Biological Psychology 149 (2020) 107775

Contents lists available at ScienceDirect

Biological Psychology

journal homepage: www.elsevier.com/locate/biopsycho

Dissociating the contributions of reward-prediction errors to trial-level adaptation and long-term learning

K.R. Lohse[a,b,*], M.W. Miller[c,d], M. Daou[c,e], W. Valerius[c], M. Jones[f]

[a] University of Utah, Department of Health, Kinesiology, and Recreation, United States
[b] University of Utah, Department of Physical Therapy and Athletic Training, United States
[c] Auburn University, School of Kinesiology, United States
[d] Auburn University, Center for Neuroscience, United States
[e] Coastal Carolina University, Department of Kinesiology, United States
[f] University of Colorado, Department of Psychology and Neuroscience, United States

ARTICLE INFO

Keywords:
EEG
Reinforcement learning
RewP
Adaptation

ABSTRACT

Reward positivity (RewP) is an EEG component reflecting reward-prediction errors. Using multilevel models, we measured single-trial RewP amplitude from trial-to-trial, while reward and prediction varied during learning. Sixty participants completed a category-learning task in either engaging or sterile conditions with the RewP time-locked to feedback. Sequential analysis of single-trial RewP showed its relationship to current and previous accuracy, and the probability of changing one's response to subsequent stimuli. Simulations show these effects can be explained in detail by the dynamics of participants' expectations according to principles of reinforcement learning. The single-trial RewP findings were consistent with previous literature linking RewP to reward-prediction error under reinforcement-learning theory. In contrast, the aggregate RewP was unrelated to the engagement manipulation or to delayed retention performance. Thus the present results provide a detailed computational account how RewP relates to acute adaptation, but suggest RewP plays little role in long-term learning.

1. Introduction

Reinforcement learning theory posits that individuals adjust their behavior in order to obtain rewards and avoid punishments (Sutton & Barto, 1998; Thorndike, 1927). These behavioral adjustments are driven by reward-prediction errors—the degree to which an actual reward differs from the learner's expectation (Schultz, 2017). Reward-prediction errors can result from rewards being better (positive) or worse (negative) than predicted, and both positive and negative prediction errors influence future behavior. Positive reward-prediction errors act as a signal within the brain to 'stamp in' a behavior, making the preceding action more likely to be selected in the future for a given state of the system (Holroyd & Krigolson, 2007; Palidis, Cashaback, & Gribble, 2018). Conversely, negative reward-prediction errors act as a signal within the brain to 'stamp out' a behavior, so that it can be avoided in the future.

In many studies of human learning, researchers measure reward-prediction errors through the proxy measure of the reward positivity (RewP) component derived from event-related potential (ERP) waveforms in electro-encephalograms (EEGs; Proudfit, 2015; Sambrook & Goslin, 2015). (This component has gone by other names including the feedback negativity, feedback-related negativity, feedback error-related negativity, and feedback correct-related positivity, but it is operationally and conceptually the same component as the RewP [Sambrook & Goslin].) Operationally, the RewP is a positive deflection in the ERP waveform that occurs 250 to 350 ms following positive feedback relative to negative feedback. The RewP exhibits a fronto-central scalp distribution, likely generated by anterior cingulate cortex. Conceptually, the RewP is believed to reflect a reward-prediction error (Holroyd & Coles, 2002). The positive prediction errors are likely transmitted to anterior cingulate cortex by a phasic increase in dopamine from the midbrain (Krigolson, 2018). In current practice, most researchers operationalize the RewP as a difference wave, specifically by averaging together ERPs from all trials with positive feedback, all trials with negative feedback, and then subtracting the latter average from the former. Alternatively, some researchers have looked at the RewP on a trial-by-trial level, operationalizing it as the voltage following feedback (Collins & Frank, 2018; Frömer, Stürmer, & Sommer,

* Corresponding Author at: 250 S 1850 E, Rm 258, Salt Lake City, UT, 84112, United States.
E-mail address: rehabinformatics@gmail.com (K.R. Lohse).

https://doi.org/10.1016/j.biopsycho.2019.107775
Received 22 January 2019; Received in revised form 23 September 2019; Accepted 23 September 2019
Available online 26 September 2019

2016). We refer to these two different measures as the aggregate RewP and single-trial RewP, respectively.

Converging evidence implicates both the aggregate and single-trial RewP as indices of reward-prediction error (Sambrook & Goslin, 2015). The difference-wave approach reliably yields a deflection in the waveform, indicating a strong sensitivity to the valence of feedback. The aggregate RewP has also been shown to be greater in response to larger rewards (e.g., $0.5 vs. $5.0). Thus, RewP is affected by both the sign and the magnitude of reward. Aggregate RewP is also sensitive to the participant's predictions (i.e., expectation of reward), being greater for unexpected outcomes than for expected outcomes (Holroyd & Krigolson, 2007; Sambrook & Goslin, 2015). The combination of positive dependence on reward and negative dependence on predictions implicates the single-trial RewP as a neural correlate of their difference, that is, reward-prediction error.

Currently, there is a limited understanding of how RewP relates to behavior over different timescales. Although a number of recent electrophysiological studies have investigated trial-by-trial dynamics of reward-prediction error (Chase, Swainson, Durham, Benham, & Cools, 2011; Collins & Frank, 2018; Fischer & Ullsperger, 2013; Frömer et al., 2016; Pedroni, Langer, Koenig, Allemand, & Jancke, 2011; Philiastides, Biele, Vavatzanidis, Kazzer, & Heekeren, 2010; Sambrook & Goslin, 2014, 2016; Sambrook, Hardwick, Wills, & Goslin, 2018), most work has focused on block- or session-level manipulations (for representative examples, see Bellebaum & Daum, 2008; Reinhart & Woodman, 2014; van der Helden, Boksem, & Blom, 2009). Moreover, there has been little to no research on how RewP relates to delayed retention or transfer, as opposed to immediate performance. The fact that much of the extant literature does not consider trial-by-trial dynamics or delayed retention and transfer tests is problematic for at least two reasons. First, analysis of trial-level (i.e., sequential) effects can be highly informative regarding the details of reinforcement learning mechanisms (Jones, Love, & Maddox, 2006; Jones, Curran, Mozer, & Wilder, 2013; Philiastides et al., 2010). Second, the relationship between the acute adaptation mechanisms (as postulated by reinforcement learning theory) and long-term learning is poorly understood. From the perspective of reinforcement learning theory, positive reward-prediction errors during practice drive adaption toward better performance. Reinforcement learning explains long-term learning as the accumulation of these adaptations (e.g., Sutton & Barto, 1998). These adaptations are also assumed to underlie generalization to novel stimuli, via overlap in stimulus representations (e.g., Jones et al., 2006). Thus, a straightforward prediction is that achieving a high level of performance during practice should be associated with better performance on subsequent retention and transfer tests. However, behavioral studies have shown that these two measures of performance are often uncorrelated or even negatively correlated (Kantak & Winstein, 2012; Pashler & Baylis, 1991).

The present experiment investigated the neurophysiological correlates of adaptation and learning in a perceptual categorization task. The primary aim was to evaluate the separate impacts of reward and prediction on the aggregate RewP and the single-trial RewP. We experimentally manipulated the value of reward between participants using a motivational game manipulation designed to enhance reward processing (Lohse, Boyd, & Hodges, 2016). To do this, we adapted a category learning task using complex visual stimuli called greebles (Gauthier & Tarr, 1997). In the "sterile" group, described in detail below, participants had to learn which of several responses corresponded to each family of greebles through trial-and-error categorization. In the "game" group, participants had to complete the same task, but rather than as a cognitive psychology experiment the task was framed as a game, "Goblin Quest", in which participants had to learn which of several "weapons" (responses) corresponded to each "goblin" (family of greebles). Building on previous work indicating that motivation enhances learning (Wulf & Lewthwaite, 2016), we hypothesized that increased motivation from the game manipulation would magnify representations of reward, yielding stronger reward-prediction errors (and thus magnified RewP).

In addition to the empirical study, we present and test a detailed analysis of computational reinforcement-learning models for trial-level dynamics of reward-prediction errors. Previous research on RewP has primarily used random feedback or stairstep procedures to hold reward probability to predetermined values (e.g., Holroyd & Krigolson, 2007). However, it is important to recognize that the probability of reward is not constant in natural learning environments. As learners become more skillful and knowledgeable in a task, success is more likely, and the accuracy of their predictions also increases. To the extent that RewP indexes prediction error, as opposed to just reward, it should be impacted by changes in learners' expectancy of success. Indeed, studies that have investigated trial-by-trial dynamics when learners' predictions are allowed to evolve freely over the course of learning reveal intriguing temporal dynamics of the single-trial RewP (Collins & Frank, 2018; Fischer & Ullsperger, 2013; Frömer et al., 2016; Philiastides et al., 2010; Sambrook & Goslin, 2016; Sambrook et al., 2018). For instance, Frömer et al. (2016) showed how the single-trial RewP displays characteristics of both a reward measure, being greater following accurate performances, and an expectancy measure, getting lesser as cumulative accuracy gets better. In the present study, we add to this research and extend it by (A) using mixed-effect regression analyses to show the relationship between the single-trial RewP and both preceding and subsequent behavior, (B) using simulations to show how the dynamics of the RewP can be explained in fine detail by a reinforcement learning model, and (C) including delayed retention and transfer tests to investigate whether observed short-term learning dynamics carry over to explain long-term learning.

1.1. Separating dynamics of prediction and reward in RewP

Because reward-prediction error is a difference between actual and expected reward, both factors are relevant to the interplay between prediction errors and learning. To the extent that RewP is sensitive to variation on the *reward* side, the most straightforward prediction is that single-trial RewP is more positive on correct trials than on incorrect trials. The corresponding property of the aggregate RewP, that the difference wave is positive, is well established and is the original basis for its interpretation as a correlate of reward-prediction error (Holroyd & Coles, 2002). However, reward-side effects on RewP also lead to several more detailed predictions. First, stronger subjective reward should produce larger reward-prediction errors, leading to greater behavioral adaptation (e.g., Cashaback, McGregor, Mohatarem, & Gribble, 2017). This is the assumption underlying the game manipulation: the gamified task would increase subjective reward, magnifying the RewP, and thus speeding learning. Similarly, to the extent that there are individual differences in strength of subjective reward, this assumption also predicts a positive correlation between aggregate RewP and training performance across participants (Grand, Daou, Lohse, & Miller, 2017; Holroyd & Krigolson, 2007). Moreover, if reinforcement learning mechanisms are important for long-term learning, and not just short-term adaptation, then the same correlation should be obtained for post-test performance (e.g., Abe et al., 2011). Finally, there is the parallel prediction at the within-participant level: If subjective reward varies across trials (separately for positive and negative feedback), there should be a positive correlation between the single-trial RewP on the current trial and the probability of repeating the current response the next time a stimulus from the same category is presented. This is because a greater RewP indicates the subjective value of that response given the stimulus will be more greatly increased (positive feedback) or more weakly diminished (negative feedback).

To the extent that RewP is sensitive to variation on the *prediction* side, three somewhat counterintuitive predictions derive directly from reinforcement learning theory. First, participants who have learned the task better will have greater expectation of reward when they choose a correct response and lesser expectation when they choose an incorrect

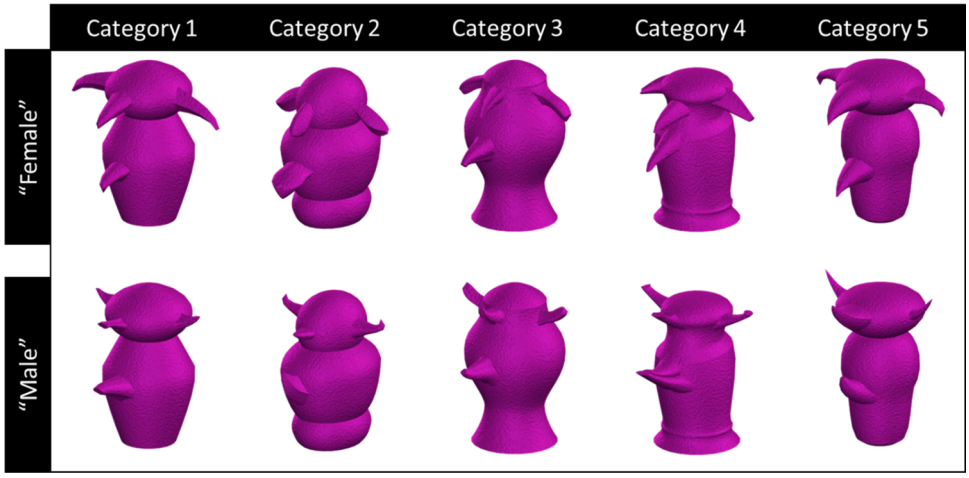

Fig. 1. Example greebles/goblins for each category. Note participants made their categorization based on the family of the greebles and ignored the "gender" of the greebles.

response, leading to a smaller aggregate RewP. This leads to the prediction of a negative correlation between aggregate RewP and training performance across participants (directly opposing the prediction above from reward-side variation). Second, a similar mechanism applies at the within-participant level: Because predictions are more accurate when the learner has better knowledge of the current category, the difference in RewP between correct and incorrect trials should decline as participants learn. Finally, controlling for the valence of the feedback, a greater prediction error indicates the participant had less confidence in the chosen response. In this way, single-trial RewP should correlate negatively with the probability of repeating the response the next time a stimulus from the same category is presented, separately for correct and for incorrect trials (again, in contrast to the prediction above from reward-side variation).

Contrasting the empirical predictions that derive from the reward versus prediction sides of the reward-prediction error also bears on the causal role of the neural signal underlying RewP in learning. The predictions above from reward-side variation are primarily based on the assumption that this signal causally drives adaptation: Variation in reward representation should lead to variation in how responses are strengthened or weakened, leading to positive correlations of aggregate RewP with practice and retention performance at the between-subjects level, and of single-trial RewP with response repetition at the within-subject level. In contrast, the predictions from the prediction side are consistent with the RewP as more of an epiphenomenal correlate of the reward-prediction error, reflecting adaptations that have already taken place. That is, if participants have already updated their internal representations of the task, they should show a smaller aggregate RewP, leading to a negative correlation with performance. Likewise, greater single-trial RewP would be associated with changing responses because it indicates lower confidence in the chosen response.

To preview our results, we find a complex pattern of relationships between single-trial RewP and choice behavior at the trial level, and between aggregate RewP and performance averaged over the training session, that are remarkably well explained by a simple reinforcement-learning model. In particular, we find that RewP tracks prior experience as predicted, but it predicts future behavior only in a correlative (not causal) manner, as anticipated by the prediction-side dynamics of the model. We also find that the game manipulation reliably impacted both training and retention performance, but with the sterile group outperforming the game group (opposite what was predicted). However, RewP was unaffected by the game manipulation and did not predict retention performance. We conclude that reinforcement-learning theory provides an excellent computational account of RewP at both aggregate and single-trial levels of analysis, but also that the local adaptation mechanisms do not account for long-term retention and transfer of information.

2. Methods

2.1. Participants

Sixty right-handed young adults (41 females), aged between 18 and 25 years (M_{age} = 20.9 years, SD = 1.40 years), participated in the study after consenting to a protocol approved by the Auburn University Institutional Review Board (#17-009 EP 1702). Participants were recruited from university courses via an online recruitment system and were compensated with course credit. Sample size was determined by an a priori power calculation using GLIMMPSE 2.2.4 (Kreidler et al., 2013). A total sample size of N = 54 was determined necessary to achieve 80% power for a 20% difference in posttest accuracy (SD = 40%), assuming an alpha of 0.05, a 1:1 allocation ratio between groups, and a correlation of 0.8 between repeated measures. To account for participant dropout and poor EEG recording, the decision was made to recruit N = 60 participants. Participants were randomly assigned to experimental groups, balanced separately within males and females.

2.2. Task and procedure

2.2.1. Day 1 (Practice)

After consenting to participate in the experiment, which was referred to as a "game" for the game group and a "task" for the sterile group, participants completed the Edinburgh Handedness Inventory to verify right-handedness while being prepared for EEG recording. Next, participants were seated 60 cm from the center of a computer monitor (38.5 cm screen). Participants adjusted the computer keyboard so that they could comfortably reach the 1, 2, 3, 4, and 5 keys in the number row near the top of the keyboard with their right index finger. Next, participants were given instructions designed to facilitate good EEG recording (e.g., "do not move your head") and had 3 min of resting EEG recorded. After the resting recording, participants were given game group or sterile group instructions via PowerPoint slides (see Supplementary Appendix i), which they advanced at their own pace. Both sets of instructions asked participants to learn to categorize complex visual stimuli (greebles; Gauthier & Tarr, 1997). Each greeble has a vertically oriented central component, the shape of which varies depending on to which one of five families the greeble belongs (see Fig. 1). Each greeble also has four protrusions, which originate in a similar configuration along the central component of all greebles. If the protrusions point upward, the greeble is of one gender, and if the

protrusions point downward, the greeble is of another gender. Participants were instructed to categorize greebles according only to their central component (family), not the protrusions (gender).

The game instruction slides included background music and described a game titled "Goblin Quest". In this game, participants were described a scenario in which they awoke in a dungeon and needed to fight goblins in order to escape. Participants were told they would encounter five types of goblins, and there would be sixteen of each type. Participants were told each type of goblin was sensitive to one of five weapons, so the participant needed to learn which weapon to use to "kill" a particular type of goblin. Participants selected a weapon by choosing its corresponding key on the keyboard: 1. Sword, 2. Nunchucks, 3. Throwing Star, 4. Axe, or 5. Slingshot. Participants were given two pieces of advice: (1) try to focus on the goblin's body shape rather than the protrusions from the goblin's body, and (2) initially try to use the same weapon until you figure out which goblin is sensitive to it. (We included this advice based on pilot data that suggested participants struggled to learn the categories otherwise.)

The sterile instruction slides described a task titled "Stimulus Categorization". In this task, participants were described a scenario in which they needed to categorize five types of complex stimuli with sixteen of each type. Participants were told different types of stimuli go in different categories, so the participant needed to learn in which category to place each stimulus. Participants selected a category by choosing its corresponding key (1, 2, 3, 4, or 5 on the keyboard). Participants were given the same two pieces of advice: (1) try to focus on the body shape of the stimulus, and (2) initially try choosing the same category until you figure out which stimulus belongs in that category.

After the participant viewed the instruction slides, the experimenter asked the participant whether he or she had any questions and answered the questions. To ensure that participants understood the instructions, participants completed two trials of the game/task before proceeding to the practice trials. These warm-up trials followed the same procedure as those used during actual practice; the goblins/stimuli used for warm-up were not repeated during practice. The game/task, which is depicted in Fig. 2, presented participants with a fixation cross for 2000 ms, followed by the goblin/stimulus. For the game group, the "goblin" was accompanied by the sound of breathing and below each response number was a weapon icon. The goblin/stimulus remained until the participant made a response by pressing the 1, 2, 3, 4, or 5 key with their right index finger. Next, a fixation square was presented for 2000 ms, followed by a checkmark indicating the participant had killed the goblin or correctly categorized the stimulus, or an X indicating the goblin survived or the stimulus was incorrectly categorized. The feedback was followed by a blank screen for 500 ms, and then the next trial began. Participants completed 80 trials (16 goblins/stimuli for each family—the goblins/stimuli were randomly ordered, and the same random order was used for all participants) while having their EEG recorded. Thus, every stimulus appeared exactly once, with no repetitions. Finally, at the end of practice, participants completed a language-adapted version of the User Engagement Scale (O'Brien & Toms, 2008).

2.2.2. Day 2 (1-week Post-tests)

One week after Day 1, participants returned to the laboratory and provided verbal consent to continue in the experiment. Participants completed game and sterile post-tests in counterbalanced order. Prior to each post-test, participants were given game or sterile post-test instructions via PowerPoint slides. These instructions were similar to the instructions given before the practice phase with only minor differences (also shown in Supplementary Appendix i). First, the instructions asked participants to try to remember what weapon/category goes with the goblin/stimulus "…based on what [they] learned the prior week". Second, the instructions did not indicate that participants would receive any feedback (since participants did not receive feedback during post-tests). Third, the instructions did not provide hints about focusing on body shape rather than protrusions or sticking with a single response until correct feedback was received (again, because no feedback was given during post-tests).

After the participant viewed the instruction slides, the experimenter asked the participant whether he or she had any questions and provided clarifications as necessary. Participants then completed the post-tests, which were similar to the practice game/task. Specifically, a fixation cross was presented for 2000 ms, followed by a goblin/stimulus. (As in the practice phase, in the game test the goblin was accompanied by sound representing heavy goblin breathing.) The stimulus remained until the participant made a response by pressing the 1, 2, 3, 4, or 5 key with their right index finger. Next, a blank screen appeared for 500 ms, and then the next trial began. Participants completed 30 trials (6 goblins/stimuli for each family) in each post-test. Half of the goblins/stimuli in each family were rotated by 15 degrees about the vertical axis

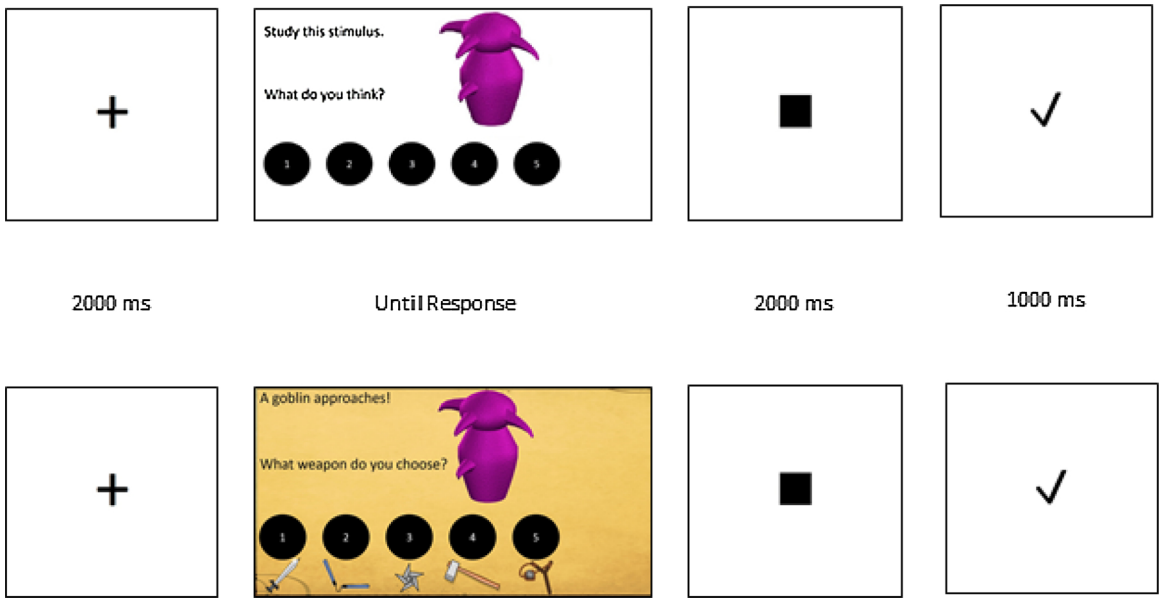

Fig. 2. Example stimulus displays from the "sterile" condition (top) and the "game" condition (bottom) showing the time-course of the stimuli and correct feedback in both conditions.

in three dimensions. The game post-test was the same for every participant in terms of the goblins that were presented and the order in which they were presented. Likewise, the sterile post-test was the same for every participant in terms of the stimuli that were presented and the order in which they were presented. The game and sterile post-tests contained different greebles, but all greebles had been presented during practice.

2.3. EEG recording and processing

2.3.1. EEG recording

Scalp EEG was collected from 31 channels of an EEG cap housing a 64-channel BrainVision actiCAP system (Brain Products GmbH, Munich, Germany) labeled in accord with an extended international 10–20 system (Oostenveld & Praamstra, 2001). EEG data were online-referenced to the left earlobe, and a common ground was employed at the FPz electrode site. Electrode impedances were maintained below 25 kΩ throughout data collection and a high-pass filter was set at 0.016 Hz with a sampling rate of 250 Hz. The EEG signal was amplified and digitized with a BrainAmp DC amplifier (Brain Products GmbH) linked to BrainVision Recorder software (Brain Products GmbH).

2.3.2. EEG processing

EEG data processing was conducted with BrainVision Analyzer 2.1 software (Brain Products GmhB). Data were re-referenced to an average ears montage and then filtered between 0.1 and 30 Hz with 4th-order roll-offs as well as a 60-Hz notch filter. Next, the data were inspected and prominent artifacts were marked in order to facilitate the subsequent independent component analysis (ICA)-based ocular correction. This function searches for an ocular artifact template in a designated channel (FP2) and then finds ICA-derived components that account for a user-specified amount (70%) of variance in the template-matched portion of the signal in the designated channel. The components are then removed from the EEG signal, which is reconstructed for further processing.

2.3.3. Reward positivity

The reward positivity (RewP) was obtained by segmenting the EEG data into epochs beginning 100 ms before and ending 1000 ms after the onset of feedback (correct [✓] or incorrect [X]). These epochs were then baseline corrected from −100 to 0 ms. Next, epochs were automatically rejected if they contained a > 50 μV/ms change, a > 100 μV change in a moving 200-ms window, or a < 0.5 μV change in a moving 200-ms window at electrodes of interest (Fz, Cz, and Pz). On average, 99.1% (SD = 3.38% across participants) of epochs were retained, and each participant had at least 60 epochs retained. The remaining epochs (M = 31.2, SD = 16.4, minimum = 11) for correct feedback were averaged, and the remaining epochs (M = 48.0, SD = 16.2, minimum = 9) for incorrect feedback were averaged separately. Next, a difference wave was formed by subtracting the incorrect average from the correct average. The positive peak in the difference wave showed considerable latency variability between participants, so an adaptive mean amplitude window technique was employed (Clayson, Baldwin, & Larson, 2013). Specifically, each participant was associated with a 40-ms time window centered around the time of maximum amplitude between 245 and 345 ms at electrode Fz in that participant's difference wave. The choices of range of possible time windows (245–345 ms) and electrode (Fz) were based on the positive peak in the difference wave grand average across all participants in both groups. Single-trial RewP was quantified by taking the mean amplitude on each trial in the subject's 40-ms time window, and aggregate RewP was quantified for each subject by taking the mean of that subject's difference wave over the same time window. Consequently, each subject's aggregate RewP is equal to the mean single-trial RewP on correct trials minus the mean on error trials. (Additional EEG components, the encoding stimulus preceding negativity and feedback stimulus preceding negativity, were extracted for exploratory analyses. Details of these analyses are presented in Supplemental Appendix II.)

3. Data analysis

Prior to statistical analyses, we visually inspected density plots for the single-trial RewP (and encoding/feedback preceding negativity). For the single-trial RewP, 99% of the data fell between −30 μV and 41 μV, but the full set of values ranged from −558 to 421 μV. To reduce the influence of these outlying scores, we excluded trials with single-trial RewP < −50 μV or > +50 μV. These criteria led to 1.5% of trials being excluded from subsequent analyses of the single-trial RewP. Analyses were conducted in R using a combination of linear and logistic mixed-effects regressions, depending on the nature of the dependent variable (Bates, Maechler, Bolker, & Walker, 2015; Kuznetsova, Brockhoff, & Christensen, 2017). In all mixed-effects regressions, we first contrast-coded all categorical variables (such that the intercept corresponds to the mean) and mean-centered all continuous variables. Trial was coded as a linear contrast in all analyses. All processed data and code for statistical analyses are available online (https://github.com/keithlohse/cat_learn/). Analyses for the assessment of learning were pre-registered prior to data collection (https://aspredicted.org/xt2zp.pdf). Analyses for the sequential effects of RewP during practice, although theoretically driven, were developed after data were collected.

3.0.1. Improvement during practice

To measure changes in response accuracy during practice, we used a logistic mixed-effects regression with fixed effects of Group (game/sterile), Trial number, and their interaction. These models included random effects of Participant, Participant × Trial, Stimulus Category, and Participant × Stimulus Category. The dependent variable was the whether or not the current response was correct.

Additionally, to measure the aggregate RewP, we used a one-sample t-test to test whether the grand-average RewP was different from zero and an independent-samples t-test to test whether the aggregate RewP differed between the groups. Finally, to assess the relationship between the aggregate RewP and practice performance we used two different linear models. First, we regressed overall accuracy onto the aggregate RewP (from the difference wave), controlling for group. Second, we regressed overall accuracy onto the average correct trial amplitude and average incorrect trial amplitude (as separate variables), controlling for group.

3.0.2. Sequential effects during practice

Reward-prediction errors, indexed by RewP, should be affected by accuracy on preceding trials and may affect accuracy on subsequent trials. To assess how the single-trial RewP was affected by the prior experience, we used a linear mixed-effects regression with fixed effects of Group (game/sterile), Trial number, Current Accuracy (response accuracy on the current trial, correct or incorrect), Previous Accuracy (defined as response accuracy for the most recent stimulus from the same category) and their interactions. The model included random effects of Participant, Participant × Trial, Stimulus Category, and Participant × Stimulus Category. The dependent variable was the single-trial RewP on the current trial. The first instance of each category was excluded from this model, because Previous Accuracy is undefined for those trials.

To measure how the RewP predicts future behavior, we used logistic mixed-effects regression with fixed effects of Group (game/sterile), Current Accuracy, the current single-trial RewP, Trial number, and their interactions. The model included random effects of Participant, Participant × Trial, Stimulus Category, and Participant × Stimulus Category. The dependent variable was whether or not participants

switched their response the next time they saw a stimulus from the same category. The last instance of each category was excluded from this model, because the dependent variable of response switching is undefined for those trials.

3.0.3. Assessment of long-term learning

First, to measure long-term learning and the effect of the game manipulation, our analysis focused on delayed post-test performance with a 2 (Group: Game/Sterile) × 2 (Post-test Condition: Retention/Transfer) × 2 (Stimulus Orientation: Old/Rotated) fixed-effects design. Given the positive skew in the number of correct responses on the post-test, we square-root transformed the dependent variable for analysis. Second, using the same 2 × 2 × 2 design, we also analyzed median response time for correctly classified stimuli on the post-test. Third, to test whether the RewP explained performance on the delayed retention and transfer tests, we fit linear models regressing average performance on the two post-tests (retention and transfer) onto the aggregate RewP, both alone and controlling for practice performance and group.

4. Results

4.1. Practice performance

Results of the analysis of improvement during practice are presented in Table 1. As explained in the Data Analysis section, this analysis predicted trial-by-trial classification accuracy during training as a function of group, trial number, and their interaction. There were main effects of trial number ($p < .001$) and group ($p = .045$), and a Group x Trial interaction ($p = .002$) as shown in Fig. 3. Although the probability of making a correct response increased across trials in both groups, this increase was much more substantial in the Sterile group.

Analysis of aggregate RewP showed that it was significantly positive and did not reliably differ between groups. Specifically, statistical analysis of the difference wave (Fig. 4A & B) showed that the average amplitude difference for the period between 245 and 345 ms after feedback onset at electrode Fz was significantly different from zero ($M = 4.55$ μV, $t(56) = 7.16$, $p < .001$). Furthermore, the RewP had a fronto-central scalp topography consistent with previous work (Proudfit, 2015), as seen in Fig. 4C. The average RewP tended to be larger for the game group ($M = 5.25$ μV) than for the sterile group ($M = 3.86$ μV), but this difference was not statistically significant, $t(56) = 1.09$, $p = .279$. Similarly, there was not a significant difference in self-reported engagement between the game group ($M = 3.63$) and the sterile group ($M = 3.57$), $t(56) = 0.37$, $p = .709$.

Regression analyses of overall performance across the practice session (Table 2, top) showed a negative relationship between the aggregate RewP and accuracy, $p = .012$. Controlling for the RewP, there remained a statistically significant advantage for the Sterile group, $p = .042$. To assess the separate contributions of positive and negative prediction errors to this relationship, an additional analysis was run that separated each participant's RewP into mean amplitude on correct trials and mean amplitude on incorrect trials (Table 2, bottom). This analysis revealed a negative relationship between correct-trial amplitude and accuracy, $p = .012$. Incorrect-trial amplitude was positively, but not significantly, related to accuracy, $p = .295$. As the simulations reported below demonstrate, the directions of these associations are consistent with effects of prior learning on RewP via the prediction side of the reward-prediction error (i.e., better category knowledge leads to accurate performance and smaller prediction errors, both positive and negative) and not with effects of RewP on subsequent learning via the reward side (i.e., stronger reward representations yield faster learning; this would predict the opposite correlations).

Table 1
Fixed and Random Effects for the model of performance during practice.

Random Effects *Group*	*Effect*	*Variance*	*SD*	*Corr*
Stimulus Category:Participant	Intercept	0.637	0.790	
Participant	Intercept	1.184	1.088	
	Trial	0.063	0.251	0.97
Stimulus Category	Intercept	0.135	0.367	

Fixed Effects *Effect*	*β*	χ^2	*df*	*p-value*
Intercept	−0.470	4.300	1	0.038
Group	0.312	4.018	1	0.045
Trial	0.183	23.693	1	< 0.001
Group × Trial	0.115	9.407	1	0.002

Note: All continuous variables were mean-centered and all categorical variables were contrast-coded prior to analysis. Group was coded as sterile = 1; game = −1.

4.2. Sequential effects on the single-trial RewP

To analyze the effects of past experience on the RewP, we modeled single-trial RewP as a function of group, trial number, accuracy on the current trial, and previous accuracy (for the most recent stimulus from the same category). This analysis, summarized in Table 3, revealed a significant main effect of the accuracy of the current trial ($p < .001$), with greater mean amplitude following correct feedback than incorrect feedback (see Fig. 5A). However, this effect was complicated by significant two-way interactions of Trial with both Current Accuracy ($p = .009$) and Previous Accuracy ($p = .020$), and a three-way interaction of Trial x Current Accuracy x Previous Accuracy ($p = .035$). As shown in Fig. 5B, when the response to the last trial of a stimulus from the same category was *incorrect*, the difference in amplitude between correct and incorrect responses remained relatively constant throughout the course of learning. However, when the response to the last trial of a stimulus from the same category was *correct,* the difference in single-trial RewP between correct and incorrect responses diminished over time. As the simulations reported below demonstrate, this complex pattern is well predicted by principles of reinforcement learning (see explanation in the Simulation section).

4.3. Retaining or changing responses

The analysis of RewP's impact on subsequent behavior used group, single-trial RewP, current accuracy, and trial number to predict whether the subject repeated or changed responses the next time a stimulus from the same category was presented. The results of this analysis are summarized in Table 4. As shown in Fig. 6A, participants partially learned the appropriate categories over the course of the practice phase, with the rate of changing one's response reducing over time, controlling for the other factors in the model (main effect of trial, $p < .001$). Additionally, the analysis showed that there was a negative relationship between identifying the stimulus correctly on the current trial and the log-odds of changing one's response for the next stimulus of the same category (main effect of current accuracy, $p < .001$). Importantly, there was also a positive effect of single-trial RewP on the current trial ($p = .010$), controlling for the other factors in the model. Thus, greater single-trial RewP on the current trial was associated with a higher probability of changing one's answer for the next stimulus of the same category. The simulations below show all three of these effects follow from principles of reinforcement learning. In particular, lesser RewP indicates stronger category knowledge (i.e., greater expectation of reward), which predicts a greater chance of selecting the same response in the future.

However, these main effects were complicated by Group × Trial × Current Accuracy ($p = .008$) and Group × Single-Trial RewP × Current Accuracy interactions ($p = .005$), shown in Fig. 6B and C. We do not have explanations for the group differences,

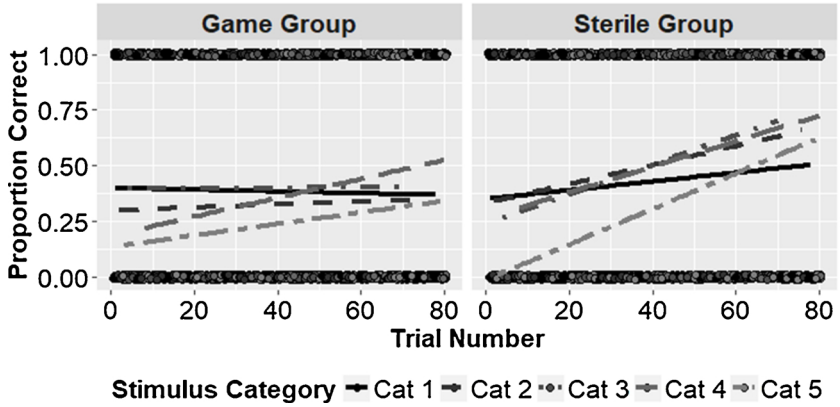

Fig. 3. The proportion of correct responses during practice (averaging across participants) as a function of group, stimulus category, and trial number. For ease of interpretation, these graphs plot fits of a linear model of response probability, rather than log-odds as in the statistical analyses reported in the text.

but the Simulation section below shows that a simple reinforcement-learning model explains the Trial × Current Accuracy and Single-Trial RewP × Current Accuracy interactions obtained by marginalizing over Group. Specifically, the effect of current accuracy is greater on later trials, because a correct response becomes more likely to indicate the participant has learned this category (as opposed to a lucky guess). Similarly, the effect of single-trial RewP is greater on correct trials than on incorrect trials, because single-trial RewP is more indicative of the strength of a participant's category knowledge in the former case.

4.4. Assessment of long-term learning

The first analysis of learning from training to testing predicted post-test performance as a function of training condition, post-test condition, and stimulus orientation. As shown in Fig. 7A, performance on the post-test was low overall, but the sterile group performed better on average (M = 10.54 correct out of 30 trials per post-test) than the game group (M = 6.46 out of 30). This difference was confirmed by a main effect of Group, p = .021 (see Table 5). There was also a significant Group by Test-Type interaction, p = .002, such that participants in the game group tended to perform better in the condition in which they trained, whereas the sterile group showed more comparable levels of performance on retention and transfer tests. None of the other effects were statistically significant (ps > .130).

The corresponding analysis of correct response time is summarized in Fig. 7B. There was considerable variability in the median response time for correct responses within groups and conditions, with the average median response time M = 2.52 s. The only statistically significant effect, however, was a main effect of group (p = .041), such that participants who practiced in the sterile group were faster with their correct responses (M = 2.16 s) than were participants in the game group (M = 2.88 s). There was no significant main effect of Test-Type

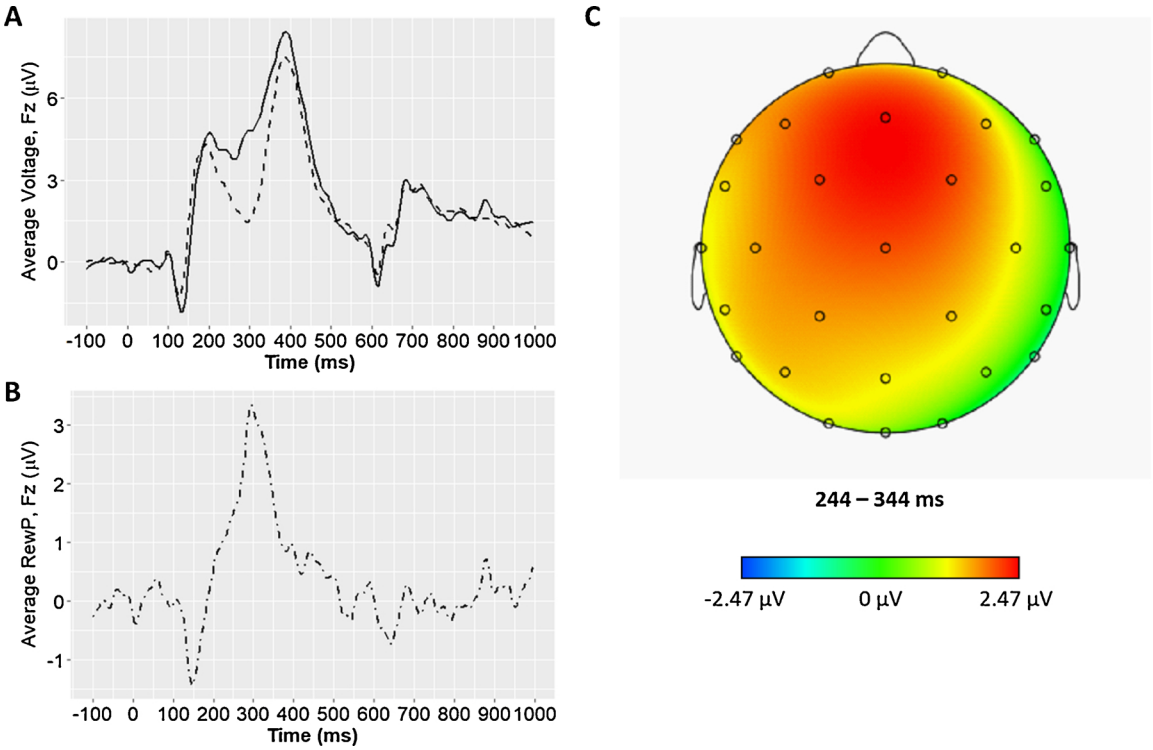

Fig. 4. (A) Average waveforms for correct (solid lines) and incorrect responses (dashed lines) during practice. Time = 0 indicates the onset of feedback. (B) The difference wave (Correct – Incorrect) showing the grand-average aggregate RewP during practice. (C) The scalp distribution for the aggregate RewP was maximal fronto-centrally, consistent with previous work.

Table 2
Models of Practice Performance.

Using RewP Difference Measure *Effect*	*β*	*t*	*df*	*p-value*
Intercept	5.868	24.83	55	< 0.001
Group	0.359	2.079	55	0.042
RewP	−0.093	−2.591	55	0.012

Separating Correct and Incorrect Trials *Effect*	*β*	*t*	*df*	*p-value*
Intercept	5.940	23.979	54	< 0.001
Group	0.387	2.212	54	0.031
Correct-Trial Amplitude	−0.094	−2.614	54	0.012
Incorrect-Trial Amplitude	0.056	1.057	54	0.295

Note that the dependent variable of number of correct responses was square-root transformed for this analysis to create a more normal distribution of residuals. Further, for the group variable, sterile = 1 and game = −1, so a positive beta indicates better performance for the sterile practice group.

Table 3
Fixed and Random Effects for the analysis of sequential effects on the single-trial RewP.

Random Effects *Group*	*Effect*	*Variance*	*SD*	*Corr*
Stimulus Category x Participant	Intercept	0.477	0.690	
Participant	Intercept	15.418	3.927	
	Trial	0.094	0.307	0.03
Stimulus Category	Intercept	0.128	0.358	

Fixed Effects *Effects*	*β*	χ^2	*df*	*p-value*
Intercept	3.745	43.040	1	< 0.001
Group	0.779	2.020	1	0.155
Trial	−0.102	1.237	1	0.266
Current Accuracy (CA)	2.152	129.114	1	< 0.001
Previous Accuracy (PA)	0.145	0.607	1	0.436
Trial x Group	−0.114	1.575	1	0.209
Group x CA	−0.006	0.001	1	0.977
Trial x CA	−0.215	6.841	1	0.009
Group x PA	−0.148	0.637	1	0.425
Trial x PA	−0.190	5.404	1	0.020
CA x PA	−0.158	0.732	1	0.392
Group x Trial x CA	−0.091	1.228	1	0.268
Group x Trial x PA	−0.003	0.001	1	0.971
Group x CA x PA	−0.021	0.013	1	0.909
Trial x CA x PA	−0.173	4.448	1	0.035
Group x Trial x PA x CA	0.080	0.945	1	0.331

Note: All continuous variables were mean-centered and categorical variables were contrast coded prior to analysis. Group as coded as sterile = 1; game = −1. Current Accuracy was coded as correct = 1; incorrect = −1. Previous Accuracy refers to the most recent stimulus from the same category and was coded as correct = 1; incorrect = −1. Stim = Stimulus Category.

(p = .656), Orientation (p = .337), nor any statistically significant interactions (ps > .312).

To assess the dependence of post-test performance on RewP and the game manipulation, we regressed the post-test accuracy onto practice accuracy, group, and the aggregate RewP. Controlling for the other factors, there was not a reliable effect of group, $t(54) = 1.33$, $p = .190$, nor a reliable effect of RewP, $t(54) = 0.70$, p = .487. However, there was a statistically significant positive relationship between practice accuracy and post-test accuracy, $t(54) = 5.48$, p < .001. Importantly, we also regressed post-test accuracy onto aggregate RewP alone, finding no significant relationship, $t(56) = -0.86$, p = .395. Thus, group had a relationship to post-test accuracy (as shown by the analysis above, in Table 5) that was mediated by performance during practice, whereas the RewP was not significantly related to post-test accuracy. These results indicate RewP did not impact retention and did not play a role in the observed group difference in retention, and that the group difference in retention performance is explained by the group difference in training performance.

4.5. Summary of results

In summary, the present experiment demonstrates a robust RewP that is closely tied to behavioral adaptation during practice. At the session level, aggregate RewP was negatively related to subjects' overall training performance. At the trial level, there are complex relationships both between subjects' past experience and performance and their current trial-level RewP, and between trial-level RewP and subsequent adaptation of behavior. Despite these strong findings from the training session, we saw no relationship between RewP and retention performance.

We see two main implications for these results. First, the detailed relationships between RewP and behavior during training (at both trial and session levels) provide valuable targets for modeling within the reinforcement learning framework. We take up this challenge in the next section. Second, the decoupling of RewP from retention performance suggests important limitations to the contribution of adaptation through reward prediction errors to long-term learning. We consider this latter implication in the Discussion section.

5. Simulation of trial-level dynamics

We simulated a simple Reinforcement Learning model to test whether it can explain the pattern of results found above. The model assumes that the participant learns associations between the features defining the categories and the response options. Recall the participants were explicitly instructed on which feature of the stimuli to use for categorization (the shape of the central component), so they had only to learn to associate these feature values to the categories. For each feature value f and response, the participant learns a value $Q(f, r)$ representing the expected reward for choosing r when the stimulus has value f (Watkins & Dayan, 1992). Thus the optimal values, for a participant who has learned perfectly, would be $Q(i, i) = 1$ for $1 \leq i \leq 5$ and $Q(i, j) = 0$ for $i \neq j$.

The model assumes responses are chosen probabilistically by a softmax rule:

$$\Pr[r|f] \propto e^{Q(f,r)/T} \quad (1)$$

where T is a choice temperature parameter.

Learning is driven by prediction error,

$$\delta = R - Q(f, r), \quad (2)$$

where, r is the participant's chosen response, and R is the feedback received, coded as 1 for correct and − 1 for incorrect. The value for the chosen response is updated according to

$$\Delta Q(f, r) = \alpha\delta, \quad (3)$$

where, α is a learning rate parameter.

Finally, we assume that the observed single-trial RewP on each trial is a noisy measure of the prediction error,

$$RewP \sim \mathcal{N}(\delta, \sigma^2_{EEG}), \quad (4)$$

where, σ^2_{EEG} is a parameter determining the error variance of this neural measure.

The model was simulated 10,000 times, each time with the same number of trials that participants experienced during practice, with parameters $\alpha = 0.1$, $T = 0.1$, and $\sigma^2_{EEG} = 0.5$. These values correspond to relatively slow learning and deterministic responding, in line with the general trends in the data. The goal was to demonstrate how the qualitative patterns in the data follow from RL theory, and thus no effort was made for quantitative parameter tuning. The simulated data from the 10,000 virtual participants were analyzed identically to the real

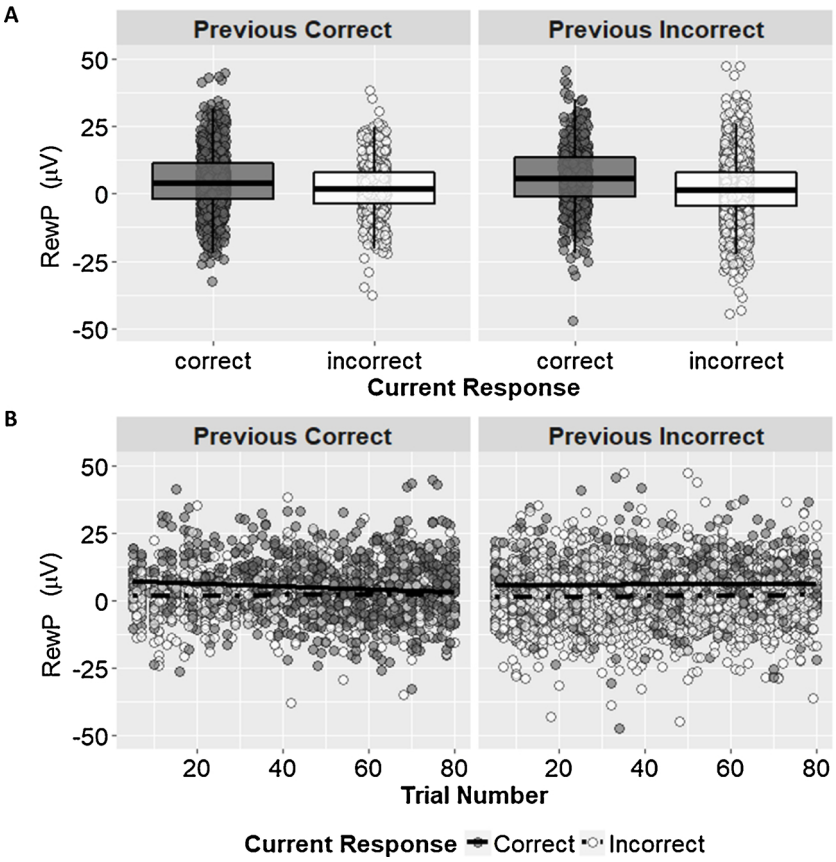

Fig. 5. (A) Single-trial RewP as a function of accuracy on the current trial and the accuracy of the previous response to a stimulus of the same category. (B) Single-trial RewP as a function of accuracy on the current trial, accuracy of the previous response to a stimulus of the same category, and the trial number.

Table 4
Fixed and Random Effects for the analysis of participants' switching behavior.

Random Effects *Group*	*Effect*	*Variance*	*SD*	*Corr*
Stimulus Category x Participant	Intercept	0.207	0.455	
Participant	Intercept	0.774	0.880	
	Trial	0.025	0.158	0.85
Stimulus Category	Intercept	0.006	0.078	

Fixed Effects *Effects*	β	χ^2	*df*	*p-value*
Intercept	0.336	6.516	1	0.011
Group	−0.189	2.216	1	0.137
Trial	−0.147	23.749	1	< 0.001
Single-trial RewP	0.011	6.618	1	0.010
Current Accuracy (CA)	−0.642	191.168	1	< 0.001
Group x Trial	−0.020	0.439	1	0.508
Group x Single-Trial RewP	−0.003	0.665	1	0.415
Trial x Single-Trial RewP	−0.001	0.488	1	0.485
Group x CA	−0.222	22.988	1	< 0.001
Trial x CA	−0.057	6.461	1	0.011
Single-Trial RewP x CA	0.001	0.059	1	0.808
Group x Trial x Single-Trial RewP	−0.004	3.645	1	0.056
Group x Trial x CA	−0.058	6.939	1	0.008
Group x Single-Trial RewP x CA	0.011	7.784	1	0.005
Trial x Single-Trial RewP x CA	0.002	1.054	1	0.305
Group x Trial x Single-Trial RewP x CA	−0.002	0.802	1	0.371

Note: All continuous variables were mean-centered and categorical variables were contrast coded prior to analysis. Group was coded as sterile = 1; game = −1. Current Accuracy was coded as correct = 1; incorrect = −1. Stim = Stimulus Category.

data from the experiment with the exception of the game manipulation, which was omitted. Results are shown in Fig. 8.

5.1. Relationship between RewP and performance

Across simulated subjects, the model predicts a negative association between aggregate RewP and number of correct responses, as shown in Fig. 8a. This prediction arises for the reason given earlier: On runs when the model learns more accurate reward values, $Q(f, r)$, it generates more correct responses as well as smaller prediction errors. Smaller prediction errors entail single-trial RewP is lower on correct trials and higher on incorrect trials, resulting in a smaller aggregate RewP. This predicted association also decomposes in the manner found in the human data. That is, accuracy correlates negatively with mean single-trial RewP on correct trials (via the accuracy of values learned for correct responses, $Q(i, i)$), and it correlates positively with mean single-trial RewP on incorrect trials (via the accuracy of values learned for incorrect responses, $Q(i, j)$, $i \neq j$).

5.2. Sequential effects on the single-trial RewP

In the analysis of sequential effects on the single-trial RewP, the model correctly reproduces all aspects of the interactions among accuracy on the current trial, accuracy on the previous instance of the same category, and trial number, as shown in Fig. 8b (compare to empirical results in Fig. 5b).

First, there is a main effect of current accuracy (compare black curves to red curves), which the model explains as a direct effect of reward: The difference in reward on correct trials ($R = 1$) and on incorrect trials ($R = -1$) translates directly to a difference in prediction error (δ) and thence to the single-trial RewP.

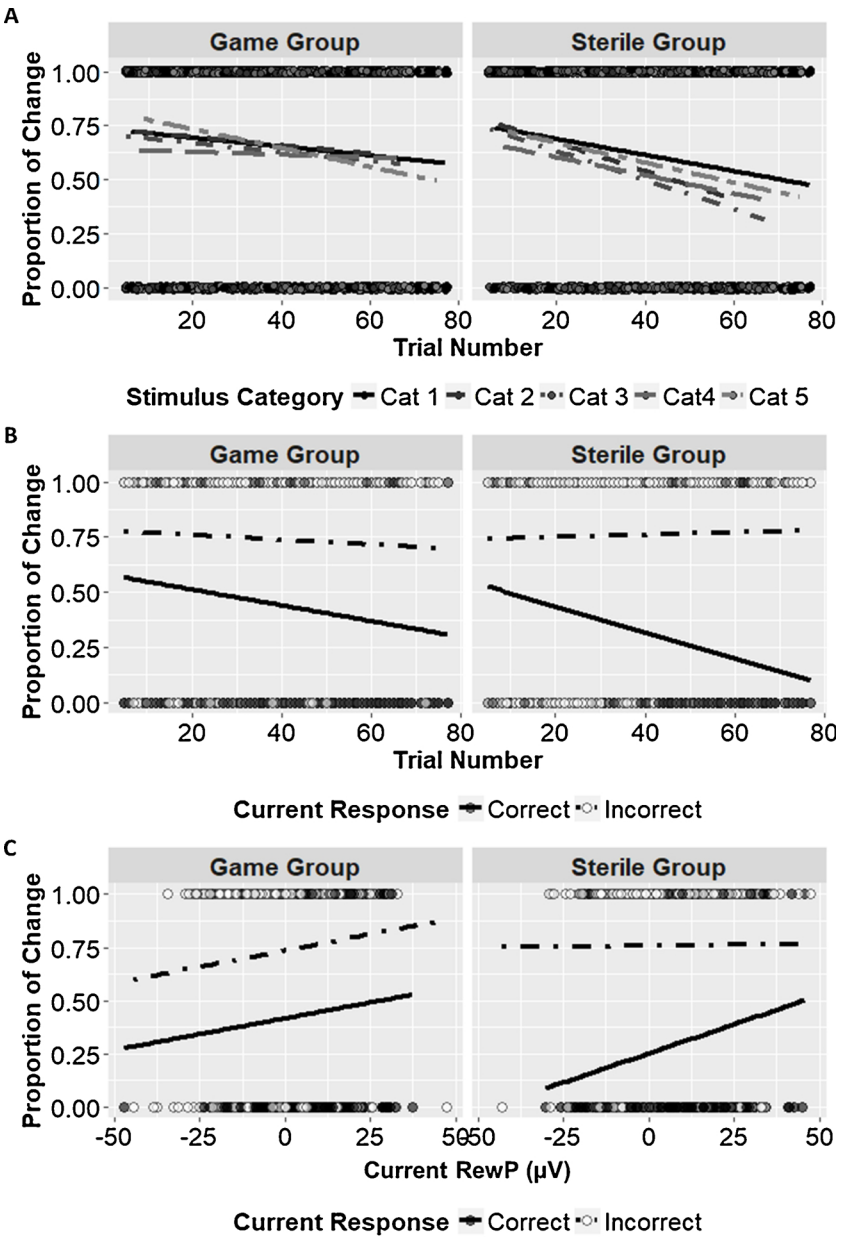

Fig. 6. (A) The proportion of changed responses on the next trial on which a stimulus from the same category is presented (averaging across participants) as a function of the stimulus category and the current trial number. Note that '1′ on this scale indicates that participants changed their response the next time they saw a stimulus from the same category, relative to the current trial. Conversely, a '0′ indicates that participants made the same response the next time they saw a stimulus from the same category as they did on the current trial. (B) The proportion of changed responses as a function of current response accuracy and current trial number. (C) The proportion of changed responses as a function of current response accuracy and current single-trial RewP. For ease of interpretation, these graphs plot fits of a linear model of change probability, rather than log-odds as in the statistical analyses reported in the text.

Second, when the previous response to the same category was correct (solid curves), the effect of current accuracy decreases with trial number. The model's explanation is based in its expected reward, $Q(f, r)$. Put simply, as the model learns the correct Q-values, prediction errors go away, and thus the initial difference in the single-trial RewP on correct versus incorrect trials closes over time. A more detailed explanation is as follows. Early in learning, the Q-values are approximately the same for all responses (correct and incorrect). Thus, the difference in δ between correct and incorrect trials is nearly the same as the corresponding difference in rewards (1 vs. − 1). As learning progresses, the Q-value for the correct response to each category increases toward 1, and thus the prediction error declines toward 0 (solid black line). Likewise, the Q-values for incorrect responses decline toward − 1, and thus the negative prediction error that arises when the model happens to pick an incorrect response increases toward 0 (solid red line). (Note that, because the model assumes stochastic, exploratory responding, it will occasionally select an incorrect response even after it has learned that response to be inferior.) The first of these effects is stronger than the second, because the model learns more about the correct response for each category than it does about any particular incorrect response.

Third, the interaction just described between trial number and current accuracy is mostly absent when the previous response to the same category was incorrect (dashed curves). This is because an incorrect previous response indicates the model has likely not yet learned the correct Q-values for this category. This in turn implies the model's expectation is roughly the same for the correct and incorrect responses, which in turn implies that the large gap in prediction error between correct and incorrect trials is maintained.

In summary, the model predicts a large effect of current feedback on prediction error, except when it has already learned what rewards to expect for the correct and incorrect responses. These expectations are developed primarily later during practice and when recent responses to the same category have been correct. Thus, our reinforcement learning model explains all the details of the 3-way interaction observed in the data of current accuracy, previous accuracy, and trial number on single-trial RewP.

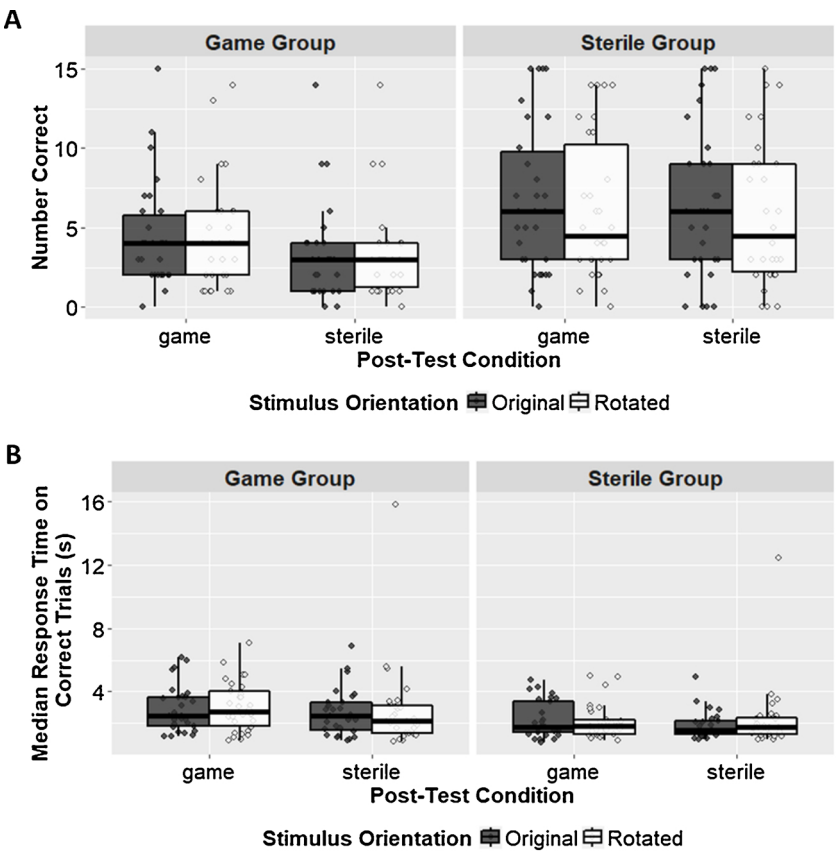

Fig. 7. (A) Post-test accuracy as a function of group, post-test condition, and stimulus orientation during the post-test. (B) Median response time for correct responses on the post-test as a function of group, post-test condition, and stimulus orientation. These boxplots show the median and the interquartile range (IQR). The whiskers go out to last data point within 1.5*IQR. Note that there were two outlying participants in both the game group and the sterile group for response time, but the statistical significance of the results was not changed whether these participants were excluded or not.

Table 5
Fixed and Random Effects for the Model of Post-test Performance.

Random Effects			
Group	*Effect*	*Variance*	*SD*
Orientation x Participant	Intercept	< 0.001	< 0.001
Test-Type x Participant	Intercept	0.080	0.282
Participant	Intercept	0.131	0.363

Fixed Effects				
Effects	*β*	χ^2	*df*	*p-value*
Intercept	2.047	360.882	1	< .001
Group	0.248	5.322	1	.021
Test-Type	0.053	2.290	1	.130
Orientation	−0.024	1.060	1	.303
Group x Test-Type	−0.106	9.198	1	.002
Group x Orientation	−0.027	1.347	1	.246
Test-Type x Orientation	−0.004	0.026	1	.872
Group x Test-Type x Orientation	0.013	0.305	1	.581

Note: The dependent variable of number of correct responses was square-root transformed for this analysis to create a more normal distribution of residuals. Group was coded as sterile = 1; game = −1. Orientation was coded as rotated = 1; original = −1. Test-type was coded as retention = 1; transfer = −1.

5.3. Retaining or changing responses

In the analysis of response switching, the model correctly reproduces the interaction between current accuracy and trial number, as shown in Fig. 8c (compare to empirical results in Fig. 6b). The immediate result of learning from current feedback leads a correct response to be more likely in the future and an incorrect response to be less likely. Additionally, correct responses tend to be high-probability, non-exploratory choices under the softmax choice rule (and hence are likely to be repeated), whereas incorrect responses tend to be low-probability choices that are unlikely to be repeated. These two facts explain the main effect of current accuracy (more switching following an incorrect response). Moreover, because of accumulated learning across trials, a correct response later in learning is more likely to reflect a high learned Q-value for that response, and hence foretells an even lower switching probability than it does early in learning (negative slope of black curve). Likewise, an incorrect response later in learning is more likely to reflect random exploration of a response that has low learned value and thus is even less likely to be repeated (positive slope of red curve). As above, the second effect is weaker than the first because there is more learning about the correct response than about any one incorrect response.

The model also correctly reproduces the positive relationship between current single-trial RewP and switching probability, as shown in Fig. 8d (compare to empirical results in Fig. 6c). As above, the explanation of this effect is based on learning the correct Q-values across trials. First, consider trials on which the current response is correct. The reward is the same on all of these trials, so a larger prediction error indicates a lower expectation of reward. That is, the correct response is not yet well-learned and has a relatively low Q-value. Thus, it is less likely to be repeated (higher switching probability). Conversely, a smaller prediction error on a correct trial indicates a greater reward expectation, meaning a well-learned response with a larger Q-value. In this case the response is more likely to be repeated (lower switching probability). Thus, prediction error, and hence single-trial RewP, is positively correlated with switching probability following correct responses (positive slope of black curve). The same argument applies following incorrect responses, although once again the effect is weaker: A higher single-trial RewP indicates a lower expectation of reward, meaning an incorrect response that has already been learned to be wrong (and was selected only because of the randomness in the decision process), which is especially unlikely to be repeated.

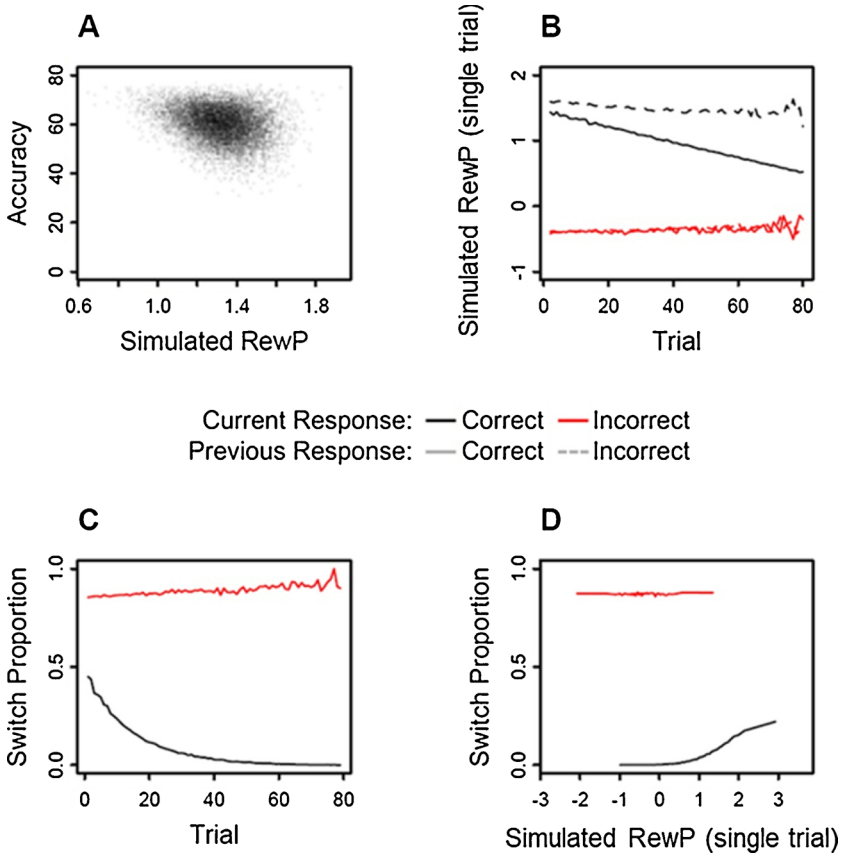

Fig. 8. Predictions of simple reinforcement-learning model. A: Scatterplot of aggregate RewP versus practice performance across simulated participants. B: Mean single-trial RewP as a function of trial, accuracy of the current response (black = correct, red = incorrect), and accuracy of the response to the previous instance of the current category (solid = correct, dashed = incorrect). Compare to empirical results in Fig. 5b. C: Probability of switching responses the next time a stimulus from the current category is presented, as a function of trial and accuracy of the current response. Compare to empirical results in Fig. 6b. D: Probability of switching responses the next time a stimulus from the current category is presented, as a function of current single-trial RewP and accuracy of the current response. Compare to empirical results in Fig. 6c. (For interpretation of the references to colour in this figure legend, the reader is referred to the web version of this article).

5.4. Variability in reward processing

Finally, we implemented a variant of the model in which the magnitude of reward varied from trial to trial:

$$R \sim \begin{cases} \mathcal{N}(1, \sigma^2_{\text{Rew}}) & \text{correct response} \\ \mathcal{N}(-1, \sigma^2_{\text{Rew}}) & \text{incorrect response} \end{cases} \quad (5)$$

Here σ^2_{Rew} is a parameter determining the amount of random variability in reward representation. One could also introduce variability at the participant level, but trial-level variability turns out to generate enough participant-level variability (via sampling error) to illustrate the effects of both. The purpose of this model variant was to illustrate how effects involving aggregate and single-trial RewP would be impacted by variation in the reward side of the prediction-error equation (Eq. 2) as opposed to variation in the prediction side.

As shown in Fig. 9, with $\sigma^2_{\text{Rew}} = 2$, the primary impact of this reward variability is to reverse the correlations between aggregate RewP and performance (Fig. 9a) and between single-trial RewP and switch probability (Fig. 9d). The correlation at the participant level arises because a larger aggregate RewP now tends to indicate a stronger average reward signal (R more positive for correct feedback and more negative for incorrect feedback), which would speed learning. Similarly, the correlation at the trial level arises because a higher single-trial RewP now tends to indicate a greater subjective reward value on that trial, which will result in the chosen response being reinforced more (if correct) or punished less (if incorrect). Thus, the chosen response is more likely to be repeated when single-trial RewP is higher, if we assume more trial-level variability in rewards. Additionally, the negative correlation between single-trial RewP and switch probability is stronger for incorrect responses (red curve), because in that case the countervailing effect from variation in expectations is weak, as explained above in the discussion of Fig. 8d.

The contrasting predictions of the RL model with and without random variation in reward strength illustrate the complementary contributions of reward and prediction to reward-prediction errors. To the extent that variation in the prediction error (as indexed by aggregate RewP across participants or single-trial RewP across trials) reflects variation in the reward signal, aggregate RewP will correlate positively with performance and single-trial RewP will correlate positively with response repetition. To the extent that variation in prediction error reflects variation in the participant's expectation of reward, these correlations will be reversed. The fact that the data support the latter prediction suggests that participant-level and trial-level variation in the strength of reward processing or of subjective reward are negligible in this task. This conclusion is congruent with the finding that the game manipulation had no impact on RewP.

6. Discussion

The present results provide several conclusions regarding the relationship between reinforcement learning, electrophysiology, motivation, and long-term learning. Specifically, our findings strengthen the connection of RewP to reward-prediction error, and they also show limitations to the functional role of this signal. Our formal modeling approach affords a separation of hypothesized effects involving reward-prediction error into (a) reward-side effects rooted in variation in reward representation after feedback and (b) prediction-side effects rooted in participants' expectation of reward based on the stimulus and chosen response on each trial. Testing of these predictions was aided by our approach of analyzing trial-by-trial variation and sequential effects in the single-trial RewP. Additionally, the manipulation of task framing (game vs. sterile) and the inclusion of a one-week retention test enabled assessment of reward-prediction error's dependence on motivational factors and dissociation of its role in short-term versus long-term learning.

The simulation work reported here demonstrates that reinforcement

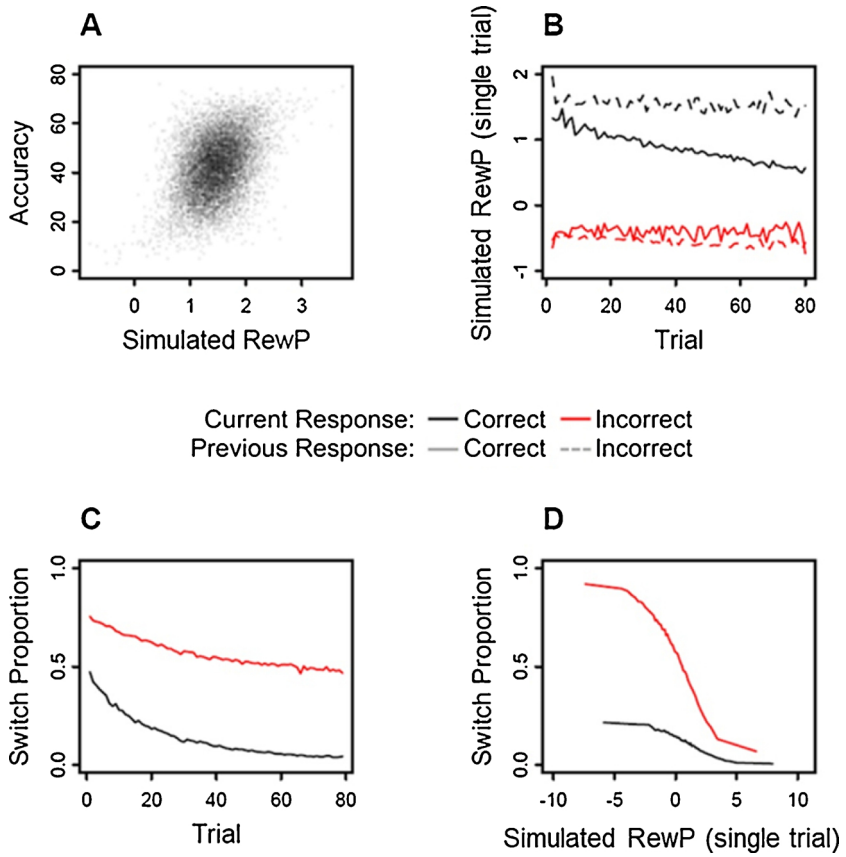

Fig. 9. Predictions of reinforcement-learning model with random variability in subjective reward strength. See Fig. 8 caption for details.

learning theory makes a number of counterintuitive predictions when participants' reward predictions are allowed to evolve freely over the course of learning, as in naturalistic settings. The human data supported all the predictions of this model, including its complex interactions and sequential effects due to dynamics of learned expectations and the simple reward-side effect of greater prediction error following correct versus incorrect feedback. However, there was no evidence for more complex reward-side effects that arise from augmenting the basic model with variation in subjective reward. These findings also accord with the non-significant results of the game manipulation on the RewP. Although the direction was opposite that predicted (Lohse et al., 2016), the game manipulation had a reliable impact on learning. The aggregate RewP did not explain this group difference, nor did the RewP explain participant-level variation in retention performance, whether controlling for acquisition performance or not. In conclusion, reinforcement learning theory and the identification of single-trial RewP with reward-prediction error provide a remarkably accurate account of learning or adaptation in the short term (trial and session levels), but they appear to be less relevant to explaining broader phenomena such as task framing effects and long-term learning.

6.1. Trial-level dynamics during training

The single-trial RewP exhibited a complex relationship with the valence of the current feedback (correct vs. incorrect), as well as previous responses, future responses, and trial number. This detailed pattern is remarkably well-fit by a simple reinforcement learning model. This agreement provides evidence for the popular hypothesis that RewP reflects neural signals of reward-prediction error (Proudfit, 2015; Sambrook & Goslin, 2015). Further, this agreement illustrates the utility of the trial-level approach adopted here, which decomposes the aggregate (difference wave) RewP into an EEG amplitude reflecting reward-prediction error on each individual trial.

In the analysis of sequential effects on the single-trial RewP, we found that not only were amplitudes in the 245–345 ms post-feedback window higher for correct versus incorrect feedback, but the magnitude of this deviation also depended on the accuracy of the previous response to the same category and on the current trial number (Fig. 5b). When the previous response to the current category was incorrect (suggesting this category was not yet learned) the average difference in single-trial RewP between correct and incorrect responses remained relatively constant across trials. However, when the previous response was correct, this difference reduced over time. Our explanation of these results, in terms of the learning dynamics of reward expectation or response confidence, also explains similar results obtained by previous researchers. For example, in their virtual throwing study, Frömer et al. (2016) observed larger single-trial RewPs when participants hit the target less frequently, suggesting that the single-trial RewP changed as participants acquired the correct movement pattern because they originally had lower expectations of success. Moreover, this effect was more pronounced following trials where participants hit the target rather than missed it, suggesting that high-performing subjects were better at discriminating successful from unsuccessful throws prior to feedback delivery.

The dynamic sequential effect on the single-trial RewP in the present study is also explained by the simple reinforcement learning model (Fig. 8b), in terms of changes in participants' confidence in correct and incorrect responses over time. More formally, reward-prediction error is assumed to equal $R - P$, where the reward R depends on the valence of the feedback (correct or incorrect) and the prediction P is equal to $Q(r, f)$, the participant's subjective expected value for the chosen response (r) given the current category-defining feature (f). The main effect of current feedback operates via the reward side of this expression in a straightforward manner. The effects of trial number and

previous response operate via the prediction side, in a more nuanced way, arising from the differential dynamics of $Q(r, f)$ for correct versus error responses over the course of learning.

The relationship between the single-trial RewP and response confidence was also shown in the analysis relating single-trial RewP to future response choices. Controlling for response accuracy, higher amplitude was associated with a greater tendency for participants to change their response the next time they saw a stimulus from the same category. This result is consistent with Philiastides et al. (2010), who observed that a larger RewP-like ERP component was associated with an increased likelihood of a participant changing their response after positive feedback. Under the conceptualization of the single-trial RewP as prediction error, $R - Q(r, f)$, this result implies that most of the variation in this signal lies in the prediction side (again controlling for response accuracy). Conversely, lower single-trial RewPs reflect well-learned responses that the participant was confident would be correct (i.e., $Q(r, f)$ close to 1) and that hence were more likely to be repeated. A similar explanation applies to incorrect responses. The model explains this effect, including the fact that the relationship is stronger for correct responses than for incorrect responses. Moreover, additional simulations showed that the relationship reverses if the strength of reward representations is assumed to vary across trials (i.e., if variation in the single-trial RewP comes from the R side of the difference computation). This bolsters the conclusion that variation in reward-prediction errors is driven by objective feedback and by variation in subjective expectations (e.g., response confidence). The finding that reward processing (i.e., the mapping from objective feedback to subjective reward) is relatively stable is also consonant with the finding that the game manipulation had no impact on RewP.

6.2. Long-term learning outcomes and aggregate analyses

Counter to our predictions and prior results (Lohse et al., 2016), analysis of the post-test showed that training in the sterile condition was better for long-term retention and transfer than training in the game condition. Participants in the sterile group were also faster, on average, in making correct responses than participants who trained in the game condition. Interestingly, there was no significant difference in self-reported engagement between the game group and the sterile group. Thus, our goal to manipulate intrinsic motivation was not successful and we failed to induce greater engagement in the game group. Instead, the additional features (more complicated graphics, sound, and narrative) might have distracted participants from relevant information in the task and created extraneous cognitive load (Sweller, 2010), leading to better learning for the less distracted “sterile” group.

Overall, there was a reliable RewP evident in the correct minus incorrect difference wave, and this RewP exhibited a fronto-central scalp distribution. Although the amplitude of the difference wave RewP was higher in the game group than in the sterile group, this contrast was not statistically significant.

Finally, regression analyses did not show any statistically significant relationships between engagement or RewP and post-test performance, controlling for group and performance during practice. Thus, although the sterile group learned the stimulus categories much better than the game group, none of the instrumental variables we collected explained the learning benefit. Overall, aggregate analyses showed that participants exhibited neural activity reflecting reward-prediction errors (RewP). However, this neural activity was not modulated by group assignment nor did individual differences in aggregate neural measures explain individual differences in learning.

This lack of a relationship between the RewP and long-term retention and transfer of the knowledge acquired during practice is an important and interesting finding. Although this null result must be treated with caution, it is worthwhile to note that there were strong differences in learning between the two groups (the sterile group showed superior retention and transfer), but no statistical difference in the RewP between groups, nor were individual differences in post-test performance explained by individual differences in the RewP observed during practice. Thus, there was no evidence for a direct relationship between the RewP and long-term retention or transfer in the present study, even though the single-trial RewP was highly related to trial-by-trial adaptations in behavior, consistent with reinforcement learning principles. This dissociation is also an important consideration for EEG studies that look for mechanistic accounts of learning in a single session of practice (Bellebaum & Daum, 2008; Reinhart & Woodman, 2014; van der Helden et al., 2009). Models need to account for additional mechanisms to explain long-term retention and transfer of information beyond short-term adjustments in behavior (Kantak & Winstein, 2012).

7. Conclusions

Consistent with past research, our experiment provided strong evidence that the ERP signal underlying the RewP, quantified here as single-trial RewP, is sensitive to reward-prediction errors. In the current experiment, we used mixed-effect regressions that allowed us to explore the trial-by-trial relationship between the single-trial RewP and behavioral adaptations, and to compare these with the predictions of reinforcement learning theory. These data showed a complex pattern of dependencies between single-trial RewP and prior choice behavior, current feedback, and future choice behavior, extending previous studies that adopted a trial-level approach. These patterns were matched in remarkable detail by the reinforcement learning model. These trial-level empirical and modeling results go beyond previous aggregate approaches in demonstrating the intricate and dynamic role of single-trial RewP in learning from feedback.

Although single-trial RewP was linked with behavioral dynamics during practice, aggregate RewP amplitude averaged across trials did not explain long-term changes in behavior, as measured by delayed retention and transfer tests. Indeed, few EEG studies examining the relationship between RewP and behavior have assessed it during delayed post-tests (i.e., long-term learning). Those that have find RewP amplitude predicts behavior only during practice (Grand et al., 2017) as we did in the current study. These findings are also consistent with behavioral data showing that short-term performance is often uncorrelated (or even negatively correlated) with long-term learning (Kantak & Winstein, 2012; Pashler & Baylis, 1991). We believe these findings speak to the complexity of human learning: Long-term retention and transfer no doubt rely on local adaptation mechanisms of reinforcement learning, but they also rely on higher-level processes such as memory consolidation, feature selection (Goldstone & Steyvers, 2001), attention (Kruschke, 2001), and other strategic or control processes (Rieskamp & Otto, 2006). Computational models of learning should strive to incorporate the psychological mechanisms that operate at both time-scales of learning: changes from trial to trial related to reward-prediction errors and changes over weeks or months.

Declaration of Competing Interest

The authors received no funding specifically to pursue this work and have no conflicts of interest to declare.

Acknowledgments

MJ was supported by grant FA9550-14-1-0318 from the Air Force Office of Scientific Research.

Appendix A. Supplementary data

Supplementary material related to this article can be found, in the online version, at https://doi.org/10.1016/j.biopsycho.2019.107775.

References

Abe, M., Schambra, H., Wassermann, E. M., Luckenbaugh, D. A., Schweighofer, N., & Cohen, L. G. (2011). Reward improves long-term retention of a motor memory through induction of offline memory gains. *Current Biology, 21*, 557–562.

Bates, D., Maechler, M., Bolker, B., & Walker, S. (2015). *lme4: Linear mixed-effects models using Eigen and S4.* R package version 1.1–72014.

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

Cashaback, J. G. A., McGregor, H. R., Mohatarem, A., & Gribble, P. L. (2017). Dissociating error-based and reinforcement-based loss functions during sensorimotor learning. *PLoS Computational Biology, 13*, e1005623. https://doi.org/10.1371/journal.pcbi.1005623.

Chase, H. W., Swainson, R., Durham, L., Benham, L., & Cools, R. (2011). Feedback-related negativity codes prediction error but not behavioral adjustment during probabilistic reversal learning. *Journal of Cognitive Neuroscience, 23*, 936–946. https://doi.org/10.1162/jocn.2010.21456.

Clayson, P. E., Baldwin, S. A., & Larson, M. J. (2013). How does noise affect amplitude and latency measurement of event-related potentials (ERPs)? A methodological critique and simulation study. *Psychophysiology, 50*, 174–186.

Collins, A. G. E., & Frank, M. J. (2018). Within- and across-trial dynamics of human EEG reveal cooperative interplay between reinforcement learning and working memory. *Proceedings of the National Academy of Sciences, 115*, 2502–2507. https://doi.org/10.1073/pnas.1720963115.

Fischer, A. G., & Ullsperger, M. (2013). Real and fictive outcomes are processed differently but converge on a common adaptive mechanism. *Neuron, 79*, 1243–1255. https://doi.org/10.1016/j.neuron.2013.07.006.

Frömer, R., Stürmer, B., & Sommer, W. (2016). The better, the bigger: The effect of graded positive performance feedback on the reward positivity. *Biological Psychology, 114*, 61–68. https://doi.org/10.1016/j.biopsycho.2015.12.011.

Gauthier, I., & Tarr, M. J. (1997). Becoming a "Greeble" expert: Exploring mechanisms for face recognition. *Vision Research, 37*, 1673–1682.

Goldstone, R. L., & Steyvers, M. (2001). The sensitization and differentiation of dimensions during category learning. *Journal of Experimental Psychology General, 130*, 116–139.

Grand, K. F., Daou, M., Lohse, K. R., & Miller, M. W. (2017). Investigating the mechanisms underlying the effects of an incidental choice on motor learning. *Journal of Motor Learning and Development, 5*, 207–226.

Holroyd, C., & Coles, M. G. H. (2002). The neural basis of human error processing: Reinforcement learning, dopamine, and the error-related negativity. *Psychological Review, 109*, 679–709. https://doi.org/10.1037//0033-295X.109.4.679.

Holroyd, C., & Krigolson, O. E. (2007). Reward prediction error signals associated with a modified time estimation task. *Psychophysiology, 44*, 913–917. https://doi.org/10.1111/j.1469-8986.2007.00561.x.

Jones, M., Curran, T., Mozer, M. C., & Wilder, M. H. (2013). Sequential effects in response time reveal learning mechanisms and event representations. *Psychological Review, 120*, 628–666.

Jones, M., Love, B. C., & Maddox, W. T. (2006). Recency effects as a window to generalization: Separating decisional and perceptual sequential effects in category learning. *Journal of Experimental Psychology Learning, Memory, and Cognition, 32*, 316–332.

Kantak, S. S., & Winstein, C. J. (2012). Learning–Performance distinction and memory processes for motor skills: A focused review and perspective. *Behavioural Brain Research, 228*, 219–231. https://doi.org/10.1016/j.bbr.2011.11.028.

Kreidler, S. M., Muller, K. E., Grunwald, G. K., Ringham, B. M., Coker-Dukowitz, Z. T., Sakhadeo, U. R., & Glueck, D. H. (2013). GLIMMPSE: Online power computation for linear models with and without a baseline covariate. *Journal of Statistical Software, 54*, i10.

Krigolson, O. E. (2018). Event-related brain potentials and the study of reward processing: Methodological considerations. *International Journal of Psychophysiology, 127*, 62–72.

Kruschke, J. K. (2001). Toward a unified model of attention in associative learning. *Journal of Mathematical Psychology, 45*(6), 812–863. https://doi.org/10.1006/jmps.2000.1354.

Kuznetsova, A., Brockhoff, P. B., & Christensen, R. H. B. (2017). lmerTest package: Tests in linear mixed effects models. *Journal of Statistical Software, 82*. https://doi.org/10.18637/jss.v082.i13.

Lohse, K. R., Boyd, L. A., & Hodges, N. J. (2016). Engaging environments enhance motor skill learning in a computer gaming task. *Journal of Motor Behavior, 48*, 172–182.

O'Brien, H. L., & Toms, E. G. (2008). What is user engagement? A conceptual framework for defining user engagement with technology. *Journal of the American Society for Information Science and Technology, 59*, 938–955.

Oostenveld, R., & Praamstra, P. (2001). The five percent electrode system for high-resolution EEG and ERP measurements. *Clinical Neurophysiology, 112*, 713–719.

Palidis, D. J., Cashaback, J., & Gribble, P. (2018). Neural signatures of reward and sensory prediction error in motor learning. *bioRxiv,* 262576.

Pashler, H., & Baylis, G. (1991). Procedural learning: 2. Intertrial repetition effects in speeded choice tasks. *Journal of Experimental Psychology: Learning, Memory & Cognition, 17*, 33–48.

Pedroni, A., Langer, N., Koenig, T., Allemand, M., & Jancke, L. (2011). Electroencephalographic topography measures of experienced utility. *Journal of Neuroscience, 31*, 10474–10480. https://doi.org/10.1523/Jneurosci.5488-10.2011.

Philiastides, M. G., Biele, G., Vavatzanidis, N., Kazzer, P., & Heekeren, H. R. (2010). Temporal dynamics of prediction error processing during reward-based decision making. *Neuroimage, 53*, 221–232. https://doi.org/10.1016/j.neuroimage.2010.05.052.

Proudfit, G. H. (2015). The reward positivity: From basic research on reward to a biomarker for depression. *Psychophysiology, 52*, 449–459. https://doi.org/10.1111/psyp.12370.

Reinhart, R. M., & Woodman, G. F. (2014). Causal control of medial–frontal cortex governs electrophysiological and behavioral indices of performance monitoring and learning. *Journal of Neuroscience, 34*, 4214–4227.

Rieskamp, J., & Otto, P. E. (2006). SSL: A theory of how people learn to select strategies. *Journal of Experimental Psychology General, 135*, 207–236.

Sambrook, T. D., & Goslin, J. (2014). Mediofrontal event-related potentials in response to positive, negative and unsigned prediction errors. *Neuropsychologia, 61*, 1–10. https://doi.org/10.1016/j.neuropsychologia.2014.06.004.

Sambrook, T. D., & Goslin, J. (2015). A neural reward prediction error revealed by a meta-analysis of ERPs using great grand averages. *Psychological Bulletin, 141*, 213–235.

Sambrook, T. D., & Goslin, J. (2016). Principal components analysis of reward prediction errors in a reinforcement learning task. *Neuroimage, 124*, 276–286. https://doi.org/10.1016/j.neuroimage.2015.07.032.

Sambrook, T. D., Hardwick, B., Wills, A. J., & Goslin, J. (2018). Model-free and model-based reward prediction errors in EEG. *Neuroimage, 178*, 162–171. https://doi.org/10.1016/j.neuroimage.2018.05.023.

Schultz, W. (2017). Reward prediction error. *Current Biology, 27*, R369–R371.

Sutton, R. S., & Barto, A. G. (1998). *Introduction to reinforcement learning.* Cambridge, MA: MIT Press https://doi.org/10.1.1.32.7692.

Sweller, J. (2010). Element interactivity and intrinsic, extraneous, and germane cognitive load. *Educational Psychology Review, 22*, 123–138.

Thorndike, E. L. (1927). The law of effect. *The American Journal of Psychology, 39*, 212–222. https://doi.org/10.2307/1415413.

van der Helden, J., Boksem, M. A., & Blom, J. H. (2009). The importance of failure: Feedback-related negativity predicts motor learning efficiency. *Cerebral Cortex, 20*, 1596–1603.

Watkins, C. J. C. H., & Dayan, P. (1992). *Q-learning, vol. 8*, Machine Learning279–292.

Wulf, G., & Lewthwaite, R. (2016). Optimizing performance through intrinsic motivation and attention for learning: The OPTIMAL theory of motor learning. *Psychonomic Bulletin & Review, 23*, 1382–1414. https://doi.org/10.3758/s13423-015-0999-9.