Computational Brain & Behavior (2020) 3:189–199
https://doi.org/10.1007/s42113-019-00069-4

ORIGINAL PAPER

Reward Prediction Errors Reflect an Underlying Learning Process That Parallels Behavioural Adaptations: A Trial-to-Trial Analysis

Chad C. Williams[1] · Cameron D. Hassall[1] · Talise Lindenbach[1] · Olave E. Krigolson[1]

Published online: 18 October 2019

Abstract
Reinforcement learning can lead to rapid changes in performance. Computational accounts of reinforcement learning align with classic learning theory, as reported by Sutton and Barto (1998, 2018) and suggest that trial-to-trial changes in performance follow rapid but decelerating learning curves. Although there is some support for a link between changes in behavioural and neural data, evidence has been inconclusive. Here, we had a computational model and human participants learn a novel language through trial-and-error while recording electroencephalographic data. By conducting linear mixed-effects models of trial-to-trial analyses, we sought to determine whether neural signals were indicative of a learning process and whether they were related to changes in behaviour. We found that neural measures did diminish with trial-to-trial changes in performance and that they were predictive of behavioural adaptations in both simulated and empirical data. These neural signals are theorised as reward prediction errors—the computational difference between expectations and outcomes—and here we provide compelling evidence that they reflect an underlying learning process that parallels behavioural adaptation.

Keywords Reinforcement learning · Reward prediction errors · Reward positivity · Feedback error-related negativity · Behavioural adaptation · Electroencephalography

Introduction

Changes in performance during the early stages of learning have long been recognised to follow non-linear trends (Newell and Rosenbloom 1981; Rosenbloom and Newell 1987). Performance curves demonstrate rapid changes in the initial stages of acquisition that depreciate with succeeding trials (i.e. the power law of practice: Newell and Rosenbloom 1981; Rosenbloom and Newell 1987). It is hypothesised that rapid, early performance improvements occur due to the development of high-level representations of knowledge within which associations are made between the constituents of what is being learned (Newell and Rosenbloom 1981; Rosenbloom and Newell 1987). The non-linear nature of the rapid, early changes seen in performance is robust as it is observed across learning domains—perception, motor, and cognitive (Johnson et al. 2003)—as well as, across species (e.g. Klaus et al. 2011). Indeed, a wealth of research has begun to probe the speed to which reinforcement learning can occur (Botvinick et al. 2019). Reinforcement learning is the process whereby associations are facilitated by performance feedback in a trial-and-error fashion (Sutton and Barto 1998, 2018). When considering small sets of items, learning has been demonstrated to be incredibly quick (Krigolson et al. 2014; see also Botvinick et al. 2019).

Changes paralleling those seen in behaviour are also hypothesised within the brain. For instance, the seminal computational modelling work of Schultz et al. (1997) and others (e.g. FitzGerald et al. 2015) simulated changes of dopamine—a neurotransmitter tied to learning mechanisms—and determined that diminishing effects of dopamine at reward delivery followed parabolic change akin to the aforementioned changes in behaviour (see also Botvinick et al. 2019; Mnih et al. 2015). More recently, a similar pattern of results has been observed in studies with humans using electroencephalography (EEG). In a seminal study, Holroyd and Coles (2002) proposed that frontal brain signals reflect the arrival of a

Electronic supplementary material The online version of this article (https://doi.org/10.1007/s42113-019-00069-4) contains supplementary material, which is available to authorized users.

✉ Chad C. Williams
ccwillia@uvic.ca

[1] Centre for Biomedical Research, University of Victoria, P.O. Box 17000 STN CSC, Victoria, British Columbia V8W 2Y2, Canada

dopaminergic reward prediction error signals from the basal ganglia to the anterior cingulate cortex via the mesolimbic dopamine system. Reward prediction errors have been proposed to reflect the computational difference between one's expectation and the actual outcome (Krigolson 2018; Proudfit 2015). When learning, humans are able to tune their expectations to better predict outcomes and reward prediction errors diminish (Bellebaum and Daum 2008; Eppinger et al. 2008; Krigolson et al. 2009, 2014; Luft 2014; Luque et al. 2012; Sailer et al. 2010; Williams et al. 2018).

For instance, Krigolson et al. (2009) had participants learn to categorise two families of complex shapes through trial-and-error. Within each trial, participants indicated whether the presented shape and family name matched and were presented with valid feedback as to their performance (i.e. there was a correct response). Unbeknownst to participants, a third family of blobs existed which were not tied to responses but instead provided equiprobable feedback (i.e. the outcome was random). Krigolson et al. (2009) found that reward prediction errors diminished for learnable stimuli, but not for unlearnable stimuli, indicating that the modulation of these signals was reflective of a learning process. It may be logical to assume that these signals would then reflect behavioural adaptations, yet this has not always been the case (see Luft 2014; Walsh and Anderson 2012). A potential cause of these inconsistencies is that the evidence is majorly derived from investigations comparing stages of learning—e.g. comparing the beginning of learning to the end of learning—rather than across trials. As the speed of learning can be rapid (e.g. Krigolson et al. 2014), aggregating data across trials may distort the true associations between neural changes and behavioural adaptations. As such, a key issue with the existing body of literature is that previous research has not focused on trial-to-trial changes in neural prediction errors to see if they align with behavioural adaptations.

Here, we sought to determine whether reward prediction errors were indicative of trial-to-trial changes in performance and whether these learning-related changes paralleled behavioural adaptations. To accomplish this, we derived theoretical predictions via a reinforcement learning computational model and collected empirical data from human participants. Specifically, a computational model and human participants learned sixty words of a novel language by developing symbol-word associations. We then conducted trial-to-trial analyses by averaging across items (i.e. symbols) while preserving single trials. If reward prediction errors (measured via the reward positivity event-related potential component; see Krigolson 2018; Proudfit 2015) were indicative of a learning process, we would expect to see them diminish across trials. Furthermore, if there existed a relationship between behavioural and neural changes, we would be able to predict behavioural adaptations from neural changes with linear mixed-effects models. Accordingly, we hypothesised that the reward positivity would diminish across trials and that trial-to-trial changes in the amplitude of this component would be concomitant to behavioural adaptations.

Methods

Participants

Thirty undergraduate students (19 female; mean age, 20 years old [19 years old, 21 years old]) from the University of Victoria participated in the experiment. All participants had normal or corrected-to-normal vision, no known neurological impairments, and volunteered for extra course credit in a psychology course. One participant was removed from analyses due to missing behavioural data, while another was removed due to technical errors in EEG data collection. All participants provided informed consent approved by the Human Research Ethics Board at the University of Victoria.

Apparatus and Procedure

Participants performed an experimental task while seated in a sound dampened room and in front of a 19" LCD computer monitor. To make responses during performance of the task, they used a handheld 5-button RESPONSEPixx controller (VPixx, Vision Science Solutions, Quebec, Canada). The experimental task was programmed in MATLAB (Version 9.3, Mathworks, Natick, USA) with the Psychophysics Toolbox extension (Version 3.0.14, Brainard 1997).

Within the task, participants learned a novel language through a trial-and-error shaping process (see Fig. 1). Specifically, participants learned an English translation for each symbol presented. It is important to note that although they were learning one-to-one translations of a novel language (i.e. a vocabulary), this is only one aspect to learning a language, and the task did not involve learning other features of a language (e.g. grammar). The experiment began with a training block wherein participants learned the symbol-response relationship for six words. This was followed by the main experiment where participants were to learn three new words each block which were mixed in with three randomly selected old words (that they had already learned). There were 19 blocks and participants learned a total of 60 words (6 in the training block and 54 in the main experiment blocks). The order of which nouns were included in each block was randomly determined across participants so that no two participants were presented the same order of nouns. The trials within the training and main phases were identical otherwise.

The trials began on a grey background [RGB: 64 64 64] where a black [RGB: 0 0 0] fixation cross appeared for 700 to 1000 ms, followed by a symbol written in black [RGB: 0 0 0] that was 4.5 cm^2 in size. The symbols used were taken from

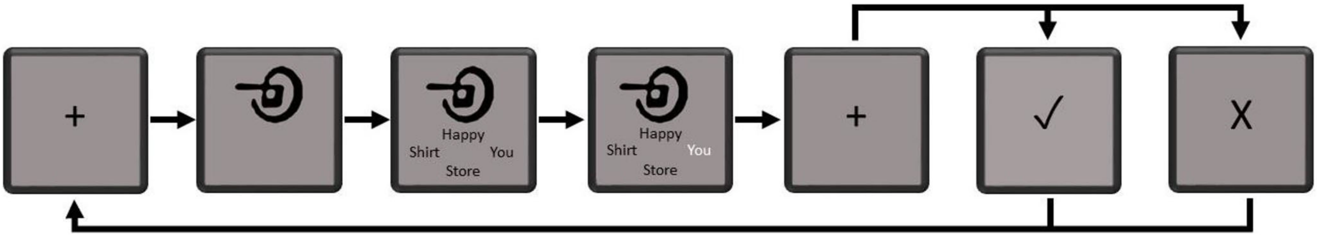

Fig. 1 A single trial of the experiment. After a fixation cross a symbol is presented, followed by four response options. Once a response is selected, it is highlighted in white. After a fixation cross, simple correct and incorrect feedback is presented

Tamil and Manipuri alphabets and were all paired with an English word meaning (see Online Resource 1, Table S1). On each trial, the target symbol was randomly chosen from the set of six symbols for the block, 500 ms after symbol presentation, four black [RGB: 0 0 0] English words appeared in the arrangement of a fixation cross (top, bottom, right, and left) below the symbol. One of the choices was the correct answer, and the three distractor words (incorrect answers) were randomly chosen from the remaining five words. The locations of all four words were randomly determined. Participants were then prompted to make a response by pressing one of the buttons on the RESPONSEPixx controller (VPixx, Vision Science Solutions, Quebec, Canada). Once a selection was made, the selected word turned white [RGB: 255 255 255] for 500 ms, the screen changed to a fixation for 700 to 1000 ms, and then a feedback stimulus appeared for one second. The feedback stimulus was a black [RGB: 0 0 0] '✓' or 'X' corresponding to correct and incorrect, respectively. Participants were informed of the respective feedback-outcome interpretations. Feedback was deterministic and did not carry any external motivators such as monetary or point accumulations. If a selection was not made within 6 s, an exclamation mark would appear to signify that the participant took too long to respond, and these trials were removed from analyses. Each block contained ten trials and would repeat until participants received 90% or higher accuracy. Once participants achieved 90%, they would progress to the next block.

Additionally, before and after the word learning phase of each block, participants saw three sentences and were to determine what they said through a multiple-choice selection. Specifically, sentences were composed of three symbols, each presented one at a time for 1000 ms and followed by a black [RGB: 0 0 0] fixation cross for 700 to 1000 ms. Half of the sentences were congruent (e.g. I went store) while the other half were incongruent (e.g. I went shirt). The words involved in the sentences included the new words introduced in the respective blocks. Four sentences (response options) would then appear in the arrangement of a fixation cross, and participants were to select the correct sentence. No feedback of their accuracy was given. As this was not the focus of the current article, this data was discarded from all analyses. After each block, participants were provided a self-paced rest period.

Data Acquisition and Processing

EEG data were recorded from a 64-electrode ActiCAP system (Brain Products GmbH, Munich, Germany) using the Brain Vision Recorder software (Version 1.10, Brain Products GmbH, Munich, Germany). All electrodes were referenced to a common ground, and electrode impedances were kept below 20 kΩ. EEG data were sampled at 500 Hz, amplified through the ActiCHamp amplifier (Revision 2, Brain Products GmbH, Munich, Germany), and filtered using an antialiasing low-pass filter of 245 Hz. To ensure temporal accuracy, EEG markers and stimuli were aligned via a DATAPixx synchronisation unit (VPixx, Vision Science Solutions, Quebec, Canada).

All EEG data were processed using Brain Vision Analyzer software (Version 2.1.1, Brain Products GmbH, Munich, Germany). First, electrodes that were excessively noisy or damaged were removed from the analysis. The data was down-sampled to 250 Hz, re-referenced to an average mastoid reference, and put through a dual-pass Butterworth filter (pass band, 0.1 to 30 Hz) and a notch filter centred at 60 Hz. Segments were then extracted from the EEG data from 1000 ms before to 2000 ms after events of interest (i.e. feedback stimulus onset). This segment width (3000 ms) was chosen to facilitate the correction of eye blinks and movement artefacts via independent component analysis (ICA). Following data segmentation, a restricted fast ICA with classic PCA sphering was used. Components reflecting ocular artefacts were manually selected and rejected via inspection of component head maps and correspondence between their related factor loadings and the related EEG data. Following component selection and removal, the EEG data were reconstructed with the remaining ICA components. Electrodes that were initially removed were next interpolated via the methods of spherical splines. Data was then re-segmented (− 200 to 600 ms around feedback stimulus onset) into six conditions (*Incorrect*, *Correct 1*, *Correct 2*, *Correct 3*, *Correct 4*, and *Correct 5*). The incorrect condition refers to all trials within which the participant selected the wrong word respective of the symbol. The Correct 1 to Correct 5 conditions refers to the first, second, third, fourth, and fifth time the participants saw a specific symbol and correctly identified the associated word. For example, across six trials, a participant may have the following feedback for the word ‘Happy’: *incorrect, correct, correct, correct, correct, correct*, and the following feedback for the word ‘Shirt’: *correct, incorrect, correct, correct, correct, correct.* Happy would then be considered as: *incorrect, correct 1, correct 2, correct 3, correct 4, correct 5*, and shirt would be considered as *correct 1, incorrect, correct 2, correct 3, correct 4, correct 5.* Next, all segments were baseline corrected using the 200 ms prior to the feedback stimulus onset following which all segments were processed with an artefact rejection algorithm that rejected trials that violated a 15 μV/ms gradient or a 150 μV absolute difference criteria. After pre-processing, event-related potential (ERP) conditional waveforms were created by averaging the segmented data for each electrode for each condition. A difference waveform was then created by subtracting the Incorrect waveform (which included all incorrect trials) from the Correct 1 waveform. The first correct (Correct 1) waveform was used for the construction of the different waveform because the amplitude of the reward positivity diminished over the experiment (as participants learned, see the ‘Results’ section). Grand average difference (Correct 1 − Incorrect) and conditional (Incorrect, Correct 1, Correct 2, Correct 3, Correct 4, and Correct 5) waveforms were created by averaging corresponding ERP waveforms across participants.

A 50-ms window surrounding the averaged peak time of the difference waveform across participants was used to extract the peak of the reward positivity and all conditions for each participant at electrode FCz, consistent with previous work (Krigolson 2018). The peak time for each participant was determined as the maximum deflection of the reward positivity of different waveforms between 250 and 350 ms following feedback stimulus onset. Although the reward positivity is traditionally measured as the difference between correct and incorrect waveforms, the reward positivity amplitude is theorised to be modulated by reward signals and not loss signals (Foti et al. 2011; Holroyd et al. 2008; Proudfit 2015). Thus, we indirectly measured the reward positivity amplitude across the experiment solely using the raw correct feedback waveforms (and not the difference between these correct waveforms and incorrect waveforms). This approach is in line with previous work from our laboratory (Krigolson et al. 2014; Krigolson 2018; Williams et al. 2018), wherein there have not been sufficient incorrect outcomes to quantify the reward positivity as the difference between correct and incorrect outcomes.

Reinforcement Learning Computational Model

Within this study, a reinforcement learning computational model simulated task performance in terms of trial-to-trial changes in accuracy, reaction times, and reward prediction errors. The computational model was used to determine parallels between reinforcement learning theory (i.e. Sutton and Barto 1998, 2018) and the observed behavioural and neural data. On each trial, the model was presented with one symbol and four possible responses. The model would then select one of four responses with the goal of learning the correct stimulus-response associations. Each stimulus-response association carried a value, $V(s, a)$, which the model used to select responses via a SoftMax equation:

$$P(s,a_i) = \frac{e^{V(s,a_i)/\tau}}{\sum_{j=1}^{4}\Sigma e^{V(s,a_j)/\tau}}. \quad (1)$$

where $P(s,a_i)$ is the probability that each action (response) is selected, $V(s, a_i)$ is the value of the current stimulus with the selected response i, and $V(s, a_j)$ is the value of the current stimulus with each possible response j. i refers to the selected response of the possible four responses in the current trial, and these four responses are a subset of the 60 responses (one for each stimulus) across the experiment. j refers to each of the four possible responses in the current trial. τ is the temperature. Here, the denominator is summed over all present responses.

The response was then compared with the target response which elicited a prediction error:

$$\delta = r - V(s,a_i) \quad (2)$$

where δ is the prediction error and r is the reward on the current trial (+ 1 for correct, − 1 for incorrect; we selected + 1 and − 1 to comply with mathematical computations and not to imply reward and punishment, respectively). Stimulus-response values were adjusted as a factor of this prediction error and a learning rate (α):

$$V(s,a_i) = V(s,a_i) + \delta\alpha. \quad (3)$$

Additionally, on correct trials, the same adjustment was used to devalue all other responses (including those not currently presented) to this symbol as well as all other symbols to this response (i.e. counterfactual learning; see Fischer and Ullsperger 2013):

$$V(s_k,a_l) = V(s_k,a_l) - \delta\alpha. \quad (4)$$

where s_k refers to all stimuli not presented on the current trial and a_l refers to all responses that were not selected. This was because during this task, a new symbol may be presented with new or old responses. If a response had been learned previously, then the participant (or model) should know not to select this response for the new symbol. For example, if you know that Ø means ‘shirt’, you also know that it does not mean ‘plane’ and that ∂ does not mean ‘shirt’. As with participants, once the model reached a 90% or higher accuracy rate in 10 trials, it would move to the next block where three new symbols were introduced and three old symbols were randomly selected from all previously learned symbols. This process continued until the model successfully learned the symbol-response associations.

The model was tuned 28 times (matching our sample size) in terms of three free parameters: learning rate, temperature, and initial weight. To determine these parameter values, the native MATLAB ‘fmincon’ function was used. The parameters were constrained with the boundaries of 0.001 to 1.0 for learning rate, 0.001 to 10 for temperature, and − 1.0 to 1.0 for initial weight values for stimulus-response pairings. The boundaries for learning rate were to include only values where learning was possible (i.e. not a learning rate of 0), and to include extreme cases of learning (i.e. a learning rate of 1.0 is where the previous trial would completely determine one’s expectations of the current trial). The lower boundary of the temperature was determined to always include exploration (i.e. not a temperature of 0), and the upper boundary was determined through experimentation of the model. Specifically, the upper temperature boundary was selected because early explorations of different models (where we were determining which parameters to include) never elicited temperature values exceeding 10. The boundaries of the initial weight were selected to match the delivered reward value when the model was incorrect (− 1) and correct (+ 1). The negative summed absolute correlation between the model simulation data and the corresponding participant data for all measures (accuracy, reaction time, and reward positivity; each model tuning considered all three measures simultaneously) was used to determine best fit parameters (learning rate, 0.47 [0.42, 0.55]; temperature, 5.82 [5.18, 6.46]; initial weight, 0.04 [− 0.02, 0.09]). Finally, we ran the model with the best fitting parameters (i.e. from each participant) to generate model data used within analyses.

Accuracy rates were measured as the proportion of times the model was correct across words. Reaction times were measured as the averaged difference between the probability of the correct response being selected relative to the average of the other three incorrect response probabilities. Thus, although simulated reaction times were proportional to empirical reaction times, they were not in units of milliseconds but rather constrained to values from 0 to 1. Prediction errors were computed as described above.

Statistical Procedures

All descriptive statistics, *t* tests, effect sizes (effsize package; Torchiano 2017), and linear mixed-effects models (lme4 package: Bates et al. 2015; see Winter 2013 and www.bodowinter.com/tutorials.html for a tutorial) were computed using R (Version 3.3.0, The R Foundation, Vienna, Austria) and R Studio (Version 0.99.902, RStudio Inc., Boston, USA). All plots were created using the ggplot2 R package (Wickham 2016).

To determine the function of trial-to-trial learning curves, we computed linear mixed-effects models for all measures of empirical and model data. These tests compared linear:

$$Measure \sim \beta_0 + \beta_1 Trial + (Trial|Participant) + \varepsilon \quad (5)$$

exponential:

$$\log(Measure) \sim \beta_0 + \beta_1 Trial + (Trial|Participant) + \varepsilon \quad (6)$$

and power:

$$\log(Measure) \sim \beta_0 + \beta_1 \log(Trial) + (Trial|Participant) + \varepsilon \quad (7)$$

functions via R^2_{GLMM} values (variability explained for generalised linear mixed models as computed by the MuMIn R package—Bartoń 2018; see Johnson 2014; Nakagawa et al. 2017; Nakagawa and Schielzeth 2013) where ‘*Measure*’ corresponded to accuracy, reaction times, or reward positivity amplitudes and ‘*Trial*’ referred to the first five trials (for behavioural analyses) or the first five trials that resulted in a correct response (for reward positivity amplitude analyses) for each word. Here, we only analysed the first five trials for each word because participants rapidly learned a large quantity of words, and we were thus only able to present each word a small number of times. Thus, we focused on five trials to ensure more included stimuli and increased statistical power. Please note, as there are some limitations to comparing linear mixed-effects models with different error distributions and with different model complexities, we have supplied additional analyses using generalised linear mixed models in Supplemental Material.

Additionally, we computed linear mixed-effects models to determine any trial-to-trial relationships between behavioural data and the reward positivity for both empirical and model data. Specifically, we analysed whether the reward positivity amplitudes were able to predict accuracy rates and reaction times:

$$\begin{aligned} BehaviouralMeasure \sim & \beta_0 + \beta_1 RewardPositivity \\ & + (Trial|Participant) + \varepsilon \end{aligned} \quad (8)$$

The assumptions of linearity, homoskedasticity, and normality of residuals were met for all empirical data. The reinforcement learning computational model data violated the assumptions of linearity, and on some occasions, the assumptions of homoskedasticity and normality of residuals. As the linear mixed-effects model formulas corresponded with learning curves and the computational models were consequent of empirical data, no corrections were made. We have here focused on effect sizes as a method for model evaluation (Cumming 2013). Specifically, we compared and contrasted models using measures of fit and specified the model with the largest variability explained to best describe our data. Although we have focused our interpretation of the data on effect sizes, we have also included null hypothesis significance tests with an alpha criterion of 0.05.

Results

The Computational Model Produced Power Learning Curves

The reinforcement learning computational model was implemented in order to examine the theoretically predicted trial-to-trial changes in behavioural and neural activity. We found that accuracy rates increased across learning (see Fig. 2a), reaction times decreased across learning (see Fig. 2b), and reward prediction errors decreased across learning (see Fig. 2c). The changes in these measures were all best accounted for by power law functions rather than exponential or linear trends (see Fig. 3, Table 1).

Empirical Measures Adhered to Power Learning Curves

To determine whether empirical data paralleled theoretical predictions of the computational models, trial-to-trial analyses on behavioural and neural data were conducted. Across learning, accuracy rates increased (see Fig. 2a) and reaction times decreased (see Fig. 2b). Congruent to model findings, changes in these measures were best accounted for by power learning curves rather than exponential or linear trends (see Fig. 3 and Table 1).

To determine neural effects of learning, the reward positivity was analysed. First, we determined that feedback stimuli elicited a reward positivity as indicated by the difference between the correct and incorrect conditions, $M_d = 5.19\ \mu\text{V}$ [$2.82\ \mu\text{V}$, $7.56\ \mu\text{V}$], $t(27) = 4.49$, $p = 0.0001$, $d = 0.85$ [0.29, 1.41] (see Fig. 4a). Next, we observed a diminishment in the amplitude of the reward positivity across learning (see Figs. 2c and 4b). In conjunction with the reinforcement learning computational model, the change in reward positivity amplitudes were best accounted for by power law functions rather than exponential or linear trends (see Fig. 3 and Table 1).

Neural Learning Signals Indicated Behavioural Adaptations

Finally, we determined whether any relationships existed between behavioural and neural measures of learning for both model and empirical data (see Fig. 5). Reward prediction errors as produced by the computational model were strongly associated with accuracy rates and reaction times (see Table 2). Congruently, empirical reward positivity amplitudes were strongly associated with accuracy rates and reaction times (see Table 2).

Discussion

Here, we provide compelling evidence that the reward positivity reflects an underlying learning process responsible for behavioural adaptations. Our investigations first focused on theoretical predictions as derived by a reinforcement learning computational model which demonstrated that reward prediction errors diminished across trials and that these learning-related changes were strongly predictive of behavioural changes in accuracy rates and reaction times. Indeed, empirical analyses also demonstrated a trial-to-trial diminishment of the reward positivity which was strongly predictive of behavioural adaptations.

A major implication of these findings is that the reward positivity ERP component does, in fact, reflect a learning process. This supports the multi-faceted nature of processes that elicit reward prediction errors (Weinberg et al. 2014). Particularly, the reward positivity has also been shown to reflect the processing of value-based outcomes (e.g. Brush et al. 2018). Hedonic influences are often investigated using a two-door paradigm where participants are to select one of two doors to achieve an award (see Proudfit 2015). In this type of paradigm, responses often have no influence on outcomes in that performance feedback is equiprobable (i.e. it is unlearnable), and outcomes are translated into rewards (monetary or

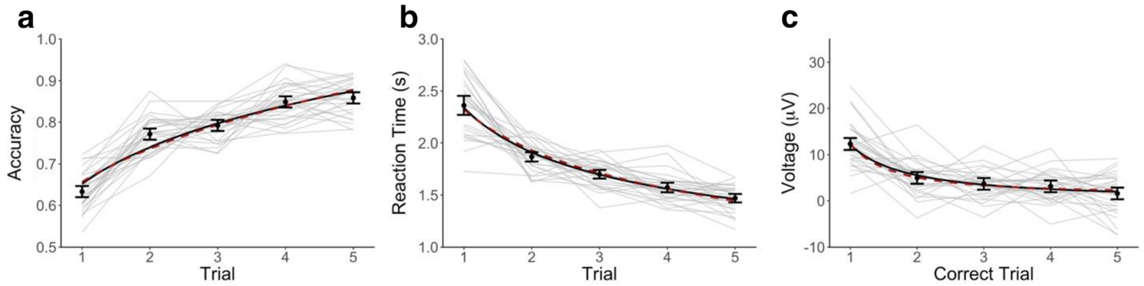

Fig. 2 Behavioural and neural data adhere to power learning curves. **a** Accuracy rates across first five trials, **b** reaction times across first five trials, **c** reward positivity amplitudes across first five correct trials. Each grey line represents an individual participant's data corrected for between subject variability. Black points reflect grand averaged participant data with 95% within-subject credible intervals (Nathoo et al. 2018). Black lines are the best-fit power trend line. Red dashed lines are the best-fit power law functions of the reinforcement learning computational model. 'Trials' refers to the averaged first-five trials for each word regardless of whether the response was correct or incorrect, while 'Correct Trials' refers to the averaged first-five trials for each word where the response was correct

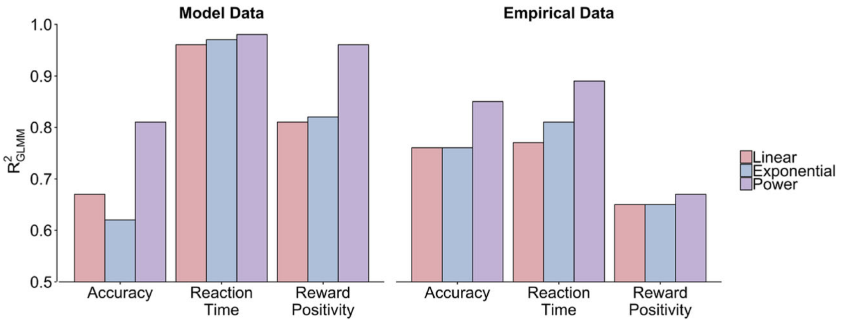

Fig. 3 Model simulated and empirical fits indicate power learning curves better describe changes in behavioural and neural data than linear and exponential trends. Three trend fits for accuracy, reaction time, and reward positivity measures. Goodness of fit measures reflects R^2_{GLMM} values. See also Table 1

points). Indeed, external rewards have been shown to impact reward prediction error signals (Brush et al. 2018). Here, we instead implemented a learnable task without any external rewards and found an impact on the reward positivity. Thus, to fully understand reward prediction errors and their biological correlates—the ventral striatum and the anterior cingulate cortex (ACC; Holroyd and Coles 2002; Holroyd and McClure 2015; Holroyd and Yeung 2012)—it is necessary to consider both learning and the processing of hedonic outcomes. Recent work by Holroyd and colleagues (Holroyd and Yeung 2012; Holroyd and McClure 2015) integrated learning with hedonic outcomes when they theorised that the ACC is the core to learning the values of high-level behaviours. Specifically, they posited that learning leads to efficiency by balancing cognitive effort with future reward—or, in other words, that learning-related signals within the ACC drive adaptive behaviour.

Nonetheless, whether there even exists a relationship between neural learning signals and behavioural adaptations has been controversial. Although there is evidence linking changes in reward prediction errors to behavioural adaptations, there are similarly findings challenging these notions (see Luft 2014; Walsh and Anderson 2012). For example, Holroyd and Krigolson (2007) provided evidence of a link between neural learning signals and reaction times when they were investigating the effects of reward expectancy in a time estimation task. They found that both reward prediction error amplitudes and reaction times were larger for the unexpected condition than the expected condition, and that there was a positive association between behaviour and neural signals. By contrast, Walsh and Anderson (2011) found that reward prediction errors diminished with learning even when accuracy rates remained constant. In this study, participants were presented with three stimuli, each with a unique probability of reward. On each trial, they were to select between two of the three stimuli with the goal of maximising rewards. In one of their conditions, they explicitly told the participants of the stimulus probabilities, thus accuracy rates were near perfect and unchanging, and yet still found a shaping of neural learning signals. They concluded that independent learning systems exist within the brain, those that guide behaviour and those that track reward prediction errors.

The inconsistency of the aforementioned findings may be in part due to an artefact of comparing aggregated data across

Table 1 Statistical outcome of trend analyses for accuracy rates, reaction times, and reward positivity amplitudes for model and empirical data

		Model data			Empirical data		
Source	Trend	*F* value	*p* value	R^2_{GLMM}	*F* value	*p* value	R^2_{GLMM}
Accuracy	Linear	280.57	< 0.0001	0.6687	163.50	< 0.0001	0.7597
	Exponential	227.70	< 0.0001	0.6209	118.40	< 0.0001	0.7634
	Power	435.53	< 0.0001	0.8140	151.92	< 0.0001	0.8475
Reaction time	Linear	390.56	< 0.0001	0.9641	167.29	< 0.0001	0.7747
	Exponential	359.40	< 0.0001	0.9692	213.56	< 0.0001	0.8095
	Power	344.57	< 0.0001	0.9771	190.79	< 0.0001	0.8881
Reward positivity	Linear	587.64	< 0.0001	0.8128	44.07	< 0.0001	0.6543
	Exponential	626.68	< 0.0001	0.8244	40.88	< 0.0001	0.6540
	Power	963.08	< 0.0001	0.9596	53.34	< 0.0001	0.6679

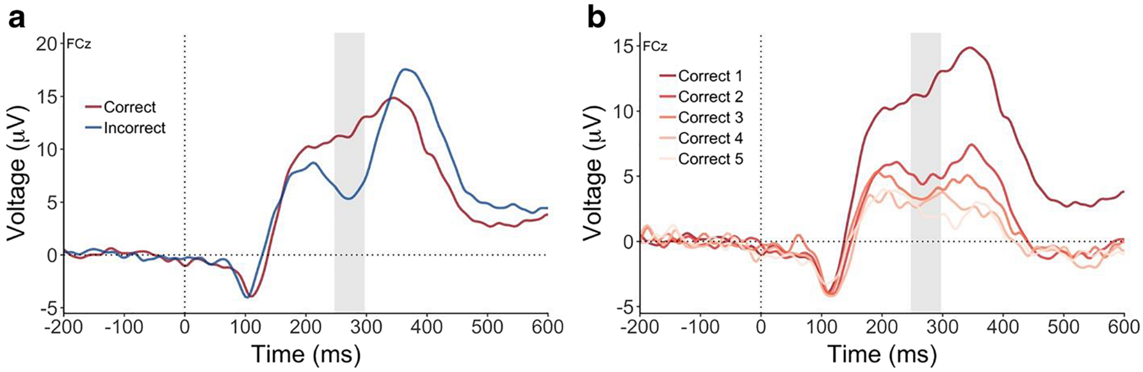

Fig. 4 Reward positivity waveforms. **a** Conditional waveforms of incorrect feedback and first correct trial feedback indicating that feedback elicited neural learning signals. **b** Conditional correct feedback waveforms across correct trials indicating that the reward positivity amplitude diminishes across trials of learning. Grey bar indicates range in which mean peaks were computed. Positive voltages are plotted up

conditions (e.g. early and late blocks). Reinforcement learning is a quick process that can take as little as one trial (Krigolson et al. 2014), thus averaging over trials may distort data and blur the true relationship between neural learning signals and behavioural adaptations. Our analyses here, however, focused on trial-to-trial changes across learning which best captures the quick process that is reinforcement learning with small state spaces. Specifically, by conducting linear mixed-effects modelling, we determined similar and associated learning curves across behavioural and neural measures. Our computational simulations suggested that reward prediction errors were heavily influential on both accuracy rates and reaction times, and our empirical findings provided compelling evidence for a strong relationship between neural learning signals and behavioural adaptations. In other words, changes in neural learning signals were indicative of behavioural adaptations.

The implications of these findings are multi-faceted, but here, we focus on the potential of real-time monitoring and future predictions of performance. In the present study, we were able to capture changes in neural learning signals at a trial-to-trial level—a technique that could theoretically be implemented in real time. The caveat is that we averaged across

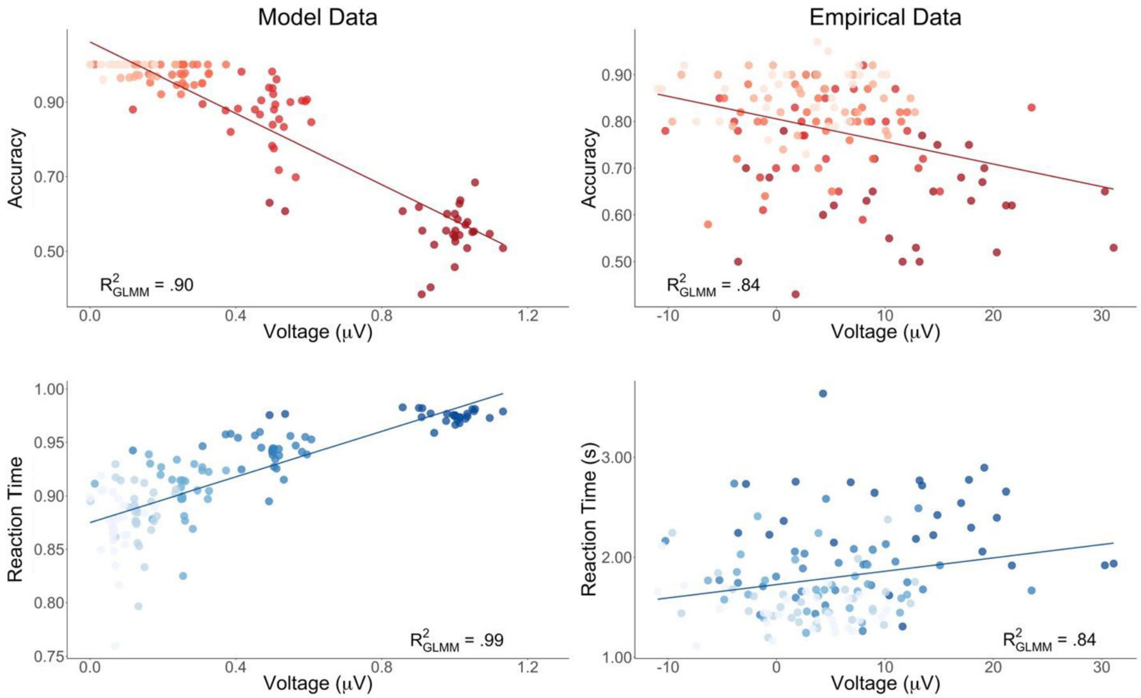

Fig. 5 A strong relationship between neural and behavioural measures for all model simulated and empirical data. Top: relationship between reward positivity amplitudes and accuracy rates, bottom: relationship between reward positivity amplitudes and reaction times. Lines reflect grand averaged linear regression for participant and model data. Each participant and simulation contributed five data points for each plot—one for each of the first five trials. The intensity of colours scales to trial in that trial 1 is the darkest colours and trial 5 is the lightest colours. See also Table 2

Table 2 Statistical outcomes of associations between the reward positivity and behavioural measures for model and empirical data

	Model data			Empirical data		
Measure	*F* value	*p* value	R^2_{GLMM}	*F* value	*p* value	R^2_{GLMM}
Accuracy	690.03	< 0.0001	0.8965	8.87	0.0035	0.8368
Reaction time	150.10	< 0.0001	0.9912	13.33	0.0004	.8594

items (i.e. words) and so, this cannot directly transfer to real-time monitoring. It is feasible, however, to further our findings by applying computational techniques such as machine learning in order to monitor performance via neural signals within single trials. Machine learning is an approach previously applied to other domains of neuroscience, such as when Müller et al. (2008) had participant type characters by deciphering single-trial EEG. The results would be twofold. First, we would achieve real-time access to an unbiased account of knowledge. We say unbiased because neural signals of learning are not influenced by many factors that distort behavioural measures, such as guessing. Second, we would be able to use these signals to predict success on future performance. The implications of these outcomes extend across environments ranging from education to the workplace and with the exponential rise of portable EEG systems (e.g. Krigolson et al. 2017), these applications are already possible.

Limitations

A possible alternative explanation to our findings, however, is rooted in the frequency within which correct feedback is delivered. Indeed, the reward positivity is sensitive to frequency effects (Holroyd 2004; Krigolson 2018). For example, it has been demonstrated that the reward positivity is larger for unexpected events relative to expected events (e.g. Williams et al. 2017). Here, as participants learn to correctly associate symbols with their English translations, the number of correct responses (and corresponding feedback) increases, making this event more expected which may diminish the reward positivity. This alternative is unlikely, however, given that our computational model corresponded with empirical findings yet does not have a mechanism that reflects frequency effects. That is not to say that frequency effects are not at all an influence in our findings, but that frequency effects are possibly an engrained component that facilitates changing reward prediction errors and learning. What remains is whether frequency effects have any influence on learning and, if so, to what degree. Future research would need to dissociate frequency effects from learning to determine trial-to-trial changes in the reward positivity. Hassall and Krigolson (2013) developed a two arm bandit task which would be suitable to answer this question. In their task, participants were presented with two coloured squares, one of which results in win feedback 60% of the time and the other 10% of the time. The task was to select the square that more often resulted in wins. Thus, learning occurred, but due to the probabilistic nature of each square, win and loss outcomes were near 50% (see also Krigolson et al. 2017; Krigolson 2018).

A limitation to the current research involves creating analogues between the simulated and empirical data. Indeed, empirical findings replicated those derived by our computational model; however, this does not necessarily indicate that human reward processing involves the exact mechanisms as within our reinforcement learning model. Although there is a long history of reinforcement learning computational modelling which provides promise that there are tight similarities between human processing and computational models (see Sutton and Barto 1998), we did not investigate how different mechanisms of the models affected trial-to-trial variations in the data and what implications this may have on what we understand of human processing. In order to further explore this, it would be necessary to systematically lesion the model (either via removal or disruption of mechanisms/parameters such as learning rate) to determine how this affects accuracy rates, reaction times, and reward prediction errors. Further, to fully understand the link between these models and learning-related trial-to-trial human processing, it would also be necessary to conduct research within clinical populations where these same mechanisms are disrupted. Indeed, there is ample research investigating clinical populations and reward prediction errors within a computational modelling context (see Holroyd and Umemoto 2016); however, this field research has yet to explore trial-to-trial learning-related variations.

Conclusion

In summary, our goal was to determine whether reward prediction errors were indicative of a learning process that parallels behavioural adaptation. For both theoretical predictions as derived by a computational reinforcement learning model and empirical data from human participants, reward prediction errors diminished with learning. Additionally, both computational simulations and empirical data indicated that neural signals of learning were predictive of accuracy rates and reactions times. In other words, reward prediction errors reflect a learning process that is indicative of behavioural adaptation. This research demonstrates potential to expand into the real-time monitoring and predictions of future performance within environments ranging from education to the workplace.

Acknowledgements Chad Williams and Cameron Hassall would like to acknowledge support from the Neuroeducation Network (grant number: 007) and Olave Krigolson would like to acknowledge support from the Natural Sciences and Engineering Research Council of Canada (RGPIN 2016-094).

Author Contribution C.W. designed the experiment, collected and analysed the data, conducted statistical analyses, and wrote the manuscript. C.H. provided technical expertise on computational modelling. T.L. collected and supported analysis of the data. O.K. is the senior author on the project.

Compliance with Ethical Standards

All participants provided informed consent approved by the Human Research Ethics Board at the University of Victoria.

Conflict of Interest The authors declare that they have no conflict of interest.

References

Bartoń, K. (2018). MuMIn: multi-model inference. R package version 1.42.1. https://CRAN.R-project.org/package=MuMIn. Accessed 22 Jun 2018

Bates, D., Maechler, M., Bolker, B., & Walker. (2015). Fitting linear mixed-effects models using lme4. *Journal of Statistical Software, 67*(1), 1–48. https://doi.org/10.18637/jss.v067.i01.

Bellebaum, C., & Daum, I. (2008). Learning-related changes in reward expectancy are reflected in the feedback-related negativity. *European Journal of Neuroscience, 27*(7), 1823–1835.

Botvinick, M., Ritter, S., Wang, J. X., Kurth-Nelson, Z., Blundell, C., & Hassabis, D. (2019). Reinforcement learning, fast and slow. *Trends in Cognitive Sciences, 23*(5), 408–422.

Brainard, D. H. (1997). The psychophysics toolbox. *Spatial Vision, 10*(4), 433–436.

Brush, C. J., Ehmann, P. J., Hajcak, G., Selby, E. A., & Alderman, B. L. (2018). Using multilevel modeling to examine blunted neural responses to reward in major depression. *Biological Psychiatry: Cognitive Neuroscience and Neuroimaging, 3*(12), 1032–1039.

Cumming, G. (2013). *Understanding the new statistics: effect sizes, confidence intervals, and meta-analysis*. Routledge.

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

Fischer, A. G., & Ullsperger, M. (2013). Real and fictive outcomes are processed differently but converge on a common adaptive mechanism. *Neuron, 79*(6), 1243–1255.

FitzGerald, T. H., Dolan, R. J., & Friston, K. (2015). Dopamine, reward learning, and active inference. *Frontiers in Computational Neuroscience, 9*, 136.

Foti, D., Weinberg, A., Dien, J., & Hajcak, G. (2011). Event-related potential activity in the basal ganglia differentiates rewards from nonrewards: temporospatial principal components analysis and source localization of the feedback negativity. *Human Brain Mapping, 32*(12), 2207–2216.

Hassall, C.D., and Krigolson, O.E. (2013). Wake up and smell the shifting probabilistic outcomes. *Psychology and Neuroscience 39th Annual Graham Goddard In-House Conference*, Dalhousie University, Halifax, NS.

Holroyd, C. (2004). A note on the oddball N200 and the feedback ERN. *Neurophysiology, 78*, 447–455.

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

Holroyd, C. B., & Krigolson, O. E. (2007). Reward prediction error signals associated with a modified time estimation task. *Psychophysiology, 44*(6), 913–917.

Holroyd, C. B., & McClure, S. M. (2015). Hierarchical control over effortful behavior by rodent medial frontal cortex: a computational model. *Psychological Review, 122*(1), 54.

Holroyd, C. B., & Umemoto, A. (2016). The research domain criteria framework: the case for anterior cingulate cortex. *Neuroscience & Biobehavioral Reviews, 71*, 418–443.

Holroyd, C. B., & Yeung, N. (2012). Motivation of extended behaviors by anterior cingulate cortex. *Trends in Cognitive Sciences, 16*(2), 122–128.

Holroyd, C. B., Pakzad-Vaezi, K. L., & Krigolson, O. E. (2008). The feedback correct-related positivity: sensitivity of the event-related brain potential to unexpected positive feedback. *Psychophysiology, 45*(5), 688–697.

Johnson, P. C. D. (2014). Extension of Nakagawa & Schielzeth's R_GLMM^2 to random slopes models. *Methods in Ecology and Evolution, 5*, 44–946.

Johnson, E. J., Bellman, S., & Lohse, G. L. (2003). Cognitive lock-in and the power law of practice. *Journal of Marketing, 67*(2), 62–75.

Klaus, A., Yu, S., & Plenz, D. (2011). Statistical analyses support power law distributions found in neuronal avalanches. *PLoS One, 6*(5), e19779.

Krigolson, O. E. (2018). Event-related brain potentials and the study of reward processing: methodological considerations. *International Journal of Psychophysiology, 132*(B), 175–183.

Krigolson, O. E., Pierce, L. J., Holroyd, C. B., & Tanaka, J. W. (2009). Learning to become an expert: reinforcement learning and the acquisition of perceptual expertise. *Journal of Cognitive Neuroscience, 21*(9), 1833–1840.

Krigolson, O. E., Hassall, C. D., & Handy, T. C. (2014). How we learn to make decisions: rapid propagation of reinforcement learning prediction errors in humans. *Journal of Cognitive Neuroscience, 26*(3), 635–644.

Krigolson, O. E., Williams, C. C., Norton, A., Hassall, C. D., & Colino, F. L. (2017). Choosing MUSE: validation of a low-cost, portable EEG system for ERP research. *Frontiers in Neuroscience, 11*, 109.

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

Luque, D., López, F. J., Marco-Pallares, J., Càmara, E., & Rodríguez-Fornells, A. (2012). Feedback-related brain potential activity complies with basic assumptions of associative learning theory. *Journal of Cognitive Neuroscience, 24*(4), 794–808.

Mnih, V., Kavukcuoglu, K., Silver, D., Rusu, A. A., Veness, J., Bellemare, M. G., et al. (2015). Human-level control through deep reinforcement learning. *Nature, 518*(7540), 529–533.

Müller, K. R., Tangermann, M., Dornhege, G., Krauledat, M., Curio, G., & Blankertz, B. (2008). Machine learning for real-time single-trial EEG-analysis: from brain–computer interfacing to mental state monitoring. *Journal of Neuroscience Methods, 167*(1), 82–90.

Nakagawa, S., & Schielzeth, H. (2013). A general and simple method for obtaining R^2 from Generalized Linear Mixed-effects Models. *Methods in Ecology and Evolution, 4*, 133–142.

Nakagawa, S., Johnson, P. C. D., & Schielzeth, H. (2017). The coefficient of determination R^2 and intra-class correlation coefficient from generalized linear mixed-effects models revisited and expanded. *Journal of the Royal Society Interface, 14*, 20170213.

Nathoo, F. S., Kilshaw, R. E., & Masson, M. E. (2018). A better (Bayesian) interval estimate for within-subject designs. *Journal of Mathematical Psychology, 86*, 1–9.

Newell, A., & Rosenbloom, P. S. (1981). Mechanisms of skill acquisition and the law of practice. *Cognitive Skills and Their Acquisition, 1*(1981), 1–55.

Proudfit, G. H. (2015). The reward positivity: from basic research on reward to a biomarker for depression. *Psychophysiology, 52*(4), 449–459.

Rosenbloom, P., & Newell, A. (1987). Learning by chunking: a production system model of practice. *Production System Models of Learning and Development*, 221–286.

Sailer, U., Fischmeister, F. P. S., & Bauer, H. (2010). Effects of learning on feedback-related brain potentials in a decision-making task. *Brain Research, 1342*, 85–93.

Schultz, W., Dayan, P., & Montague, P. R. (1997). A neural substrate of prediction and reward. *Science, 275*(5306), 1593–1599.

Sutton, R. S., & Barto, A. G. (1998). *Reinforcement learning: an introduction*. MIT Press.

Sutton, R. S., & Barto, A. G. (2018). *Reinforcement learning: an introduction* (2nd ed.). The MIT Press.

Torchiano, M. (2017). effsize: Efficient effect size computation. R package version 0.7.1. https://CRAN.R-project.org/package=effsize. Accessed 21 Mar 2017.

Walsh, M. M., & Anderson, J. R. (2011). Modulation of the feedback-related negativity by instruction and experience. *Proceedings of the National Academy of Sciences, 108*(47), 19048–19053.

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

Weinberg, A., Riesel, A., & Proudfit, G. H. (2014). Show me the money: the impact of actual rewards and losses on the feedback negativity. *Brain and Cognition, 87*, 134–139.

Wickham, H. (2016). *ggplot2: Elegant graphics for data analysis*. New York: Springer-Verlag.

Williams, C. C., Hassall, C. D., Trska, R., Holroyd, C. B., & Krigolson, O. E. (2017). When theory and biology differ: The relationship between reward prediction errors and expectancy. *Biological Psychology, 129*, 265–272.

Williams, C. C., Hecker, K. G., Paget, M. K., Coderre, S. P., Burak, K. W., Wright, B., & Krigolson, O. E. (2018). The application of reward learning in the real world: Changes in the reward positivity amplitude reflect learning in a medical education context. *International Journal of Psychophysiology, 132*(B), 236–242.

Winter, B. (2013). Linear models and linear mixed-effects models in R with linguistic applications. arXiv preprint arXiv:1308.5499.