Psychophysiology, 44 (2007), 905–912. Blackwell Publishing Inc. Printed in the USA.
Copyright © 2007 Society for Psychophysiological Research
DOI: 10.1111/j.1469-8986.2007.00567.x

It's worse than you thought: The feedback negativity and violations of reward prediction in gambling tasks

GREG HAJCAK,[a,]* JASON S. MOSER,[b,]* CLAY B. HOLROYD,[c] AND ROBERT F. SIMONS[b]

[a]Department of Psychology, Stony Brook University, Stony Brook, New York, USA
[b]Department of Psychology, University of Delaware, Newark, Delaware, USA
[c]Department of Psychology, University of Victoria, Victoria, British Columbia, Canada

Abstract

The reinforcement learning theory suggests that the feedback negativity should be larger when feedback is unexpected. Two recent studies found, however, that the feedback negativity was unaffected by outcome probability. To further examine this issue, participants in the present studies made reward predictions on each trial of a gambling task where objective reward probability was indicated by a cue. In Study 1, participants made reward predictions following the cue, but prior to their gambling choice; in Study 2, predictions were made following their gambling choice. Predicted and unpredicted outcomes were associated with equivalent feedback negativities in Study 1. In Study 2, however, the feedback negativity was larger for unpredicted outcomes. These data suggest that the magnitude of the feedback negativity is sensitive to violations of reward prediction, but that this effect may depend on the close coupling of prediction and outcome.

Descriptors: Feedback negativity, Predictions, ERP, ERN, Reinforcement learning

Event-related brain potential (ERP) studies have consistently reported the presence of a medial frontal negative deflection peaking at approximately 250 ms when participants receive negative compared to positive performance feedback (Holroyd & Coles, 2002; Luu, Tucker, Derryberry, Reed, & Poulsen, 2003; Nieuwenhuis, Holroyd, Mol, & Coles, 2004; Nieuwenhuis et al., 2002; Ruchsow, Grothe, Spitzer, & Kiefer, 2002). A similar feedback negativity has been reported following the presentation of stimuli indicating monetary loss or nonreward compared to reward (Gehring & Willoughby, 2002; Hajcak, Holroyd, Moser, & Simons, 2005; Hajcak, Moser, Holroyd, & Simons, 2006; Holroyd, Hajcak, & Larsen, 2006; Yeung, Holroyd, & Cohen, 2005; Yeung & Sanfey, 2004). In fact, Nieuwenhuis, Yeung, Holroyd, Schurger, and Cohen (2004) found that a feedback negativity could be elicited by *either* monetary *or* performance information when feedback conveyed information about both dimensions simultaneously; the aspect of the feedback that elicited the feedback negativity was determined by the aspect of the feedback that was emphasized. Nieuwenhuis et al. argued that monetary losses and negative performance feedback are functionally equivalent insofar as both reflect outcomes along a good–bad dimension. In terms of the neural generator of the feedback negativity, studies that utilize source-localization suggest that the feedback negativity is generated near the anterior cingulate cortex (ACC; Gehring & Willoughby, 2002; Miltner, Braun, & Coles, 1997; Nieuwenhuis, Yeung, et al., 2004); consistent results implicating the ACC in processing negative feedback have been reported using fMRI (Nieuwenhuis, Heslenfeld, et al., 2005; Nieuwenhuis, Schweizer, Mars, Botvinick, & Hajcak, 2007).

Holroyd and Coles (2002) have argued that the feedback negativity reflects the activity of a reinforcement learning system. This reinforcement learning theory is predicated on animal research that implicates the basal ganglia and the midbrain dopamine system in reward prediction and reinforcement learning. In particular, this research indicates that when events are better or worse than anticipated, the basal ganglia induce a phasic increase or decrease, respectively, in the activity of midbrain dopamine neurons (Barto, 1995; Montague, Dayan, & Sejnowski, 1996; Schultz, 2002). The reinforcement learning theory proposes that the amplitude of the feedback negativity is determined by the impact of this phasic dopamine signal on the ACC (Holroyd & Coles, 2002), such that unexpected negative feedback is associated with a large negativity and unexpected positive feedback is associated with a small negativity.

Recent studies have tested the reinforcement learning theory's contention that the feedback negativity reflects a reward prediction error signal. For instance, evidence consistent with the reinforcement learning theory has been reported in studies where participants were required to learn stimulus–response mappings based on feedback; on some trials, participants received feedback

*Denotes shared first-authorship. The order of the first two authors is alphabetical. This research was supported in part by National Institutes of Mental Health (NIMH) predoctoral fellowship MH069047 to G.H. Portions of this article were presented at the 44th annual meeting of the Society for Psychophysiological Research, Santa Fe, New Mexico, October, 2004.

Address reprint requests to: Greg Hajcak, Ph.D., Department of Psychology, Stony Brook University, Stony Brook, NY, 11794-9500, USA. E-mail: greg.hajcak@stonybrook.edu that was inconsistent with learned stimulus–response mappings. In these studies, unexpected negative compared to positive feedback has been found to elicit the largest feedback negativity (Gibson, Krigolson, & Holroyd, 2006; Holroyd & Coles, 2002; Nieuwenhuis, Nielen, Mol, Hajcak, & Veltman, 2005; Nieuwenhuis et al., 2002). Additionally, Holroyd, Nieuwenhuis, Yeung, and Cohen (2003) manipulated feedback frequency in a gambling task to induce expectations regarding feedback valence and found that the feedback negativity was larger when monetary losses were infrequent.

A recent study by Hajcak et al. (2005), however, failed to find an influence of reward probability on the magnitude of the feedback negativity in a gambling task. Participants in Experiment 1 of the Hajcak et al. study performed a gambling task in which they chose one from among four doors following a cue that indicated the probability of reward on each trial (25%, 50%, or 75%). Based on the reinforcement learning theory, we predicted that the feedback negativity would be largest for nonrewarding feedback delivered on the 75% trials because such occurrences should violate subjects' expectations set by the predictive cue. Results revealed that the magnitude of the feedback negativity elicited by nonrewarding feedback was, in fact, the same across all levels of probability, suggesting that the feedback negativity is insensitive to objective reward probability. There are, however, potential methodological reasons why this study yielded results that were not supportive of the reinforcement learning theory and contrasted with findings from previous reinforcement learning studies.

Most previous feedback negativity studies that did find effects of expectancy (Gibson et al., 2006; Holroyd & Coles, 2002; Nieuwenhuis, Nielen, et al., 2005; Nieuwenhuis et al., 2002) employed reinforcement learning tasks in which subjects were required to learn the probability of reward throughout the experiment, and the learning process may have induced greater attentional engagement and kept reward expectations closely linked to the actual reward probability on each trial. Possibly, the Hajcak et al. (2005) participants developed expectations that were only loosely tied to the probability of reward indicated by the predictive cue. Although the success of the experimental manipulation was supported by self-report following the experiment—participants indicated that they received the most rewards when the cue indicated that a reward was probable and the fewest rewards when the cue indicated that a reward was improbable—these self-reports were retrospective and global (see also Baker, Krigolson, & Holroyd, 2006). Thus, on a trial-by-trial basis, participants' expectations might have been inconsistent with the predictive cue or not particularly strong. The distinction being drawn here between reinforcement learning tasks that may involve a strong coupling between subjective expectations and reward probability on the one hand, and gambling tasks that may involve a relatively more loose coupling of subjective expectations and reward probability on the other, is similar to Kahneman and Tversky's (1982) suggestion that there exist "passive" or automatic expectations built up through experience (e.g., during a reinforcement learning task) that are stronger and differ from more "conscious anticipations" that are subject to doubt and reconsideration (e.g., during a gambling task). After receiving nonrewards on two or three trials in a row in the gambling task, for example, participants may have succumbed to the "gambler's fallacy" and expected a reward even if the chance of reward was objectively low (25%). Or, it is possible that despite the retrospective reports of differential expectations, these expectations were relatively uniform or weak across the three levels of predictive cue: Participants may have expected rewards on 50%, 55%, and 60% of the trials, even though the probability of reward was 25%, 50%, and 75%, respectively.

To further examine the sensitivity of the feedback negativity to expectancy violations, participants in the current experiments performed a gambling task similar to that used in the Hajcak et al. (2005) study in which the odds of reward and nonreward were manipulated on a trial-by-trial basis via a predictive cue. Going beyond previous gambling studies, however, participants were asked to indicate prospectively whether they believed they would receive a reward on each trial. Thus, we were able to explicitly examine feedback negativity magnitude elicited by predicted and unpredicted outcomes as determined by the participants on a trial-by-trial basis and could therefore provide a more direct measure of the influence of expectancy on the magnitude of the feedback negativity. We expected this design to control for any inconsistencies between objective probability indicated by the predictive cue and subjective expectations developed through the course of the task, as we obtained subjects' own predictions on a trial-by-trial basis. If the feedback negativity reflects a reward prediction error, as Holroyd and Coles (2002) suggest, then unpredicted nonrewards should elicit a larger feedback negativity compared to predicted nonrewards.

We also evaluated the effect of prediction on the P3, as it has been shown to be sensitive to expectancies in numerous studies (Courchesne, Hillyard, & Courchesne, 1977; Duncan-Johnson & Donchin, 1977; Johnson & Donchin, 1980). It was, therefore, hypothesized that unpredicted outcomes would elicit larger P3s than predicted outcomes and would allow us to examine the effects of prediction on multiple stages of feedback processing.

EXPERIMENT 1

Method

Participants

Seventeen undergraduate students (8 women) in an upper-level psychology class at the University of Delaware participated in the current experiment for extra credit. Additionally, participants were told that they could earn between $0.00 and $24.00 in bonus money based on their performance. All participants were paid $12.00.

Task

The task was administered on a Pentium III class computer, using Presentation software (Neurobehavioral Systems, Inc.) to control the presentation and timing of all stimuli. Subjects' primary objective on each trial was to guess which of four doors presented horizontally in a color graphic hid a prize by pressing the left or right "ctrl" or "alt" key. At the beginning of each trial, a white "1," "2," or "3" appeared on the screen for 1000 ms to inform the participants how many doors contained prizes; therefore, "1," "2," and "3" cues indicated that the probability of reward on the upcoming trial was .25, .50, and .75, respectively. One and a half seconds after the offset of the cue, the question: "Do you think you will win on this trial?" appeared on the screen and remained there until participants indicated yes or no using the left and right "ctrl" and "alt" keys. Thus, participants were first presented with a cue that conveyed the objective probability of reward on the upcoming trial; then, participants predicted whether or not *they* thought they would choose correctly. Immediately following their subjective prediction, the graphic of the doors appeared until the participant chose a door. Five hundred milliseconds following their choice, a feedback stimulus appeared on the screen for 1000 ms: a green "+" feedback indicated a correct guess, and a green "o" feedback indicated an incorrect guess. All other stimuli were presented in white font against a black background; all stimuli were positioned in the center of the screen. The cue and feedback stimuli occupied approximately 2° of visual angle horizontally, and 2° vertically. A fixation mark (+) was presented in the intertrial interval. The interval between offset of the feedback stimulus and the following cue was 1000 ms.

Participants were informed that they would earn $.05 for each correct guess. Unbeknownst to the participants, the outcome of each trial was predetermined and pseudorandom such that overall the participants received exactly 50% correct feedback; negative feedback was delivered on 75% of 1-cue trials, 50% of 2-cue trials, and 25% of 3-cue trials.

Procedure

After a brief description of the experiment, EEG sensors were attached and the participant was given detailed task instructions. To familiarize participants with the task, each was given a practice block consisting of 40 trials and was instructed to guess which door hid a prize. The experiment consisted of 12 blocks of 40 trials (480 total trials) with each block initiated by the participant. The experimenter entered the room every 160 trials to inform the participant how much money he or she had earned. Because all participants received the same number of rewarding and nonrewarding feedback, all participants were paid a uniform amount for their participation.

Psychophysiological Recording, Data Reduction, and Analysis

The electroencephalogram (EEG) was recorded using a Neurosoft Quik-Cap. Recordings were taken from three locations along the midline: Frontal (Fz), Central (Cz), and Parietal (Pz). In addition, Med-Associates tin electrodes were placed on the left and right mastoids (A1 and A2, respectively). During the recording, all activity was referenced to Cz. The electrooculogram (EOG) generated from blinks and vertical eye movements was also recorded using Med-Associates miniature electrodes placed approximately 1 cm above and below the participant's right eye. The right earlobe served as a ground site. All EEG/EOG electrode impedances were below 10 kΩ, and the data from all channels were recorded by a Grass Model 7D polygraph with Grass Model 7P1F preamplifiers (bandpass = 0.05–35 Hz).

All bioelectric signals were digitized on a laboratory microcomputer using VPM software (Cook, 1999). The EEG was sampled at 200 Hz. Data collection began with the participants' response (500 ms prior to feedback), and continued for 1500 ms. Off-line, the EEG for each trial was corrected for vertical EOG artifacts using the method developed by Gratton, Coles, and Donchin (1983; Miller, Gratton, & Yee, 1988) and then re-referenced to the average activity of the mastoid electrodes. Trials were rejected and not counted in subsequent analysis if there was excessive physiological artifact (i.e., 25 ms of invariant analog data on any channel or A/D values on any channel that equaled that converters minimum or maximum values). Single-trial EEG data were lowpass filtered at 20 Hz with a 19 weight FIR digital filter as per Cook and Miller (1992).

As the primary aim of the current study was to evaluate the effect of the participant's predictions on the feedback-locked ERPs, grand average waveforms were created for predicted and unpredicted nonrewards and predicted and unpredicted rewards, collapsed across the objective probability factor (i.e., 1-cue, 2-cue, and 3-cue trials). The average activity in the 200-ms prestimulus window served as a baseline. One outstanding issue in feedback negativity (FN) research is whether variance in FN amplitude between rewarding and nonrewarding feedback results mainly from activity elicited by nonrewards, by rewards, or by both. As argued at length by Luck (2005), the absolute amplitudes of ERP components are meaningless in and of themselves because, for example, an apparent decrease in the amplitude of a component can result from the superposition of a component with opposite polarity (see his rule #4, p. 56). For these reasons, we have used a difference-wave approach to isolate the valence-related variance in the ERP in a manner that is independent of the source of the variance (nonrewards, rewards, or both; Holroyd, 2004; Luck, 2005). Specifically, difference waves were created by subtracting the ERPs observed following rewards from the ERPs observed following nonrewards. These difference waves were created separately based on the predictions made by participants: predicted (predicted nonrewards minus predicted rewards feedback) and unpredicted (unpredicted nonrewards minus unpredicted rewards feedback). FNs were then defined as the maximum negative amplitude of these difference waves within a window between 200 and 500 ms following feedback at each electrode site. This procedure controlled for the main effect of stimulus probability and prediction on the ERP, ensuring that the ERP measure was sensitive to the interaction of feedback expectations and valence (Holroyd, 2004).

The P3 was evaluated for each outcome and prediction at all sites. The P3 was defined as the most positive peak in the 200–600-ms window following feedback onset. The FN and P3 were statistically evaluated using SPSS (version 13.1) General Linear Model software with the Greenhouse–Geisser correction applied to p values associated with multiple df repeated measures comparisons.

Results

Behavioral Results

On average, participants predicted that they would receive a reward on 15.81% ($SD = 18.63$), 73.93% ($SD = 22.31$), and 94.67% ($SD = 6.63$) of 1-, 2-, and 3-cue trials, respectively. An ANOVA on number of predicted rewards at each level of cue confirmed the impression that the cue influenced participants' predictions, $F(2,32) = 91.52, p < .001$. Post hoc analyses indicated that participants predicted rewards on more 3-cue trials than both 2- and 1-cue trials, $t(16) = 3.84$, $p < .01$ and $t(16) = 13.85$, $p < .001$, respectively; additionally, rewards were predicted more on 2- than 1-cue trials, $t(16) = 8.39$, $p < .001$. Thus, the behavioral data established that the predictive cue influenced participants' predictions regarding the subsequent feedback. It is interesting to note the presence of a general positive bias: Participants predicted rewards on more than 50% of 2-cue trials, $t(16) = 4.42$, $p < .001$, and more than 75% of 3-cue trials, $t(16) = 12.23$, $p < .001$; on the other hand, reward prediction on 1-cue trials did not differ reliably from 25%, $t(16) = 2.03$, $p > .05$.

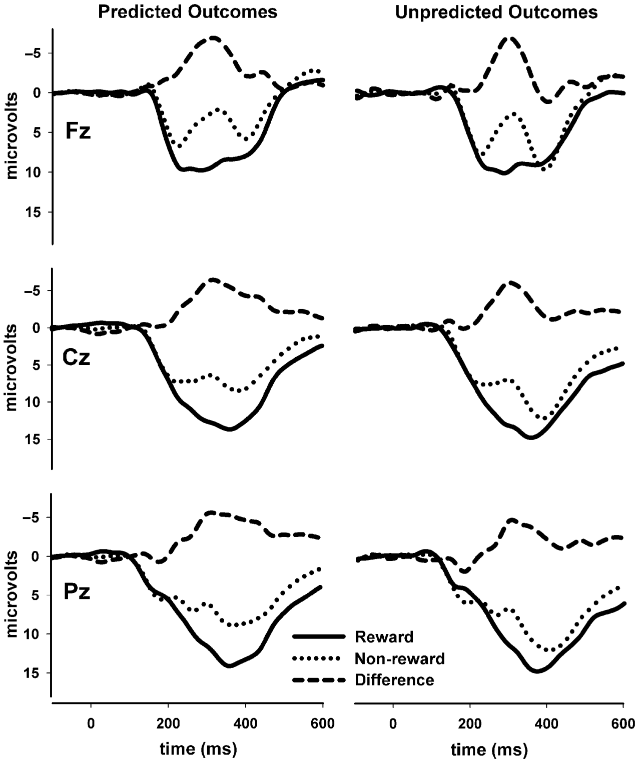

Figure 1. ERPs elicited by predicted (left) and unpredicted (right) rewarding and nonrewarding feedback, as well as the nonreward minus reward difference waveform, at Fz (top), Cz (middle), and Pz (bottom) from Experiment 1. Feedback onset occurred at 0 ms.

Table 1. *Mean (M) and Standard Deviation (SD) for FN and P3 Magnitudes in Experiment 1 (Left) and Experiment 2 (Right)*

	Study 1				Study 2			
	Predicted		Unpredicted		Predicted		Unpredicted	
Measure	*M*	*SD*	*M*	*SD*	*M*	*SD*	*M*	*SD*
FN								
Fz	– 8.61	4.71	– 8.36	5.79	– 6.64	3.31	– 9.02	4.87
Cz	– 8.39	5.22	– 8.27	6.25	– 6.54	3.28	– 9.37	4.96
Pz	– 8.07	5.59	– 7.06	5.44	– 5.39	3.37	– 8.33	4.87
P3								
Reward								
Fz	12.51	5.21	13.00	6.27	7.51	4.48	11.14	6.45
Cz	15.66	6.36	16.72	8.31	12.94	5.33	17.72	7.57
Pz	15.47	6.77	16.22	8.84	16.88	5.84	20.78	7.97
Nonreward								
Fz	8.97	4.21	11.26	5.05	7.61	5.91	9.33	4.79
Cz	11.13	4.91	13.31	5.40	11.29	6.75	14.81	6.57
Pz	11.23	5.47	13.17	5.20	14.60	5.96	17.78	6.45

The FN

Figure 1 presents feedback-locked ERP averages for rewarding and nonrewarding feedback for predicted (left) and unpredicted (right) outcomes at Fz (top), Cz (middle), and Pz (bottom). Nonrewards were associated with a frontally maximal negative-going deflection that peaked approximately 300 ms following feedback. Figure 1 also presents the difference wave obtained by subtracting reward from nonreward for predicted and unpredicted outcomes (predicted nonreward minus predicted reward and unpredicted nonreward minus unpredicted reward) at Fz (top), Cz (middle), and Pz (bottom), and Table 1 presents the average FN amplitudes at each recording site. A 3 (Location) × 2 (Prediction) repeated measures ANOVA indicated a trend toward larger FNs at frontal-central recording sites, $F(2,32) = 2.91$, $p < .09$. Additionally, the FN on predicted and unpredicted feedback did not differ, $F(1,16) < 1$, and the interaction between location and prediction did not reach significance, $F(2,32) < 1$.[1]

Based on the behavioral data, it appeared that the strength of subjects' predictions varied considerably with the predictive cue. To further characterize the effect of prediction on FN magnitude, then, we conducted a 2 (Prediction: predicted vs. unpredicted) × 2 (Prediction Strength: strong vs. weak) ANOVA on a subset of the data. This follow-up analysis only included 11 participants because the other 6 did not have the appropriate minimum number of trials for ERP analysis (20) in the weak prediction condition. For weak predictions, the predicted and unpredicted difference waves were calculated by subtracting rewards from nonrewards on 2-cue trials. For strong predictions, the predicted difference wave was calculated by subtracting predicted rewards on 3-cue trials from predicted nonrewards on 1-cue trials; the unpredicted difference wave was calculated by subtracting unpredicted rewards on 1-cue trials from unpredicted nonrewards on 3-cue trials. We conducted the analysis in this manner because the number of predicted rewards on 1-cue trials and predicted nonrewards on 3-cue trials were insufficient to yield stable waveforms. The analysis, conducted at Cz, where the FN was numerically largest, indicated that the FN did not vary with respect to prediction, $F(1,10) < 1$, or prediction strength, $F(1,10) < 1$, and prediction strength did not interact with prediction, $F(1,10) < 1$. Thus, the magnitude of the FN was insensitive to predictions when they were both strong ($M = -8.05$, $SD = 4.06$ and $M = -8.52$, $SD = 5.52$ for predicted vs. unpredicted outcomes, respectively) and weak ($M = -6.37$, $SD = 5.40$ and $M = -6.09$, $SD = 3.83$ for predicted vs. unpredicted outcomes, respectively).

The P3

The P3 amplitudes following predicted and unpredicted feedback are presented in Table 1. Consistent with the impression suggested by Figure 1, a 2 (Prediction) × 2 (Outcome) × 3 (Location) repeated measures ANOVA confirmed that the P3 became larger at more parietal recording sites, $F(2,32) = 4.89$, $p < .05$. Additionally, the P3 was larger following unpredicted than predicted feedback, $F(1,16) = 7.92$, $p < .05$, and larger for rewards than nonrewards, $F(1,16) = 15.87$, $p < .001$. The three-way interaction and all two-way interactions failed to reach statistical significance (all *p*s > .20).

Discussion

The current experiment evaluated the effect of perceived reward probability on feedback processing in a manner similar to that reported by Hajcak et al. (2005). The current experiment extended the methodology of the Hajcak et al. study, however, by asking participants to indicate whether or not *they* thought they would receive a reward on a trial-by-trial basis. This modification controlled for possible inconsistencies between the objective probabilities indicated by the predictive cues and the partici-

[1] Even when the FN was evaluated at Cz, as in Holroyd et al. (2003), prediction did not influence the magnitude of the FN, $F(2,32) < 1$.

pants' subjective expectations. In line with the retrospective reports of the participants in the Hajcak et al. study, participants' predictions in the present experiment were reliably modulated by the predictive cue such that they predicted the most rewards on 3-cue trials and the fewest rewards on 1-cue trials. Thus, we were able to build on our previous study in two important ways: (a) We were able to provide a behavioral measure to verify that the predictive cue induced appropriate expectations on a trial-by-trial-basis and (b) we were able to compare directly predicted versus unpredicted rewards and nonrewards as perceived by the participants on each trial.

The results of the current study indicated that the feedback negativity was not larger for unpredicted outcomes as perceived by the participants on each trial. Further analyses of predicted versus unpredicted feedback as a function of cue (i.e., predictions on 2-cue trials versus those made on 1- and 3-cue trials) similarly indicated that there was no effect of prediction on the feedback negativity.

Interestingly, the results from the present study did confirm that subjective predictions differed substantially from objective probability. That is, participants underestimated rewards somewhat on 1-cue trials, but demonstrated a positive bias on 2- and 3-cue trials, predicting rewards on approximately 74% and 95% of trials, respectively. Overall, these data indicate that predictive cues induced appropriate expectations overall, but that participants' predictions differed from objective probability in a number of cases.

In Experiment 1, participants were asked to make their predictions immediately after the predictive cue was presented but prior to choosing a door. Therefore, their predictions were made prior to their choice in the gambling task, but after having knowledge about the likelihood of reward. Research on "magical thinking"—whereby individuals believe that their predictions have some influence over a probabilistic outcome—suggests that making predictions before actions might actually strengthen expectations (Rothbart & Snyder, 1970). Specifically participants in the Rothbart and Snyder study were more confident in being correct and wagered more money when predicting an outcome before engaging in a probabilistic task (i.e., rolling dice). Thus, we believed that this design would lead to high confidence and strong expectations about predicted rewards.

This design choice, however, was made on the assumption that expectations induced by the cue would remain constant throughout the duration of each trial. This might not have been the case, however. Because participants had time to reconsider their predictions at multiple time points during each trial, it is possible that the strength of the prediction made after the cue faded or even changed over the course of the trial. For instance, participants could have felt confident that they would receive a reward based on a cue (e.g., on a 3-cue trial), but changed their expectations after their gambling choice. These possibilities are consistent with Kahneman and Tversky's (1982) analysis of how most individuals commit themselves to expect a range of possible outcomes in a task rather than an exact probability estimate, and given enough time to reconsider their predictions in the current study, subjects might have had fairly weak (i.e., wide ranged) expectations at the time of feedback presentation.

To explore these possibilities, participants in Experiment 2 performed a similar gambling task. In this case, however, subjective predictions were made *following* responses and just prior to the presentation of feedback. We hypothesized that this design might strengthen expectations by having participants make reward predictions just after their response and immediately before receiving feedback. In this way, the time between prediction and outcome was reduced, and this, in turn, would reduce the possibility that participants might second guess or change their predictions and thereby increase attention to the action–outcome pairs. We again predicted that the feedback negativity should be enhanced for unpredicted outcomes.

EXPERIMENT 2

Method

Participants

Seventeen different undergraduate students (15 women) in a separate upper-level psychology class at the University of Delaware participated in the current experiment for extra credit. As in Experiment 1, participants were told that they could earn between $0.00 and $24.00 in bonus money based on their performance; all participants were paid $12.00.

Task and Procedure

The task and procedures and data analysis strategies for Experiment 2 were identical to Experiment 1 except that the subjective prediction question—"Do you think you will win on this trial?"—was presented *after* participants chose a door.

Results

Behavioral Results

Participants predicted a reward on 21.5% ($SD = 18.4$), 69.0% ($SD = 19.5$), and 94.9% ($SD = 6.9$) of 1-, 2-, and 3-cue trials, respectively. An ANOVA on number of predicted rewards at each level of cue confirmed that the cue influenced participants' predictions, $F(2,32) = 114.14$, $p < .001$. Post hoc analyses indicated that participants predicted rewards on more 3-cue trials than both 2-cue trials, $t(16) = 6.98$, $p < .001$, and 1-cue trials, $t(16) = 15.27$, $p < .001$; additionally, rewards were predicted more on 2- than 1-cue trials, $t(16) = 7.91$, $p < .001$. Thus, the behavioral data established that the predictive cue influenced participants' predictions regarding the subsequent feedback. As in Experiment 1, participants exhibited a positive bias in their predictions and predicted rewards on more than 50% and 75% of 2- and 3-cue trials, respectively, $t(16) = 4.02$, $p < .001$, and $t(16) = 11.89$, $p < .001$, respectively; however, participants did not predict fewer than 25% of rewards on 1-cue trials, $t(16) = .79$, $p > .40$.[2]

The FN

Figure 2 presents feedback-locked ERP averages for rewarding and nonrewarding feedback for predicted (left) and unpredicted (right) outcomes at Fz, Cz, and Pz. As in Experiment 1, nonrewarding feedback elicited a negative deflection approximately 300 ms following feedback onset. Figure 2 also presents the difference wave obtained by subtracting rewards from nonrewards for predicted and unpredicted outcomes at each recording

[2] To examine whether the percentage of predicted rewards varied between experiments, we conducted a 3 (Cue) × 2 (Experiment) repeated measures ANOVA on percent of predicted rewards for each type of cue. Although the percentage of predicted rewards varied as a function of cue, $F(2,64) = 202.15$, $p < .001$, there was no effect of experiment, $F(1,32) < 1$, and Experiment did not interact with Cue, $F(2,64) < 1$. Thus, subjective probabilities were not affected by when participants made their prediction.

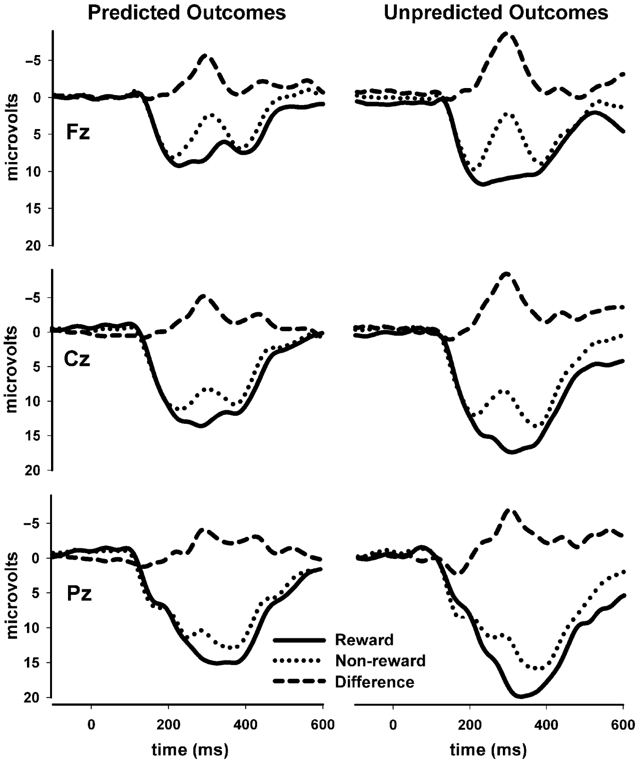

Figure 2. ERPs elicited by predicted (left) and unpredicted (right) rewarding and nonrewarding feedback, as well as the nonreward minus reward difference waveform, at Fz (top), Cz (middle), and Pz (bottom) from Experiment 2. Feedback onset occurred at 0 ms.

site, and Table 1 presents the average FN amplitudes at each recording site. Unlike Experiment 1, however, the critical 3 (Electrode Site) × 2 (Prediction) repeated measures ANOVA on FN amplitude revealed that it *was* larger following unpredicted outcomes, $F(1,16) = 6.31, p < .05$. Although the FN was larger at frontal electrode sites, $F(2,32) = 3.78, p < .05$, the interaction of electrode site and prediction did not reach significance, $F(2,32) < 1$.[3]

As in Experiment 1, we sought to determine whether prediction strength modulated the effect of prediction on the FN and compared the FN elicited by predicted and unpredicted feedback on 2-cue trials (i.e., the weak prediction condition) with that on 1- and 3-cue trials (i.e., the strong prediction condition). Like Experiment 1, this follow-up analysis was conducted on a subsample of 11 participants who had at least 20 trials of each type in the weak prediction condition. At Cz, where the FN was largest, the FN did not differ as a function of prediction in this subset of the data, $F(1,10) < 1$. Critically, however, the main effect of prediction strength, $F(1,10) = 4.96, p < .05$, was qualified by an interaction with prediction, $F(1,10) = 6.07, p < .05$, indicating that the FN did, indeed, differ as a function of prediction on strong trials ($M = -5.43$, $SD = 2.99$ for predicted and $M = -10.43$, $SD = 4.80$ for unpredicted outcomes), $F(1,10) = 7.23$, $p < .05$, but not weak trials ($M = -10.88$, $SD = 6.61$ for predicted and $M = -10.04$, $SD = 6.77$ for unpredicted outcomes), $F(1,10) < 1$. Thus, the interaction between prediction and prediction strength seems to be driven by the relatively small FN on predicted outcomes when predictions are strong.

The P3

The average P3 amplitudes elicited by predicted and unpredicted feedback are presented in Table 1. A 2 (Prediction) × 3 (Outcome) × 3 (Electrode Site) repeated measures ANOVA indicated that the P3 was larger at more parietal recording sites, $F(2,32) = 35.31, p < .001$. The P3 was again larger following unpredicted compared to predicted feedback, $F(1,16) = 42.86$, $p < .001$, and, like Experiment 1, the P3 was also larger following rewards than following nonrewards, $F(1,16) = 12.17, p < .01$. Consistent with the depiction in Figure 2, the influence of outcome (i.e., the FN) on the P3 was largest at parietal recording sites, $F(2,32) = 13.82$, $p < .001$. The three-way interaction and other two-way interactions did not reach significance (all $ps > .05$).

Consistent with our FN results, which demonstrated an effect of prediction strength in this experiment, an analysis of the P3 at Pz, where it was largest, showed that the difference between unpredicted and predicted feedback was larger in Experiment 2 ($M = 3.53$ μV, $SD = 2.50$) than Experiment 1 ($M = 1.34$ μV, $SD = 2.09$), $t(32) = 2.77, p < .01$.

General Discussion

The reinforcement learning theory holds that feedback negativity amplitude is determined by an interaction between feedback valence and expectedness, such that unexpected feedback induces greater variance in feedback negativity amplitude relative to expected feedback. This prediction has been confirmed in several experiments (Gibson et al., 2006; Holroyd & Coles, 2002; Nieuwenhuis, Nielen, et al., 2005; Nieuwenhuis et al., 2002). However, in a recent study we found that feedback negativity amplitude appeared insensitive to the objective probability of feedback in two gambling tasks (Hajcak et al., 2005). A possible explanation for this discrepancy is that the overt probabilities in these experiments only loosely corresponded to the participants' actual expectations. Thus, if a person does not believe that trials associated with 50% reward will in fact lead to reward half of the time, then feedback negativity amplitude will not reflect a prediction error associated with 50% reward. In the present study we investigated this possibility by modifying the Hajcak et al. task so that participants were queried about their subjective reward prediction either before (Experiment 1) or after (Experiment 2) their response. We found that feedback negativity amplitude varied with subjective expectation only when predictions were made following the response. Furthermore, this modulation of the feedback negativity was apparent when predictions were presumably strongest (i.e., on 1- and 3-cue trials but not 2-cue trials). These results provide additional support for the reinforcement learning theory but indicate that the methods for inferring subjective expectancies should be carefully evaluated in future experiments, and further, that participants' expectancies may have to be relatively strong and closely coupled with action-outcome pairs to influence feedback negativity amplitude.

[3] Because the sample in Experiment 2 was comprised largely of women, we reanalyzed the data from Experiment 1 using gender (8 women, 9 men) as a between-groups variable to investigate whether gender might explain the expectancy effect found in Experiment 2. Across all sites, and at Fz only, there was no effect of prediction on the magnitude of the FN, $F(1,15) < 1$, the magnitude of the FN did not vary by gender, $F(1,15) < 1$, and the interaction between prediction and gender did not approach significance, $F(1,15) < 1$. Thus, it is unlikely that the difference between Experiments 1 and 2 was a function of the different gender distributions.

Why was feedback negativity amplitude associated with reward expectancy only when the expectation was queried after the response, but not before? Given the data at hand, it seems reasonable to assume that participants' predictions fluctuated over the course of each trial, solidifying only after participants were committed to a response. Presumably, the process of action selection involved an evaluation of the likelihood of each action–outcome pair, followed by selection of the action with the highest subjective expected value. The number of potential outcomes would also have been smaller after the response (when there were only two action–outcome possibilities), compared to before each response (when then there were a total of eight action–outcome possibilities). For these reasons, participants would have confronted fewer possibilities, and would have evaluated these more thoroughly, following each response relative to before it. Thus it seems that participants would have been most confident in their predictions at the end of each trial. Importantly, the P3 data are consistent with this position: The effect of expectancy violation on the P3 was more than twice as large in Experiment 2 as in Experiment 1 (partial eta squared = .72 vs. .33, respectively; see also the Results section for Experiment 2). In contrast, participants' behavior did not differ between the two experiments (i.e., percentage of predicted rewards for each cue), suggesting that when participants made their predictions did not alter their gambling habits altogether, but rather changed how strongly they felt about their chances. Given that we did not collect self-reported confidence on each trial, future studies will have to further evaluate this possibility.

Putting the current findings in the context of the reinforcement learning theory and results of previous reinforcement learning tasks, it appears that the feedback negativity is highly sensitive to task demands that make the connections between predictions and action–outcome pairs more or less salient. We are further examining this issue in a series of ongoing experiments (Baker et al., 2006; Gibson et al., 2006; Lee, Krigolson, & Holroyd, 2006), and our preliminary results suggest that the role of predictions depends on depth of processing: Participants appear to make more concrete predictions in trial-and-error learning tasks in which they are required to attend more deeply to the relationship between feedback and behavior compared to gambling or guessing tasks where the contingencies are less apparent. Thus, reinforcement learning tasks seem to involve more "passive expectations" that are quite strong and built up through experience whereas gambling tasks seem to involve more "conscious anticipations" that are more susceptible to doubt and reconsideration (cf. Kahneman & Tversky, 1982). In other words, the differences between reinforcement learning tasks and gambling tasks might be understood in terms of the former relying on more basic "bottom up" processes that might become strong and automatic because of heightened attention and the latter relying on more complex "top down" processes involving multiple evaluations and considerations of actions and outcome that are more flexible because of limited attentional capture by the task (cf. Holroyd et al., 2006).

In this respect, it is interesting to note that Yeung et al. (2005) found that the magnitude of the feedback negativity was smaller in two gambling tasks that did not require overt responses. Participants in this study rated the gambling tasks that involved making overt responses as more interesting; additionally, participants had larger self-reported affective responses to nonrewarding versus rewarding outcomes during tasks that required a response, despite the fact that the objective value of feedback was identical. Yeung et al. interpreted differences in the feedback negativity between gambling tasks that require a response and those that do not as reflecting differences in participants' level of interest. Alternatively, the present studies suggest that (even implicit) reward predictions following responses may explain the larger feedback negativity in tasks that require a response.

We recognize that, by asking the participants to make a dichotomous prediction on each trial, the meaning of the feedback is actually changed. For instance, participants would often have correctly predicted negative feedback; in this way, the negative feedback could simultaneously indicate that a participant's prediction was correct, and that he or she did not receive a reward. Interestingly, the feedback negativity appeared to track the monetary value of the feedback and was not related to its performance-related aspect. This observation is consistent with a recent study by Nieuwenhuis and colleagues, who found that the feedback negativity could reflect either utilitarian- or performance–related information conveyed by feedback, depending on which aspect of the feedback is emphasized (Nieuwenhuis et al., 2004). These data suggest that the feedback negativity is *not*, however, elicited simply by unpredicted outcomes (Oliveira, 2005)—as even predicted outcomes elicited a feedback negativity in both Experiments 1 and 2.

We have noted that the P3 was larger for unpredicted compared to predicted feedback in Experiments 1 and 2. These data are consistent with previous studies that have demonstrated sensitivity of the P3 to expectancy violations (Courchesne et al., 1977; Duncan-Johnson & Donchin, 1977; Johnson & Donchin, 1980). However, the P3 also was larger following rewards than nonrewards in both Experiments 1 and 2, which might suggest that the variance in amplitude of the feedback negativity in the present study resulted from component overlap with the P3. However, the opposite appears to be the case, as the variance in P3 amplitude appears to have resulted from component overlap with the feedback negativity. This conclusion holds because feedback negativity amplitude (measured as a difference wave) was frontal-centrally distributed for both Experiments 1 and 2: If variance in the size of the difference wave resulted from an increase in P3 amplitude across conditions, then the difference wave would exhibit a posterior scalp distribution (Holroyd, 2004). Further, both the feedback negativity and P3 were larger following unpredicted nonrewards in Experiment 2 relative to Experiment 1; if just one component were driving both effects in Experiment 2, one would expect a larger feedback negativity to relate to a smaller P3, or vice versa. For these reasons we conclude that the changes in feedback negativity amplitude observed in the present study did not result from component overlap with the P3.

Overall, the present study provides further insight into electrocortical activity related to processing rewards and nonrewards. In particular, the feedback negativity was clearly sensitive to violations of reward prediction when a prediction was made *following* a response and *immediately before* receiving feedback in the gambling task. These data suggest that the magnitude of error signals are modulated by participants' predictions, but that these predictions solidify only after the participants have committed to a particular action, perhaps because of increased attention to or confidence in the action–outcome associations. Future studies can test this possibility, for example by asking participants about their attention to or confidence in their predictions in addition to asking them about the predictions themselves.

REFERENCES

Baker, R., Krigolson, O. E., & Holroyd, C. B. (2006). Examining the feedback error-related negativity using predictive stimuli. *Psychophysiology*, *43*, S22.

Barto, A. G. (1995). Adaptive critics and the basal ganglia. In J. Houk, J. Davis, & D. Beiser (Eds.), *Models of information processing in the basal ganglia* (pp. 215–232). Cambridge, MA: MIT Press.

Cook, E. W., 3rd. (1999). *VPM reference manual*. Birmingham, AL: Author.

Cook, E. W., 3rd, & Miller, G. A. (1992). Digital filtering: Background and tutorial for psychophysiologists. *Psychophysiology*, *29*, 350–367.

Courchesne, E., Hillyard, S. A., & Courchesne, R. Y. (1977). P3 waves to the discrimination of targets in homogeneous and heterogeneous stimulus sequences. *Psychophysiology*, *14*, 590–597.

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

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

Gibson, J., Krigolson, O. E., & Holroyd, C. B. (2006). Sensitivity of the feedback error-related negativity to reward probability. *Psychophysiology*, *43*, S41.

Gratton, G., Coles, M. G., & Donchin, E. (1983). A new method for off-line removal of ocular artifact. *Electroencephalography and Clinical Neurophysiology*, *55*, 468–484.

Hajcak, G., Holroyd, C. B., Moser, J. S., & Simons, R. F. (2005). Brain potentials associated with expected and unexpected good and bad outcomes. *Psychophysiology*, *42*, 161–170.

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*, 148–154.

Holroyd, C. B. (2004). A note on the oddball N200 and feedback ERN. In M. Ullsberger & M. Falkenstein (Eds.), *Errors, conflicts, and the brain: Current opinions on response monitoring*. Leipzig: MPI of Cognitive Neuroscience.

Holroyd, C. B., & Coles, M. G. (2002). The neural basis of human error processing: Reinforcement learning, dopamine, and the error-related negativity. *Psychological Review*, *109*, 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*, 93–101.

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

Johnson, R., Jr., & Donchin, E. (1980). P300 and stimulus categorization: Two plus one is not so different from one plus one. *Psychophysiology*, *17*, 167–178.

Kahneman, D., & Tversky, A. (1982). Variants of uncertainty. *Cognition*, *11*, 143–157.

Lee, K., Krigolson, O. E., & Holroyd, C. B. (2006). Modulation of the fERN amplitude by reward expectancy. *Psychophysiology*, *43*, S57.

Luck, S. J. (2005). *An introduction to the event-related potential technique*. Cambridge, MA: The MIT Press.

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*, 47–53.

Miller, G. A., Gratton, G., & Yee, C. M. (1988). Generalized implementation of an eye movement correction procedure. *Psychophysiology*, *25*, 241–243.

Miltner, W. H., Braun, C. H., & Coles, M. G. H. (1997). Event-related brain potentials following incorrect feedback in a time-estimation task: Evidence for a ''generic'' neural system for error detection. *Journal of Cognitive Neuroscience*, *9*, 788–798.

Montague, P. R., Dayan, P., & Sejnowski, T. J. (1996). A framework for mesencephalic dopamine systems based on predictive Hebbian learning. *Journal of Neuroscience*, *16*, 1936–1947.

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

Nieuwenhuis, S., Holroyd, C. B., Mol, N., & Coles, M. G. (2004). Reinforcement-related brain potentials from medial frontal cortex: Origins and functional significance. *Neuroscience and Biobehavioral Reviews*, *28*, 441–448.

Nieuwenhuis, S., Nielen, M. M., Mol, N., Hajcak, G., & Veltman, D. J. (2005). Performance monitoring in obsessive-compulsive disorder. *Psychiatry Research*, *134*, 111–122.

Nieuwenhuis, S., Ridderinkhof, K. R., Talsma, D., Coles, M. G., Holroyd, C. B., Kok, A., et al. (2002). A computational account of altered error processing in older age: Dopamine and the error-related negativity. *Cognitive, Affective & Behavioral Neuroscience*, *2*, 19–36.

Nieuwenhuis, S., Schweizer, T. S., Mars, R. B., Botvinick, M. M., & Hajcak, G. (2007). Error-likelihood prediction in the medial frontal cortex: A critical evaluation. *Cerebral Cortex*, *17*, 1570–1581.

Nieuwenhuis, S., Yeung, N., Holroyd, C. B., Schurger, A., & Cohen, J. D. (2004). Sensitivity of electrophysiological activity from medial frontal cortex to utilitarian and performance feedback. *Cerebral Cortex*, *14*, 741–747.

Oliveira, F. T. P. (2005). *Electrophysiological correlates of performance monitoring and error detection in response to augmented feedback*. Vancouver, BC: Simon Fraser University.

Rothbart, M., & Snyder, M. (1970). Confidence in the prediction and postdiction of an uncertain outcome. *Canadian Journal of Behavioural Science*, *2*, 38–43.

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*, 203–206.

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

Yeung, N., Holroyd, C. B., & Cohen, J. D. (2005). ERP correlates of feedback and reward processing in the presence and absence of response choice. *Cerebral Cortex*, *15*, 535–544.

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

(Received October 18, 2006; Accepted June 6, 2007)