Social Cognitive and Affective Neuroscience Advance Access published March 12, 2016

Running head: BINARY SENSITIVITY FOR GAIN AND LOSS

Binary sensitivity of theta activity for gain and loss when monitoring parametric prediction errors

Denise J.C. Janssen, Edita Poljac, & Harold Bekkering

Radboud University, Donders Institute for Brain, Cognition and Behaviour, 6500 GL Nijmegen, the Netherlands

Address correspondence to:

Denise Janssen, Department of Cognitive Psychology, Donders Centre of Cognition, Radboud University, Montessorilaan 3, 6525 HR Nijmegen, the Netherlands. Email: d.janssen@donders.ru.nl

Telephone: 024-3612635

Abstract

Several theories have been proposed to account for the medial frontal activity that is elicited during the evaluation of outcomes. Respectively, these theories claim that the medial frontal response reflects (1) the absolute deviation between the value of an outcome and its expected value (i.e. an absolute prediction error); (2) the deviation between actual and expected outcomes, with a specific sensitivity to outcomes that are worse than expected (i.e. a negative prediction error); (3) a binary evaluation in terms of good and bad outcomes. In the current electroencephalography study, participants were presented with cues that induced specific predictions for the values of trial outcomes (a gain or loss of points). The actual outcomes occasionally deviated from the predicted values, producing prediction errors with parametrically varying size. Analysis of the medial frontal theta activity in response to the outcomes demonstrated a specific sensitivity to the occurrence of a loss of points when a gain had been predicted. However, the absolute deviation with respect to the predicted value did not modulate the theta response. This finding is consistent with the idea that outcome monitoring activity measured over medial frontal cortex is sensitive to the binary distinction between good and bad outcomes.

Keywords: EEG, FRN, medial frontal cortex, outcome evaluation, performance monitoring

Total number of words: 4944

Many decisions that we make, as well as events that we observe, can have beneficial or detrimental consequences. Research into the neural basis of outcome evaluation has shown that the medial frontal cortex plays a critical role here (Cohen and Ranganath, 2007; Donamayor *et al.*, 2011; Muller *et al.*, 2005; Nieuwenhuis *et al.*, 2005b; Ruchsow *et al.*, 2002). From this brain region, outcome monitoring activity manifests itself as the feedback-related negativity (FRN) when measured in the time domain and as theta oscillations when measured in the time-frequency domain (Cavanagh *et al.*, 2010; Cohen *et al.*, 2007; Crowley *et al.*, 2014; Luu *et al.*, 2003; Marco-Pallares *et al.*, 2008).

A number of models have been proposed to account for the outcome monitoring activity measured over medial frontal cortex. Holroyd and Coles presented the influential theory that it reflects a signed prediction error (Holroyd and Coles, 2002). Specifically, medial frontal activity was thought to be elicited in response to outcomes that are worse than expected. Other authors have suggested that any surprising outcome, whether better or worse than expected, will evoke medial frontal activity that scales with the degree of surprise (Alexander and Brown, 2011; Jessup *et al.*, 2010; Oliveira *et al.*, 2007; Pfabigan *et al.*, 2015). Still others have proposed that the medial frontal activity reflects a categorical good-bad evaluation (Hajcak *et al.*, 2006; Holroyd *et al.*, 2006; Sato *et al.*, 2005; Toyomaki and Murohashi, 2005; Yeung and Sanfey, 2004). The differences between these theories stem from discrepancies among research findings, which might have been driven by methodological issues such as confounds in the designs, or by variability in quantification of the neural response, see (Martin, 2012; Sambrook and Goslin, 2015).

To date, perhaps the best indication as to which model most accurately represents the outcome monitoring activity follows from a recent meta-analysis (Sambrook and Goslin, 2015). Deriving a "great grand average" from FRN studies, this meta-analysis found significant modulating effects of the valence, probability and magnitude of outcomes, which seems to be compatible with a signed prediction error model. Yet, this meta-analysis could only use binary factors to assess the effects of probability and magnitude (i.e. likely versus unlikely and high versus low). To further examine the validity of the proposed models, it is important to also take into account the hypothesized parametric character of the prediction error models (Rushworth and Behrens, 2008; Yacubian *et al.*, 2006).

For the current study, we developed a simple decision making task with gains and losses of several magnitudes to induce prediction errors with parametrically varying size. Of interest was the outcome monitoring activity in response to the deviations between the predicted and actual outcomes. Our analysis focused primarily on medial frontal activity in terms of theta power, as the time-frequency representation provides a richer and more robust alternative to the inconsistent quantification of the FRN (Cohen *et al.*, 2011; Crowley *et al.*, 2014). By analyzing outcome monitoring activity according to the rationale of the three abovementioned models, we aimed to establish which of these would provide the best fit. Thus, we tested if outcome monitoring activity over medial frontal cortex displays a sensitivity for unsigned quantitative, signed quantitative, or categorical good-bad deviations between predictions and outcomes.

Materials and Methods

Participants

Twenty-five healthy volunteers (eighteen females, 19-34 years of age) participated in this study. Participants gave written informed consent and received financial compensation for their time. Three participants were excluded from further data analyses: one after reporting not to have attended the stimuli at all times and the other two due to early termination for personal reasons unrelated to the experiment. The data of twenty-two participants (sixteen females) were eventually included in the analyses. The study was approved by the Radboud University institutional ethical review board (ECG2012-0910-058 DCC-NWO-EUea-Bekkering).

Stimuli and task

We designed a novel paradigm to induce precise predictions of the value of trial outcomes (+30, +10, -10, or -30 points), as well as precise perception of the deviations between predicted and actual outcomes (+60, +40, +20, 0, -20, -40, or -60 points). To emphasize the parametric nature of the design, predictive cues and actual outcomes were presented by means of levels on a vertical score bar (see the left panel in Figure 1 for an overview of the predictive and outcome stimuli). By default, the score bar was filled up to 50% of its height. A predictive cue indicated the likely outcome of the trial, which was a change in level on the score bar. These cues consisted of arrows whose length corresponded to a number of points. Starting from the midline level at 50%, an arrow pointing upward could either indicate the score bar level of 60% (predicting a gain of 10 points) or the score bar level of 80% (predicting a gain of 30 points). Similarly, arrows pointing downward predicted a loss of 10 or 30 points, respectively. At the end of the trial, the actual outcome was presented by means of a change in the filling of the score bar: 80% filled (+30 points), 60% filled (+10 points), 40% filled (-10 points), or 20% filled (-30 points). Predictive cues had a validity of .75, meaning that deviations between predicted and actual outcomes occurred in 25% of the trials. The four predictive stimuli and the four outcome stimuli occurred with equal frequencies. Different stimulus features were chosen for predictive cues and outcomes (i.e. arrows versus level changes) in order to prevent perceptual mismatch from modulating the ERPs, see (Folstein and Van Petten, 2008).

(please enter Figure 1 around here)

As passive settings are not optimal for eliciting outcome monitoring activity (Itagaki and Katayama, 2008; Martin and Potts, 2011), we introduced a task to the participants. The task was moreover chosen to provide a context for the predictive and outcome stimuli. Specifically, participants interacted with a virtual pet, which could either give or take points. Participants were instructed that the predictive cues reflected the intention of the virtual pet at the onset of a trial.

While it was likely that the pet would act according to its intention, there was a small chance that it would change its mind, leading to a different outcome. Participants were not informed about the exact probabilities. The task was either to feed or to cuddle the pet by pressing a corresponding button, which was specified based on the posture of the virtual pet. Two of the four postures were associated with feeding and the other two with cuddling. After the participant interacted with the pet by pressing the specified button, the pet would make its final decision, as indicated by the outcome stimulus (a gain or loss of points) that then appeared on the screen. Importantly, the task was programmed such that the outcome of a trial was independent of the performance on that trial. Although participants were not informed about this beforehand, most of them reported afterwards that they had soon realized that their task performance did not affect the outcomes. This indicates that participants knew that the predictive cues contained all the relevant information needed for monitoring, which allowed them to anticipate the outcomes of the trials.

To make sure that participants paid attention to the score bar, they were instructed to a) keep track of the predictive cues and the outcomes, and to b) report on their performance in terms of gain and loss of points after each block. Specifically, after each block of trials, the participants were asked to estimate if their overall number of points had increased, decreased or stayed the same over the course of that block. Participants were instructed not to explicitly calculate, but rather to develop a gut feeling regarding their overall performance during a block. Note however that the actual amount of win and loss was equal in each block, as all prediction-outcome combinations were randomized block-wise. As blocks were lengthy and gains and losses came in two magnitudes, performance estimations proved to be challenging for participants and as such required their active participation. To further enhance the perceived significance of the score bar, participants were told that the total number of points won (or lost) at the end of the experiment would affect their monetary compensation. Afterwards the experimenter explained that the outcomes were predetermined and summed to zero. The monetary compensation was therefore 20 Euros for each participant.

Procedure and design

The onset of a trial was marked by the appearance of the score bar in the right half of the screen, showing the predictive cue. The score bar remained visible for the full duration of a trial. After 500 ms, one of the four postures of the virtual pet was presented in the middle of the screen, with the two options “feed” and “cuddle” to its left and its right. The positions of “feed” and “cuddle” were interchanged randomly over trials and the correct button press (left or right) corresponded to the current position of the correct option. Once the participant had pushed the correct button, a frame appeared around the virtual pet, indicating that the outcome stimulus would follow shortly. Incorrect button presses did not evoke any effect, meaning that participants needed to correct their errors. The predictive arrow disappeared from the score bar 500 ms after the appearance of the frame, at which time the level of the score bar changed, showing the outcome of the trial. After an interval of 1000 ms, a white screen with a fixation cross was shown for 500 ms, bridging the inter-trial interval. This sequence of trial events is illustrated in the right panel of Figure 1.

Before the start of the experiment, the participant was explicitly informed about the stimulus-response mappings of the virtual pet task. Subsequently, the experimenter guided the participant through 10 practice trials and explained how the levels of the score bar mapped onto the numerical outcome values. During the practice part, the participant was instructed to keep the eye gaze fixated on the centre of the screen, and to attend the score bar in a covert manner, in order to prevent saccades. In addition, the participant was asked to try to limit eye blinks from occurring outside of inter-trial intervals. The experiment itself consisted of a total of 1200 trials, divided over 13 blocks (12 blocks with 96 trials and the last block with 48 trials). After each block, the participant took a short break (self-paced). In addition, there were breaks of one minute duration after blocks 3, 5, 8 and 10. After finishing the experiment, participants filled in a questionnaire, which was designed to assess their rating of the various prediction-outcome combinations in terms of satisfaction. All 16 combinations were listed in a randomized order and were rated on a 7-point scale, where the left-most option corresponded to very unsatisfying, the middle option corresponded to a neutral feeling, and the right-most option corresponded to very satisfying. The total duration of the participant's visit to the lab was approximately two hours.

Electrophysiological Recording and Processing

EEG was recorded with active Ag/AgCl electrodes (ActiCap), Brain Amp DC and Brain Vision Recorder software (Brain Products GmbH, Germany). The signal was obtained from 28 scalp sites according to the international 10-10 system: Fp1, Fp2, F7, F3, Fz, F4, F8, FC5, FC1, FCz, FC2, FC6, T7, C3, Cz, C4, T8, CP5, CP1, CP2, CP6, P7, P3, Pz, P4, P8, O1, O2 and the right mastoid, with AFz as the common ground and the left mastoid as the online reference. In addition, electrooculography (EOG) was measured from the outer canthi of both eyes and above and below the left eye. The data were obtained with a sampling rate of 500 Hz and filtered online with a low cut-off at 0.016 Hz and a high cut-off at 125 Hz.

Pre-processing of the EEG was done using the Fieldtrip toolbox (Oostenveld *et al.*, 2011) for Matlab (MathWorks, Natick, MA). Electrophysiological activity related to the processing of the predictive cues and the outcome stimuli was extracted in segments starting 3000 ms before stimulus onset and 3000 ms after stimulus onset. The data were re-referenced to the averaged mastoids and the interval between 500 and 200 ms preceding the stimulus was used for baseline correction. By means of an independent component analysis, electrophysiological activity related to eye movements was identified and corrected for (Lee *et al.*, 1999). Remaining sources of noise were removed based on summary plots of the variance in all 1200 trials and 28 scalp electrodes, which allowed us to identify extreme data points.

We continued to use the Fieldtrip toolbox to acquire time-frequency representations, transforming the time series data with Morlet wavelets. The power spectrum was assessed for frequencies ranging from 2 to 30 Hz in the time interval between 500 ms preceding the stimulus and 1000 ms following stimulus onset. Using 7 cycles per wavelet, the spectral bandwidth at a given frequency F was equal to F/14 Hz. The outcome monitoring activity was derived from the power in the theta band (4-8 Hz), measured at electrode FCz between 200 and 500 ms following stimulus onset, which was based on previous studies (Luft, 2014; van de Vijver *et al.*, 2011). Though not of primary interest, we also measured outcome monitoring as reflected by the delta frequency band (2-4 Hz at electrode FCz). This was done because it has been suggested that activity in the theta and delta frequency bands may play different roles in outcome processing (Foti *et al.*, 2015). To complement the time-frequency data, outcome monitoring was also measured in the time domain. After applying a low-pass filter at 9 Hz, the peak-to-peak amplitude of the FRN, typically associated with outcome monitoring, was assessed by measuring the negative peak in the time interval from 250 ms to 350 ms, relative to the preceding positive peak. In addition to the outcome stimuli, we also analyzed the neural responses following the predictive cues. As we had no a priori information regarding the time-frequency characteristics of outcome anticipation, the analysis of predictive cue processing was done with a data-driven approach. Visual inspection of the time-frequency representation of electrode FCz revealed activity in the delta frequency range (2-4 Hz) between 300 and 1000 ms (see Figure 2, left panel), which we further inspected in the statistical analysis. As the prediction-outcome combinations had different occurrence rates (i.e. predicted versus unpredicted outcomes), we randomly sampled 25 trials from each of the frequent conditions to keep the number of trials equal in all conditions.

(please enter Figure 2 around here)

Models for outcome monitoring

The collected data were analyzed according to the rationale of three different theoretical approaches. The first approach, which we will refer to as the *unsigned quantitative model*, assumed that theta power would scale with the size of the absolute prediction error. This equals the absolute difference between the actual outcome *R* and the expected value, the latter being determined by all possible outcomes *x* and their respective probabilities *p* (Yacubian *et al.*, 2006). The unsigned prediction error δ_u thus follows from the following formula:

$$\delta_u = [R - \sum_{i=1}^{n}(x_i * p_i)].$$

The second model, which we will call the *signed quantitative model*, was slightly more complex, as it assumed that negative prediction errors had a different impact than positive prediction errors. This model therefore added a slope term S_0 and an intercept S_1 to account for the sign of the prediction error δ_s:

$$\delta_s = [R - \sum_{i=1}^{n}(x_i * p_i)] * S_0 + S_1.$$

Finally, the *categorical model* assumed that outcomes would be processed according to a binary good-bad distinction. Irrespective of the actual size of the deviation, this model categorized prediction errors δ_c as “negative” when a predicted gain was followed by a loss, and as “positive” when a predicted loss was followed by a gain. Gains followed by gains and losses followed by losses were assumed to not elicit prediction errors. The corresponding model contained separate intercepts for positive categorical prediction errors ($C_1 \neq 0$ if $R > 0$ and $EV < 0$) and negative categorical prediction errors ($C_2 \neq 0$ if $R < 0$ and $EV > 0$):

$$\delta_c = C_1 + C_2.$$

Statistical Analyses

Our main analysis focused on outcome monitoring as reflected by medial frontal theta activity. With a generalized linear mixed model analysis, we compared the fit of the unsigned quantitative model, signed quantitative model, and categorical model to the theta response following outcomes. Aside from the model-specific factors (described in the paragraph above), each analysis included parameters for the intercept of the fixed effect, for the subject-based intercept, and for the repeated measures. This resulted in a total of four parameters for the unsigned quantitative model, six parameters for the signed quantitative model (including the interaction term for the prediction error and the sign), and five parameters for the categorical model (since two parameters were required to model the categorical prediction errors). A robust estimation of covariances was applied to handle potential violations of model assumptions. To evaluate the explanatory value of the three models, the corrected Akaike Information Criterion (AIC) was used to calculate the models' likelihood with respect to one another (Burnham *et al.*, 2011; Glover and Dixon, 2004). Please note that the AIC measure corrects for model complexity, such that the assessment of model fit is not biased by the number of parameters (Burnham *et al.*, 2011). To complement the results for the theta response, we repeated the same analysis for outcome-locked delta activity and the FRN.

In addition, we performed another two analyses. As we expected the neural activity following predictive cues to reflect the generation of outcome predictions, we were interested in finding out whether participants selectively attended the valence or magnitude information that was provided by the predictive cues. Inspection of the time-frequency representation revealed activity in the medial frontal delta band (2-4 Hz). We entered the medial frontal delta power from 300 to 1000 ms after the onset of predictive cues into a mixed model analysis that tested the effects of Magnitude (10, 30), Valence (gain, loss) and their interaction. Lastly, we examined whether the subjective experience of the outcomes was analogous to the processing characteristics of the theta frequency band. To this end, we coded the satisfaction ratings on a numerical scale from -3 (most unsatisfying) to +3 (most satisfying) and entered these scores in three mixed model analyses. Again, we derived the corrected AIC to assess if the unsigned quantitative model, the signed quantitative model, or the categorical model provided the best fit.

The superior models were analyzed in more detail to illustrate the neural processing and subjective experience of the different types of outcomes. Bonferroni correction was applied for multiple comparisons. All statistical analyses were performed with IBM SPSS statistics, version 19 (Armonk, NY: IBM corp.).

Results

Theta activity following outcomes

We evaluated the three theoretical models for outcome processing according to the degree to which they accounted for the theta activity in response to deviations between predicted and actual outcomes in our task (see Table). The lowest corrected AIC was observed for the categorical model (5836.92), followed by the signed quantitative model (5852.40) whereas the unsigned quantitative model produced the highest value (5882.24). As lower information criterion values indicate a better model fit, it follows that the categorical model most accurately represented the data. In fact, when we derived the Akaike Weights to assess the relative likelihood of the models, the categorical model turned out to have 99.96% probability of providing the best fit (i.e. $p < .001$). The signed quantitative model followed with a relative likelihood of .04%.

In a subsequent analysis, we took a closer look at the effects described by the categorical model. As illustrated in the top panel of Figure 3, theta power differed significantly depending on the categorical type of prediction error ($F(2, 349) = 7.90$, $p < .001$). Negative categorical prediction errors evoked a stronger theta response compared to both positive categorical prediction errors ($t(349) =$

3.02, *p* = .005) and the absence of categorical prediction errors (*t*(349) = 3.97, *p* < .001). See the right panel of Figure 2 for the time-frequency representation of this sensitivity to unpredicted losses. In response to positive categorical prediction errors, there was also a stronger theta response than when categorical prediction errors were absent, but this effect fell short of reaching statistical significance (*t*(349) = 1.90, *p* = .059). Together, these findings demonstrate that the role of theta oscillations in outcome monitoring is strongly related to categorical negative surprise.

(please enter Figure 3 around here)

Delta activity and FRN following outcomes

For completeness, we also analyzed outcome monitoring as reflected by delta activity and the FRN. Importantly, similar to the theta response, delta activity and the FRN were predicted most accurately by the categorical model (see Table). Subsequent inspection of the categorical model revealed that the categorical type of prediction error significantly modulated the delta response *F*(2, 349) = 3.64, *p* = .027. Specifically, negative categorical prediction errors elicited a stronger delta response than both positive categorical prediction errors (*t*(349) = 2.60, *p* = .029) and the absence of categorical prediction errors (*t*(349) = 2.26, *p* = .049). The modulation of the FRN by categorical prediction errors did not, however, reach significance (*F*(2, 349) = 1.54, *p* = .22). So, even though the numerical effects were in the same direction for the FRN as for the delta and theta oscillations, our data indicate that the FRN was less sensitive in distinguishing categorical prediction errors than the time-frequency counterparts. As the ERP waveforms in Figure 4 seem to suggest, time jitter in the occurrence of the FRN peak as well as potential effects of component overlap could have contributed to this lack of sensitivity (Sambrook and Goslin, 2015).

(please enter Figure 4 around here)

(please enter Table 1 around here)

Delta activity following predictive cues

The processing of outcomes described above was preceded by an interval initiated by the predictive cues. Our explorative analysis of the power in the delta frequency band during this interval revealed a specific sensitivity to the impact of the predicted values. As illustrated in the left panel of Figure 2, large magnitudes (+30 and -30 points) elicited a stronger response than small magnitudes (+10 and -10 points, $F(1, 84) = 6.09$, $p = .016$). Interestingly, delta activity in response to predicted gains did not differ from the activity in response to predicted losses ($F(1, 84) < 1$), nor did the predicted valence modulate the effect of predicted magnitude ($F(1, 84) < 1$). As such, the activity following predictive cues was qualitatively different from the activity following outcomes.

Subjective experience of outcomes

The three theoretical models for outcome monitoring were also fitted to the subjective experience that participants reported after finishing the computer task. This time, the lowest corrected AIC value was observed for the signed quantitative model (1146.89), whereas the categorical model and the unsigned quantitative model provided a worse fit (1222.97 and 1386.31, respectively). With a relative likelihood of more than 99.99%, the signed quantitative model was most accurate in explaining the satisfaction ratings (i.e. $p < .001$). Further analysis of this model showed that satisfaction ratings decreased as a function of the size of negative prediction errors ($\beta = -0.049$, $SE = .005$, $t = -9.32$, $p < .001$) and increased with the size of positive prediction errors ($\beta = 0.020$, $SE = .008$, $t = 2.42$, $p = .017$). From this, it follows that the impact of negative prediction errors on satisfaction ratings was 2.46 times as large as the impact of positive prediction errors. See Figure 5 for a graphical overview of the subjective experience of the prediction-outcome combinations.

(please enter Figure 5 around here)

Discussion

In this study, participants observed a broad range of predicted values and (occasionally) deviating outcomes, which served as a test for evaluating the adequateness of three models in explaining the neural correlates of outcome processing. Despite the parametric nature of the deviations between predictions and outcomes, the theta response was elicited in a categorical manner, being specifically sensitive to the occurrence of a loss when a gain had been predicted. This finding lends support to the notion that the medial frontal outcome monitoring system evaluates outcomes according to a binary good-bad distinction (Hajcak *et al.*, 2006; Holroyd *et al.*, 2006; Sato *et al.*, 2005; Toyomaki and Murohashi, 2005; Yeung and Sanfey, 2004).

Detail of outcome predictions

At a general level, medial frontal activity in response to gains and losses is conceptualized as reflecting prediction errors (Walsh and Anderson, 2012). When comparing the proposed theoretical accounts, the differences seem to lie in the complexity inherent to the predictions. In their most complex form, models for reward prediction errors take into account possible outcome values, valence and probabilities (Rushworth and Behrens, 2008). In contrast, a more simple account assumes a binary distinction between gain and loss (Hajcak *et al.*, 2006). The activity in the theta frequency band observed in the current study is most accurately interpreted with the latter, categorical approach. However, this leaves the question whether the findings described here are an inherent feature of medial frontal monitoring activity, or rather the result of task-specific outcome processing.

A possible answer to this question is that outcome monitoring occurred at different levels of complexity in different neural assemblies. Consistent with this explanation, an fMRI study observed that some brain regions processed outcomes in a graded manner whereas other regions showed evidence for binary outcome processing (Nieuwenhuis *et al.*, 2005a). In that study, the binary sensitivity was particularly evident for the posterior cingulate gyrus, which is one of the regions thought to contribute to the outcome monitoring activity measured over medial frontal cortex (Cohen and Ranganath, 2007; Muller *et al.*, 2005; Nieuwenhuis *et al.*, 2005b). According to this logic, processes other than the medial frontal theta response would have elicited a more graded evaluation of the outcomes.

Alternatively, theta power during the monitoring of outcomes could have been modulated by the context of the task: neural monitoring was measured while participants covertly observed predictions and outcomes to develop a rough estimate of their score. It is possible that the ongoing medial frontal theta activity had been tuned according to the strategy that participants applied to perform the task. Specifically, they could have kept track of their score by intentionally using “quick and dirty” good-bad distinctions. A different task strategy that focused on the subtle differences between prediction-outcome combinations, which was evidently used when participants filled out the questionnaire, might have elicited a more graded theta response. Accordingly, it has been observed that outcome monitoring activity over medial frontal cortex is highly context dependent, being scaled according to the range of possible outcomes (Holroyd *et al.*, 2004). An interesting aim for future studies would be to examine whether this context-dependency is evident in the amount of monitored detail as well.

The role of negative surprise

Like many preceding studies (see (Walsh and Anderson, 2012) for a review), we observed that the medial frontal activity was driven by negative surprise. To explain the apparent importance of negative outcomes, it has been suggested that such events often indicate the need for behavioral adaptation (Cavanagh *et al.*, 2010). Accordingly, the medial frontal activity in response to negative outcomes could be related to the engagement of motor regions in the brain (Cohen *et al.*, 2011). Although the involvement of volitional action does indeed enhance medial frontal activity (Itagaki and Katayama, 2008; Martin and Potts, 2011), the effect of valence remains even after controlling for behavioral adaptation (von Borries *et al.*, 2013). An alternative explanation for the strong influence of negative outcomes emphasizes their affective significance (Gehring and Willoughby, 2002; Luu *et al.*, 2003; Moser and Simons, 2009). That is, people are also subjectively more sensitive to losses than to gains of equivalent size, with the impact of losses being roughly twice that of gains (Tom *et al.*, 2007). We observed a similar finding in the current study, with the impact of negative prediction errors on satisfaction ratings being 2.46 times as large as the impact of positive prediction errors.

Delta activity following predictive cues

As theta activity in response to outcomes reflected a valence-based binary prediction error, we had expected that the activity following predictive cues would also have been generated at that binary level. In other words, it would make sense if the binary value of the outcome (gain or loss) were compared to the binary value of the predictive cue (gain or loss). In contrast to this logic, the medial frontal delta activity that we observed following predictive cues did not demonstrate any sensitivity to valence. Rather, it was sensitive to the degree of impact that a predicted outcome would have on the score, whether positive or negative. The delta activity possibly reflected the time-frequency counterpart of the *stimulus preceding negativity* (SPN), a slow wave in the event-related potential that is known to be modulated by the affective salience of anticipated stimuli (Brunia *et al.*, 2011). Predicted outcomes with a large impact could accordingly have elicited a stronger affective response in anticipation. Even though we did not find evidence for valence-based anticipatory activity, we expect that such predictions were reflected by a different neural correlate not identified in the current study.

Together, the findings in the delta and theta frequency bands indicate that the predictive cues and outcomes were processed along two orthogonal binary axes: small versus large impact, and gain versus loss. The finding of activity specific to magnitude information has an important implication for the binary sensitivity to valence that was observed for the theta response. Namely, it indicates that the valence-specific response cannot be accounted for by a behavioral strategy which ignored magnitude information.

Conclusions

In the current study, medial frontal theta activity in response to outcomes demonstrated a specific sensitivity to negative surprise, irrespective of the quantitative deviation from predicted values. This finding is consistent with the suggestion that outcome monitoring activity over medial frontal cortex can be tuned to binary sensitivity and may in fact reflect a binary distinction between good and bad outcomes.

References

Alexander, W. H.Brown, J. W. (2011). Medial prefrontal cortex as an action-outcome predictor. *Nature Neuroscience*, 14(10), 1338-U1163.

Brunia, C. H., Hackley, S. A., van Boxtel, G. J., Kotani, Y.Ohgami, Y. (2011). Waiting to perceive: reward or punishment? *Clinical Neurophysiology*, 122(5), 858-868.

Burnham, K. P., Anderson, D. R.Huyvaert, K. P. (2011). AIC model selection and multimodel inference in behavioral ecology: some background, observations, and comparisons. *Behavioral Ecology and Sociobiology*, 65(1), 23-35.

Cavanagh, J. F., Frank, M. J., Klein, T. J.Allen, J. J. B. (2010). Frontal theta links prediction errors to behavioral adaptation in reinforcement learning. *Neuroimage*, 49(4), 3198-3209.

Cohen, M. X., Elger, C. E.Ranganath, C. (2007). Reward expectation modulates feedback-related negativity and EEG spectra. *Neuroimage*, 35(2), 968-978.

Cohen, M. X., Mimes, K. A.van de Vijver, I. (2011). Cortical electrophysiological network dynamics of feedback learning. *Trends Cogn Sci*, 15(12), 558-566.

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

Crowley, M. J., van Noordt, S. J. R., Wu, J., Hommer, R. E., South, M., Fearon, R. M. P., et al. (2014). Reward feedback processing in children and adolescents: Medial frontal theta oscillations. *Brain Cogn*, 89, 79-89.

Donamayor, N., Marco-Pallares, J., Heldmann, M., Schoenfeld, M. A.Munte, T. F. (2011). Temporal Dynamics of Reward Processing Revealed by Magnetoencephalography. *Hum Brain Mapp*, 32(12), 2228-2240.

Folstein, J. R.Van Petten, C. (2008). Influence of cognitive control and mismatch on the N2 component of the ERP: A review. *Psychophysiology*, 45(1), 152-170.

Foti, D., Weinberg, A., Bernat, E. M.Proudfit, G. H. (2015). Anterior cingulate activity to monetary loss and basal ganglia activity to monetary gain uniquely contribute to the feedback negativity. *Clinical Neurophysiology*, 126(7), 1338-1347.

Gehring, W. J.Willoughby, A. R. (2002). The medial frontal cortex and the rapid processing of monetary gains and losses. *Science*, 295(5563), 2279-2282.

Glover, S.Dixon, P. (2004). Likelihood ratios: A simple and flexible statistic for empirical psychologists. *Psychonomic Bulletin & Review*, 11(5), 791-806.

Hajcak, G., Moser, J. S., Holroyd, C. B.Simons, R. F. (2006). The feedback-related negativity reflects the binary evaluation of good versus bad outcomes. *Biological Psychology*, 71(2), 148-154.

Holroyd, C. B.Coles, M. G. H. (2002). The neural basis. of human error processing: Reinforcement learning, dopamine, and the error-related negativity. *Psychological Review*, 109(4), 679-709.

Holroyd, C. B., Hajcak, G.Larsen, J. T. (2006). The good, the bad and the neutral: electrophysiological responses to feedback stimuli. *Brain Research*, 1105(1), 93-101.

Holroyd, C. B., Larsen, J. T.Cohen, J. D. (2004). Context dependence of the event-related brain potential associated with reward and punishment. *Psychophysiology*, 41(2), 245-253.

Itagaki, S.Katayama, J. (2008). Self-relevant criteria determine the evaluation of outcomes induced by others. *Neuroreport*, 19(3), 383-387.

Jessup, R. K., Busemeyer, J. R.Brown, J. W. (2010). Error effects in anterior cingulate cortex reverse when error likelihood is high. *Journal of Neuroscience*, 30(9), 3467-3472.

Lee, T. W., Girolami, M.Sejnowski, T. J. (1999). Independent component analysis using an extended infomax algorithm for mixed subgaussian and supergaussian sources. *Neural Comput*, 11(2), 417-441.

Luft, C. D. (2014). Learning from feedback: The neural mechanisms of feedback processing facilitating better performance. *Behavioural Brain Research*, 261, 356-368.

Luu, P., Tucker, D. M., Derryberry, D., Reed, M.Poulsen, C. (2003). Electrophysiological responses to errors and feedback in the process of action regulation. *Psychological Science*, 14(1), 47-53.

Marco-Pallares, J., Cucurell, D., Cunillera, T., Garcia, R., Andres-Pueyo, A., Munte, T. F., et al. (2008). Human oscillatory activity associated to reward processing in a gambling task. *Neuropsychologia*, 46(1), 241-248.

Martin, L. E.Potts, G. F. (2011). Medial frontal event-related potentials and reward prediction: Do responses matter? *Brain Cogn*, 77(1), 128-134.

Martin, R. S. (2012). Event-related potential studies of outcome processing and feedback-guided learning. *Frontiers in Human Neuroscience*, 6.

Moser, J. S.Simons, R. F. (2009). The neural consequences of flip-flopping: The feedback-related negativity and salience of reward prediction. *Psychophysiology*, 46(2), 313-320.

Muller, S. V., Moller, J., Rodriguez-Fornells, A.Munte, T. F. (2005). Brain potentials related to self-generated and external information used for performance monitoring. *Clinical Neurophysiology*, 116(1), 63-74.

Nieuwenhuis, S., Heslenfeld, D. J., von Geusau, N. J., Mars, R. B., Holroyd, C. B.Yeung, N. (2005a). Activity in human reward-sensitive brain areas is strongly context dependent. *Neuroimage*, 25(4), 1302-1309.

Nieuwenhuis, S., Slagter, H. A., von Geusau, N. J. A., Heslenfeld, D. J.Holroyd, C. B. (2005b). Knowing good from bad: differential activation of human cortical areas by positive and negative outcomes. *European Journal of Neuroscience*, 21(11), 3161-3168.

Oliveira, F. T. P., McDonald, J. J.Goodman, D. (2007). Performance monitoring in the anterior Cingulate is not all error related: Expectancy deviation and the representation of action-outcome associations. *Journal of Cognitive Neuroscience*, 19(12), 1994-2004.

Oostenveld, R., Fries, P., Maris, E.Schoffelen, J. M. (2011). FieldTrip: Open source software for advanced analysis of MEG, EEG, and invasive electrophysiological data. *Comput Intell Neurosci*, 2011, 156869.

Pfabigan, D. M., Seidel, E. M., Paul, K., Grahl, A., Sailer, U., Lanzenberger, R., et al. (2015). Context-sensitivity of the feedback-related negativity for zero-value feedback outcomes. *Biological Psychology*, 104, 184-192.

Ruchsow, M., Grothe, J., Spitzer, M.Kiefer, M. (2002). Human anterior cingulate cortex is activated by negative feedback: evidence from event-related potentials in a guessing task. *Neuroscience Letters*, 325(3), 203-206.

Rushworth, M. F. S.Behrens, T. E. J. (2008). Choice, uncertainty and value in prefrontal and cingulate cortex. *Nature Neuroscience*, 11(4), 389-397.

Sambrook, T. D.Goslin, J. (2015). A Neural Reward Prediction Error Revealed by a Meta-Analysis of ERPs Using Great Grand Averages. *Psychological Bulletin*, 141(1), 213-235.

Sato, A., Yasuda, A., Ohira, H., Miyawaki, K., Nishikawa, M., Kumano, H., et al. (2005). Effects of value and reward magnitude on feedback negativity and P300. *Neuroreport*, 16(4), 407-411.

Tom, S. M., Fox, C. R., Trepel, C.Poldrack, R. A. (2007). The neural basis of loss aversion in decision-making under risk. *Science*, 315(5811), 515-518.

Toyomaki, A.Murohashi, H. (2005). Discrepancy between feedback negativity and subjective evaluation in gambling. *Neuroreport*, 16(16), 1865-1868.

van de Vijver, I., Ridderinkhof, K. R.Cohen, M. X. (2011). Frontal oscillatory dynamics predict feedback learning and action adjustment. *J Cogn Neurosci*, 23(12), 4106-4121.

von Borries, A. K. L., Verkes, R. J., Bulten, B. H., Cools, R.de Bruijn, E. R. A. (2013). Feedback-related negativity codes outcome valence, but not outcome expectancy, during reversal learning. *Cognitive Affective & Behavioral Neuroscience*, 13(4), 737-746.

Walsh, M. M.Anderson, J. R. (2012). Learning from experience: event-related potential correlates of reward processing, neural adaptation, and behavioral choice. *Neurosci Biobehav Rev*, 36(8), 1870-1884.

Yacubian, J., Glascher, J., Schroeder, K., Sommer, T., Braus, D. F.Buchel, C. (2006). Dissociable systems for gain- and loss-related value predictions and errors of prediction in the human brain (vol 26, pg 9530, 2006). *Journal of Neuroscience*, 26(39).

Yeung, N.Sanfey, A. G. (2004). Independent coding of reward magnitude and valence in the human brain. *Journal of Neuroscience*, 24(28), 6258-6264.

Tables

Electrophysiological measure	**Model**	**Number of parameters**	**Akaike Information Criterion (corrected)**	**Akaike Weights**
Theta	Unsigned quantitative	4	5882.24	$1.44 * 10^{-8}$ %
	Signed quantitative	6	5852.40	.04 %
	Categorical	5	5836.92	99.96 %
Delta	Unsigned quantitative	4	5901.54	$1.84 * 10^{-5}$ %
	Signed quantitative	6	5877.36	3.28 %
	Categorical	5	5870.59	96.72 %
FRN	Unsigned quantitative	4	1493.06	.24 %
	Signed quantitative	6	1492.42	.33 %
	Categorical	5	1480.99	99.43 %

Table. Evaluation of the models according to the Akaike Information Criterion procedure, for three electrophysiological measures. The Akaike Weights denote the relative likelihood of the models.

Figure captions

Figure 1. Left: Predictive cues, outcome stimuli and the probabilities of their combined occurrence. Right: An example of the sequence of events in a trial, in which the predictive cue indicates a loss of 10 points, followed by an actual gain of 30 points.

Figure 2. Time-frequency representations of medial frontal activity as measured at electrode FCz. Left: response to predictive cues, contrasting large values and small values. Right: response to outcomes, contrasting unexpected losses and unexpected gains.

Figure 3. Theta power as a function of predictive cues and actual outcomes. The gray bars highlight unpredicted losses, which elicited the strongest theta response. The lower panel illustrates the signed quantitative prediction errors for each of the prediction-outcome combinations in the panel above.

Figure 4. Upper graph: event-related potentials while processing the three types of categorical prediction errors, as measured at electrode FCz. Lower graph: the corresponding peak-to-peak amplitudes of the FRN.

Figure 5. Subjective experience as a function of predictive cues and actual outcomes. The gray bars highlight outcomes that were worse than predicted, which had the strongest impact on subjective experience. The lower panel illustrates the signed quantitative prediction errors for each of the prediction-outcome combinations in the panel above.

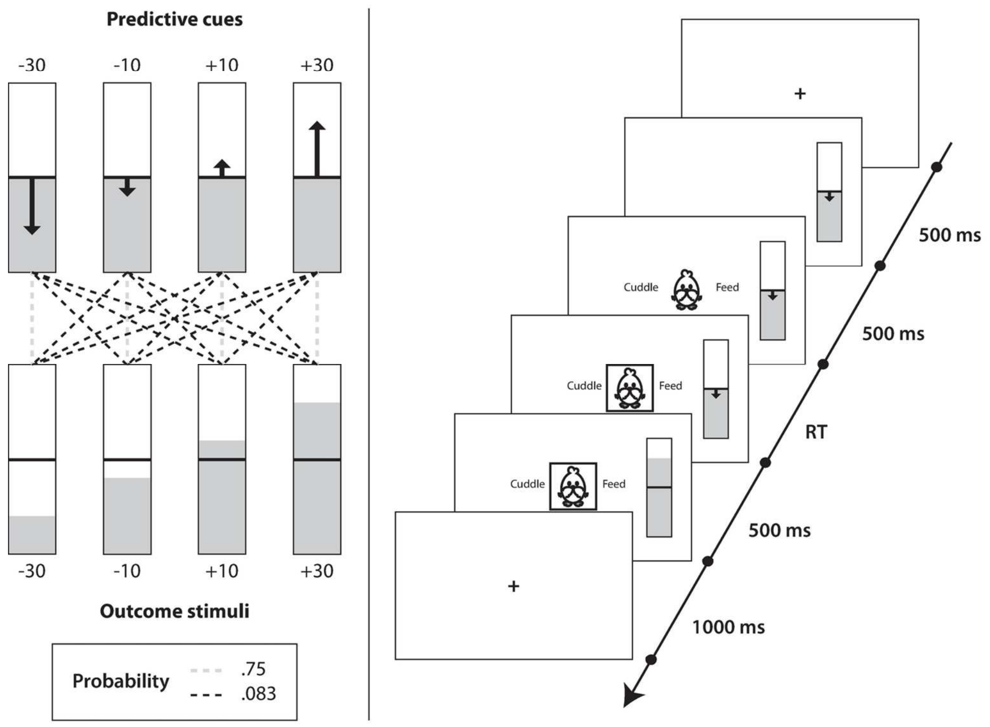

Figure 1. Left: Predictive cues, outcome stimuli and the probabilities of their combined occurrence. Right: An example of the sequence of events in a trial, in which the predictive cue indicates a loss of 10 points, followed by an actual gain of 30 points.
134x105mm (300 x 300 DPI)

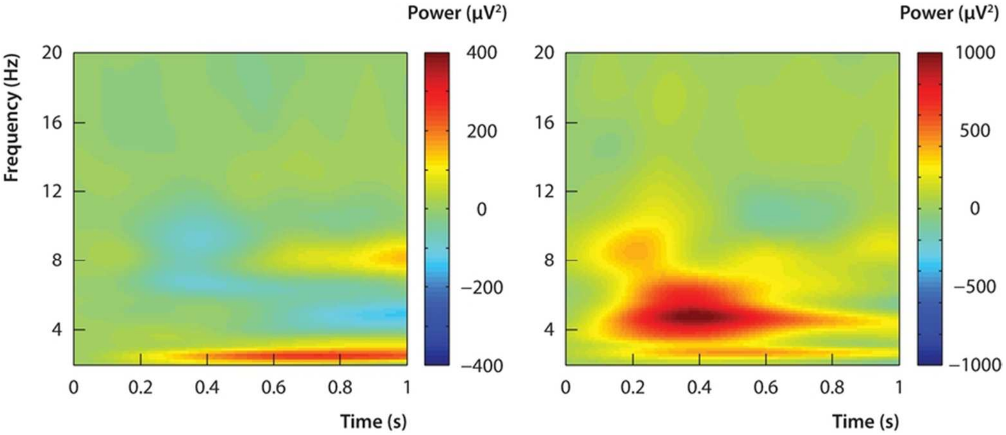

Figure 2. Time-frequency representations of medial frontal activity as measured at electrode FCz. Left: response to predictive cues, contrasting large values and small values. Right: response to outcomes, contrasting unexpected losses and unexpected gains.
75x32mm (300 x 300 DPI)

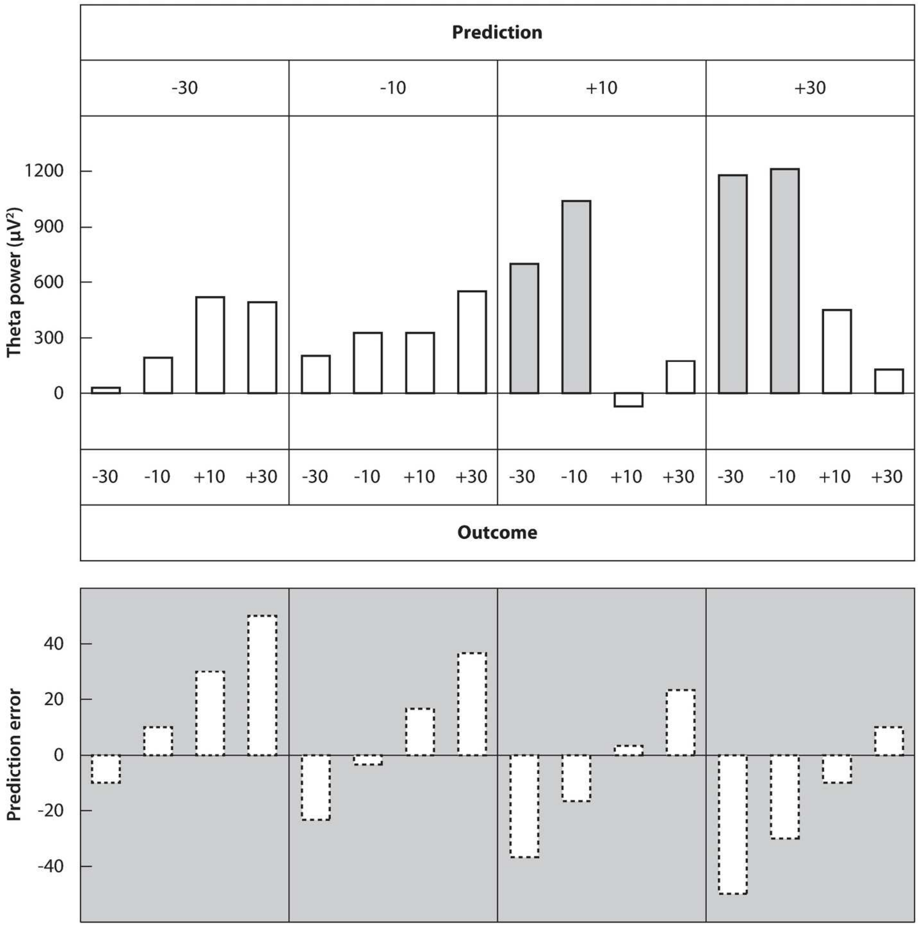

Figure 3. Theta power as a function of predictive cues and actual outcomes. The gray bars highlight unpredicted losses, which elicited the strongest theta response. The lower panel illustrates the signed quantitative prediction errors for each of the prediction-outcome combinations in the panel above.
184x190mm (300 x 300 DPI)

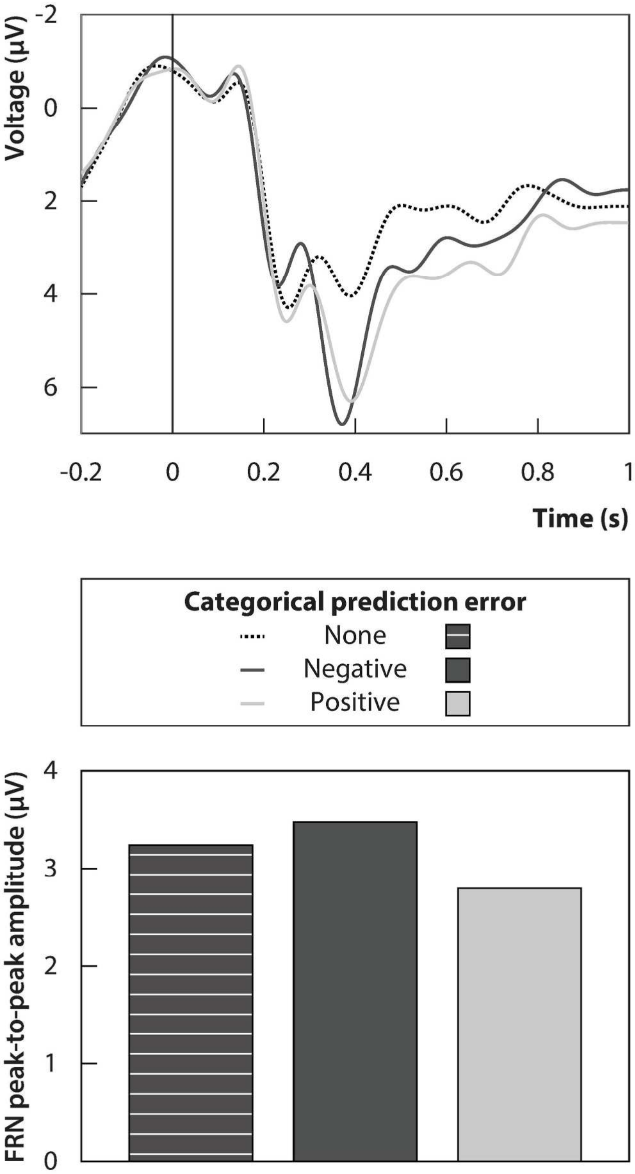

Figure 4. Upper graph: event-related potentials while processing the three types of categorical prediction errors, as measured at electrode FCz. Lower graph: the corresponding peak-to-peak amplitudes of the FRN. 152x258mm (300 x 300 DPI)

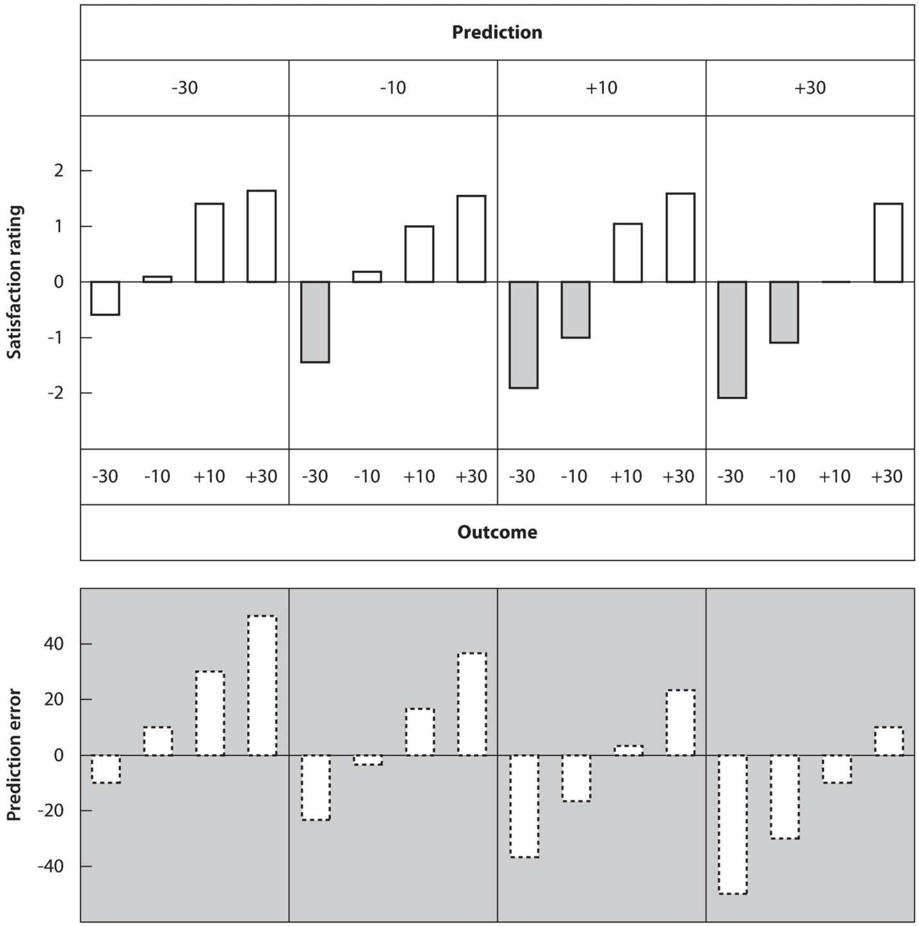

Figure 5. Subjective experience as a function of predictive cues and actual outcomes. The gray bars highlight outcomes that were worse than predicted, which had the strongest impact on subjective experience. The lower panel illustrates the signed quantitative prediction errors for each of the prediction-outcome combinations in the panel above.
184x190mm (300 x 300 DPI)