. Despite both cues being paired with an allergic reaction the same number of times during the training trials, the participants responded allergy to the predictive cue (P) more often than to the blocked cue (B) (Figure 3). These results are congruent with previous studies of predictive and causal learning in humans (see López & Shanks, 2008 for a review). Therefore, these results indicate that the present study design is a successful adaptation of a standard human predictive learning paradigm that allows for the recording of neurophysiological data. Compared with more traditional testing protocols, the

duration of the task was either significantly increased or the test for the cue-interaction effect was embedded within a probabilistic task.

With regard to the ERP data, the results from the feedback-locked components showed a decrease in FRN magnitude across trials, which was defined as the difference between negative and positive feedback. This finding supports the hypothesis that FRN generation is related to neural processes that are involved in gradual learning (e.g., Holroyd & Coles, 2002). More specifically, the independent analyses performed for positive and negative feedback showed that the decrease in FRN magnitude was due to a learning effect that resulted in diminished amplitude of feedback-locked ERPs following positive feedback signals. Interestingly, no learning effect was shown following negative feedback signals, which was similar to the findings of previous studies (Eppinger et al., 2008; Eppinger, Mock, & Kray, 2009).

This asymmetry between positive and negative feedbacks was consistent with associative learning models suggesting that the FRN is a neural correlate to the error term. For example, the RW model assumes two different parameters for the salience of the presence and absence of the outcome (β^+_j and β_j, respectively) involved in learning, and it is generally assumed that $\beta^+_j > \beta_j$ (Rescorla & Wagner, 1972). In situations where cues and outcomes are highly correlated, we can also assume that positive feedback is usually indicative of a positive error term across trials, as long as learning is pre-asymptotic (i.e., the outcome is present even though it is not fully predicted). Conversely, negative feedback (i.e., the predicted outcome is not present) is usually indicative of a negative error term. Importantly, the error term in the RW model is defined as the difference between the presence or absence of the outcome and the system's expectation. Because each of these error terms is weighted by its corresponding β salience parameter, learning-related changes in these error terms will be greater in trials with the outcome than in trials without the outcome. Therefore, if feedback related brain potentials

reflect an error signal similar to that proposed by associative learning models, such as RW, a larger learning effect would be expected for correct trials than for incorrect trials. Although there is no direct evidence to support these assumptions about different parameters, this theory is congruent with the finding that positive feedback was more relevant than negative feedback (with respect to learning effects).

An unexpected result from this account of the FRN was the lack of Cues effect in the learning phase results. Note that learning curves for AB and CP trials were different and so a different pattern should have been found in the FRN. However, the FRN analysis did not show such effect. Although it is difficult to know why this effect could not be detected, it is possible that we did not have enough signal-to-noise ratio for its detection. In the Cues analysis, we divided by four the number of trials in each condition, which could have diminished the sensitivity of this test. Additionally, it is important to notice that the Cues effect in the behavioral analysis was weaker than the other main learning effects.

The behavioral main effect of Blocks reported (i.e., participants performed worse on the initial than on the final blocks of trials) cannot be explained exclusively on the basis of current learning theories. A possible explanation of these results is that the task became more familiar with practice. Thus, errors related with procedural factors of the task (e.g., participants' confusion with response buttons) diminished with practice. It may be assumed that this increased familiarity with the task somewhat increased participants' learning rate of the task.

Holroyd, Pakzad-Vaezi, and Krigolson (2008) recently proposed that the negativity obtained following negative feedback was another ERP component, namely the N200. The N200 is a frontocentral negativity that is elicited by infrequent stimuli (Pritchard, Shappell, & Brandt, 1991). Thus, the N200 would be elicited by both positive and negative infrequent feedback. FRN, however, would be caused exclusively by a positivity related to positive feedback, which overlaps the default N200. Thus, positive feedback may be more effective

than negative feedback in decreasing N200. In extreme cases, positive feedback may actually abolish the N200, whereas negative feedback may increase the N200 amplitude.

The theory proposed by Holroyd et al. (2008) is supported by the present results concerning acquisition effects and other recently published studies. For example, a specific frequency electrophysiological modulation in the beta band range has been shown to be associated with positive feedback processing (Cohen, Elger, & Fell, 2008; Marco-Pallares et al., 2008). In these studies, Holroyd et al. (2008)'s proposal could be detailed by establishing that processing following positive feedback complies with the principles of associative models and is in agreement with the theory that feedback-related positivity reflects a reduction of positive reward prediction errors during learning (see also Potts, Martin, Burton, & Montague, 2006; Holroyd et al., 2008).

The learning effect observed following positive feedback could also be caused by a P300 (specifically, P3b) modulation that was correlated with learning. Studies have established that the P300 component is modulated by the probability of stimuli (i.e., it is increased by infrequent stimuli). Because positive feedback was more frequent across trials, it is possible to attribute the decrease in feedback positivity to variations in the P300 magnitude. Thus, as positive feedback becomes more frequent, we would also expect to observe a decrease in the P3b component. Regarding our results, note that the decreasing activity associated with learning shown in the topography of Figure 6 is larger at frontocentral locations. For a P300 account of this learning effect, this topographic distribution was not expected because of the standard parietal topography of the probability-modulated positive component (Duncan-Johnson & Donchin, 1977).

The learning effects on P300 positivity was expected considering the information theory of the P300 component, which states that the amplitude increases depending on the amount of information that is extracted from the feedback stimulus (Donchin & Coles, 1988;

Johnson, 1988). As individuals learn an association, no more information is needed from the feedback stimuli, which should reduce the amplitude of the positive component over time. In agreement with this interpretation, Rose, Verleger and Wascher (2001) performed an associative learning study and showed an increase in the amplitude of the P300 component for the conditioned stimulus (S1) and a decrease for the imperative stimulus (S2). An interpretation of these learning effects might be related to a decrease in the amount of information processed in the feedback condition: As learning proceeds, the amount of required external feedback information is reduced.

If the P300 account is true, it is expected a learning effect in the positive feedbacks (as it is explained above) and also in the negative feedbacks. Because negative feedback becomes more infrequent across trials, there should be a time-dependent increase in the P3b parietal component following negative feedback; however, no learning effect was found after negative feedback. A possible interpretation of these findings is that participants were more attentive to negative than positive feedbacks during the learning phase. Thus, participants would pay attention to the negative feedback during the learning blocks regardless of whether they have correctly learned the association. Notice however, that we have no independent measure of attention to rule out this interpretation. This interpretation was in agreement with the idea that the amplitude of P300 is dependent on the information extracted from the feedback (Donchin & Coles, 1988; Johnson, 1988). Due to the probabilistic nature of the task used in the present study, negative feedback was given in 16% of the trials regardless of whether participants gave the correct response, which may have induced an automatic orientation response and prompted participants to update and shift their learned stimulus-reward contingency. Indeed, several authors have proposed that the P300 amplitude reflects the obligatory allocation of attention to task-relevant events (Strayer & Kramer, 1990). In a way, the present situation was similar to a reversal learning paradigm (i.e., a learned stimulus-reward association followed by negative feedback indicates that a change in the association is required) (set-shifting, see Cools, Clark,

Owen, & Robbins, 2002). This theory was also in agreement with the prediction error signal, which would be large in our study because the expected feedback for the response was positive (Cohen & Ranganath, 2007 and Chase, Swainson, Durham, Benham, & Cools, 2011). In a probabilistic learning task, Bellebaum and Daum (2008) found a larger impact of positive feedback on the amplitude of the P3 at the beginning of the training phase (when negative feedback would be expected) and a reduction of the P3 amplitude as learning progressed (when the positive feedback was less informative). This was consistent with the present findings regarding the selective modulation of the amplitude of the P3 component in positive feedback trials.

Interestingly, recent experiments have suggested that learning effects on FRN and the P300 components reflect different processes. In a recent reversal learning study (Chase, Swainson, Durham, Benham, & Cools, 2011), the authors provided participants with explicit rules for changes in the pattern of responses. The Chase et al. study found that P300 modulations indexed behavioral adjustments on the basis of the explicit rules, whereas the FRN component reflected an error measure of an associative nature (see also Bellebaum and Daum, 2008; Bellebaum, Polezzi, & Daum, 2010). Thus, the results of the Chase et al. study were consistent with the *reinforcement learning theory* of FRN and cast doubt on the explanation of the FRN solely in terms of a P300 modulation.

In summary, modulation of the positive component involved in FRN seems to be sensitive to the amount of error and the amount of information required by the feedback to learn the different programmed associations. We suggest two possible accounts to the different learning effects found in positive and negative feedbacks. The first account comes from associative learning theory. Following these models, it is expected a lower salience in negative than in positive feedbacks (Rescorla & Wagner, 1972) and, hence, the learning effect would be larger in the positive than in the negative feedback trials. On the other hand, the P300 account would assume that as learning progresses, the amplitude of the more recent positive component

decreases in the positive feedback condition. In negative feedback trials, however, participants might be reacting to negative information throughout all trials, due to the probabilistic nature of the task and the triggering of set-shifting processes aimed to remap the learned stimulus-reward associations. Note that these two accounts are to a certain extent incompatible because it is widely assumed that salience and attention are two factors that are positively correlated. An interesting future question is in which degree negative feedbacks receive more or less attention and how this factor modulates learning.

****

Please insert Figure 6 around here

****

In contrast to previous studies, the relationship between FRN and associative models of learning was also evaluated by a cue-interaction test. Cue-interaction effects are a central prediction of modern associative models of learning and are considered the "[…] canonical paradigm for assessing the role of prediction error in learning" (Waelti et al., 2001, p. 43). Our results showed that the FRN was modulated by a cue-interaction manipulation. Specifically, when participants responded allergy, the FRN was larger in the trials that included a highly predictive cue compared with the trials that included a blocked cue (Figure 5).

Previous studies have used cue-interaction phenomena as a marker of the implication of associative models in different psychological processes, such as causal learning (e.g., Dickinson et al., 1984), categorization (Shanks, 1991) or social psychology (Cramer, Weiss, Steigleder, & Balling, 1985; Van Overwalle, 2007). The same strategy was used by Waelti et al. (2001) to determine whether the phasic activity of neurons in the midbrain dopaminergic system compute an error signal equivalent to the error term proposed by associative models of learning. Waelti et al. analyzed the activity of single midbrain dopamine neurons in primates

during a cue-interaction experiment using a classical conditioning paradigm and found that the activity of dopamine neurons recorded during reward (or positive feedback) in predictive cue trials was lower than baseline activity following negative feedback (i.e., trials in which the reward did not occur). Moreover, the inhibition of dopamine activity was not observed in blocked cue trials. Thus, a 'negative' error signal was computed by dopamine neurons when the reward was expected but absent (as it is the case for the predictive cue). Because the reward was not expected in the case of blocked cue trials, very little error was computed during negative feedback, and no inhibition was found. This finding was in line with the predictions of the associative models. The results of the present study may then be understood as an extension of the results of a study by Waelti et al. (2001). Similar to the results of Waelti et al., we found a larger impact of negative feedback (compared with positive feedback) in the predictive cue compared with the blocked cue. This pattern of results is consistent with the idea that FRN reflects an error signal or mismatch between an actual outcome and an expectation concerning that outcome. Importantly, this expectation was generated by a mechanism that is sensitive to cue-interaction manipulations, as it is proposed by modern associative learning models.

Wills et al. (2007) assessed the extent of the involvement of attentional processes in cue-interaction effects and showed an early attentional cue-interaction effect in the N1, which supported attentional models of associative learning (e.g., Mackintosh, 1975; Pearce & Hall, 1980). Attentional models of associative learning explain cue interactions on the basis of changes in the attention paid to target cues. According to attentional models, participants would pay more attention to predictive cues (cue P) rather than blocked cues (cue B) during learning. Thus, associative links involving predictive cues would be stronger than those involving blocked cues, hence the cue-interaction effect. The possible implication of early attentional processes in the present results must be acknowledged. It is important to note, however, that all of the associative models predict the same cue-interaction effect regardless of their specific computational details. One of the goals of the present study was to show that the

FRN component was sensitive to the cue-interaction effect, which would be predicted by the associative learning theory. We were not concerned about determining which of the associative proposals better explained the overall pattern of the reported results, and the methodological strategy that we followed could not discriminate between associative proposals.

The cue interaction reflected in the FRN component was compatible with the *reinforcement learning theory* of FRN, which proposes that the expectation of the outcome (or the reward) is formed by an associative mechanism equivalent to the RW model (Holroyd and Coles, 2002). Specifically, Holroyd & Coles suggested that the midbrain dopaminergic system conveys an error signal to the ACC that is equivalent to the *temporal difference error*. This is an error term calculated by a reinforcement learning algorithm called the *method of temporal differences* (Sutton, 1988), which is an extension of the RW model algorithm to the continuous time domain. This error signal is conveyed to the ACC to reinforce the most successful response, which results in the FRN component (see Holroyd & Coles, 2002 for further details on the model). Thus, the cue-interaction phenomenon is a direct prediction of the *method of temporal differences,* and its effect on FRN is a direct prediction of the *reinforcement learning theory* of FRN.

Results of published studies are not conclusive with regard to whether an associative-like error signal is actually computed in midbrain areas and then conveyed to the ACC, which is assumed by the reinforcement learning theory. According to the cue-interaction FRN results obtained in the present study, we propose that an error signal that is sensitive to cue-interaction manipulations is conveyed from the midbrain to the ACC. On the one hand, the study performed by Waelti et al. (2001) provided evidence that supports this hypothesis in non-human primates. On the other hand, although experiments with humans that have shown that associative error-driven learning is somehow associated with midbrain areas (e.g., Duzel et al., 2009), the evidence is sparse concerning cue-interaction experiments. In an event-related fMRI experiment, Turner et al. (2004) measured the BOLD activation associated with another cue-

interaction effect (i.e., super learning; see Aitken, Larkin, & Dickinson, 2000) and found a reliable relationship between lateral prefrontal cortex activation and the magnitude of the cue-interaction effect. No effect was observed in the striatum, which was unexpected considering the error computation neural network proposed by Holroyd and Coles (2002). Interestingly, Tobler and colleagues (2006) performed an fMRI cue-interaction experiment (using a similar experimental cue-interaction design to the one used here) and showed that the activation of the ventral putamen and orbitofrontal cortex was decreased in trials with a blocked cue compared with trials involving a predictive cue. Although it seems clear that prefrontal areas are involved in cue-interaction effects, it is still unclear whether midbrain regions participate in these effects in humans. More research in humans is needed to draw firmer conclusions on the neural paths that underlie associative error processing in cue-interaction designs.

Previous ERP experiments assessing the *reinforcement learning theory* of FRN have manipulated the probability of reward (i.e., positive feedback) and revealed that the higher the probability of reward, the higher the magnitude of FRN in nonrewarded trials (Cohen et al., 2007; Eppinger et al., 2008; Holroyd & Coles, 2002). This pattern of results may be incorrectly interpreted to mean that the FRN is sensitive to the absolute number of cue-reward pairings. The results of the present study show that FRN is not a *direct function of outcome or reward probability; rather, it is a function of how much informative value the cue can provide about the outcome prediction*. Importantly, the absolute probability of the outcome of both target cues (i.e., the predictive and the blocked cue) was the same: 0.83 and 0.5 during learning and test trials, respectively. Although these values were the same, these cues elicited very different feedback-related potentials during testing, which was expected based on the *reinforcement learning theory* of FRN.

In summary, the present results showed that feedback-locked ERPs may be a correlate of error computation and are closely related to the predictions derived from modern associative models. The results of the present study were similar to recent studies that have shown that

variations in the magnitude of the expectation-outcome mismatch are related to variations in FRN (Bellebaum & Daum, 2008; Bellebaum, Polezzi, & Daum, 2010; Holroyd, Krigolson, Baker, Lee, & Gibson, 2009). In addition, the results regarding the blocking effect extend the interpretation that the expectation-formation mechanism underlying FRN complies with the assumptions of modern associative models of learning.

References

Aitken, M. R., Larkin, M. J., & Dickinson, A. (2000). Super-learning of causal judgements. *Quarterly Journal of Experimental Psychology, 53B,* 59–81.

Bellebaum, C., & Daum, I. (2008). Learning-related changes in reward expectancy are reflected in the feedback-related negativity. *European Journal of Neuroscience, 27,* 1823–1835.

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

Belouchrani, A., Abed-Meraim, K., Cardoso, J. F., & Moulines, E. (1993). Second-order blind source separation of correlated sources. *Proceedings of the International Conference on Digital Signal Processing* (pp. 346–351). Available at: http://cloe.ucsd.edu/adel/.

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.

Càmara, E., Rodríguez-Fornells, A., Ye, Z., Münte, T. F. (2009). Reward networks in the brain as captured by connectivity measures. *Frontiers in Neuroscience, 3*(3), 1–11.

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.

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.

Cohen, M. X., Elger, C. E., & Fell, J. (2008). Oscillatory activity and phase-amplitude coupling in the human medial frontal cortex during decision making, *Journal of. Cognitive Neuroscience, 21*, 390–402.

Cohen, M. X., Elger, C. E., Ranganath, C. (2007). Reward expectation modulates feedback-related negativity and EEG spectra. *Neuroimage, 35*, 968–978.

Cohen, M. X., & Ranganath, C. (2007). Reinforcement learning signals predict future decisions. *Journal of Neuroscience, 27*(2), 371–378.

Cools, R., Clark, L., Owen, A. M., Robbins, T. W. (2002). Defining the neural mechanisms of probabilistic reversal learning using event-related functional MRI. *Journal of Neuroscience, 22*(11), 4563–4567.

Cramer, R. E., Weiss, R. F., Steigleder, M. K., & Balling, S. S. (1985). Attraction in context: Acquisition and blocking of person-directed action. *Journal of Personality and Social Psychology, 49,* 1221–1230.

Dickinson, A., & Burke, J. (1996). Within-compound associations mediate the retrospective revaluation of causality judgements. *Quarterly Journal of Experimental Psychology, 49B,* 60–80.

Dickinson, A., Shanks, D. R., & Evenden, J. L. (1984). Judgment of act-outcome contingency: the role of selective attribution. *Quarterly Journal of Experimental Psychology, 36A,* 29–50.

Donchin, E., & Coles, M. G. H. (1988). Is the P300 component a manifestation of cognitive updating? *The Behavioral and Brain Sciences, 11,* 357–427.

Duncan-Johnson, C. C., & Donchin, E. (1977). On quantifying surprise: The variation in event-related potentials with subjective probability. *Psychophysiology, 14,* 456–467.

Duzel, E., Bunzeck, N., Guitart-Masip, M., Wittmann, B., Schott, B. H., & Tobler, P. N. (2009). Functional imaging of the human dopaminergic midbrain. *Trends in Neuroscience, 32,* 321–328.

Eppinger, B., Kray, J., Mock, B., Mecklinger, A. (2008) Better or worse than expected? Aging, learning, and the ERN. *Neuropsychologia, 46*(2), 521–539.

Eppinger, B., Mock, B., & Kray, J. (2009) Developmental differences in learning and error processing: Evidence from ERPs. *Psychophysiology, 46*(5), 1043–1053.

Fletcher, P. C., Anderson, J. M., Shanks, D. R., Honey, R., Carpenter, T. A., Donovan, T., Papadakis, N., & Bullmore, E. T. (2001). Responses of human frontal cortex to surprising events are predicted by formal associative learning theory. *Nature Neuroscience, 4,* 1043–1048.

Gehring, W. J., & Willoughby, A. R. (2002). The medial frontal cortex and the rapid processing of monetary gains and losses. *Science, 295,* 2279–2282.

Haber, S. N. & Knutson, B. (2010). The Reward Circuit: Linking Primate Anatomy and Human Imaging. *Neuropsychopharmacology Reviews, 35*, 4–26.

Hinchy, J., Lovibond, P. F., & Ter-Horst, K. M. (1995). Blocking in human electrodermal conditioning. *Quarterly Journal of Experimental Psychology, 48B,* 2–12.

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.

Holroyd, C. B., Krigolson, O. E, Baker, R., Lee, S., & Gibson, J. (2009). When is an error not a prediction error? An electrophysiological investigation. *Cognitive, Affective, & Behavioral Neuroscience, 9,* 59–70.

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

Johnson, R., Jr. (1988). The amplitude of the P300 component of the event-related potential: Review and synthesis. *Advances in Psychophysiology, 3,* 69–137.

Kamin, L. J. (1968). "Attention-like" processes in classical conditioning. In M. R. Jones (Ed.), *Miami Symposium on the Prediction of Behavior, 1967: Aversive Stimulation* (pp. 9–31). Coral Gables, Florida: University of Miami Press.

López, F. J., & Shanks, D. R. (2008). Models of animal learning and their relations to human learning. In R. Sun (Ed.), *Handbook of computational cognitive modelling* (pp. 589–611). Cambridge, MA: Cambridge University Press.

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

Marco-Pallares, J., Cucurell, D., Cunillera, T., García, R., Andres-Pueyo, A., Münte, T.F., Rodríguez-Fornells., A. (2008). Human oscillatory activity associated to reward processing in a gambling task. *Neuropsychologia, 46,* 241–248.

Marco-Pallares, J., Müller, S.V., & Münte, T. F. (2007). Learning by doing: an fMRI-study of feedback-related brain activations. *Neuroreport, 18,* 1423–1426.

Maris, E., & Oostenveld, R. (2007). Nonparametric statistical testing of EEG- and MEG-data. *Journal of Neuroscience Methods, 164,* 177–190.

Matute, H., Arcediano, F., & Miller, R. R. (1996). Test question modulates cue competition between causes and between effects. *Journal of Experimental Psychology: Learning, Memory, and Cognition, 22,* 182–196.

Müller, S. V., Möller, J., Rodríguez-Fornells, A., & Münte, T. F. (2005). Brain potentials related to internal and external information used for performance monitoring. *Clinical Neurophysiology, 116,* 63–74.

O'Doherty, J. P., Dayan, P., Friston, K., Critchley, H., & Dolan, R. J. (2003). Temporal difference models and reward-related learning in the human brain. *Neuron, 38,* 329–337.

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

Potts, G. F., Martin, L. E., Burton, P., & Montague, P. R. (2006). When things are better or worse than expected: the medial frontal cortex and the allocation of processing resources. *Journal of Cognitive Neuroscience, 18,* 1112–1119.

Pritchard, W.S., Shappell, S.A & Brandt, M.E. (1991). Psychophysiology of N200/N400: A review and classification scheme. In P. K Ackles, J. R. Jennins, & M. G. H. Coles, (Eds.), *Advances in Psychophysiology (vol. 4)* (pp. 43–106). Greenwich: JAI Press.

Rose, M., Verleger, R., & Wascher, E. (2001). ERP correlates of associative learning. *Psychophysiology, 18,* 271–282.

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 and theory* (pp. 64–99). New York: Appleton.

Schultz, W. (2002). Getting formal with dopamine and reward. *Neuron, 36,* 241–263.

Shanks, D. R. (1991) Categorization by a connectionist network. *Journal of Experimental Psychology: Learning, Memory, and Cognition, 17,* 433–443.

Shanks, D. R. (2010). Learning: From association to cognition. *Annual Review of Psychology, 61,* 273–301.

Shanks, D. R., & López, F. J. (1996). Causal order does not affect cue selection in human associative learning. *Memory & Cognition, 24*, 511–522.

Strayer, D. L., & Kramer, A. F. (1990). Attentional requirements of automatic and controlled processing. *Journal of Experimental Psychology: Learning, Memory, and Cognition, 16*(1), 67–82.

Sutton, R. S. (1988). Learning to predict by the methods of temporal differences. *Machine Learning, 3*, 9 – 44.

Sutton, R. S., & Barto, A. G. (1981). Toward a modern theory of adaptive networks: Expectation and prediction. *Psychological Review, 88,* 135–140.

Tobler, P. N., O'Doherty, J. P., Dolan, R. J., & Schultz, W. (2006). Human neural learning depends on reward prediction errors in the blocking paradigm. *Journal of Neurophysiology, 95,* 301–310.

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

Vadillo, M. A., Castro, L., Matute, H., & Wasserman, E. A. (2008). Backward blocking: The role of within-compound associations and interference between cues trained apart. *Quarterly Journal of Experimental Psychology, 61*, 185–193.

Van Hamme, L. J., & Wasserman, E. A. (1994). Cue competition in causality judgments: The role of nonpresentation of compound stimulus elements. *Learning and Motivation, 25*, 127–151.

Van Overwalle, F. (2007). *Social connectionism: A reader and handbook for simulations*. New York: Psychology Press.

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

Waelti, P., Dickinson, A., & Schultz, W. (2001). Dopamine responses comply with basic assumptions of formal learning theory. *Nature*, *412,* 43–48.

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.

Footnote.

#1. Additionally, we used a non-parametric technique to confirm that we choose the best time-window for the FRN component (Maris & Oostenveld, 2007). We applied this technique in the learning bin 1 condition because this is the condition in which a larger FRN is expected. The result of the Maris method indicated that the best time window for our procedure was 244-360 ms, what is almost identical to the time windows used for our analyses.

Table 1.

Experimental design for each experimental block.

Trial types	**Relationships programmed**	**Outcome frequency**		**Predictions of RW (Allergy responses)**
		Allergy	No Allergy	
Training trials	A-Allergy	10	2	↑
	A**B**-Allergy	10	2	↑
	C-No Allergy	2	10	↓
	C**P**-Allergy	10	2	↑
Filler trials	EF-No Allergy	4	20	↓
Test trials	**B**-Allergy/No Allergy	4	4	**B<P**
	P-Allergy/No Allergy	4	4	

Note: Letters stand for either cues or stimuli. Target cues B and P are printed in bold. The last column gives qualitative predictions from the RW model for each trial-type: upward arrows indicate strong 'cue(s) – Allergy' associations, whereas downward arrows indicate weak 'cue(s) – Allergy' associations. In the test trials, the RW model predicts a stronger 'P-Allergy' than 'B-Allergy' association. Different objects from 6 categories were pseudorandomly assigned to the different cues. Please refer to the main text for further explanation.

Figure captions

Figure 1. Task design. This figure presents a fictitious AB → Allergy trial. A) A fixation cross is shown in the center of the screen. B) The stimuli (or stimulus) that represent cue(s) in each particular trial were presented for 1.2 seconds. This figure shows two stimuli, which correspond to cues A and B. C) Response screen indicating an allergy response. D) After the response, a blank screen was presented for 1 s prior to presentation of the E) feedback for 1 second. In this example, the response was correct; therefore, the feedback was positive.

Figure 2. Proportion of allergy responses across learning. Error bars represent the standard error of the mean.

Figure 3. Behavioral effects of cue interactions. Cue interactions can be seen in the proportion of allergy responses. Specifically, participants responded allergy more often in predictive trials than in blocked trials. Error bars represent the standard error of the mean.

Figure 4. Grand-averaged ERP responses in feedback across learning. Different columns represent the learning effect across the four learning bins. The arrows indicate the ERPs for electrodes Fz, Cz and Pz. The solid and dashed lines represent ERPs following positive feedback and negative feedback, respectively. Gray rectangles indicate the time window used for statistical analyses. Differences between positive and negative feedback diminished with learning. This effect was most pronounced in the Fz electrode. The topographical distributions of the FRN are provided at the bottom of the figure (positive minus negative feedback; 40-ms interval centered on the peak amplitude value; relative scale, minimum/maximum values for each map: bin 1, −2.5/0 μV; bin 2 and bin 3, −1.5/1 μV; bin 4, −0.1/1.5 μV).

Figure 5. Cue interaction in FRN. Grand-averaged feedback-locked ERP responses in test trials. The figure depicts ERPs in blocked (cue B)/predictive (cue P) trials. The solid and

dashed lines represent ERPs following positive feedback and negative feedback, respectively. Gray rectangles indicate the time window used for statistical analyses. The topographical distributions of the FRN are provided at the bottom of the figure (positive minus negative feedback; 40-ms interval centered on the peak amplitude value; scale, minimum/maximum values for both maps: −2.8/-0.5 μV).

Figure 6. Scalp distribution of voltage and current source density (CSD) maps for positive feedback across trials. The last column gives the difference between bin 1 and bin 4, which indicates the scalp distribution for the learning effect. These maps show possible frontocentral sources (time window 350-400 ms; relative scale, minimum/maximum values for each map: mean voltage maps, bin 1, bin 2, bin 3 and bin 4 −7/12 μV, bin 1 minus bin 4 −0.1/1.5 μV; CSD maps, bin 1, bin 2, bin 3 and bin 4 −0.02/0.02 mV/mm^2, bin 1 minus bin 4 −0.007/0.007 mV/mm^2).

Figure #1

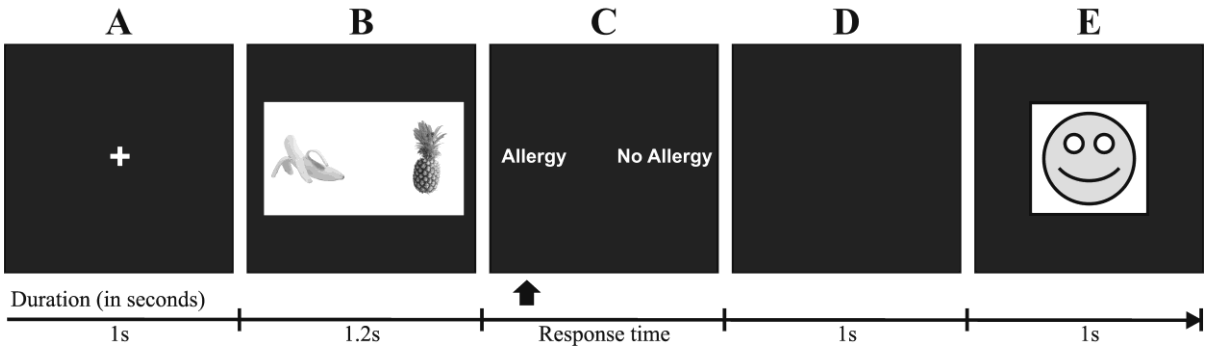

A
B
C
D
E
+
Allergy
No Allergy
Duration (in seconds)
1s
1.2s
Response time
1s
1s

Figure #2

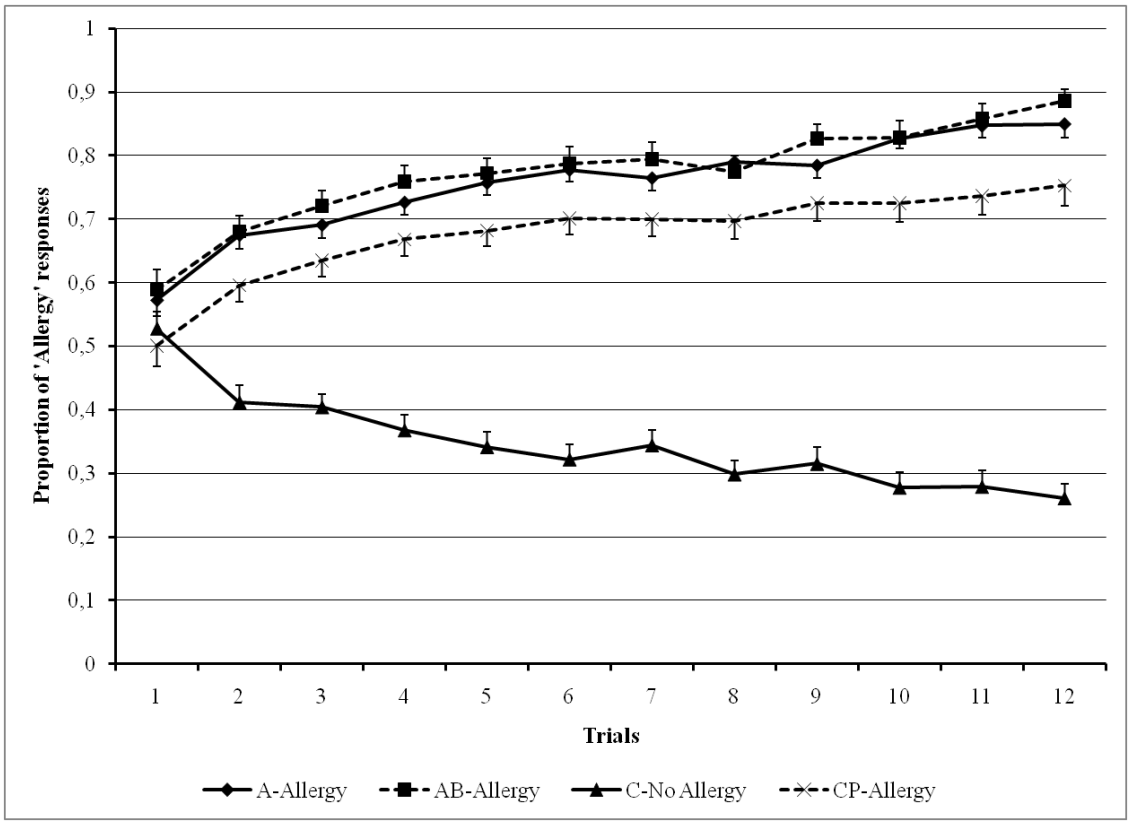

1
0,9
0,8
0,7
0,6
0,5
0,4
0,3
0,2
0,1
0
Proportion of 'Allergy' responses
1
2
3
4
5
6
7
8
9
10
11
12
Trials
A-Allergy
AB-Allergy
C-No Allergy
CP-Allergy

Figure #3

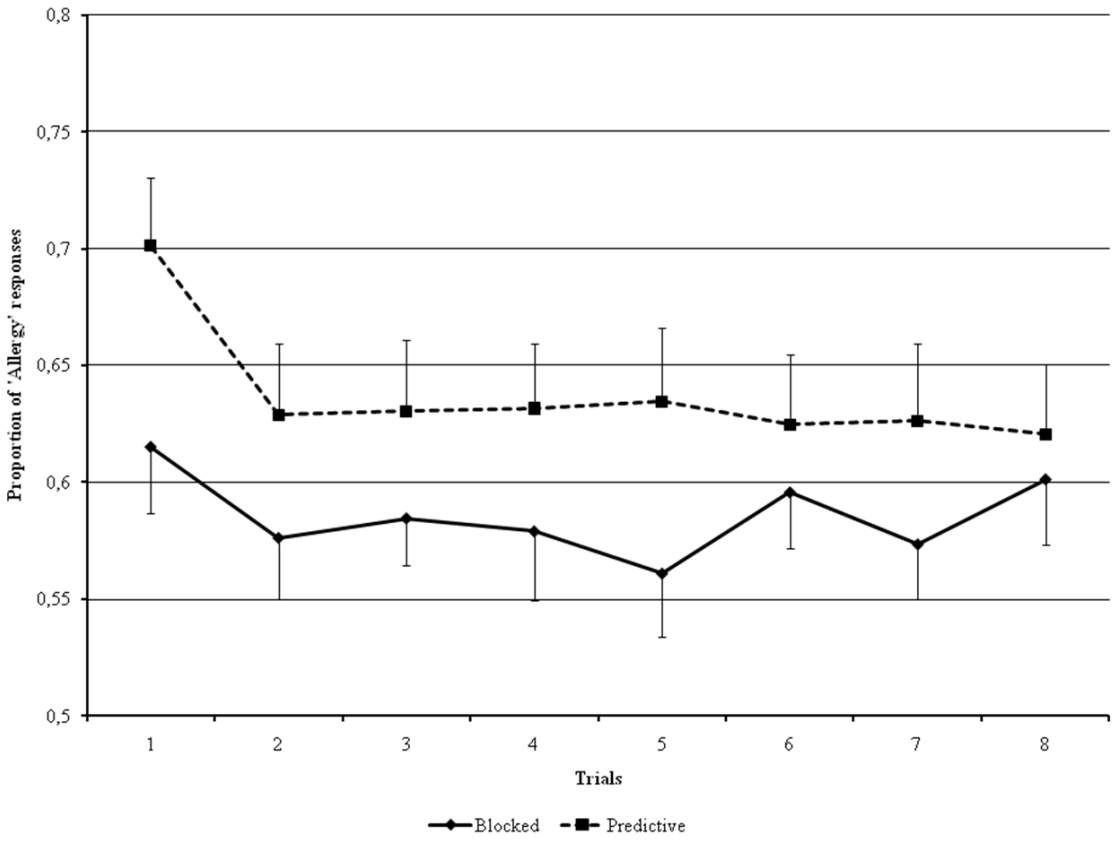
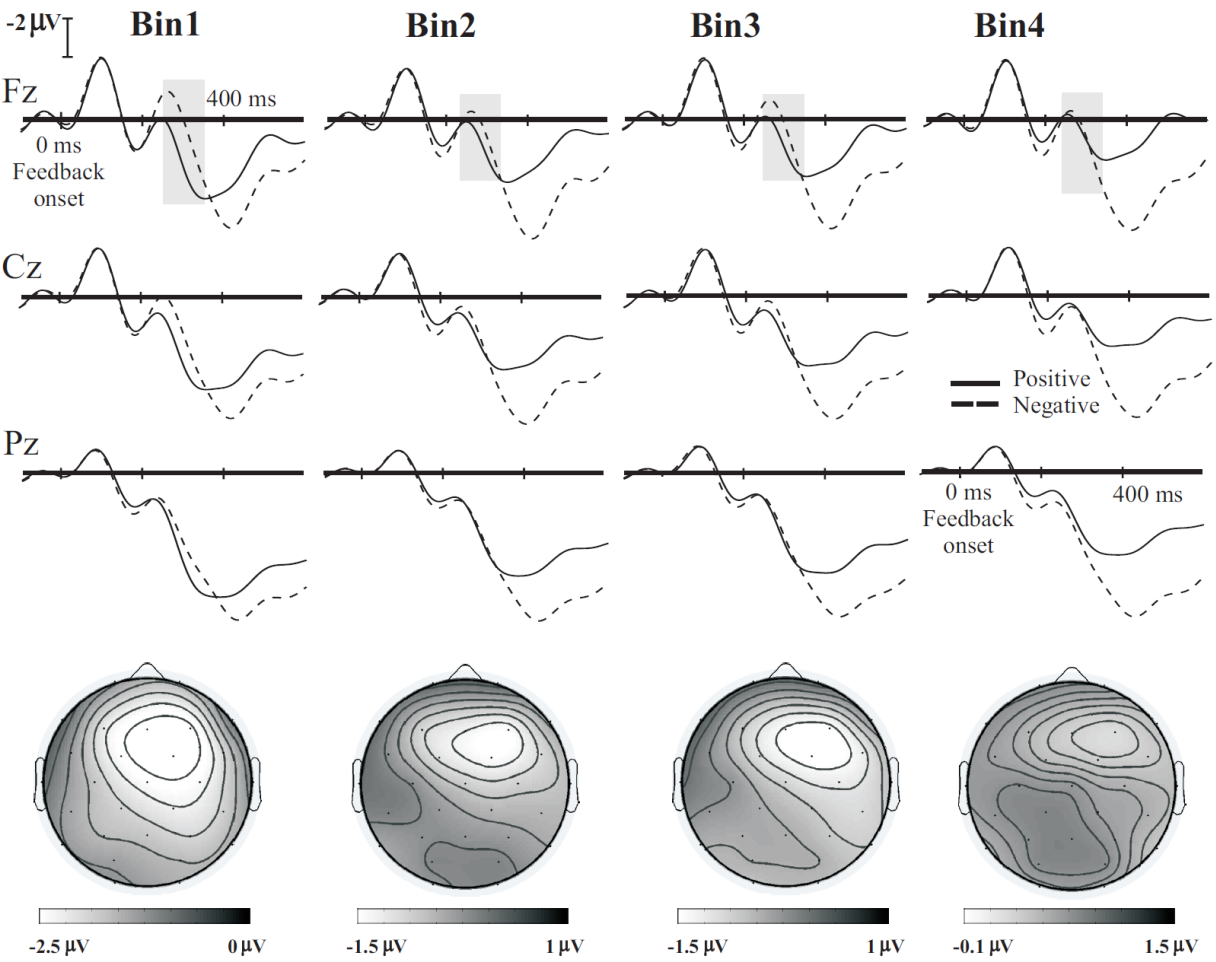

0,8
0,75
0,7
0,65
0,6
0,55
0,5
Proportion of 'Allergy' responses
1
2
3
4
5
6
7
8
Trials
Blocked
Predictive

Figure #4


-2 µV
Bin1
Bin2
Bin3
Bin4
Fz
400 ms
0 ms
Feedback
onset
Cz
Positive
Negative
Pz
0 ms
Feedback
onset
400 ms
-2.5 µV
0 µV
-1.5 µV
1 µV
-1.5 µV
1 µV
-0.1 µV
1.5 µV

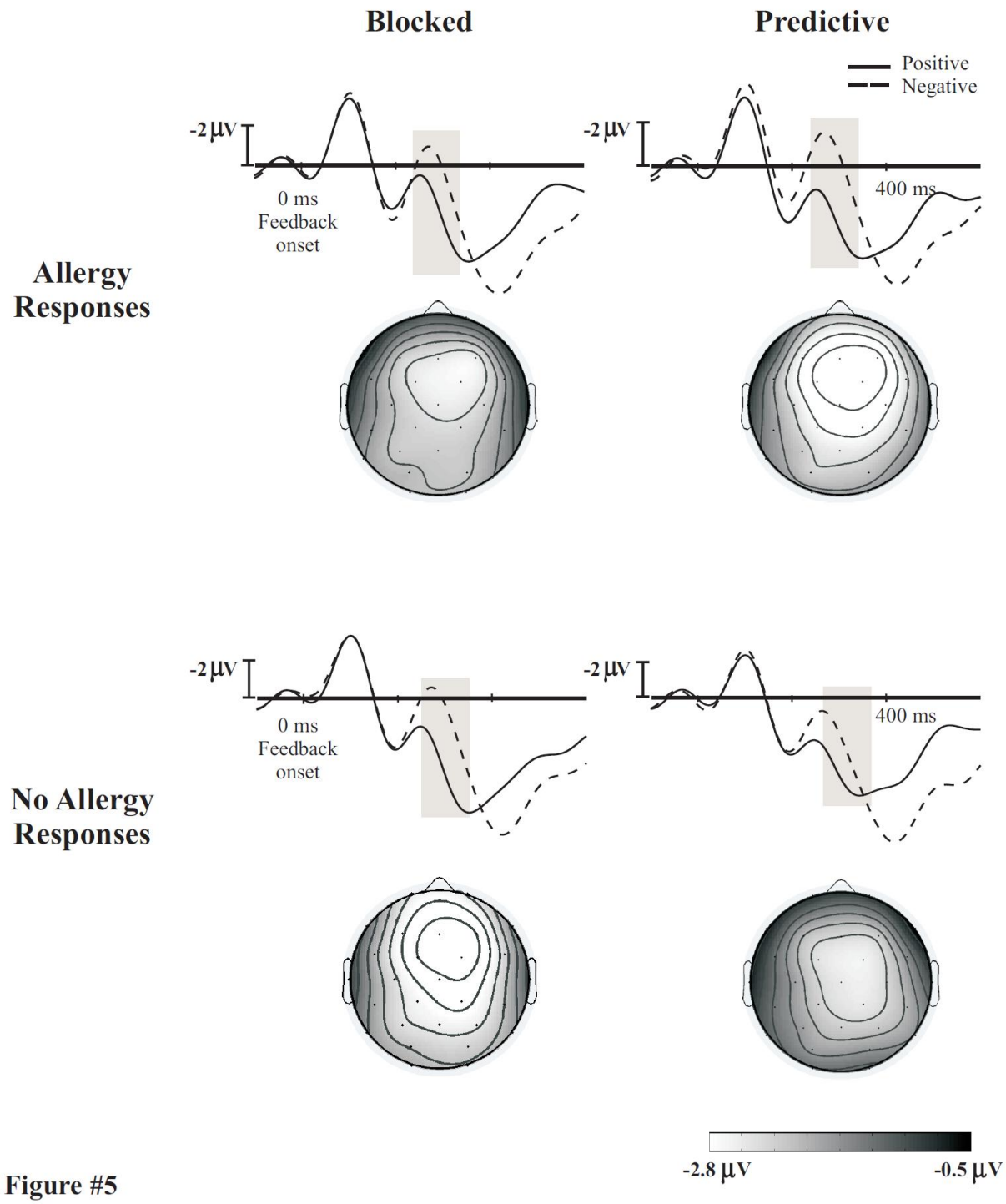

Blocked
Predictive
Positive
Negative
-2 µV
0 ms
Feedback
onset
400 ms
Allergy
Responses
No Allergy
Responses
-2.8 µV
-0.5 µV

Figure #5

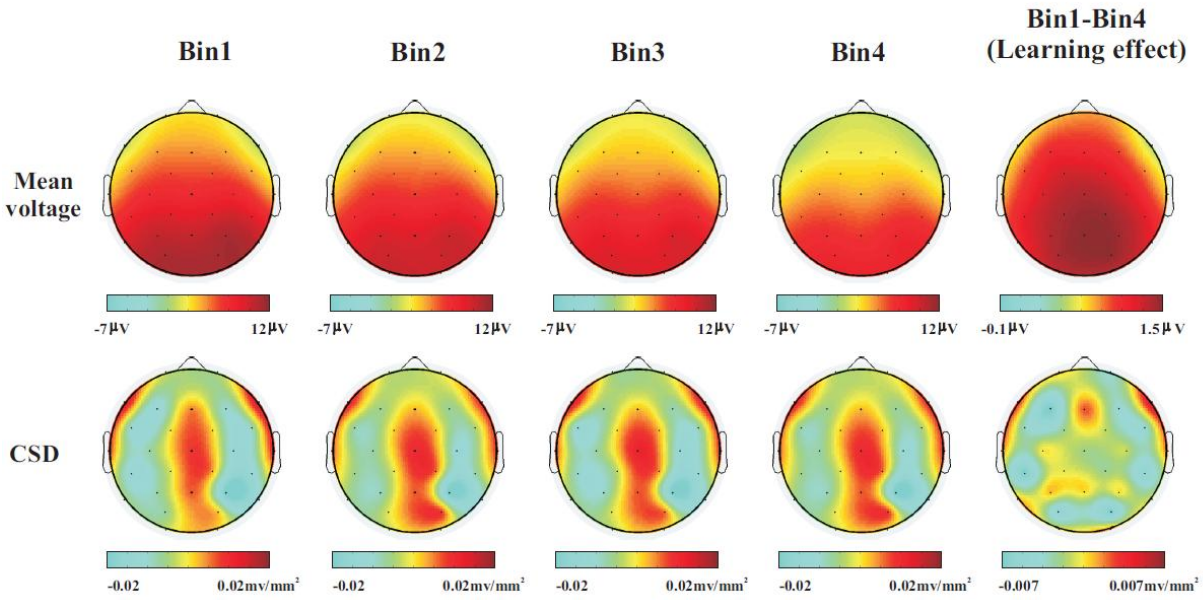

Bin1
Bin2
Bin3
Bin4
Bin1-Bin4
(Learning effect)
Mean
voltage
-7μV
12μV
-7μV
12μV
-7μV
12μV
-7μV
12μV
-0.1μV
1.5μV
CSD
-0.02
0.02mv/mm²
-0.02
0.02mv/mm²
-0.02
0.02mv/mm²
-0.02
0.02mv/mm²
-0.007
0.007mv/mm²

Figure #6