Modulation of the feedback-related negativity by instruction and experience

Matthew M. Walsh[1] and John R. Anderson[1]

Department of Psychology, Carnegie Mellon University, Pittsburgh, PA 15213

Contributed by John R. Anderson, October 18, 2011 (sent for review June 3, 2011)

A great deal of research focuses on how humans and animals learn from trial-and-error interactions with the environment. This research has established the viability of reinforcement learning as a model of behavioral adaptation and neural reward valuation. Error-driven learning is inefficient and dangerous, however. Fortunately, humans learn from nonexperiential sources of information as well. In the present study, we focused on one such form of information, instruction. We recorded event-related potentials as participants performed a probabilistic learning task. In one experiment condition, participants received feedback only about whether their responses were rewarded. In the other condition, they also received instruction about reward probabilities before performing the task. We found that instruction eliminated participants' reliance on feedback as evidenced by their immediate asymptotic performance in the instruction condition. In striking contrast, the feedback-related negativity, an event-related potential component thought to reflect neural reward prediction error, continued to adapt with experience in both conditions. These results show that, whereas instruction may immediately control behavior, certain neural responses must be learned from experience.

Reinforcement learning (RL) formalizes the notion that humans and animals learn from trial-and-error interactions with the environment (1). According to many RL models, differences between actual and expected outcomes, or reward prediction errors, provide teaching signals. These signals convey information about the magnitude and valence of the difference between actual and expected rewards. By using reward prediction errors to revise expectations, RL models increasingly select advantageous actions. Behavioral studies furnished early support for RL in the form of the "law of effect" (2). This law states that actions that are followed by rewards will be repeated. Single-cell recordings from animals provided further support by showing that responses of midbrain dopamine neurons to outcomes scale according to the differences between actual and expected rewards (3). Neuroimaging experiments have since extended this result to humans by demonstrating that blood–oxygen level-dependent (BOLD) responses in the striatum and prefrontal cortex also mirror reward prediction errors (4).

On the basis of these findings, RL has emerged as a prominent theory of behavioral adaptation and neural reward valuation. As it stands, however, RL is an incomplete theory. Individuals learn from nonexperiential sources of information as well. For example, by using language to acquire knowledge about outcome likelihoods, humans can avoid costly mistakes. This raises the question, How does information provided by instruction mediate trial-and-error learning?

Several theories seek to explain how the brain uses instruction and experience to select actions (5–8). These theories agree that instruction engages the prefrontal cortex and medial temporal lobes (PFC/MTL), whereas experience engages the basal ganglia (BG) and their dopaminergic afferents. These theories disagree, however, on whether and how instruction and experience are combined. According to some accounts, the relationship between learning systems is antagonistic. For example, activation of the PFC/MTL is sometimes associated with deactivation of the BG (5, 9–11). According to other accounts, learning systems interact. For example, knowledge representations in the PFC/MTL may bias the BG to learn what is described by instruction regardless of what is experienced (6–8). According to still other accounts, learning systems operate independently until a response is required. For example, the PFC/MTL and the BG may learn in parallel, but the system that possesses greater certainty may override the other at the moment of action selection (6, 12, 13).

In the present study, we used scalp-recorded event-related potentials (ERPs) to explore how information provided by instruction mediates trial-and-error learning. We focused on an ERP component called the feedback-related negativity (FRN), a frontocentral negativity that appears 200–350 ms after the display of negative performance feedback relative to positive feedback (14). The FRN appears sometimes as a negative deflection following losses and sometimes as a positive deflection following wins (15). Several features of the FRN indicate that it is a neural manifestation of reward prediction error. First, the FRN is sensitive to violations of reward probability and magnitude (16, 17). Second, the FRN is associated with posterror adjustments (18). Third, the amplitude of the FRN changes with experience and in a manner consistent with prediction errors produced by RL models (16). Namely, FRN amplitude, defined as the difference between waveforms following losses and wins, increases for improbable outcomes and decreases for probable outcomes. Fourth, and finally, converging methodological approaches indicate that the FRN originates from the anterior cingulate cortex (ACC) (14, 19, 20), a region implicated in cognitive control and behavioral selection (21).

These ideas have been synthesized in the reinforcement learning theory of the error-related negativity (RL-ERN) (22), which holds that midbrain dopamine neurons transmit a prediction error signal to the ACC. This signal reinforces or punishes actions that preceded outcomes. By this view, the FRN tracks operations of the BG system and its dopaminergic afferents (23). Specifically, the FRN tracks reward prediction errors generated by an RL technique such as temporal-difference learning.

We asked how instruction and feedback modulate expression of the FRN in a probabilistic learning task (Fig. 1). In each trial, participants selected from two cues that were rewarded with different probabilities, $P = \{0\%, 33\%, \text{and } 66\%\}$. They could increase their earnings by selecting the cue that was more likely to be rewarded within each pair (i.e., the 66% cue when it was paired with the 33% or the 0% cues and the 33% cue when it was paired with the 0% cue). In the *no instruction condition*, participants received feedback only about whether their choices were rewarded. In the *instruction condition*, they also viewed the cues and received a description of their associated reward probabilities before performing the task.

Author contributions: M.M.W. and J.R.A. designed research; M.M.W. performed research; M.M.W. analyzed data; and M.M.W. and J.R.A. wrote the paper.

The authors declare no conflict of interest.

[1]To whom correspondence may be addressed. E-mail: mmw187@andrew.cmu.edu or ja@cmu.edu.

This article contains supporting information online at www.pnas.org/lookup/suppl/doi:10.1073/pnas.1117189108/-/DCSupplemental.

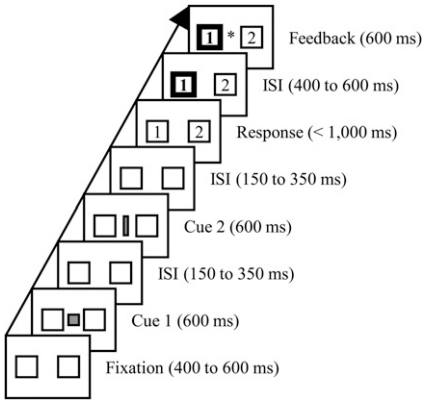

Fig. 1. Trial procedure. Participants sequentially viewed two cues followed by a response screen. After participants selected a cue, they received feedback about whether their choice was rewarded.

On the basis of the idea that the FRN tracks reward prediction errors generated by temporal-difference learning, we evaluated three hypotheses. First, if instruction engages prefrontal regions and disengages reward valuation regions that incrementally learn from experience, instruction will diminish the FRN. [This hypothesis rests on the assumption that the FRN tracks a neural learning signal (for reviews, see refs. 22 and 24), an assumption that is strengthened by the fact that the FRN is maximal when participants perceive a relationship between their actions and trial outcomes (25, 26).] Second, if instruction does not disengage reward valuation regions, and if information provided by instruction penetrates reward valuation regions, an FRN will appear in both conditions. The FRN will immediately reflect stated reward probabilities in the instruction condition, whereas the FRN will gradually change to reflect experienced probabilities in the no instruction condition. Third, if information provided by instruction neither disengages nor penetrates reward valuation regions, an FRN will appear in both conditions. The FRN will gradually change to reflect experienced reward probabilities in both conditions. We also hypothesized that temporal-difference learning would aptly characterize choice behavior when participants received only trial-and-error feedback. We expected that such a model would fail to account for behavior when participants also received information about reward probabilities, in which case they would rely on instruction rather than experience.

Results

Behavioral Results. Response accuracy, defined as the percentage of trials where participants selected the cue that was more likely to be rewarded, increased with experience in the no instruction condition. Conversely, accuracy began and remained at asymptote in the instruction condition (Fig. 2). A 2 (condition) × 2 (block half) × 3 (cue pair) ANOVA revealed significant effects of condition, $F_{1,19} = 45.02$, $P < 0.0001$, block half, $F_{1,19} = 41.08$, $P < 0.0001$, and cue pair, $F_{2,38} = 16.31$, $P < 0.0001$. The interaction between condition and cue pair was significant, $F_{2,38} = 9.45$, $P < 0.0001$, because the effect of cue pair was far greater in the no instruction condition. The interaction between condition and block half was also significant, $F_{1,19} = 32.62$, $P < 0.0001$, because behavioral learning occurred *only* in the no instruction condition. When we divided blocks into quarters, the interaction between condition and block quarter remained significant, $F_{3,57} = 22.29$, $P < 0.0001$. Response accuracy increased with block quarter in the no instruction condition, $F_{3,57} = 31.63$, $P < 0.0001$, but did not change in the instruction condition, $F_{3,57} = 0.59$, $P > 0.1$. This conclusion is strengthened by the finding that participants were more likely to select previously rewarded cues in the no instruction condition, whereas they were insensitive to the prior sequence of outcomes in the instruction condition (*SI Materials and Methods* and Figs. S1 and S2).

ERP Results. Participants displayed an FRN for improbable outcomes (losses after 66 cues minus wins after 33 cues) and probable outcomes (losses after 33 cues minus wins after 66 cues) in both conditions (Fig. 3). A 2 (condition) × 2 (outcome likelihood) × 3 (sites FCz, Cz, and CPz) ANOVA of FRN amplitude revealed significant effects of outcome likelihood, $F_{1,19} = 28.50$, $P < 0.0001$, and site, $F_{2,38} = 10.13$, $P < 0.001$, but not of condition, $F_{1,19} = 0.53$, $P > 0.1$. The FRN was greater for improbable than for probable outcomes and was maximal at frontocentral sites. No interactions involving the factor of condition approached significance (all $P > 0.1$), indicating that the FRN appeared similarly in both conditions.

We measured the FRN at site FCz and over the first and second halves of experiment blocks (Fig. 4). If the FRN was sensitive to experience, we reasoned that the FRN would increase for improbable outcomes as participants learned that those events were unlikely and that the FRN would decrease for probable outcomes as participants learned that those events were likely. A 2 (condition) × 2 (outcome likelihood) × 2 (block half) ANOVA revealed a significant effect of outcome likelihood, $F_{1,19} = 28.87$, $P < 0.0001$, but not of condition, $F_{1,19} = 0.43$, $P > 0.1$, or block half, $F_{1,19} = 0.07$, $P > 0.1$. Critically, the interaction between outcome likelihood and block half was significant, $F_{1,19} = 5.82$, $P < 0.05$, but the three-way interaction was not, $F_{1,19} = 0.03$, $P > 0.1$, because neural learning occurred in *both conditions*. When we divided blocks into quarters, the interaction between outcome likelihood and block quarter remained significant, $F_{3,57} = 3.18$, $P < 0.05$, whereas the three-way interaction again was not, $F_{3,57} = 1.16$, $P > 0.1$. To establish when the FRN for probable and improbable outcomes first differed, we applied paired t tests to the block quarter data. In the no instruction condition, curves diverged in the second quarter and remained different thereafter ($P < 0.05$, corrected). In the instruction condition, curves diverged in the third quarter and remained different thereafter ($P < 0.05$, corrected).

Model Results. Participants' choices appeared to depend on experience only in the no instruction condition. In contrast, the FRN appeared to depend on experience in both conditions. These results apply to the aggregate data. To determine whether trial-by-trial behavioral and neural responses depended on experience, instruction, or both, we explored predictions of three RL models. The *learning model* began without knowledge of the relative cue values and used reward prediction errors to learn which cues to select. The *start model* began with knowledge of the relative cue values and ignored reward prediction errors. Finally, the *full model* began with knowledge of the relative cue values, but continued to use reward prediction errors to revise expectations.

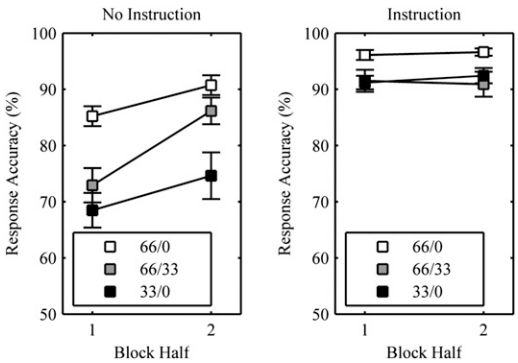

Fig. 2. Response accuracy (±1 SEM) by condition, block half, and cue pair.

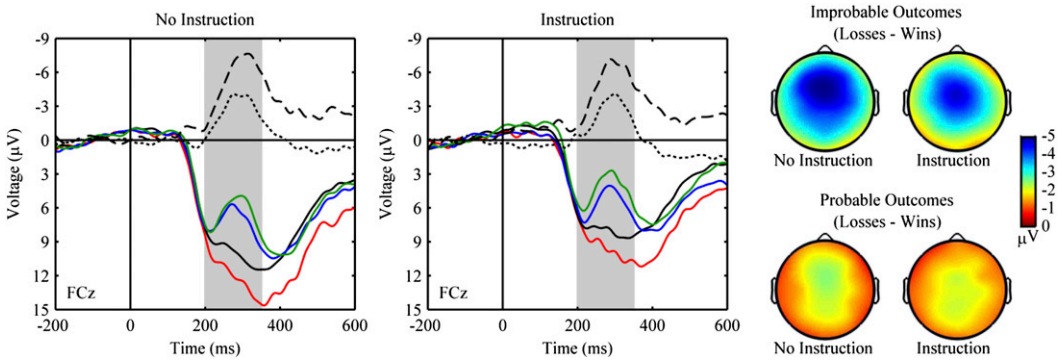

Fig. 3. ERPs for improbable losses (green), probable losses (blue), probable wins (black), and improbable wins (red) by condition at FCz. FRN (calculated as the difference between loss and win waveforms) for improbable outcomes (long-dashed line) and probable outcomes (short-dashed line) is shown. Scalp maps show topography of the FRN by condition and outcome likelihood. Time is from 200 to 350 ms with respect to feedback onset.

Table 1 contains parameter estimates and model fits to the behavioral data. When comparing models, we used the Bayesian information criterion (BIC) to account for the trade-off between goodness of fit and model complexity. Smaller BIC scores provide greater evidence in favor of a model. In the no instruction condition, BIC scores were smaller for the learning model than for the start model, $t(19) = 4.87$, $P < 0.001$, and the full model, $t(19) = 3.90$, $P < 0.001$. Thus, inclusion of learning improved the fit to the behavioral data in the no instruction condition, whereas inclusion of distinct start values did not justify the increased model complexity. In the instruction condition, BIC scores were smaller for the start model than for the learning model, $t(19) = 10.62$, $P < 0.0001$, and the full model, $t(19) = 20.25$, $P < 0.0001$. Thus, inclusion of start values improved the fit to the behavioral data in the instruction condition, whereas inclusion of learning did not justify the increased model complexity.

We considered a start model that scaled instructed values linearly and allowed only selection noise to vary ($\beta = 1.0$, $\gamma = 1.0$, and τ = free). This *fixed* start model is a nested version of the current, *free* start model. Using the likelihood-ratio test, we found that the free start model fit most participants better than the fixed start model in both conditions (instruction, 11/20; no instruction, 14/20; $P < 0.05$). Aggregating data likelihoods over participants, the free start model provided a far better fit in both conditions (instruction, $\chi^2_{18} = 106.83$, $P < 0.0001$; no instruction, $\chi^2_{18} = 272.94$, $P < 0.0001$). We also considered a learning model that initialized cues to a nonzero, uniform value (α = free, τ = free, Q_{start} = free). This *free* learning model is a generalized version of the current, *fixed* learning model. Using the likelihood-ratio test, we found that the free learning model did not fit any participant better than the fixed learning model in either condition (all $P > 0.1$). This conclusion was confirmed upon aggregating data likelihoods over participants (all $P > 0.1$).

Table 2 contains parameter estimates and model fits to the neural data. In the no instruction condition, BIC scores were smaller for the learning model than for the start model, $t(19) = 11.01$, $P < 0.0001$, and the full model, $t(19) = 75.51$, $P < 0.0001$. Likewise, in the instruction condition, BIC scores were smaller for the learning model than for the start model, $t(19) = 6.54$, $P < 0.0001$, and the full model, $t(19) = 15.52$, $P < 0.0001$. Thus, inclusion of learning improved the fit to the neural data in both conditions, whereas inclusion of distinct start values did not justify the increased model complexity.

Although differences between BIC scores among the neural models may appear modest, these differences, or log Bayes factors, approximate the average log odds in favor of a model at the level of the individual (27). The log Bayes factors provided positive to very strong evidence ($\Delta BIC > 2$) for the learning model over the start model for nearly all participants in both conditions (instruction, 18/20; no instruction, 20/20). Additionally, the log Bayes factors provided positive to very strong evidence for the learning model over the full model for all participants in both conditions. As a further test, we estimated the single set of parameters that maximized the data likelihood over all participants and for each model, but allowed the slope (b_1) and intercept (b_2) terms to vary among individuals. The aggregate log Bayes factors provided strong evidence for the learning model over the start model (instruction, 7.57; no instruction, 80.78) and the full model (instruction, 10.77; no instruction, 18.19).

In the no instruction condition, the learning model provided the best account for behavior and for the FRN. The learning rates estimated from the behavioral and neural data did not differ, $t(19) = 0.84$, $P > 0.1$. In the instruction condition, the start model provided the best account for behavior, but the learning model provided the best account for the FRN. Interestingly, the learning rates estimated from the neural data in the no instruction and instruction conditions did not differ, $t(19) = 0.97$, $P > 0.1$, and were correlated across individuals ($R^2 = 0.54$, $P < 0.001$). We investigated the effect of setting the neural learning rates in both conditions to values estimated from behavior in the no instruction condition. This method should worsen the BIC scores if the learning rates differed, but improve the scores if they were the same. When we fixed the neural learning rates to the behavioral learning rates, the BIC scores decreased in the no instruction condition, $t(19) = 2.27$, $P < 0.05$, and the instruction condition, $t(19) = 8.82$, $P < 0.0001$. Together, these results indicate that neural learning rates in both conditions coincided with the behavioral learning rates in the no instruction condition.

In fitting the models to the neural data, we estimated slope (b_1) and intercept (b_2) terms to scale model prediction errors to the observed voltages at FCz (Table 2). Prediction errors

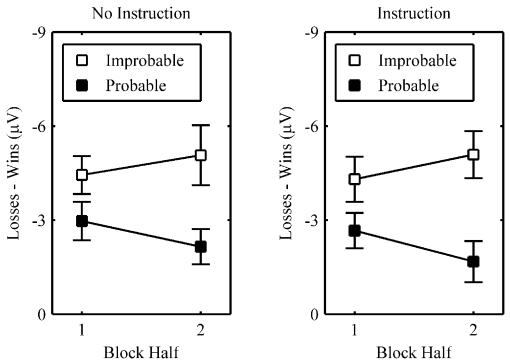

Fig. 4. FRN (±1 SEM) at FCz by condition, block half, and outcome likelihood.

Table 1. Average individual parameter estimates and model fits for behavioral data

Condition	Model	BIC	α	τ	β	γ
No instruction	Full	408	0.08	0.15	0.05	5.44
	Learning	402	0.09	0.17	0.00*	0.00*
	Start	436	0.00*	0.05*	0.09	0.91
Instruction	Full	238	0.01	0.11	3.79	1.63
	Learning	271	0.10	0.14	0.00*	0.00*
	Start	228	0.00*	0.05*	1.03	0.87

Parameters include learning rate (α), selection noise (τ), and curvature and elevation of the probability weighting function (β and γ).
*Fixed values. Setting β = 0.00 and γ = 0.00 in the learning model initializes all utility values to zero. Setting α = 0.00 in the start model prevents learning, and setting τ = 0.05 forces unique estimates of β and γ that maximize the likelihood of the data.

produced by the learning model clearly related to observed voltages: The values of b_1 were significantly greater than zero for all but one participant in both conditions. Although b_1 did not differ between conditions, $t(19) = 0.96$, $P > 0.1$, b_2 was larger in the no instruction condition, $t(19) = 4.86$, $P < 0.001$. This result is seen in the fact that the differences between waveforms were maintained across conditions, but that all waveforms were more positive in the no instruction condition (Fig. 3).

We estimated separate slope and intercept terms using the average voltage at each site and over three time windows (0–150 ms, 200–350 ms, and 400–550 ms). This method allowed us to determine whether b_1 and b_2 varied by scalp location and time (Fig. 5). A 2 (condition) × 3 (time) × 3 (sites FCz, Cz, and CPz) ANOVA of b_1 revealed a main effect of time, $F_{2,38} = 27.43$, $P < 0.0001$, but not of condition, $F_{1,19} = 1.23$, $P > 0.1$, or site, $F_{2,38} = 2.25$, $P > 0.1$. The b_1 values were maximal at frontal sites and from 200 to 350 ms in both conditions, evidenced by a significant interaction between site and time, $F_{4,76} = 30.47$, $P < 0.0001$. The topography and timing of the b_1 effect coincide with the FRN. A 2 (condition) × 3 (time) × 3 (site) ANOVA of b_2 revealed main effects of condition, $F_{1,19} = 19.85$, $P < 0.0001$, time, $F_{2,38} = 101.20$, $P < 0.0001$, and site, $F_{2,38} = 5.11$, $P < 0.05$. The b_2 values were maximal at posterior sites and from 400 to 550 ms. The topography and timing of the b_2 effect coincide with the P300, a late posterior ERP component evoked by stimulus processing (28). Unlike b_1 values, b_2 values were greater in the no instruction condition than in the instruction condition. Collectively, these results show that instruction diminished the P300 but did not affect the FRN.

Discussion

The goal of this study was to understand how instruction influences trial-and-error learning. We found that instruction eliminated participants' reliance on feedback as evidenced by their immediate asymptotic performance in the instruction condition. In striking contrast, the FRN continued to change with experience in both conditions.

Several theories seek to explain how the brain uses instruction and experience to select actions. These theories disagree on whether and how instruction and experience are combined. According to some accounts, the relationship between learning systems is antagonistic (5, 9–11). In one experiment that supports such accounts (5), the BOLD response to feedback in the nucleus accumbens and the ventromedial prefrontal cortex decreased following instruction. The decreased BOLD response in these regions was functionally correlated with an increased BOLD response in the dorsolateral prefrontal cortex, suggesting that the dorsolateral prefrontal cortex controlled the degree to which reward valuation regions processed outcomes. Additional support for such accounts comes from fMRI experiments that have revealed a negative association between MTL and striatal activation (10). Moreover, pharmacological deactivation of the MTL abolishes explicit learning and facilitates reinforcement learning supported by the striatum (11).

According to other accounts, learning systems interact (6–8, 29). In one experiment that supports such accounts (29), the BOLD response in the striatum reflected value estimates generated by a temporal-difference learning model. The BOLD response further reflected predictions generated by a model that used information about state transitions and reward probabilities, rather than stored value estimates, to prospectively calculate expected action values. This result suggests that top–down information provided by an internal world model influenced striatal reward computations. Additional support for such accounts comes from the finding that people overweigh outcomes that are compatible with instruction, indicating that knowledge representations in the PFC bias the BG to learn what is described by instruction (6, 7). Moreover, genetic polymorphisms associated with enhanced striatal dopaminergic functioning predict the degree to which people overweigh outcomes that are compatible with instruction, suggesting that striatal learning is sensitive to information contained in the PFC (8).

According to still other accounts, learning systems operate independently until a response is required (6, 12, 13). In one experiment that supports such accounts (13), the BOLD response in the striatum reflected the output of a temporal-difference learning model. In contrast, the BOLD response in the lateral prefrontal cortex reflected the output of a model that learned about state transitions and used this information to reason about probable outcomes. The latter model accounted for participants' initial behavior, whereas the former accounted for their asymptotic behavior. This result is consistent with animal conditioning studies that have shown that goal-directed and habit learning occur simultaneously in separate neural circuits and that behavior gradually transitions from the goal-directed to the habit system (30). This result is also consistent with the finding that neurons in the PFC and BG adapt at different rates during reversal learning (31).

ERP studies have been largely uninformative with respect to this issue because no study has hitherto examined the effects of instruction on EEG activity. Several ERP studies have examined the effects of experience on the FRN, however. These studies have consistently found that the FRN is sensitive to outcome valence and likelihood (14–19, 22–26). We replicated these results in the no instruction condition and showed that the FRN can evolve in the absence of behavioral change in the instruction condition. This result is consistent with accounts in which different neural systems operate independently until a response is required (12). These accounts traditionally distinguish between model-free and model-based RL. Model-free RL is mediated by the BG and uses temporal-difference learning to associate states and actions with rewards. Model-based RL, on the other hand, is mediated by the PFC/MTL and uses information about state

Table 2. Average individual parameter estimates and model fits for ERP data from FCz

Condition	Model	BIC	α	β	γ	b_1	b_2
No instruction	Full	1,990	0.14	2.37	3.84	3.95	7.32
	Learning	1,978	0.14	0.00*	0.00*	3.85	7.09
	Start	1,988	0.00*	2.72	6.60	3.68	7.43
Instruction	Full	1,974	0.18	2.57	5.10	3.71	6.08
	Learning	1,963	0.20	0.00*	0.00*	3.54	5.36
	Start	1,970	0.00*	4.81	6.43	3.54	6.11

Parameters include learning rate (α), curvature and elevation of the probability weighting function (β and γ), and the slope (b_1) and intercept (b_2) used to scale model prediction errors to observed voltages.
*Fixed values.

NEUROSCIENCE

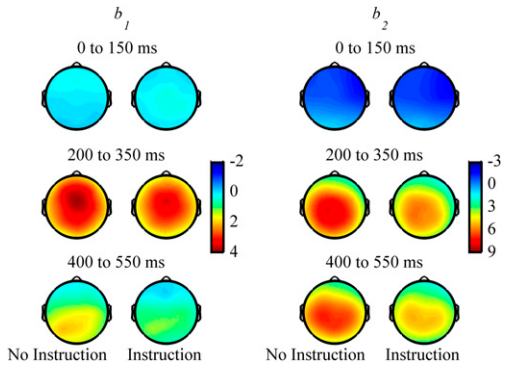

Fig. 5. Topography of regression coefficients by condition and time.

transitions and reward probabilities to prospectively calculate expected outcomes. According to one proposal, the cognitive system arbitrates between model-free and model-based controllers on the basis of the relative uncertainty of their estimates (12). In our task, instruction provided accurate information about reward probabilities. As such, the model-based system possessed greater certainty and controlled behavior in the instruction condition.

This result raises the question of whether the FRN contributes to behavioral adaptation or simply tracks a parallel neural process. In the instruction condition, the FRN quite clearly tracked a parallel neural process. A standard practice in neuroimaging research is to fit computational models to behavioral data and to then identify neural signals that correlate with latent model variables such as reward prediction error (32). This approach assumes that the same model produces the behavioral and neural results. By fitting models separately to the behavioral and neural data, we challenged this assumption. In doing so, we found that different models accounted for participants' choices and for the FRN in the instruction condition.

Establishing a relationship between the FRN and reward prediction errors is complicated by the fact that prediction errors depend on an individual's unique history of experience. We used computational models to account for the effects of experience on participants' expectations. This method allowed us to examine the trial-by-trial correspondence between EEG activity and reward prediction errors. We found that reward prediction errors were positively associated with EEG activity in both conditions. The strength of the association was maximal at frontocentral sites and from 200 to 350 ms, coinciding with the topography and timing of the FRN. This result supports the notion that the FRN tracked reward prediction error in both conditions.

EEG activity was not identical between conditions, however. All outcomes were more positive in the no instruction condition than in the instruction condition. This effect was captured by the greater intercept term in the no instruction condition (Table 2). The intercept was maximal at posterior sites and from 400 to 550 ms, coinciding with the topography and timing of the P300. Factors such as stimulus probability, stimulus significance, and attention alter P300 amplitude (28). One interpretation of this result, then, is that feedback was more significant and participants paid closer attention to feedback when it provided information about how to respond. In a related study (25), the P300 was larger when participants viewed outcomes in a choice task than when they passively viewed identical outcomes. There too, feedback may have been more significant and participants may have paid more attention to feedback when behavioral adaptation was possible.

Superficially, our results appear to contradict those of Li et al. (5), who found that the BOLD response to feedback in the nucleus accumbens and the ventromedial prefrontal cortex decreased following instruction. We can think of two reasons for the apparent difference between our results and theirs. First, our participants received instruction only at the start of epochs, whereas their participants received instruction on every trial. To the extent that the ACC is responsible for ongoing performance monitoring (21), participants may have paid less attention to feedback when exogenous signals provided information about reward contingencies on every trial. Second, Li et al. focused on the nucleus accumbens, whereas the FRN is thought to arise from the ACC (14, 22, 24). Although midbrain dopamine neurons target both structures, and although reward learning engages each (19, 33), the ACC and nucleus accumbens may respond differently to instruction.

Conclusion

We asked how information provided by instruction influences trial-and-error learning. We found that instruction about reward probabilities eliminated the effect of feedback on behavior. In striking contrast, the FRN continued to change with experience in both conditions. These results advance theories of neural reward valuation by showing that the FRN conforms to error signals produced by temporal-difference learning. More importantly, these results advance theories of learning by showing that although instruction may immediately control behavior, certain neural responses must be acquired from experience.

Materials and Methods

Participants. Twenty students participated on a paid volunteer basis (13 males and 7 females, ages ranging from 19 to 28 y, with a mean age of 23 y). All were right-handed, and none reported a history of neurological impairment.

Stimuli and Procedure. Participants sequentially viewed two cues followed by a choice screen (Fig. 1). They had 1,000 ms to select the cue that appeared first or second by choosing the box that contained the number 1 or 2. Participants responded by pressing *F* or *J* on a keyboard, using their left and right index fingers. When they responded before the deadline, the selected option turned green. Otherwise, both options turned red. At the end of the trial, feedback appeared. The symbols # and * denoted positive and negative feedback and were counterbalanced across participants. The symbol *!* appeared when participants failed to respond before the deadline (<2% of trials). In addition to receiving US$10.00, participants received performance-based payment. Positive feedback was worth 1 point, and 50 points were worth $1.00.

Each participant completed a no instruction condition and an instruction condition. The no instruction condition consisted of four epochs of 120 trials. Each epoch contained three cues that were rewarded with different probabilities, $P = \{0\%, 33\%, \text{and } 66\%\}$. Participants never received reward after choosing the 0 cue, they received reward with 33% after choosing the 33 cue, and they received reward with 66% after choosing the 66 cue. Two cues appeared in each trial, creating three unique pairs (66/33, 66/0, and 33/0) that occurred with equal frequencies. Although no cue was rewarded with 100% probability, participants could increase their earnings by selecting the cue that was more likely to be rewarded within each pair. Cues were 2D gray shapes, and no cue appeared in more than one epoch.

The instruction condition was identical to the no instruction condition with the following exception. At the start of each epoch, participants viewed the cues and received a description of their associated reward probabilities. Before continuing, they completed a quadruple dropout test to ensure they had memorized the reward probabilities. Condition order was counterbalanced, and participants completed all epochs of one condition before advancing to the next. Because condition order did not affect the behavioral or neural results, we excluded this factor from further analyses.

EEG Recording and Analysis. EEG data were recorded and processed according to standard protocols (*SI Materials and Methods*). We created feedback-locked ERPs for trials where participants selected the 66 cue or the 33 cue. Because neural responses depended *only* on the selected option (i.e., neural responses after the 66 cue did not depend on whether it was paired with the 33 or the 0 cue, and neural responses after the 33 cue did not depend on whether it was paired with the 66 or the 0 cue), we excluded the factor of cue pair from further analyses. To isolate the FRN, we compared losses and wins that were equally likely (16, 26). We created a probable outcome difference wave (losses after 33 cues minus wins after 66 cues) and an improbable outcome difference wave (losses after 66 cues minus wins after 33 cues). The FRN is typically maximal from 200 to 350 ms and at frontocentral sites (14–19, 22–26). As such, we measured the FRN as mean voltage of the

difference waves from 200 to 350 ms after feedback onset. We analyzed data from three midline sites (FCz, Cz, and CPz), and we applied the Greenhouse–Geisser correction when factors had more than two levels.

Computational Models. We compared predictions of three RL models that learned from instruction, experience, or both. In each trial, two cues appeared. The probability of selecting a cue, π_a was determined by a soft-max decision rule (1),

$$\pi_a = \frac{\exp\left(Q_a \big/ \tau\right)}{\sum_{b \in A(s)} \exp\left(Q_b \big/ \tau\right)}.$$

Selection noise (τ) controlled the degree of randomness in choices. After each outcome, r, the model computed a reward prediction error, $\delta = r - Q_a$. The model used the reward prediction error to update the utility of the selected cue, $Q_a \leftarrow Q_a + \alpha \cdot \delta$. Learning rate ($\alpha$) controlled the weighting of each outcome. The model received rewards of +1 and 0 for positive and negative feedback, respectively. This process constitutes an action-value model because it treats prediction error as the difference between reward and the value of the previous action. An alternate model, actor/critic, calculates the difference between reward and the value of the previous state (1). Supplementary analyses indicated that the FRN depended on the difference between the outcome and the previous action rather than the previous state, however (*SI Text*).

Empirical studies have shown that decision makers do not treat stated probabilities linearly (34). As such, we converted instructed probabilities to starting utility values using a two-parameter weighting function,

$$Q_a = \frac{\beta \cdot P_a^{\gamma}}{\beta \cdot P_a^{\gamma} + (1 - P_a)^{\gamma}}.$$

The γ- and β-parameters control the curvature and elevation of the weighting function, respectively. Together, γ and β determine the differences between the starting values of the three cues. Other two-parameter weighting functions would yield identical results.

We used the behavioral data to estimate separate parameter values for each participant and condition. To do so, we presented the models with the history of choices and rewards that the participant experienced. For each trial, t, we calculated the probability that the model would make the same choice as the participant, $p_k(t)$. We used the simplex optimization algorithm with multiple start points to identify parameter values that maximized the log likelihood of the observed choices, LLE = $\sum_t \ln(p_k(t))$. The full model contained four free parameters (α, τ, β, and γ), the learning model contained two free parameters (α and τ), and the start model contained two free parameters (β and γ). To compare models, we used the BIC, defined as $-2 \cdot \text{LLE} + p \cdot \ln(n)$. For each participant and condition, the model predicted ~480 data points. We calculated separate BIC scores for each participant and condition. We also considered a model with separate learning rates for wins and losses and a model with an instruction confirmation bias (6–8). Neither addition improved model performance.

We also used the neural data to estimate separate parameter values for each participant and condition. For each trial, we calculated the observed voltage at FCz from 200 to 350 ms and the reward prediction error that the model generated on the basis of the participant's selection and the outcome. We then estimated slope and intercept terms to scale model prediction errors to observed voltages, Voltage = $b_1 \cdot \text{PE} + b_2$. We found parameter values that minimized the mean squared error (MSE) between the expected and observed voltages. The full model contained five free parameters (α, β, γ, b_1, and b_2), the learning model contained three free parameters (α, b_1, and b_2), and the start model contained four free parameters (β, γ, b_1, and b_2). To compare models, we used the BIC, defined as $n \cdot \ln(\text{MSE}) + p \cdot \ln(n)$. We calculated separate BIC scores for each participant and condition.

ACKNOWLEDGMENTS. This work was supported by a Training Grant T32MH019983 fellowship (to M.W.M.) and National Institute of Mental Health Grant MH068243 (to J.R.A.).

1. Sutton RS, Barto AG (1998) *Reinforcement Learning: An Introduction* (MIT Press, Cambridge, MA).
2. Thorndike EL (1911) *Animal Intelligence: Experimental Studies* (MacMillan, New York).
3. Schultz W (1998) Predictive reward signal of dopamine neurons. *J Neurophysiol* 80: 1–27.
4. O'Doherty JP (2004) Reward representations and reward-related learning in the human brain: Insights from neuroimaging. *Curr Opin Neurobiol* 14:769–776.
5. Li J, Delgado MR, Phelps EA (2011) How instructed knowledge modulates the neural systems of reward learning. *Proc Natl Acad Sci USA* 108:55–60.
6. Doll BB, Jacobs WJ, Sanfey AG, Frank MJ (2009) Instructional control of reinforcement learning: A behavioral and neurocomputational investigation. *Brain Res* 1299:74–94.
7. Biele G, Rieskamp J, Gonzalez R (2009) Computational models for the combination of advice and individual learning. *Cogn Sci* 33:206–242.
8. Doll BB, Hutchison KE, Frank MJ (2011) Dopaminergic genes predict individual differences in susceptibility to confirmation bias. *J Neurosci* 31:6188–6198.
9. McClure SM, Laibson DI, Loewenstein G, Cohen JD (2004) Separate neural systems value immediate and delayed monetary rewards. *Science* 306:503–507.
10. Poldrack RA, et al. (2001) Interactive memory systems in the human brain. *Nature* 414: 546–550.
11. Frank MJ, O'Reilly RC, Curran T (2006) When memory fails, intuition reigns: Midazolam enhances implicit inference in humans. *Psychol Sci* 17:700–707.
12. Daw ND, Niv Y, Dayan P (2005) Uncertainty-based competition between prefrontal and dorsolateral striatal systems for behavioral control. *Nat Neurosci* 8:1704–1711.
13. Gläscher J, Daw N, Dayan P, O'Doherty JP (2010) States versus rewards: Dissociable neural prediction error signals underlying model-based and model-free reinforcement learning. *Neuron* 66:585–595.
14. Miltner WHR, Braun CH, Coles MGH (1997) Event-related brain potentials following incorrect feedback in a time-estimation task: Evidence for a "generic" neural system for error detection. *J Cogn Neurosci* 9:788–798.
15. Holroyd CB, Pakzad-Vaezi KL, Krigolson OE (2008) The feedback correct-related positivity: Sensitivity of the event-related brain potential to unexpected positive feedback. *Psychophysiology* 45:688–697.
16. Walsh MM, Anderson JR (2011) Learning from delayed feedback: Neural responses in temporal credit assignment. *Cogn Affect Behav Neurosci* 11:131–143.
17. Holroyd CB, Larsen JT, Cohen JD (2004) Context dependence of the event-related brain potential associated with reward and punishment. *Psychophysiology* 41: 245–253.
18. Cohen MX, Ranganath C (2007) Reinforcement learning signals predict future decisions. *J Neurosci* 27:371–378.
19. Holroyd CB, et al. (2004) Dorsal anterior cingulate cortex shows fMRI response to internal and external error signals. *Nat Neurosci* 7:497–498.
20. Niki H, Watanabe M (1979) Prefrontal and cingulate unit activity during timing behavior in the monkey. *Brain Res* 171:213–224.
21. Rushworth MFS, Walton ME, Kennerley SW, Bannerman DM (2004) Action sets and decisions in the medial frontal cortex. *Trends Cogn Sci* 8:410–417.
22. Holroyd CB, Coles MGH (2002) The neural basis of human error processing: Reinforcement learning, dopamine, and the error-related negativity. *Psychol Rev* 109: 679–709.
23. Santesso DL, et al. (2009) Single dose of a dopamine agonist impairs reinforcement learning in humans: Evidence from event-related potentials and computational modeling of striatal-cortical function. *Hum Brain Mapp* 30:1963–1976.
24. Nieuwenhuis S, Holroyd CB, Mol N, Coles MGH (2004) Reinforcement-related brain potentials from medial frontal cortex: Origins and functional significance. *Neurosci Biobehav Rev* 28:441–448.
25. Yeung N, Holroyd CB, Cohen JD (2005) ERP correlates of feedback and reward processing in the presence and absence of response choice. *Cereb Cortex* 15:535–544.
26. Holroyd CB, Krigolson OE, Baker R, Lee S, Gibson J (2009) When is an error not a prediction error? An electrophysiological investigation. *Cogn Affect Behav Neurosci* 9:59–70.
27. Kass RE, Raftery AE (1995) Bayes factors. *J Am Stat Assoc* 90:773–795.
28. Johnson R, Jr. (1986) A triarchic model of P300 amplitude. *Psychophysiology* 23: 367–384.
29. Daw ND, Gershman SJ, Seymour B, Dayan P, Dolan RJ (2011) Model-based influences on humans' choices and striatal prediction errors. *Neuron* 69:1204–1215.
30. Balleine BW, O'Doherty JP (2010) Human and rodent homologies in action control: Corticostriatal determinants of goal-directed and habitual action. *Neuropsychopharmacology* 35:48–69.
31. Pasupathy A, Miller EK (2005) Different time courses of learning-related activity in the prefrontal cortex and striatum. *Nature* 433:873–876.
32. Mars RB, Shea NJ, Kolling N, Rushworth MFS (2010) Model-based analyses: Promises, pitfalls, and example applications to the study of cognitive control. *Q J Exp Psychol* 29: 1–16.
33. Santesso DL, et al. (2008) Individual differences in reinforcement learning: Behavioral, electrophysiological, and neuroimaging correlates. *Neuroimage* 42:807–816.
34. Gonzalez R, Wu G (1999) On the shape of the probability weighting function. *Cognit Psychol* 38:129–166.