Mental Model Updating and Pupil Response

Sungjoon Park

Thesis Supervisor: Dr. Britt Anderson

Thesis Reviewer: Dr. James Danckert

PSYCH 499C Thesis Document – Fall 2019

Department of Psychology

University of Waterloo

Abstract

To behave adaptively, we construct mental (internal) models of the environment to inform our decisions. To change or update our mental models, we detect unexpected events in the form of surprises or prediction errors. In reinforcement learning, this prediction error is used to guide future predictions to make adaptive decisions. However, not all surprises are relevant to our mental models. Mental model updating occurs selectively for informative surprises, known as Bayesian surprise. Surprises that are uninformative, or purely information theoretic, do not contribute to our models. In addition, one must take into consideration the reliability of the surprises. In a stable environment, a surprising outcome is a reliable sign that a change has occurred. However, in an unstable environment, where the outcome is more stochastic, we can become less confident about the informativeness of the prediction error. Using pupilometry in humans solving a probabilistic learning task, we found that Bayesian surprises predict faster peak pupil dilations than information theoretic surprises. In addition, feedback reliability did not relate to greater peak pupil dilation during Bayesian surprises. Together, our results provide evidence that it may be possible to differentiate Bayesian surprise from information theoretic surprise through pupilometry.

Introduction

To behave adaptively in an ever changing and stochastic environment, we create mental models; internal representations of the world (Johnson-Laird, 2013). These models are built from our experiences and it informs many of our behaviors, decision making process and expectations (Filipowicz et al., 2016). Since the utility of a mental model depends on how accurately it predicts the future, it must be able to update itself with new information about our environment (Filipowicz et al., 2016). Reisenzein et al. (2019) claim that surprises inform us about our flawed mental models and that it initiates an attempt to explain the cause of the surprise and/or update our mental models to mend the discrepancies. In the reinforcement learning literature, this discrepancy between expectation and observation is referred to as a prediction error (PE) (Sutton & Barto, 2018). And the process of belief updating has been described to follow a delta-rule model described as:

$$B_{t+1} = B_t + \alpha_t \times \delta_t$$

Where our future belief state (B_{t+1}) depends on our current belief state (B_t) and the current PE experienced (δ_t), which is modified by the a learning rate (α_t) (Nassar et al., 2010). There have been numerous studies investigating the neural correlates of PE in both animals and humans, and many studies reference the dopamine (DA) reward hypothesis, where the DA system plays a role in computing PE (Garrison et al., 2013). However, not all PE are thought to initiate mental model updating (O'Reilly et al., 2013), and we often ignore or give less weight to surprising outcomes when we are not confident about their reliability (Meyniel & Dehaene, 2017). It has been argued that confidence may modulate learning rates and evidence thresholds, which can determine how much a PE updates a mental model (Meyniel & Dehaene, 2017). In this study we explore two types of PE, Information Theoretic (IT) surprise and Bayesian surprise (Schwartenbeck et al., 2016), and their relationship with confidence in how they relate to mental model updating and whether we can use pupilometry to observe signs of mental model updating while humans participants perform a probabilistic learning task.

We were motivated to use pupilometry because while many studies employed invasive or expensive functional imaging tools to explore the neural correlates of mental model updating, pupilometry may be an affordable alternative to study this cognitive process in light of its connections to brain areas implicated in mental model updating. Past studies show that the noradrenergic locus coeruleus (LC) is a key brain structure implicated in regulating pupil dilation (Urai et al., 2017), central arousal, and signaling uncertainties about the environment (Yu & Dayan, 2005). And, as the LC is intimately connected with the dopaminergic midbrain area, dopaminergic activities associated with belief updating, a form of mental model updating, could be associated with pupil dilation (Van Slooten et al., 2018).

Surprise Types

Schwartenbeck et al. (2016) suggest that Bayesian surprises, but not IT surprises, are associated with mental model updating and that the two have distinct neural characteristics. IT surprise is defined purely by how unexpected the observation was based on one's mental model and it does not indicate how useful the observation is to better understand the environment (Schwartenbeck et al., 2016). It corresponds to its Shannon information, the negative log probability of the observation based on the mind's representation before the observation

(O'Reilly et al., 2013). On the other hand, Bayesian surprise is defined by how meaningful it is or by its epistemic value. It corresponds to the Kullback–Leibler divergence between one's belief before and after the observation (Schwartenbeck et al., 2016). In other words, information theoretic surprise is evaluated by the probability of an observation based on one's expectations, and Bayesian surprise is evaluated by the probability of whether one's expectations will change after an observation (O'Reilly et al., 2013). In their fMRI study, Schwartenbeck et al. (2016) found that the dopaminergic activities in the midbrain area (e.g., Substantia Nigra Compacta, Ventral Tegmental Area) were greater during belief switches and they suggest that these activities play an important role in belief updating. Of interest is how these types of surprises might be reflected in pupil responses. While, O'Reilly et al. (2013) claim that pupil responds differently to Bayesian and IT surprises, it is not clear what kind of relationship they have.

Despite the uncertain relationship between belief updating and pupil response, several studies suggest that the close connection between the dopaminergic midbrain structures and the noradrenergic locus coeruleus (LC), a key brain structure associated with pupil responses (Urai et al., 2017) and uncertainty (O'Reilly et al., 2013), play a vital role in modulating the pupil during a reinforcement learning context; a form of mental model updating (Van Slooten et al., 2018). It may be possible to observe pupil responses associated with Bayesian surprises beyond what is observed with information theoretic surprises that do not contribute to mental model updating. In support of this assertion, Colizoli et al. (2018) suggest pupil response to PE varies with belief state.

In a study where participants were tasked to identify the overall movement direction of random dot kinematograms. Colizoli et al. (2018) found that "pupil responses during feedback anticipation and after reward feedback were modulated by the decision-makers' internal belief states". In other words, depending on what the observers believe they have observed, their pupil responded differently before and after observing a confirmation or a refutation of it. In addition, the magnitude of pupil response varied depending on the ambiguity of the dot direction (difficulty condition of their study). Correct identifications predicted smaller pupil response compared to incorrect identification (Colizoli et al., 2018). Their result suggests that pupil responds differently depending on whether an expectation was met or not. However, this experiment only suggests that the pupil respond to surprises (PE) in general, and does not claim whether the increased pupil response from surprises can be associated with mental model updating. Our study goes a step further to explore the potential discrepancy in pupil response between IT and Bayesian surprise. Additionally, we are also interested in the peculiarity of how we do not unconditionally accept useful information. Sometimes, we doubt its legitimacy and question how reliable it is. Do we give PE a different weight depending on how confident we feel about their validity? And how is this relationship reflected in pupil responses?

Confidence

To understand how confidence can affect mental model updating, imagine a situation where we seek medical help. When a layman friend suggests a medication it is often ignored or we wait to gather more evidence to support their suggestion. However when a medical doctor prescribes a medication, we are more likely to immediately heed their advice. In one case, new information hardly changed our belief while on the other case our belief is immediately changed. It is evident that we place greater confidence in the doctor's words since they have a history of reliable information. In this context, the confidence we give to the doctor can be viewed as the high probability that the doctor is giving a good prescription (Meyniel et al., 2015).

In the context of computational learning theories, Meyniel & Dehaene (2017) state that confidence is used to determine how much new information should be incorporated during learning; the learning rate. It is thought that the more confident we are about new information, the more it changes our beliefs and the more confident we are about our previous beliefs, the less new information changes them (Meyniel & Dehaene, 2017). Boldt et al. (2017) claim that our confidence in our beliefs are reflected by the evidence reliability that they are based on. In both Urai et al. (2017) and Colizoli et al. (2018)'s studies, the investigators manipulated evidence reliability of a random dot kinematogram by manipulating the ambiguity (task difficulty) where some dot kinematogram directions were easier to identify than others (Beck et al., 2008). They observed that pupil response to PE for more difficult dot kinematograms were smaller than for easier dot kinematograms. This suggests that the arousal to PE is weaker when one is less confident about their prediction. Since the stimulus was less reliable, one might be less invested in their prediction and be less surprised when they are wrong. Could it be the case that when the environment is unpredictable, we give PE less weight and update our beliefs less than we would if the environment was more stable and the PE more reliable?

Experiment

To explore how the epistemic value of a PE and its relationship with confidence affect the degree in which we update our mental models, our study gathered both behavioral evidence in the form of explicit confidence report, and implicit physiological responses in the form of pupil response and reaction time while participants performed a probabilistic learning task. This task involved participants observing a visual stimulus with a color and shape feature, deciding whether the color or the shape predicted an outcome, and reporting their subjective confidence in their choice, all while having their pupil diameter monitored by an eye-tracker. Participants learned which feature was the predictive contingent by experiencing PE. For instance, if a participant based their prediction on color and they experience a PE, they could decide to base their next prediction on the shape. Throughout the experiment, the correct contingent alternated between the two features. Participants experienced IT surprises when they experience PE while both color and shape lead to the same prediction, and they experienced Bayesian surprise when the two features made different predictions. To manipulate confidence, we implemented stochasticity into the prediction feedback where there was always a chance that a correct prediction would be reported as a wrong prediction. During certain blocks, participants were exposed to a reliable (stable) condition where there was a 90% chance that the feedback will be reliable and during other blocks they were exposed to an unreliable (unstable) condition where there was only a 75% chance that the feedback would be reliable.

We hypothesized that i) PE will elicit greater pupil response than correct predictions, ii) Bayesian surprise will elicit greater pupil response than IT surprise, iii) being exposed to a more stochastic environment will lead to weaker confidence, iv) being exposed to a more stochastic environment will lead to smaller pupil response to Bayesian surprise, and v) that Bayesian surprise will predict greater change in mental model in the form of explicit confidence and that its magnitude will be modulated by the reliability manipulation where changes will be greater when feedback is more reliable.

Methods

Participants

Neurologically healthy students from the University of Waterloo (N = 8, 1 Male) were recruited through the university's SONA participant pool in exchange for course credits and monetary rewards across three experimental sessions (mean age = 20.26, SD = 1.82). One participant opted not to participate in the third session. None of the participants had a history of brain or eye injury and they reported normal or corrected-to-normal vision. All participants provided informed consent prior to participation, and the research protocol was approved by the Office of Research Ethics at the University of Waterloo.

The experiment lasted three hours across three sessions (one hour per session) and each session was conducted no more than 20 days from the previous session (mean days apart = 10.3, SD = 5.8). Participants were remunerated with course credits for the first session and half of the second session at a rate of 0.5 course credits per 30 minutes of participation, and they were remunerated with cash for the second half of the second session and the whole of the third session at the rate of $7 (CAD) per 30 minutes of participation. In addition, participants were incentivized to place effort with an opportunity to receive additional monetary reward between $0 and $10, linearly spaced between 50% and 100% of task accuracy for each session (e.g., 50% = $0, 75% = $5, 100% = $10), and an additional $15 for completing all three sessions.

Materials

The Cambridge Research System ltd. LiveTrack Lightening eye tracking system was used to collect binocular pupillometric data at a sampling rate of 500Hz with an accuracy of 0.1mm. The eye tracking camera was positioned 30 cm away from the participants eyes, without obstructing the monitor screen.

Participants were seated 70 cm away from a flat-screen CRT monitor, an NEC MultiSync FP2141SB, with a refresh rate of 160Hz at a resolution of 1024 by 768. The screen was color calibrated using the Cambridge Research Systems ltd. ColorCAL MKII Colorimeter. The participants rested their head on a head-rest, fixed on a table 70 cm away from the monitor, to maintain stable viewing of the screen and for the eye tracker to maintain a stable view of the participant's eyes.

An optical mouse was placed at the right side of the table and a keyboard was placed between the eye tracking camera and the monitor which was used by the participant to navigate the instructions for the computerized experimental task and for the researcher to set-up the experiment. Lastly, a pair of stereo speakers were placed at the left and right side of the monitor which were used to play feedback tones.

Probabilistic Learning Task

As depicted in figure 1, participants were tasked to make predictions after viewing a visual stimulus which contained a color and a shape feature. They based their predictions on either the color or the shape of the stimuli while reporting how confident they were about which feature is the current predictive contingent. After making their reports, participants heard a tone that signaled whether they made a correct or incorrect prediction. The confidence report had no bearing on the result of a trial. A reverse-coded version of the task was assigned to half of the participants; this is elaborated in greater detail later. The task involved a total of 576 trials, divided into three sessions with 192 trials each. Each session contained 12 blocks which contained 10 to 22 trials, randomly chosen from a uniform distribution. Participants were given a five minutes break before the start of the sixth block where they had the option to cut their break short.

Each sessions randomly started with either a color or shape contingent and alternated every block (e.g., shape → color → shape → color....). Each session started with a random chance to be either a stable or an unstable reliability condition. In the stable condition, there was 90% chance that the prediction feedback will be reliable where there is a 10% chance that the result will signal a PE regardless of whether the participant made a correct prediction or not. In the unstable condition, there was only 75% chance that the prediction feedback will be reliable. This reliability manipulation was reassigned every two blocks, where there was a random chance that for the next pair of blocks to be in a stable or an unstable condition. There was no case where more than 70% of the session was occupied by either one of the reliability conditions.

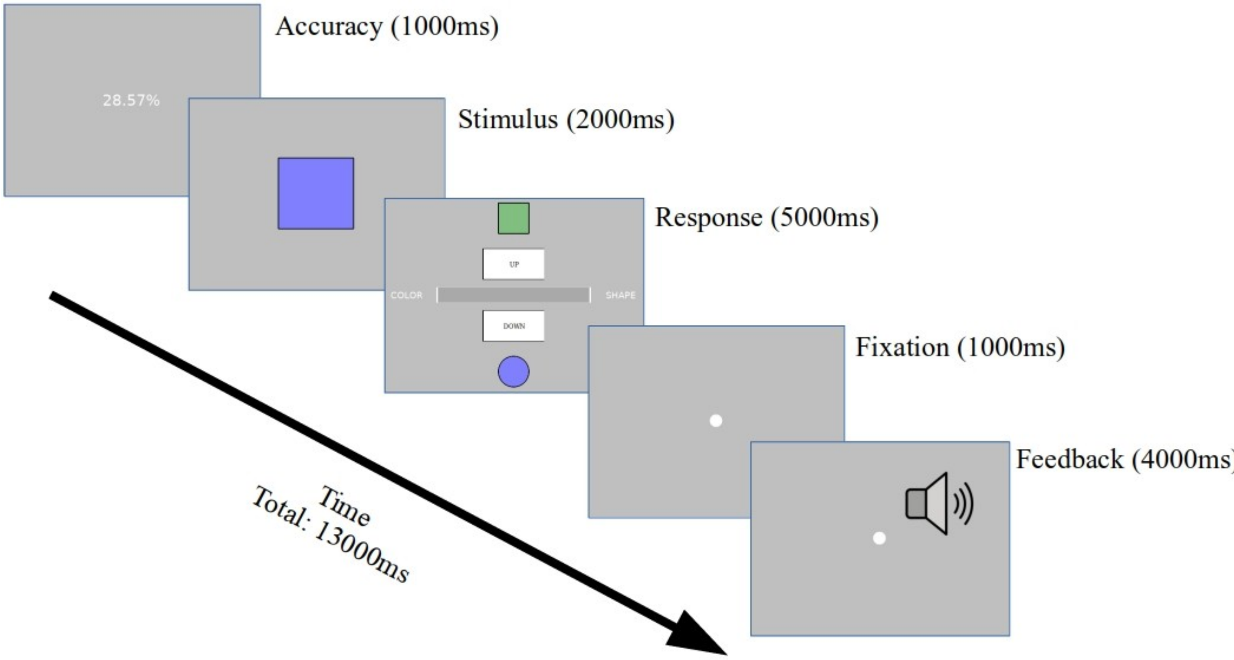

Figure 1. On each trial, i) participants start by seeing their accuracy during the session, ii) then a visual stimulus (green-square, green-circle, blue-square, blue-circle). iii) Next they see the response screen where they decide to predict 'up' or 'down' and rate how confident they feel that either color or shape is the correct contingent feature. iv) Lastly, they observe a fixation point where they were instructed to focus their gaze on while they listened to the feedback tone.

Visual Stimulus. On every trial, participants first viewed their success rate at the center of the screen for one second. Next, participants viewed a visual stimulus for two seconds with both a color and shape feature to make a prediction. The prediction involved clicking either 'up' or 'down' where one of two types of color and shape was associated with one or the other direction. When the color was the predictive contingent, green signaled 'up' and blue signaled 'down'. When the shape was the predictive contingent, square signaled 'up' and circle signaled 'down'. A reverse-coded version of the task flipped these associations where green and square signaled 'down' and blue and circle signaled 'up'. The four possible visual stimulus were green-square, green-circle, blue-square, and blue-circle.

Response. After viewing the visual stimulus, a response screen was presented for five seconds where participants made their prediction by clicking a button labeled ‘up’ or ‘down’ and giving their explicit rating of how confident they were in their belief that color (shape) was the current predictive contingent. They used a horizontal slider to indicate their confidence between the extremes of 100% color or 100% shape. The midpoint indicated equipoise. The ends were flipped in the reverse-coded task where left corresponded with shape and right with color. The participants could freely change their decisions during the five seconds window and they could not proceed to the next stage of the trial before the five seconds have elapsed. Failure to make either the direction choice or a confidence report lead to a screen that informed the participant about this fact and after three seconds, the trial ended.

Feedback. After the response screen, the participants viewed a white central fixation point for one second where they were instructed to fixate their gaze until the fixation point disappeared. This fixation point lasted until the end of the trial. Once one second had elapsed, a tone signaling successful prediction or PE was played for 3.5 seconds. The tones were hammed where the onset and offset is smoothed and linearly faded in and out for 0.5 seconds. A success (high) tone was a sine wave at 700 Hz with an amplitude of 0.5. An error (low) tone was a sine wave at 400 Hz with an amplitude of 0.8. These tones were made using the Generate → Tone option from the Audacity (R) audio recording and editing software (Version 2.3.3; Mazzoni, 2019). The tone associations were flipped in the reverse-coded task. The trial was over 0.5 seconds after the end of the tone.

Surprise Types. PE where both color and shape predicts the same outcome were IT surprises. For example, when the stimulus was a green square both the color and shape signaled 'up'. With IT surprises, one cannot make a rational inference that the predictive contingent has changed. If we previously believed that the color was the predictive contingent, we are not more convinced that shape is the predictive contingent since square also predicted 'up'. A counter-factual scenario where we believe shape is the predictive contingent does not lead to a successful trial. It it more likely that the PE was a result of the stochastic nature of the task. On the other hand, Bayesian surprises occur when PE involve color and shape making different predictions. If the stimulus was a green circle and we experience a PE, we might be able to infer that the contingent has changed. If we previously believed that color was the predictive contingent, we might consider that the contingent has changed to shape and imagine a counter-factual scenario where our prediction is based on the shape and we click 'down' and experience a successful trial.

Procedures

Participants read an information letter describing the study and decided to sign a consent form. Afterwords, participants read a detailed instruction about the task on the monitor screen. Once they informed the researcher that they finished reading the instructions, participants performed the practice trials without being timed and with no reliability manipulation. The trials were pre-set to so that each trial showed a distinct stimulus and the feature contingent switched after the fourth trial. The following eight trials were identical but timed. Once the participants indicated that they understood the task, the eye tracker calibration was performed. Participants placed their head on the head-rest and a nine-point calibration was conducted. After the calibration, the participants proceeded to perform the main experimental task. Once the fifth block was completed, participants were given a five minute break with the option to cut their break short. Once the session was over, participants received their monetary reward and they signed a receipt.

The second and third sessions were identical to the first session without reading an information letter or signing another consent form. Instead the task was verbally reviewed and they proceeded to perform the experimental task. At the end of the third session, participants received their additional reward for completing all three sessions and they were given a debrief and a feedback letter.

Pupil Preprocessing

As pupil size fluctuates with unintentional cognitive and luminance related factors, we preprocessed our raw pupilometry data to eliminate undesirable artifacts, as depicted in Figure 2, before analysis (Winn et al., 2018). All preprocessing was done using the R statistical software package (R Core Team, 2019). First, we identified blinks using a marker provided by the eye tracker that labels whether the eyes were successfully detected. The blinks were extended so that samples 100ms before and after a blink were considered to be blinks and we removed their pupil diameter samples. Next, we performed linear interpolation so that there was no missing pupil samples due to blinks. This was done using the na_interpolatetion function from the imputeTS package (Moritz & Bartz-Beielstein, 2017) and the na.fill function from the zoo package (Zeileis & Grothendieck, 2005). Next, we passed the pupil signal through a 4Hz, fourth order Butterworth low-pass filter. This was done using the butter and filter functions from the signal package (signal developers, 2014). Lastly, we down-sampled the pupil data to approximately 62Hz. This was done by removing every other sample three times.

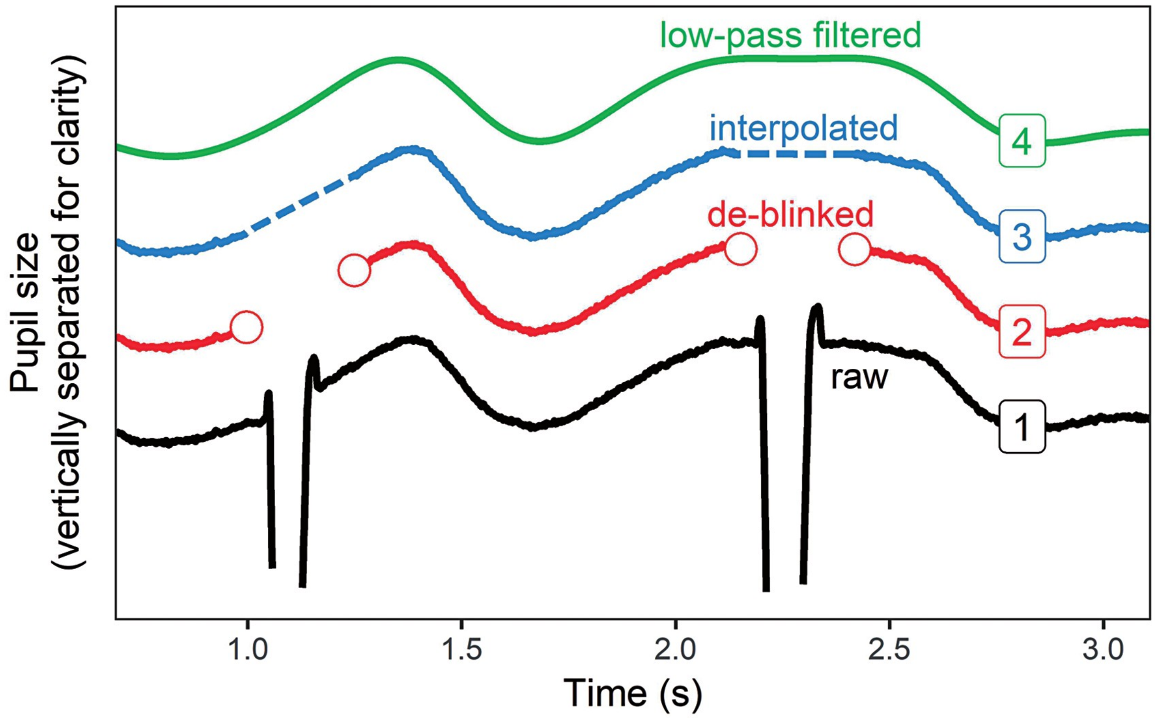

Figure 2. Illustration of preprocessing of raw pupil data. Blinks are detected and linearly interpolated. Lastly, a low-pass filter is applied to smoothen the signal, removing minor pupil artifacts (Winn et al., 2018).

Analysis

All analysis were conducted using the R statistical software package (R Core Team, 2019). Fixed effects linear regression analysis was done using the lm function in the base stats package (R Core Team, 2019). Mixed effects linear regression analysis was done using the lmer function in the lme4 package (Bates, Mächler, et al., 2015) with the bobyqa optimizer from the optimx package (Nash & Varadhan, 2011). When presenting the results, we report the F-statistics of the analysis and the standardized estimate (β.est), the standardized error (SE), and the 95% confidence intervals (CI). F-statistics was reported using the anova function from the base stats package (R Core Team, 2019) and the rest was reported using the std_beta function from the sjstats package (Lüdecke, 2019). When possible, a mixed effects regression analysis was performed. However, when the analysis faced singularity fit or convergence warnings, we formulated a more parsimonious mixed model with fewer parameters (Bates, Kliegl, et al., 2015) or we fell back to a fixed effects linear regression analysis. These issues are known to be caused by the lack of data with respect to the parameter complexity of the mixed model (Singmann & Kellen, 2017). For the behavioral task, task accuracy, explicit confidence, change in explicit confidence, and implicit confidence as reaction time (RT) was analyzed for manipulation checks of the reliability manipulation, and surprise types. For pupilometry, peak pupil dilation and the peak pupil dilation latency were analyzed for their sensitivity to PE and surprise types.

Results

Results are reported in the order of behavioral analysis then pupilmometry. For the behavioral analysis we explored our manipulations affected participants accuracy, explicit confidence, change in explicit confidence, and reaction time. For pupilometry, we explored how our manipulations affected the peak pupil dilation and the peak pupil dilation latency. Our manipulations and their descriptions can be referenced in Table 1.

Table 1. Names of manipulations and their description for reference.

Variable Name	Description
Reliability Condition	Manipulation of feedback reliability (75% and 90%).
Contingent Condition	Manipulation of predictive feature (Color and Shape).
Session	The three sessions when tasks were performed.
Stimuli associations	Task assignment where half the participants experienced reverse coded stimuli.
	Contingent feature association reversed (green/square = up → blue/circle = up)
	Location of contingent on confidence report slider reversed (left = color → right = color)
	Audio feedback tone association reversed (low tone = PE → high tone = PE)
PE	Whether a participant made a prediction error or a correct prediction.
Surprise Type	Whether a participant experienced a correct prediction, IT or Bayesian surprise.

Behavioral Analysis

Accuracy. Participant's accuracy was measured using their success rate; the fraction of successful trials among all the trials. To analyze whether participant's performance was affected by the our manipulations and whether learning was involved across the three sessions, we performed a linear mixed effects regression analysis with success rate as the dependent variable and the reliability condition, contingent condition, the interaction between the two manipulations, sessions, and stimuli associations as fixed effect independent variables with random intercepts among sessions within participants. Random slopes were only modeled for the first two fixed effects because of a convergence warning when attempting to model more random effects. In addition, this model preserves the slope variability among participants of our main manipulations. Analysis revealed the significant effect of reliability manipulation where participants performed better when feedback was more reliable ($F(1,21.7) = 45.618$, $p < 0.001$, β.est = -0.138, SE = 0.025, CI = [0.095, 0.195]). There was a significant effect of session where performance improved with successive task performance ($F(1,20.9) = 5.39$, $p = 0.03$, β.est = 0.062, SE = 0.027, CI = [0.01, 0.11]). There was an effect of reversed stimulus association where participant who experienced reversed stimuli tended to performed better ($F(1,20.8) = 10.01$, $p < 0.01$, β.est = 0.084, SE = 0.026, CI = [-0.136, -0.032]). There was no effect of contingent manipulation where accuracy was not significantly different whether color or shape was the contingent ($F(1,22) = 1.4$ $p = 0.25$, β.est = 0.027, SE = 0.023, CI = [-0.018, 0.072]). There was no significant effect of the interaction between the reliability condition and the contingent condition where the effect of reliability manipulation was not significantly different regardless of the contingent condition ($F(1,3345.8) = 0.165$, $p = 0.68$, β.est = -0.006, SE = 0.015, CI = [-0.036, 0.024]) (Figure 3).

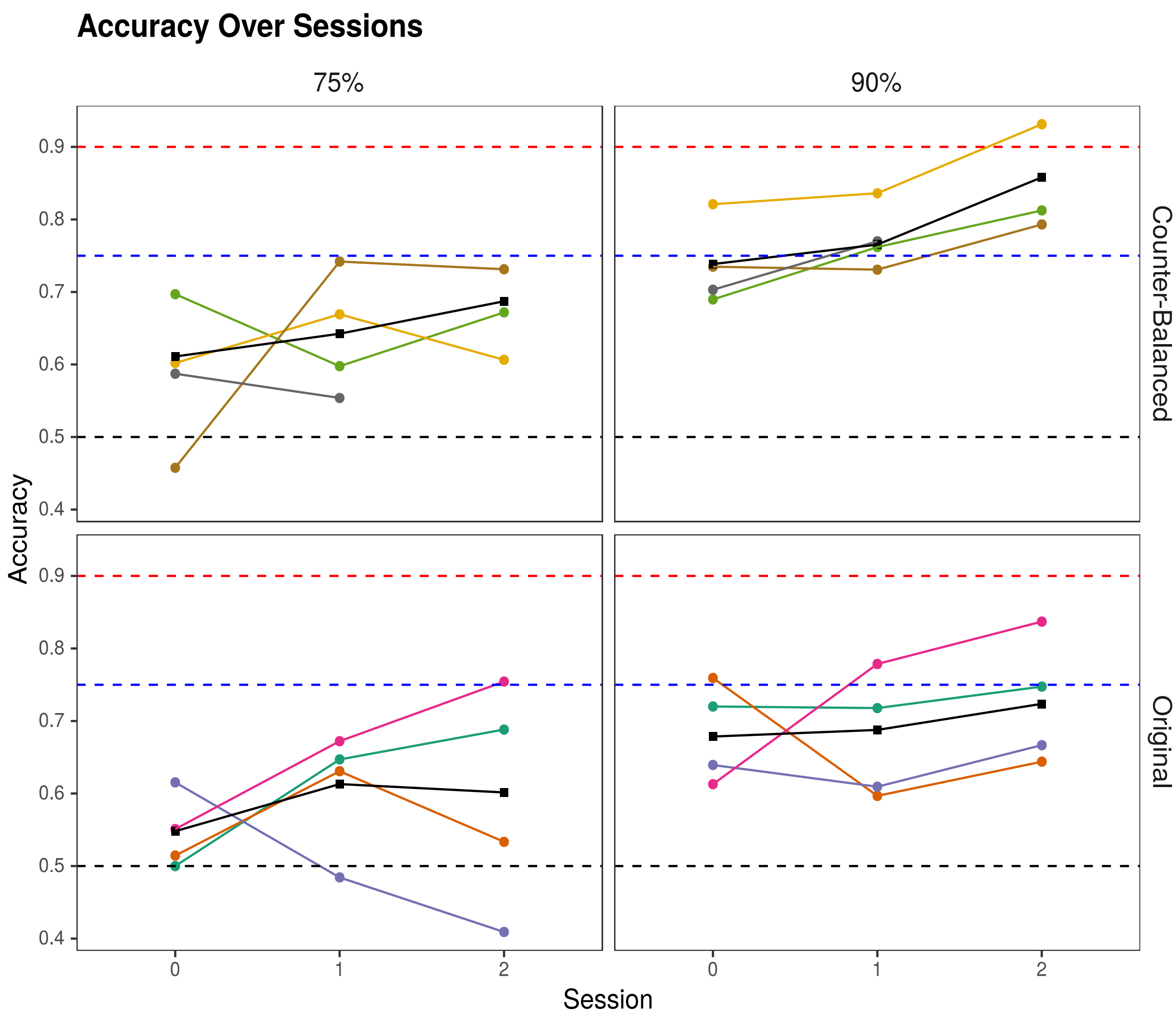

Figure 3. Change in participant's accuracy over sessions. We can observe most participants improve, especially when the feedback is more reliable. We also observe that participants who experienced reverse coded stimuli tend to perform better with more reliable feedback than participants assigned to the original task. Red dashed-line marks 90%, Blue dashed-line marks 75% respective of the reliability manipulation. Black dashed-line marks chance.

Explicit confidence. Participant's explicit confidence was measured using a slider bar during the response screen. The slider represented confidence in a range between 0 and 20, where each ends represents complete confidence in either color or shape. This scale was centered and aligned so that 0 was the middle point of the scale and -10 represented complete confidence in color. The absolute value of this confidence report was used to measure the participants explicit confidence level, regardless of their contingent belief state. To ascertain whether the reliability manipulation was effecting the participant's explicit confidence level, we performed a mixed effects linear regression with the explicit confidence rating as the dependent variable and the reliability manipulation, contingent manipulation, their interaction, stimuli associations, and session as the independent fixed effect variables with random intercepts among sessions within participants and random slopes for the reliability manipulation and contingent manipulation among sessions within participants. This model assumes the variance of intercepts and slopes for reliability across sessions within participants. The analysis suggest that there was no significant effect of reliability where absolute explicit confidence rating were similar whether the feedback was more or less reliable ($F(1,21.83) = 0.67$, $p = 0.42$, β.est = 0.003, SE = 0.024, CI = [-0.04, 0.05]). There was a significant effect of contingent condition where absolute explicit confidence report tended to be lesser when shape was the contingent feature ($F(1,21.93) = 5.28$, $p < 0.05$, β.est = -0.02, SE = 0.01, CI = [-0.04, -0.003]). There was a significant effect of the interaction between reliability condition and contingent condition where the effect of feedback reliability was significantly different when the contingent feature was different ($F(1,2212.52) = 3.97$, $p < 0.05$, β.est = 0.02, SE = 0.01, CI = [0.0004, 0.0445]). There was no significant effect of session where absolute explicit confidence report was consistent across sessions ($F(1,20.03) = 0.007$, $p = 0.93$, β.est = 0.01, SE = 0.17, CI = [-0.33, 0.36]). Lastly, there was no significant effect reversed stimulus association where absolute explicit confidence report was similar whether the stimuli associations were reversed or not ($F(1,20) = 1.44$, $p = 0.24$, β.est = -0.21, SE = 0.17, CI = [-0.55, 0.13]) (Figure 4).

Figure 4. Effect of feedback reliability on participants absolute confidence rating for each contingent feature. Each colored line represents participants. Black line represents the mean.

Explicit confidence change to PE. Change in confidence was measured as the difference in confidence rating between the current trial and the subsequent trial. For example, if the confidence report during the current trial was -10 and the confidence report during the following trial was -8, the change in confidence was +2 where confidence in color was decreased by 2 points while confidence in shape was increased by 2 points; the maximum change in confidence can be ±20. To explore how PE effect participant's change in confidence and belief, we performed a linear regression analysis with the change in confidence as the dependent variable, and PE, the reliability condition, the contingent condition and their interaction, session, and the stimuli associations as the independent variables. The analysis revealed that there was a significant effect of PE where PE related to positive confidence change towards shape ($F(1,173) = 4.2$, $p = 0.04$, β.est = 0.096, SE = 0.05, CI = [0.004, 0.19]). There was a significant effect of contingent condition where when the shape was the contingent feature, confidence change tended to go towards shape ($F(1,174) = 17.26$, $p < 0.001$, β.est = 0.28, SE = 0.07, CI = [0.15, 0.41]). There was a significant effect of the interaction between PE and contingent condition where the effect of prediction error on confidence change was significantly different depending on the contingent feature ($F(1,174) = 186.58$, $p < 0.001$, β.est = -0.91, SE = 0.07, CI = [-1.03, -0.78]). There was a significant effect of a three way interaction between PE, reliability condition and contingent condition where the effect of PE on confidence change was significantly different depending on the reliability of the feedback and the contingent feature ($F(1,174) = 15.53$, $p < 0.001$, β.est = -0.26, SE = 0.07, CI = [-0.39, -0.13]) (Figure 5, 6). The complete analysis is presented in Table 2 and the non-significant effects are reported in the next paragraph.

There was no significant effect of the reliability manipulation where confidence change was not different depending on the reliability of the feedback ($F(1,174) = 0.48$, $p = 0.49$, β.est = 0.07, SE = 0.05, CI = [-0.13, 0.28]). There was no significant effect of session where confidence change was consistent across all the sessions ($F(1,174) = 1.56$, $p = 0.21$, β.est = 0.07, SE = 0.05, CI = [-0.04, 0.17]). There was no significant effect of stimuli associations where confidence change was not different whether the task response options were reversed or not ($F(1,174) = 0.18$, $p = 0.67$, β.est = -0.02, SE = 0.05, CI = [-0.13, 0.08]). Lastly, there was no significant interaction between reliability and contingent where confidence change resulting from more reliable feedback was not different depending on the contingent feature ($F(1,174) = 0.21$, $p = 0.65$, β.est = 0.06, SE = 0.13, CI = [-0.31, 0.19]).

Table 2. F-Statistics, Standardized Estimates, Standardized Errors and Confidence Intervals (95%) from the analysis of the change in explicit confidence in response PE and to the rest of our manipulations.

	DF	F	Pr(>F)	β.est	SE	CI: Lower	CI: Upper
PE	1, 174	9.390	< 0.01	-0.32	0.11	-0.53	-0.12
Shape	1, 174	6.099	< 0.05	-0.26	0.11	-0.46	-0.05
PE:90%	1, 174	6.014	< 0.05	-0.31	0.13	-0.57	-0.06
PE:Shape	1, 174	31.542	< 0.001	0.72	0.13	0.47	0.97
PE:90%:Shape	1, 174	12.303	< 0.001	0.49	0.14	0.21	0.76
90%	1, 174	0.48	0.49	0.07	0.11	-0.13	0.28
Session	1, 174	1.56	0.21	0.07	0.05	-0.04	0.17
Stimuli association	1, 174	0.18	0.67	-0.02	0.05	-0.13	0.08
90%:Shape	1, 174	0.21	0.65	0.06	0.13	-0.31	0.19

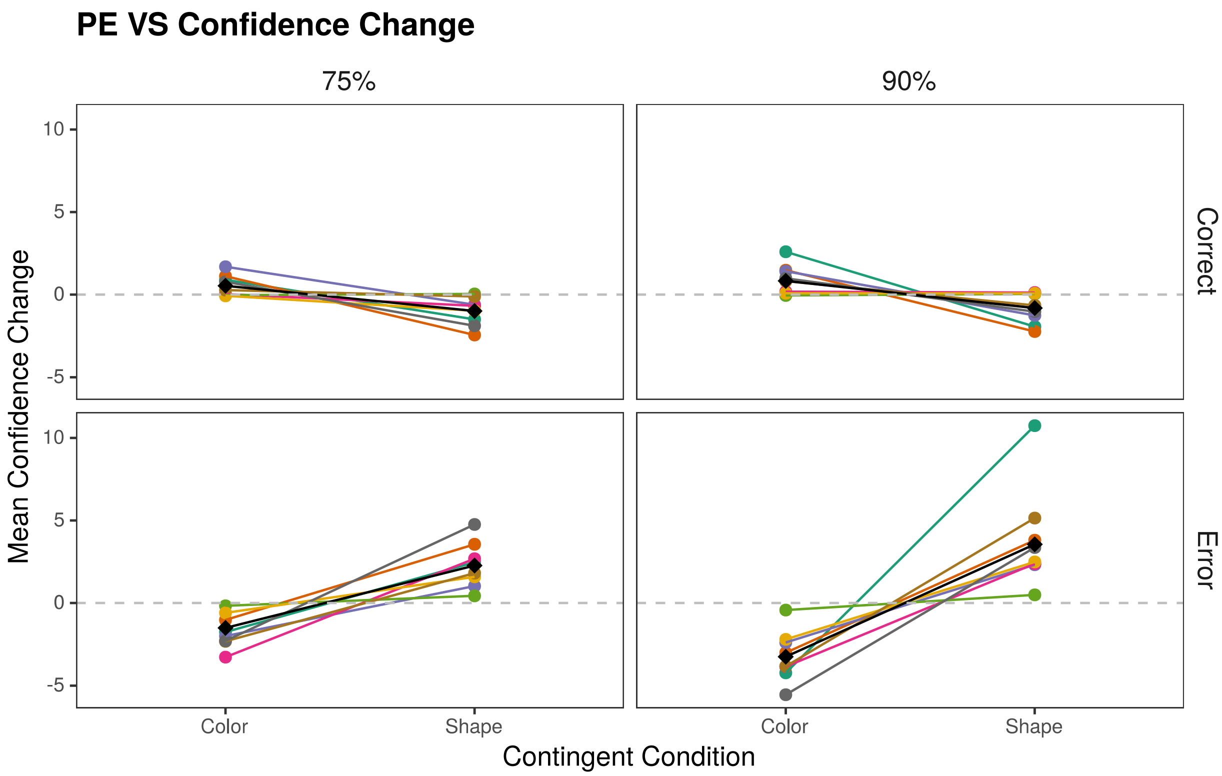

Figure 5. Effect of PE on mean confidence change per participant. Column separates reliability condition and rows separate prediction state. PE related to confidence change towards the correct contingent. We can observe the effect of reliability where confidence change was more drastic when feedback was more reliable. Negative confidence value relate to confidence change towards color and positive confidence value relate to confidence change towards shape.

Figure 6. Simplified version of Figure 4 the lines represent correct or error predictions, connecting the mean confidence change.

Explicit confidence change to surprise types. To test whether Bayesian surprise related to greater change in confidence than IT surprise, a linear fixed effect regression analysis was performed with change in confidence as the dependent variable with the surprise types, the contingent manipulation, the reliability manipulation and their interactions, sessions, and the stimuli associations as the independent variables. No linear mixed effects regression is presented because the simplest model faced singular fit warning. The analysis revealed a significant effect of surprise type were Bayesian surprise related to confidence change towards the correct contingent feature while IT surprise did not. When color was the correct contingent, Bayesian surprise predicted negative confidence change towards color ($F(2,255) = 10.67$, $p < 0.001$, β.est

= -0.37, SE = 0.12, CI = [-0.6, -0.14]) while IT surprise related to confidence change towards shape (β.est = -0.37, SE = 0.11, CI = [-0.08, 0.37]). When shape was the correct contingent, Bayesian surprise related to confidence change towards shape ($F(2,225) = 24.76$, $p < 0.001$, β.est = 0.68, SE = 0.13, CI = [0.43, 0.94]) while IT surprise did not (β.est = -0.18, 0.13, CI = [-0.43, 0.07]). We observed a marginal effect of sessions where later sessions predicted tendency for confidence change towards shape ($F(1,225) = 2.9$, $p = 0.09$, β.est = 0.09, SE = 0.05, CI = [-0.01, 0.19]). (Figure 7, 8). The complete analysis is presented in Table 3 and the non-significant effects are reported in the next paragraph.

Consistent with the previous analysis, there was no significant effect of the reliability manipulation where confidence change was not different depending on the reliability of the feedback ($F(1,174) = 0.14$, $p = 0.71$, β.est = 0.05, SE = 0.12, CI = [-0.19, 0.29]). There was no significant effect of the contingent manipulation where confidence change was not different whether color or shape was the correct contingent ($F(1,174) = 1.78$, $p = 0.18$, β.est = -0.16, SE = 0.12, CI = [-0.4, 0.08]). There was no significant effect of stimuli associations where confidence change was not different whether the task response options were reversed or not ($F(1,174) = 0.97$, $p = 0.33$, β.est = 0.05, SE = 0.05, CI = [-0.05, 0.15]). There was no significant interaction between surprise type and reliability ($F(1,174) = 1.05$, $p = 0.35$) where confidence change after IT surprise was consistent whether the feedback was more reliable or not (β.est = -0.18, SE = 0.13, CI = [-0.42, 0.07]), and where confidence change after Bayesian surprise was consistent whether the feedback was more reliable or not (β.est = -0.13, SE = 0.13, CI = [-0.38, 0.13]). There was no significant interaction between reliability and contingent condition where confidence change during the more reliable feedback was consistent whether the correct contingent was color or shape (F(1,174) = 0.06, p = 0.8, β.est = -0.04, SE = 0.15, CI = [-0.33, 0.25]). Lastly, There was no three way interaction between surprise type, reliability and contingent condition (F(1,174) = 1.16, p = 0.31) where confidence change after IT (β.est = 0.15, SE = 0.13, CI = [-0.11, 0.4]) and Bayesian surprise (β.est = 0.2, SE = 0.14, CI = [-0.07, 0.47]) was consistent regardless of reliability and the contingent feature.

Table 3. F-Statistics, Standardized Estimates, Standardized Errors and Confidence Intervals (95%) from the analysis of the change in explicit confidence in response to surprise type and the rest of our manipulations.

	DF	F	Pr(>F)	β.est	SE	CI: Lower	CI: Upper
IT	2, 225	10.67	< 0.001	0.15	0.11	-0.08	0.37
Bayes	-	-	-	-0.37	0.12	-0.6	-0.14
IT:Shape	2, 225	24.76	< 0.001	-0.18	0.13	-0.43	0.07
Bayes:Shape	2, 225	24.76	< 0.001	0.68	0.13	0.43	0.94
Session	1, 255	2.9	0.09	0.09	0.05	-0.01	0.19
90%	1, 225	0.14	0.71	0.05	0.12	-0.19	0.29
Shape	1, 225	1.78	0.18	-0.16	0.12	-0.4	0.08
Stimuli associations	1, 225	0.97	0.33	0.05	0.05	-0.05	0.15
IT:90%	2, 225	1.05	0.35	-0.18	0.13	-0.42	0.07
Bayes:90%	-	-	-	-0.13	0.13	-0.38	0.13
90%:Shape	1, 225	0.06	0.8	-0.04	0.15	-0.33	0.25
IT:90%:Shape	2, 225	1.16	0.31	0.15	0.13	-0.11	0.4
Bayes:90%:Shape	-	-	-	0.2	0.14	-0.07	0.47

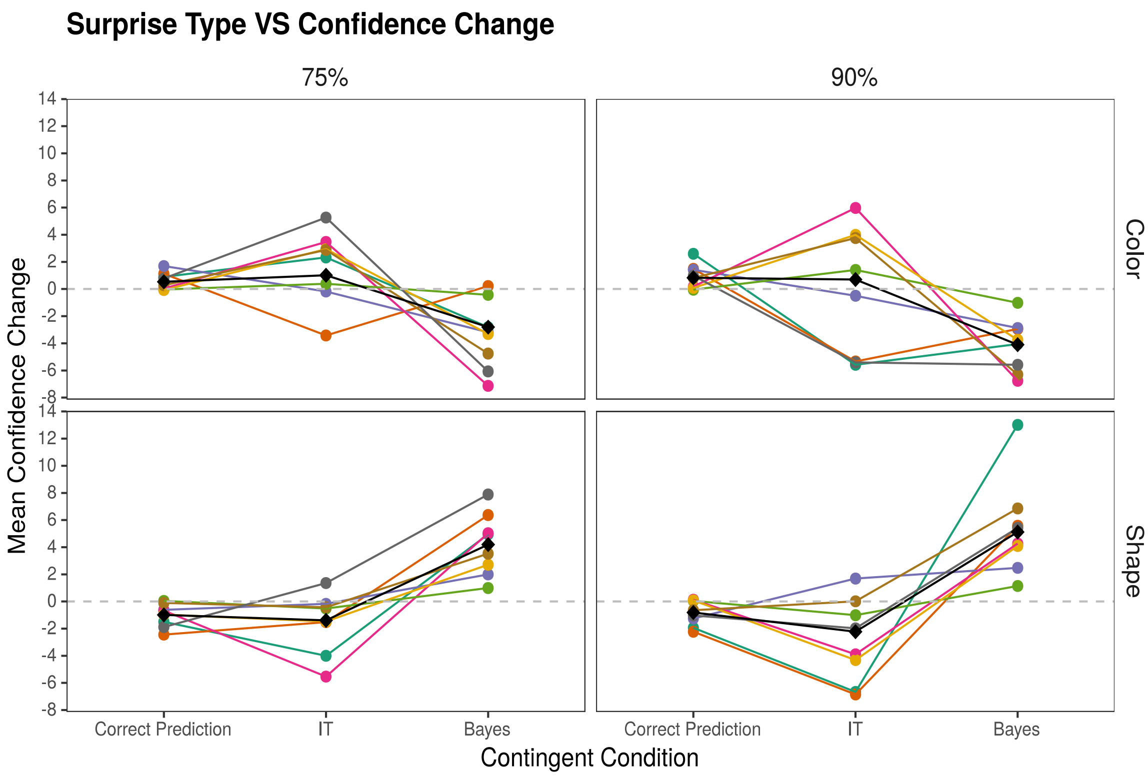

Figure 7. The mean confidence change after receiving feedback. Each color represents a participant and black represents the average across all participants. The columns separate the reliability conditions and the rows separate contingent condition. We observe that Bayesian surprise is associated with confidence change towards the correct contingent. Correct predictions does not tend to change confidence while IT surprise predict varied, sometimes erroneous, confidence change.

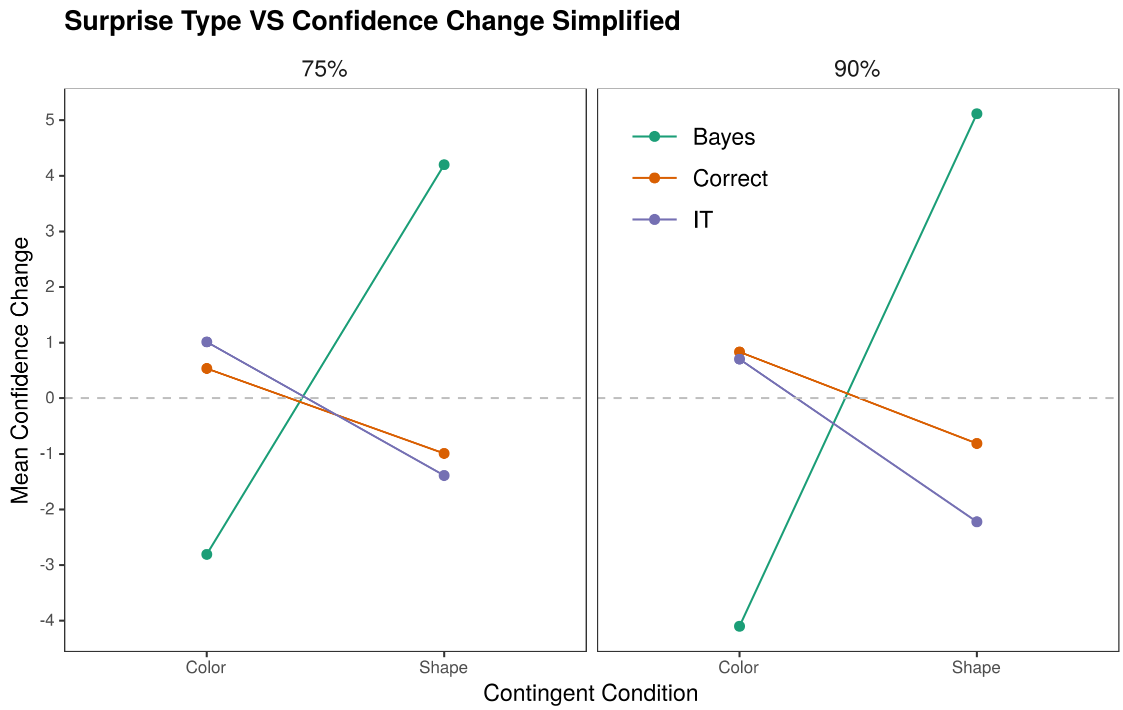

Figure 8. Simplified version of Figure 7 with no representation of participants.

Reaction time as implicit confidence. RT was measured as the time elapsed between the presentation of the response screen and the first press of the mouse on either a direction button or the confidence slider. To test whether the reliability manipulation affected implicit confidence, a linear mixed effects regression analysis was performed with RT as the dependent variable and reliability condition, contingent condition, their interactions, sessions, and stimuli associations as fixed effect independent variables, with random intercepts among sessions within participants. Only reliability condition was modeled with random slopes to avoid singular fit warning. The analysis revealed no significant effect of reliability condition on RT where RT was consistent whether the feedback was more reliable or not ($F(1,22.2) = 2.602$, $p = 0.12$, $\beta.est = 0.001$, $SE = 0.027$, $CI = [-0.05, 0.05]$). However, there was a significant interaction between reliability condition and contingent condition where RT was slower during more reliable conditions and when shape was the contingent feature ($F(1,4381.1) = 8.14$, $p < 0.01$, β.est = 0.067, SE = 0.024, CI = [0.021, 0.113]). There was a significant effect of reversed stimulus association where participants who experienced reversed stimuli associations tended to show slower RT ($F(1,20.6) = 6.61$, $p < 0.05$, β.est = 0.186, SE = 0.072, CI = [0.044, 0.327]). There was a significant effect of session where RT was quicker with successive performance of the task ($F(1,20.5) = 17.32$, $p < 0.001$, β.est = -0.3, SE = 0.07, CI = [-0.44, -0.16]). Lastly, there was no significant effect of contingent condition where RT was consistent whether color or shape was the correct contingent feature ($F(1,4375.3) = 0.016$, $p < 0.9$, β.est = -0.04, SE = 0.02, CI = [-0.07, 0.002]) (Figure 9).

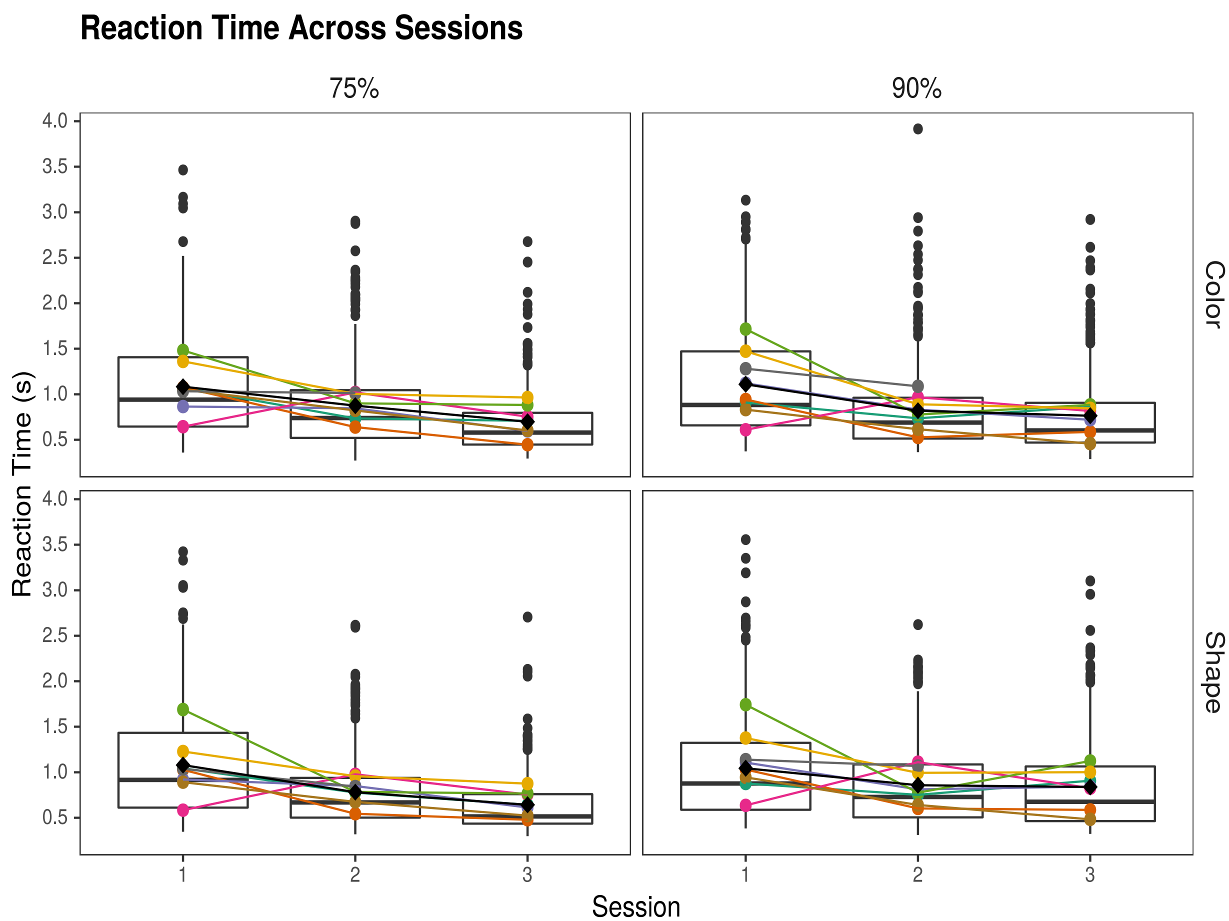

Figure 9. We observed that there was no significant difference in reaction time between reliability condition or contingent condition. We can see that RT became faster with subsequent sessions.

Pupilometry

Peak pupil dilation and its latency was recorded as the point where the pupil diameter was the largest during the feedback section before the end of a trial. Peak pupil dilation was measured as the maximum pupil diameter during the feedback screen subtracted by the average pupil diameter during a 100ms window before the onset of the feedback screen. Peak pupil dilation latency was measured as the time elapsed before the pupil dilated to its peak size. Following recommendations by Winn et al. (2018) Trials when blinks occupied more than 50%

of the fixation and feedback screen or when the eye tracking camera failed to track the eyes were removed from analysis (180/4416 trials; 4%).

PE vs peak pupil dilation size. To analyze whether the peak pupil dilation was affected by PE, we conducted a linear mixed effects regression analysis with the peak pupil dilation as the dependent variable, and PE, reliability condition, contingent condition, their interactions, baseline pupil diameter, peak pupil dilation latency, session and stimuli associations as the fixed effect independent variable with random intercepts among sessions within participants. Only PE, reliability condition, contingent condition, and baseline pupil diameter variables were modeled with random slopes to avoid convergence and singular fit warnings. The analysis revealed that there was a significant effect of PE where PE was associated with larger pupil dilation ($F(1,21.3) = 34.26$, $p < 0.001$, $\beta.est = 0.11$, $SE = 0.03$, CI = [0.05, 0.17]). There was a significant effect of baseline pupil diameter where larger baseline diameter predicted smaller pupil dilation ($F(1,22.1) = 84.07$, $p < 0.001$, $\beta.est = -0.66$, $SE = 0.07$, CI = [-0.81, -0.52]). There was a significant effect of peak dilation latency where slower dilation were related to greater dilation ($F(1,4179.9) = 832.7$, $p < 0.001$, $\beta.est = 0.38$, $SE = 0.01$, CI = [0.35, 0.4]). There was a significant effect of stimulus association where peak dilation tended to be smaller for participants who experienced the reversed stimuli associations ($F(1,19.1) = 8.59$, $p < 0.01$, $\beta.est = -0.42$, $SE = 0.14$, CI = [-0.71, -0.14]) (Figure 10). The complete analysis is presented in Table 4 and the non-significant effects are reported in the next paragraph.

There was no significant effect of the reliability manipulation where peak pupil dilation was not different depending on the reliability of the feedback ($F(1,24.2) = 0.54$, $p = 0.47$, $\beta.est = -0.004$, $SE = 0.027$, CI = [-0.057, 0.049]). There was no significant effect of the contingent manipulation where peak pupil dilation was not different whether color or shape was the correct contingent ($F(1,28.3) = 1.1$, $p = 0.3$, β.est = -0.015, SE = 0.022, CI = [-0.060, 0.028]). There was no significant effect of the sessions where peak pupil dilation was consistent across the sessions ($F(1,19.4) = 0.74$, $p = 0.4$, β.est = -0.125, SE = 0.145, CI = [-0.410, 0.160]). There was no significant interaction between PE and reliability where peak pupil dilation after PE was consistent between the reliability manipulation ($F(1,3239.6) = 2.42$, $p = 0.12$, β.est = 0.021, SE = 0.025, CI = [-0.027, 0.069]). There was no significant interaction between PE and contingent condition where peak pupil dilation after PE was consistent whether the contingent feature was color or shape ($F(1,3976.5) = 0.01$, $p = 0.93$, β.est = -0.005, SE = 0.025, CI = [-0.054, 0.044]). There was no significant interaction between reliability and contingent condition where peak pupil dilation during more reliable feedback was consistent regardless of the correct contingent feature ($F(1,2345.6) = 0.18$, $p = 0.67$, β.est = 0.001, SE = 0.025, CI = [-0.048, 0.051]). Lastly there was no significant three way interaction between PE, reliability and contingent condition where peak pupil dilation after PE was consistent regardless of its reliability and the contingent feature ($F(1,4137) = 0.14$, $p = 0.71$, β.est = 0.009, SE = 0.025, CI = [-0.040, 0.058]).

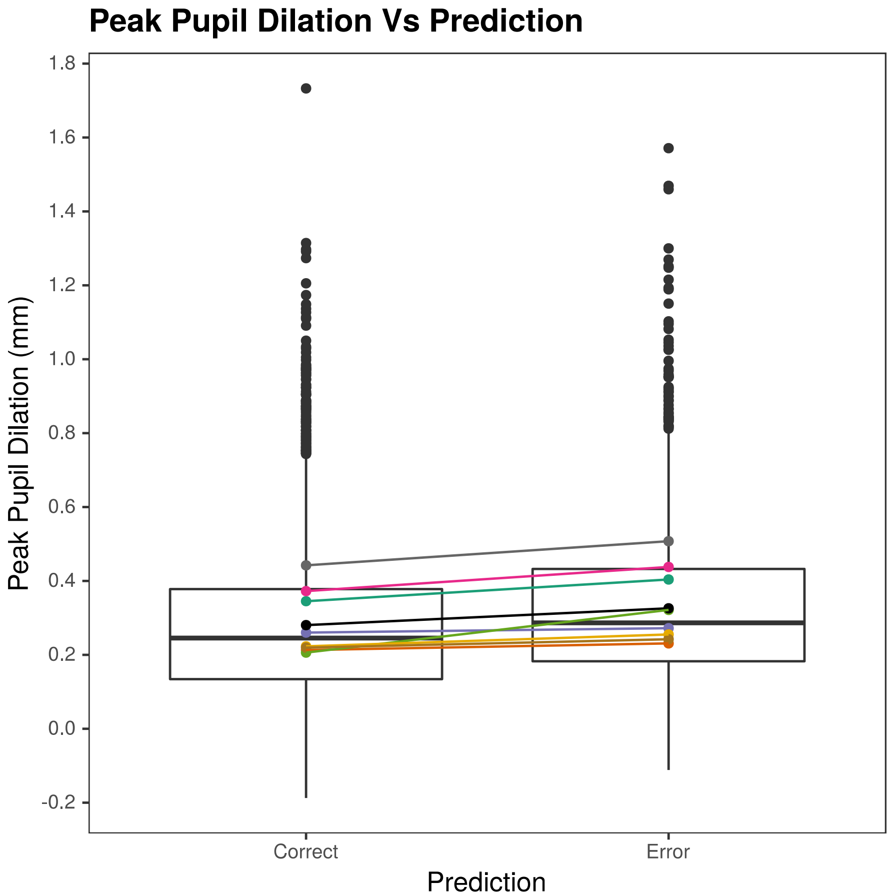

Figure 10. PE was associated with greater peak pupil dilation. Each colored lines represent a different participant and the black line represent the average over all the participants.

Table 4. F-Statistics, Standardized Estimates, Standardized Errors and Confidence Intervals (95%) from the analysis of the peak pupil dilation size in response to PE and the rest of our manipulations.

	DF	F	Pr(>F)	β.est	SE	CI: Lower	CI: Upper
PE	1, 21.3	34.26	< 0.001	0.112	0.029	0.055	0.170
Baseline	1, 22.1	84.07	< 0.001	-0.664	0.072	-0.807	-0.523
Latency	1, 4179.9	832.7	< 0.001	0.376	0.013	0.351	0.402
Stimuli associations	1, 19.1	8.59	< 0.01	-0.425	0.145	-0.709	-0.141
90%	1, 24.2	0.54	0.47	-0.004	0.027	-0.057	0.049
Shape	1, 28.3	1.1	0.3	-0.015	0.022	-0.060	0.028
Session	1, 19.4	0.74	0.4	-0.125	0.145	-0.410	0.160
PE:90%	1, 3239.6	2.42	0.12	0.021	0.025	-0.027	0.069
PE:Shape	1, 3976.5	0.01	0.93	-0.005	0.025	-0.054	0.044
90%:Shape	1, 2345.6	0.18	0.67	0.001	0.025	-0.048	0.051
PE:90%:Shape	1, 4137	0.14	0.71	0.009	0.025	-0.040	0.058

PE vs peak pupil dilation latency. Next, we performed a similar analysis with peak pupil dilation latency as the dependent variable. Other differences involved, the inclusion of peak pupil dilation as a fixed effect independent variable with random intercepts and slopes, and the elimination of random slopes for the reliability and contingent condition to avoid the singular fit warning. This left PE, baseline pupil diameter and peak dilation to be modeled with random slopes. The analysis revealed a significant effect of PE where PE related to shorter peak dilation latency ($F(1,22.4) = 12.29$, $p < 0.01$, β.est = -0.08, SE = 0.03, CI = [-0.13, -0.02]). There was a marginal effect of reliability condition where dilation latency tended to be shorter when feedback was more reliable ($F(1,4063.6) = 2.82$, $p = 0.09$, β.est = -0.01, SE = 0.02, CI = [-0.05, 0.03]). There was a significant effect of peak pupil dilation where greater pupil dilation related to slower dilation latency ($F(1,19) = 135.42$, $p < 0.001$, β.est = 0.47, SE = 0.04, CI = [0.39, 0.55]). Unlike the previous analysis, there was no significant effect of baseline pupil diameter. However, there was a tendency for dilation latency to be shorter with larger baseline pupil diameter ($F(1,18.1) = 2.28$, $p = 0.15$, β.est = -0.07, SE = 0.05, CI = [-0.17, 0.02]). Additionally, unlike the previous analysis there was no significant effect of stimulus association where dilation latency tended to be similar whether or not stimulus associations were flipped ($F(1,11.7) = 0.005$, $p = 0.95$, β.est = -0.004, SE = 0.06, CI = [-0.11, 0.11]) (Figure 11). The complete analysis is presented in Table 5 and the rest of the non-significant effects are reported in the next paragraph.

There was no significant effect of the contingent manipulation where peak dilation latency was not different whether color or shape was the correct contingent ($F(1,4162.7) = 0.45$, $p = 0.5$, β.est = -0.014, SE = 0.023, CI = [-0.06, 0.031]). There was no significant effect of the sessions where peak dilation latency was consistent across the sessions ($F(1,12) = 0.03$, $p = 0.86$, β.est = -0.01, SE = 0.057, CI = [-0.121, 0.101]). There was no significant interaction between PE and reliability where peak dilation latency after PE was consistent between the reliability manipulation ($F(1,3211.3) = 0.23$, $p = 0.63$, β.est = -0.002, SE = 0.026, CI = [-0.054, 0.049]). There was no significant interaction between PE and contingent condition where peak dilation latency after PE was consistent whether the contingent feature was color or shape ($F(1,4158) = 0.99$, $p = 0.32$, β.est = 0.026, SE = 0.027, CI = [-0.026, 0.078]). There was no significant interaction between reliability and contingent condition where peak dilation latency during more reliable feedback was consistent regardless of the correct contingent feature ($F(1,4138.9) = 0.95$, $p = 0.33$, β.est = -0.014, SE = 0.027, CI = [-0.067, 0.038]). Lastly there was no significant three way interaction between PE, reliability and contingent condition where peak dilation latency after PE was consistent regardless of its reliability and the correct contingent feature ($F(1,4155.9) = 0.14$, $p = 0.71$, β.est = -0.01, SE = 0.027, CI = [-0.062, 0.043]).

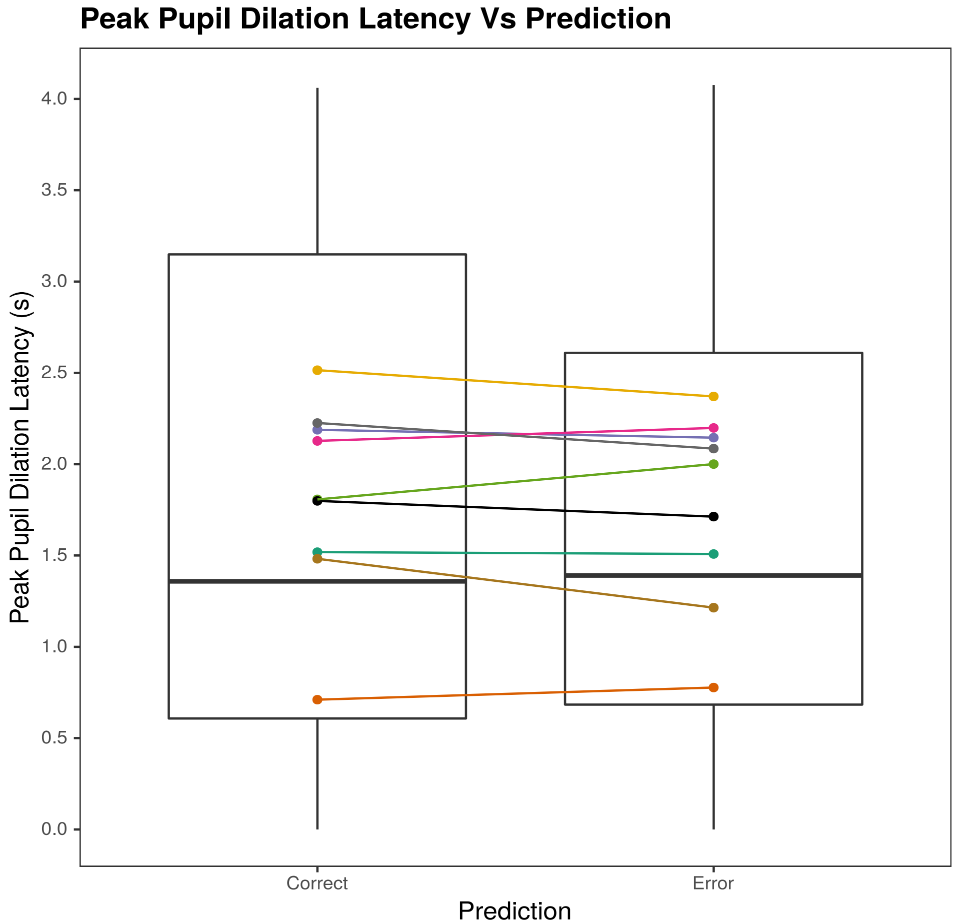

Figure 11. PE was associated with shorter peak pupil dilation latency. Each colored lines represent a different participant and the black line represent the average over all the participants.

Table 5. F-Statistics, Standardized Estimates, Standardized Errors and Confidence Intervals (95%) from the analysis of the peak pupil dilation latency in response to PE and the rest of our manipulations.

	DF	F	Pr(>F)	β.est	SE	CI: Lower	CI: Upper
PE	1, 22.4	12.29	< 0.01	-0.077	0.028	-0.132	-0.021
Pupil Dilation	1, 19	135.41	< 0.001	0.467	0.040	0.389	0.546
90%	1, 4063.6	2.82	0.09	-0.009	0.022	-0.052	0.034
Shape	1, 4162.7	0.45	0.50	-0.014	0.023	-0.060	0.031
Baseline	1, 18.1	2.28	0.15	-0.074	0.049	-0.169	0.022
Session	1, 12	0.03	0.86	-0.010	0.057	-0.121	0.101
Stimuli associations	1, 11.7	0.005	0.95	0.004	0.056	-0.107	0.114
PE:90%	1, 3211.3	0.23	0.63	-0.002	0.026	-0.054	0.049
PE:Shape	1, 4158	0.99	0.32	0.026	0.027	-0.026	0.078
90%:Shape	1, 4138.9	0.95	0.33	-0.014	0.027	-0.067	0.038
PE:90%:Shape	1, 4155.9	0.14	0.71	-0.010	0.027	-0.062	0.043

Surprise type vs peak pupil dilation size. Next, we analyzed how the two surprise type effects peak dilation. We performed a linear mixed effects regression analysis with peak dilation as the dependent variable, with surprise type, reliability condition, contingent condition and their interactions, baseline pupil diameter, peak dilation latency, session, and stimuli association as fixed effect independent variables with random intercepts among sessions within participants. We modeled random slopes for reliability condition, contingent condition, baseline pupil diameter, peak dilation latency. Surprise type and other independent variables were not able to be modeled with random slopes without facing singular fit and convergence warnings. The analysis revealed that there was a significant effect of surprise type where peak pupil dilation after experiencing an IT or Bayesian surprise was statistically different from dilation after correct predictions ($F(1,4132.7) = 54.1$, $p < 0.001$), IT surprise predicted greater pupil dilation than correct predictions (β.est = 0.06, SE = 0.02, CI = [0.02, 0.1]) and Bayesian surprise also predicted even greater pupil dilation (β.est = 0.11, SE = 0.02, CI = [0.06, 0.15]). There was a significant effect of baseline pupil diameter where peak dilation was smaller with larger baseline pupil diameter ($F(1,21.4) = 79.2$, $p < 0.001$, β.est = -0.7, SE = 0.08, CI = [-0.86, -0.55]). There was a significant effect of peak dilation latency where pupil dilation tended to be larger when dilation latency was greater ($F(1,20.7) = 92.42$, $p < 0.001$, β.est = 0.34, SE = 0.04, CI = [0.27, 0.41]). There was marginal effect of the interaction between surprise type and reliability condition where there was a marginal tendency for peak dilation after Bayesian surprise to be larger when feedback was more reliable ($F(1,4152.3) = 2.64$, $p < 0.07$, β.est = 0.03, SE = 0.02, CI = [-0.02, 0.07]) (Figure 12). The complete analysis is presented in Table 6 and the rest of the non-significant effects are reported in the next paragraph.

There was no significant effect of the reliability manipulation where peak pupil dilation was not different depending on the reliability of the feedback ($F(1,39.1) = 1.21$, $p = 0.28$, $\beta.est = -0.005$, $SE = 0.026$, $CI = [-0.056, 0.046]$). There was no significant effect of the contingent manipulation where peak pupil dilation was not different whether color or shape was the correct contingent ($F(1,68.8) = 0.2$, $p = 0.66$, $\beta.est = -0.011$, $SE = 0.022$, $CI = [-0.055, 0.033]$). There was no significant effect of the sessions where peak pupil dilation was consistent across the sessions ($F(1,19) = 0.15$, $p = 0.7$, $\beta.est = -0.053$, $SE = 0.135$, $CI = [-0.317, 0.212]$). There was no significant effect of stimuli associations where peak pupil dilation was consistent despite the response reversal ($F(1,18.9) = 2.44$, $p = 0.14$, $\beta.est = -0.210$, $SE = 0.135$, $CI = [-0.475, 0.054]$). There was no significant interaction between surprise type and contingent condition ($F(2,4092.9) = 0.94$, $p = 0.39$) where peak pupil dilation after IT surprise ($\beta.est = 0.008$, $SE = 0.021$, $CI = [-0.032, 0.049]$) or Bayesian surprise ($\beta.est = -0.029$, $SE = 0.025$, $CI = [-0.078, 0.019]$) was consistent whether color or shape was the correct contingent feature. There was no significant interaction between reliability and contingent condition where peak pupil dilation during the more reliable feedback was consistent whether the correct contingent was color or shape ($F(1,3343.9) = 0.60$, $p = 0.44$, $\beta.est = -0.001$, $SE = 0.025$, $CI = [-0.05, 0.05]$). There was no three way interaction between surprise type, reliability and contingent condition ($F(2,4152.8) = 0.36$, $p = 0.7$) where peak pupil dilation after IT ($\beta.est = 0.008$, $SE = 0.021$, $CI = [-0.033, 0.049]$) and Bayesian surprise ($\beta.est = 0.02$, $SE = 0.025$, $CI = [-0.028, 0.068]$) was consistent regardless of reliability and the contingent feature.

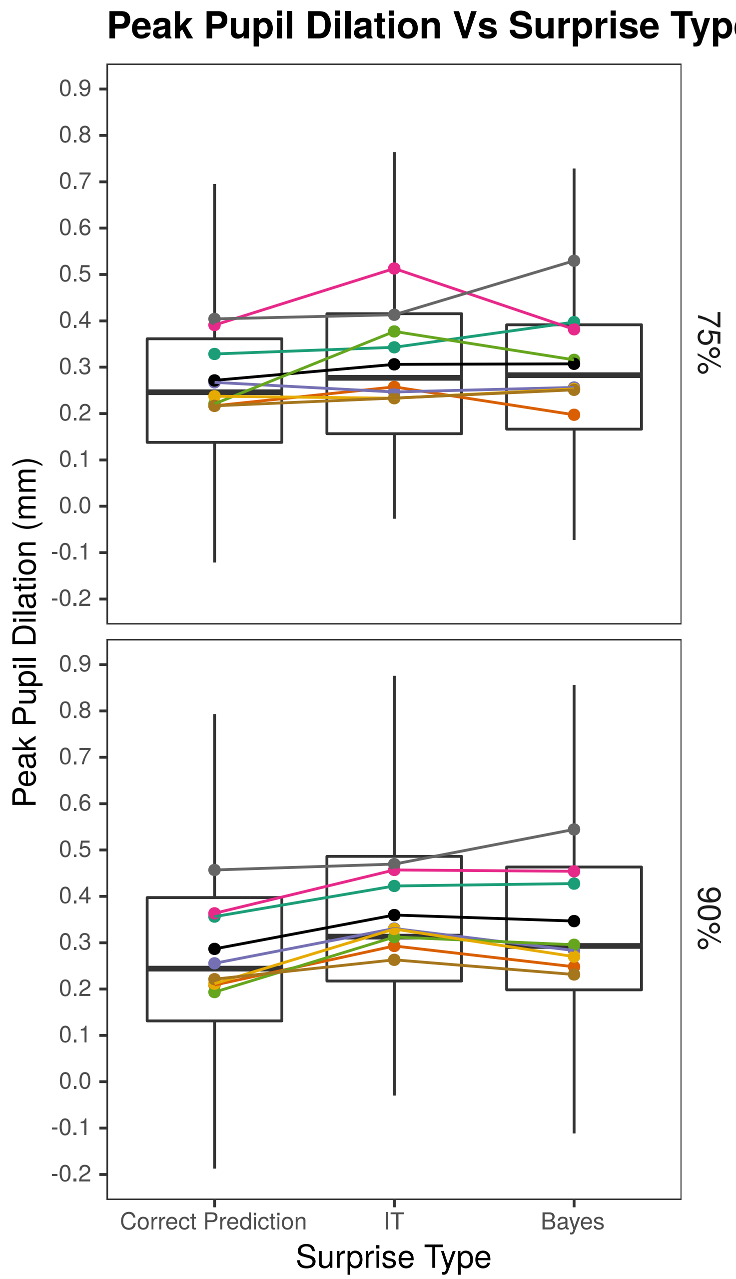

Figure 12. Both IT and Bayesian surprise predict greater pupil dilation than correct predictions. The difference in peak pupil dilation between IT and Bayesian surprise is marginal. There was a marginal interaction between Bayesian surprise and reliability where peak pupil dilation after Bayesian surprise was greater when feedback was more reliable (90% instead of 75%). Each colored lines represent a different participant and the black line represent the average over all the participants.

Table 6. F-Statistics, Standardized Estimates, Standardized Errors and Confidence Intervals (95%) from the analysis of the peak pupil dilation size in response to surprise types and the rest of our manipulations.

	DF	F	Pr(>F)	β.est	SE	CI: Lower	CI: Upper
IT	2, 4132.7	54.08	< 0.001	0.055	0.020	0.016	0.095
Bayes	-	-	-	0.105	0.023	0.059	0.151
Baseline	1, 21.4	79.21	< 0.001	-0.703	0.079	-0.858	-0.548
Dilation Latency	1, 20.7	92.42	< 0.001	0.343	0.036	0.273	0.413
IT:90%	2, 4152.3	2.64	0.07	0.004	0.020	-0.036	0.045
Bayes:90%	-	-	-	0.026	0.024	-0.022	0.073
90%	1, 39.1	1.21	0.28	-0.005	0.026	-0.056	0.046
Shape	1, 68.8	0.20	0.66	-0.011	0.022	-0.055	0.033
Session	1, 19	0.15	0.70	-0.053	0.135	-0.317	0.212
Stimuli associations	1, 18.9	2.44	0.14	-0.210	0.135	-0.475	0.054
IT:Shape	2, 4092.9	0.94	0.39	0.008	0.021	-0.032	0.049
Bayes:Shape	-	-	-	-0.029	0.025	-0.078	0.019
90%:Shape	1, 3343.9	0.60	0.44	-0.001	0.025	-0.050	0.049
IT:90%:Shape	2, 4152.8	0.36	0.70	0.008	0.021	-0.033	0.049
Bayes:90%:Shape	-	-	-	0.020	0.025	-0.028	0.068

Surprise type vs peak pupil dilation latency. Lastly, we performed the similar analysis, with peak dilation latency as the dependent variable, the peak dilation as an independent variable. We only modeled contingent condition, baseline pupil diameter and peak dilation with random slopes to avoid convergence and singular fit warnings. The analysis revealed that there was a significant effect of surprise type ($F(1,1312.55) = 5.13$, $MSE = 4.46$, $p = 0.02$) where Bayesian surprise predicted shorter peak dilation latency than correct predictions (β.est = -0.07, SE = 0.03, CI = [-0.13, -0.007]), while IT surprise did not predict significantly different peak dilation latencies than correct predictions (β.est = -0.02, SE = 0.02, CI = [-0.06, 0.02]). Consistent with the previous analysis, there was a significant effect of peak dilation size where dilation latency was slower for greater pupil dilation ($F(1,18.6) = 155.3$, $p < 0.001$, β.est = 0.46, SE = 0.04, CI = [0.39, 0.53]) and there was no significant effect of baseline pupil diameter ($F(1,18.1) = 2.45$, $p = 0.13$, β.est = -0.08, SE = 0.05, CI = [-0.17, 0.02]) (Figure 13, 14). The complete analysis is presented in Table 7 and the rest of the non-significant effects are reported in the next paragraph.

There was no significant effect of the reliability manipulation where peak dilation latency was not different depending on the reliability of the feedback ($F(1,4136.7) = 0.84$, $p = 0.36$, β.est = -0.009, SE = 0.022, CI = [-0.052, 0.034]). There was no significant effect of the contingent manipulation where peak dilation latency was not different whether color or shape was the correct contingent ($F(1,544.8) = 0.11$, $p = 0.75$, β.est = -0.016, SE = 0.023, CI = [-0.062, 0.030]). There was no significant effect of the sessions where peak dilation latency was consistent across the sessions ($F(1,12.7) = 0.001$, $p = 0.97$, β.est = 0.002, SE = 0.057, CI = [-0.109, 0.113]). There was no significant effect of stimuli associations where peak dilation latency was consistent despite the response reversal ($F(1,12.4) = 0.09$, $p = 0.76$, β.est = 0.017, SE = 0.057, CI = [-0.094, 0.128]). There was no significant interaction between surprise type and reliability condition ($F(2,4165.2) = 0.12$, $p = 0.88$) where peak dilation latency after IT surprise (β.est = -0.001, SE = 0.022, CI = [-0.044, 0.043]) and Bayesian surprise (β.est = 0.012, SE = 0.026, CI = [-0.039, 0.064]) was consistent regardless of how reliable the feedback was. There was no significant interaction between surprise type and contingent condition ($F(2,4161.7) = 1.23$, $p = 0.29$) where peak dilation latency after IT surprise (β.est = -0.003, SE = 0.022, CI = [-0.047, 0.041]) and Bayesian surprise (β.est = 0.048, SE = 0.027, CI = [-0.005, 0.1]) was consistent whether color or shape was the correct contingent feature. There was no significant interaction between reliability and contingent condition where peak dilation latency during the more reliable feedback was consistent whether the correct contingent was color or shape ($F(1,4161.7) = 0.53$, $p = 0.46$, β.est = -0.014, SE = 0.027, CI = [-0.067, 0.039]). Lastly, there was no three way interaction between surprise type, reliability and contingent condition (F(2,4169.6) = 0.64, p = 0.53) where peak dilation latency after IT (β.est = 0.007, SE = 0.023, CI = [-0.037, 0.052]) and Bayesian surprise (β.est = -0.027, SE = 0.027, CI = [-0.079, 0.025]) was consistent regardless of reliability and the contingent feature.

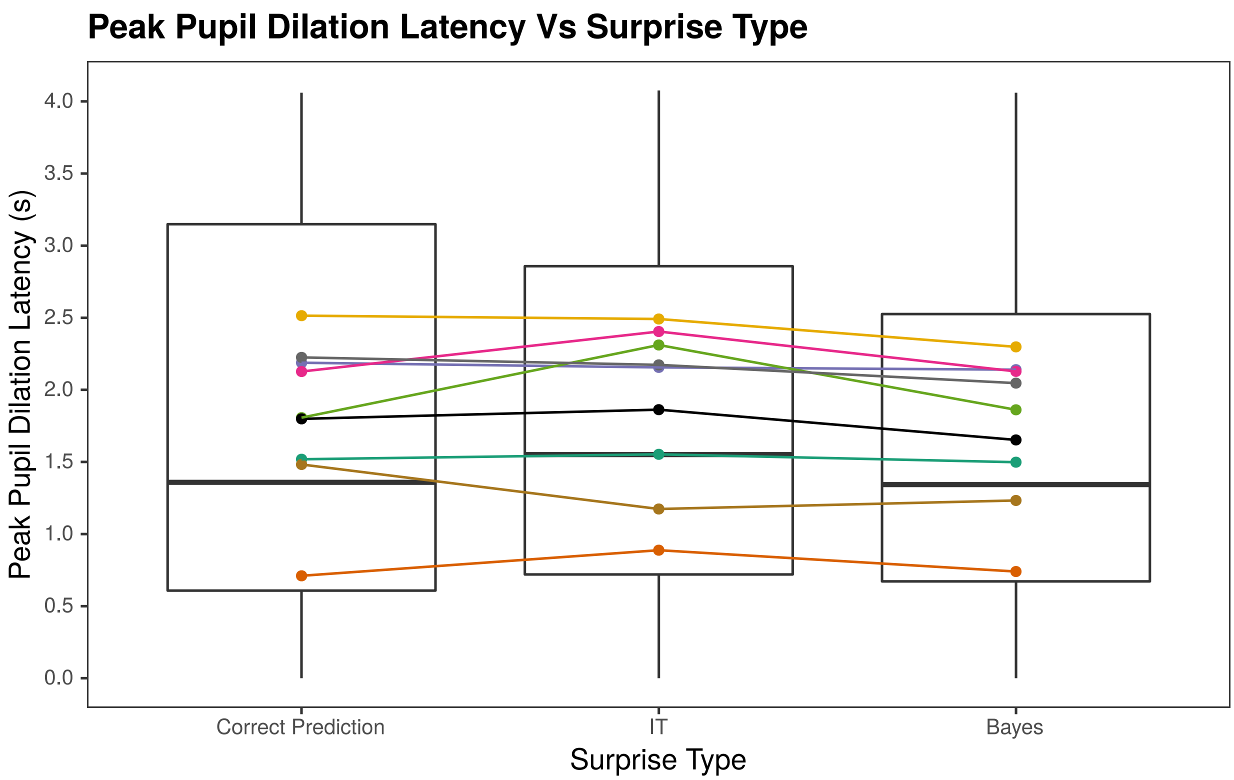

Figure 13. Unlike IT Bayesian surprise predicted faster peak pupil dilation latencies. Each colored lines represent a different participant and the black line represent the average over all the participants.

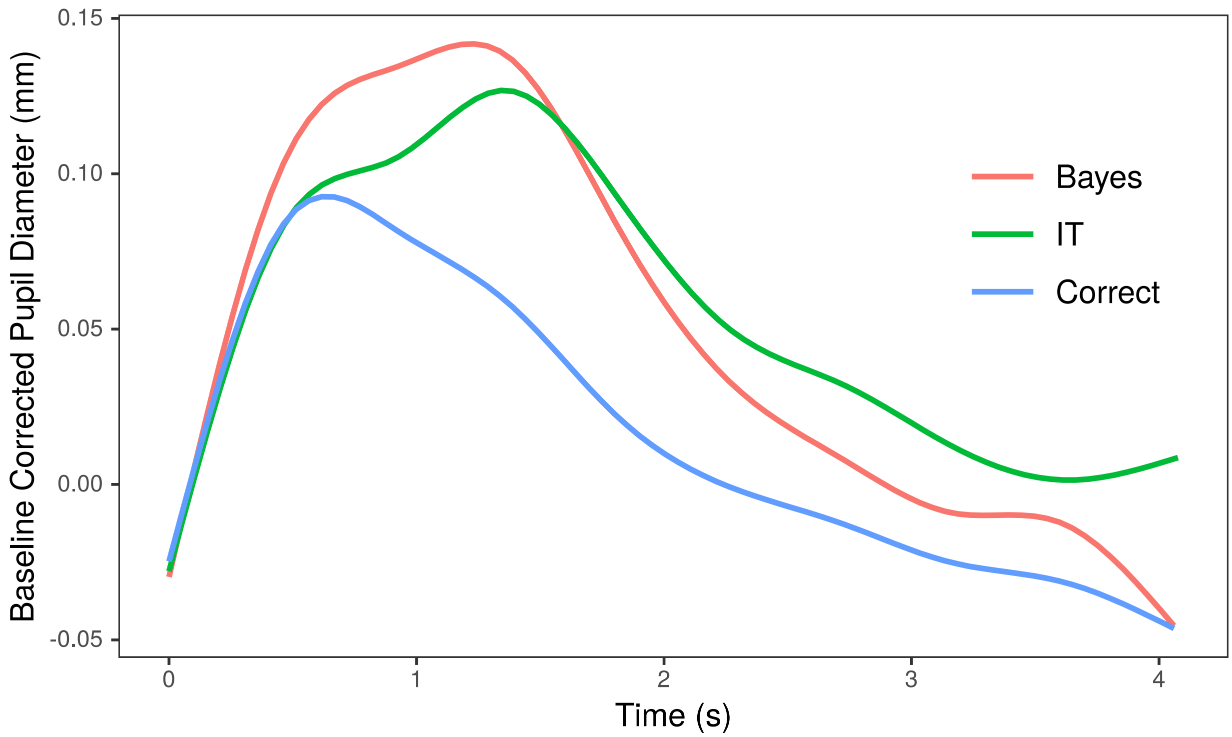

Figure 14. Time series baseline corrected pupil diameter during the feedback screen. Feedback tone was played at 0 second and each colored line represents a surprise type. We can see that Bayesian surprise is associated with a faster rise in pupil diameter.

Table 7. F-Statistics, Standardized Estimates, Standardized Errors and Confidence Intervals (95%) from the analysis of the peak pupil dilation latency in response to surprise types and the rest of our manipulations.

	DF	F	Pr(>F)	β.est	SE	CI: Lower	CI: Upper
IT	2, 4175.0	13.130	< 0.001	-0.022	0.022	-0.065	0.021
Bayes	-	0.000	-	-0.092	0.025	-0.142	-0.043
Peak Dilation	1, 18.6	155.320	< 0.001	0.465	0.037	0.392	0.538
90%	1, 4136.7	0.836	0.36	-0.009	0.022	-0.052	0.034
Shape	1, 544.8	0.105	0.75	-0.016	0.023	-0.062	0.030
Baseline	1, 18.1	2.454	0.13	-0.076	0.048	-0.170	0.019
Session	1, 12.7	0.001	0.97	0.002	0.057	-0.109	0.113
Stimuli associations	1, 12.4	0.093	0.76	0.017	0.057	-0.094	0.128
IT:90%	2, 4165.2	0.123	0.88	-0.001	0.022	-0.044	0.043
Bayes:90%	-	0.000	-	0.012	0.026	-0.039	0.064
IT:Shape	2, 4161.7	1.226	0.29	-0.003	0.022	-0.047	0.041
Bayes:Shape	-	0.000	-	0.048	0.027	-0.005	0.100
90%:Shape	1, 4161.7	0.535	0.46	-0.014	0.027	-0.067	0.039
IT:90%:Shape	2, 4169.6	0.637	0.53	0.007	0.023	-0.037	0.052
Bayes:90%:Shape	-	-	-	-0.027	0.027	-0.079	0.025

Discussion

The goal of this study was to explore the potential link between mental model updating and pupil dilation in light of the neural connections between brain regions implicated in mental model updating and pupil dilation (Van Slooten et al., 2018). In addition, we explored the relationship between mental model updating and confidence where it has been suggested that confidence plays a role in controlling the degree in which we update our mental models (Meyniel & Dehaene, 2017). We achieved this goal by implementing a probabilistic learning task where participants could inform their decisions by experiencing prediction errors (PE) in the form of information theoretic (IT) surprise or Bayesian surprise (Schwartenbeck et al., 2016). In support of past work, we hypothesized that i) PE will predict greater pupil response than correct predictions, ii) Bayesian surprise will predict greater pupil response than IT surprise, iii) being exposed to a more stochastic environment will lead to weaker explicit confidence, iv) being exposed to a more stochastic environment will lead to weaker pupil response to Bayesian surprise, v) Bayesian surprise will predict greater change in mental model in the form of explicit confidence report, and vi) the degree of confidence change will be greater when feedback is more reliable.

Pupilometry

Our results supported our hypothesis where we observed greater pupil dilation in response to PE. This result is in line with previous studies that suggest pupil dilation is sensitive to PE and that the noradrenergic LC, which is associated with pupil dilation, is sensitive to uncertainties and is responsive to unexpected events (O'Reilly et al., 2013; Urai et al., 2017). However, we were not able to support our hypothesis relating to Bayesian surprise, where we could not differentiate between pupil dilation magnitude in response to IT surprise and Bayesian surprise. While past literature suggests that pupil dilation may be sensitive to neural activities implicated in mental model updating, it may be the case that the magnitude of pupil dilation is not the component that is sensitive to those activities. Another possible contribution to this negative result may be our lack of control for participant's gaze. As pupil diameter is sensitive to luminance and saccades (Winn et al., 2018), if the participants were not focusing their gaze on the central fixation point or they made eye movements during the feedback period of our task, there may be confounding pupil responses that drown out the effect of our surprise manipulation. These effects are well documented in pupilometry studies and it is advised to control for them by filtering out data points that involve gaze that are outside an area of interest (Winn et al., 2018).

Additionally, we could not support our hypothesis that more reliable Bayesian surprises are related to greater pupil dilation. While Urai et al. (2017) and Colizoli et al. (2018) suggest that pupil dilation is sensitive to prediction errors in response to more or less difficult random dot kinometograms, we could not replicate their result using our paradigm. It is possible that the manipulation of the difficulty in a visual task is more noticeable than our manipulation of feedback reliability. If this is the case, participants performing the kinometogram task may be able to, more easily, infer that their PE was due to chance instead of questioning their ability to identify dot movement directions. Future, studies using a similar paradigm as ours may wish to record whether participants are aware of their current reliability condition or make the manipulation more noticeable. However, this bring into question whether task difficulty's effect on pupil response is reliant on the conscious awareness of the task difficulty; the reliability of a feedback in our case. It may be of value to investigate whether this is the case.

While our investigation into pupil dilation magnitude led to a negative result, we observed an unexpected marginal relationship between surprise types and pupil dilation latency where Bayesian surprise was associated with faster pupil dilation than IT surprise. Past studies pertaining to pupil dilation latency suggests that dilation latency reflects task difficulty and that it is a type of orienting response, signaling a need for greater cognitive resources. Simpson et al. (1968) observed faster pupil dilation latency during a mental imagery task when the noun being imagined were considered to be more difficult. In a different study by Koelewijn et al. (2015), participants were tasked to predict where the next speaker will speak from. Some participants attended to speech projecting from a fixed location while others attended to speech projecting from random locations. They found that pupil dilation latency was quicker in response to listening to speech projecting from random locations. They suggest that processing speech from an unexpected direction, akin to a cocktail party setting, induces greater cognitive load and difficulty. From a non-human primate (NHP) study by Hampson et al. (2010), the investigators suggest that fast pupil dilations and its neural correlates are associated with processing of valuable visual features that facilitate the performance of a delayed-match-to-sample task. They observed faster pupil dilations when NHPs were shown a sample image and when the NHPs searched for that image among distractors.

Because mental model updating is not only reliant on surprising observations but also their epistemic value, it may be possible for Bayesian surprise to predict faster pupil dilation latency. Unlike IT surprises, which are surprising but uninformative, Bayesian surprises are both surprising and contain epistemic value, and they can elicit counter-factual thoughts which requires cognitive effort. Similar responses that involve both surprisal and cognitive load should be of interest for investigating mental model updating. Additionally, future brain imaging studies may wish to investigate the correlates for faster pupil latency as its literature is relatively sparse.

Another observed feature of pupil dilation latency is its relative insensitivity of baseline pupil diameter compared to the peak dilation magnitude where if the baseline diameter is smaller the resulting dilation tends to be larger. However, dilation latency did not experience this constraint where regardless of the baseline pupil diameter, the time it took to reach the peak dilation was better explained by the prediction state or by the magnitude of the peak dilation. However, we are cautious of making any strong claims because if we observe the estimates and the confidence intervals, it may be the case that with a larger sample size we can observe a statistically significant relationship between baseline pupil diameter and peak dilation latency. This relationship is possible considering the significant relationship between peak dilation latency and peak dilation magnitude and the significant relationship between peak dilation magnitude and baseline pupil diameter.

Another unforeseen relationship was between stimuli associations and pupil dilation magnitude where pupil dilation tended to be smaller for participants who experienced the reverse coded stimuli. A notable change from the reversal is the associations of the feedback tones. Normally, participants heard a high tone (700Hz) for correct predictions and a low tone (400Hz) for PE; this association was reversed for half the participants. If it is the case that most audio feedback follow the convention where higher tones are related to positive or ‘good’ feedback and low tones are related to negative or ‘bad’ feedback, the reversal of this convention may lead to a change in the arousal related to the feedback and possibly reflect on pupil dilation. Foregoing the reverse coding of the feedback tone may be desirable for future studies.

Behavioral Paradigm

Overall, our manipulations in our experimental task performed partially as we intended and we observed behaviors that supported some of our hypotheses. We demonstrated prediction errors relate to changes in confidence towards the correct contingent feature, supporting past literature (Filipowicz et al., 2016; Nassar et al., 2010; Reisenzein et al., 2019; Schwartenbeck et al., 2016; Sutton & Barto, 2018). Additionally, we observed that the effect of IT surprise was not discernible from the that of correct predictions while Bayesian surprise was. This suggests that IT surprise have little contradictory information against the participant's past belief. These results suggest that our implementation of IT surprise and Bayesian surprise was successful. In addition, we observed that the degree of explicit confidence change increased following more reliable PE; supporting our hypothesis. We observed similar results when examining the effect of surprise type where Bayesian surprise predicted greater confidence change than IT surprise. However, we did not observe our expected effect of reliability manipulation on participant's explicit and implicit confidence level and on the degree of explicit confidence change after experiencing a Bayesian surprise. While we expected the absolute confidence ratings to be higher, RT to be faster with more reliable feedback, and explicit confidence change to be greater with more reliable Bayesian surprise, this was not the case.

Similar to our explanation as to why our reliability manipulation did not have the intended effect on the pupil dilation magnitude, it may be the case that the effect of task reliability requires a more obvious indication in order to have a similar effect seen with difficulty manipulations from other studies. However, an obvious confound exists to why our RT measure may not be sensitive to our reliability manipulation. The design of our task had it so that when the response screen appeared, the mouse cursor's position was set to the center of the screen which is also the mid-point of the confidence slider (the maximally uncertain position). As such, participant with lower confidence did not have to move their mouse far in order to make their response. This results in a RT advantage for situations when participants have low confidence. A scenario where the potential disadvantage of lower confidence is compensated by shorter mouse movement. This issue can be avoided by recording the RT as the initiation time of the movement or click of the mouse or by designing a confidence bar where every confidence level is equidistant from the cursor's starting point. In addition, participants had an ample amount of time to deliberate about their confidence and decision before the onset of the next response screen. Participants were given eight seconds after the onset of the feedback tone for the participant to decide what their confidence report will be and what contingent feature they will base their next decision on. It may be possible to observe our intended effect of confidence on RT if the time between the audio feedback and the next response screen was shorter.

Another unexpected observation was the significant effect of contingent condition on confidence rating where participants tended to have less confidence when shape was the contingent feature. A possible explanation may reside in how participants identify our visual stimuli. When participants observe our visual stimulus, they would typically identify it as shapes possessing a color, a ‘colored shape’, a green-square or a blue-circle, the shape is the subject with the adjective color. It is not often the case, as an English speaker, that the color is the subject with an adjective of shape (e.g., a square-green or a circular-blue). Since color is spoken first participants may tend to be anchored to it as their default contingent and attempt to disprove it. This may result in a situation where when color is the true contingent, participants may feel like normalcy has returned and show more confidence. However, when shape is the correct contingent they feel like it is an abnormality, leading to a more cautious report of confidence. If this is the case, it may be worth devising stimuli where the contingent features are equally salient.

Similar to our observation of pupil dilation magnitude, we observed an unexpected significant relationship between stimuli associations and RT where RT tended to be slower for participants who experienced the reverse coded stimuli. A notable change in the reversal is the polarity of the contingent feature on the confidence rating slider. In the original task, the left side of the confidence slider was associated with ‘color’. This was flipped so that the left side of the confidence slider was associated with ‘shape’. A possible explanation may be a similar one to that of the effect of contingent feature on confidence. Since English speakers often speak of the adjective color before the shape, it may be more intuitive to associate the left side of the confidence slider with color and the right side with shape. Since this is the opposite when reverse coding, the inconsistency may contribute to a lag in the planning process of the mouse movement, regardless of whether participants report direction or confidence first, contributing to the slower RT. If this is the case, it may be worth devising stimuli where the order of the spoken contingent are interchangeable. However, the linear nature of the English language may be a limiting factor. Or it may be advisable not to reverse the confidence slider and keep it consistent for all participants.

In conclusion, the results of our study provides insight into how surprises contribute to the change in pupils and to mental models. In particular, we supported the hypothesis that Bayesian surprises tend to predict changes in belief, more than IT surprises. Additionally, while we could not differentiate pupil dilation magnitude between IT and Bayesian surprises, we found that peak pupil dilation latency may be sensitive to Bayesian surprise or other underlying mental model updating processes. While more studies are needed, this study makes a preliminary suggestion that pupilometry is an accessible method for studying mental model updating.

References

Bates, D., Kliegl, R., Vasishth, S., & Baayen, H. (2015). *Parsimonious Mixed Models*.

Bates, D., Mächler, M., Bolker, B., & Walker, S. (2015). Fitting Linear Mixed-Effects Models Using lme4. *Journal of Statistical Software*, *67*(1), 1–48. https://doi.org/10.18637/jss.v067.i01

Beck, J. M., Ma, W. J., Kiani, R., Hanks, T., Churchland, A. K., Roitman, J., Shadlen, M. N., Latham, P. E., & Pouget, A. (2008). Probabilistic Population Codes for Bayesian Decision Making. *Neuron*, *60*(6), 1142–1152. https://doi.org/10.1016/j.neuron.2008.09.021

Boldt, A., de Gardelle, V., & Yeung, N. (2017). The impact of evidence reliability on sensitivity and bias in decision confidence. *Journal of Experimental Psychology: Human Perception and Performance*, *43*(8), 1520–1531. https://doi.org/10.1037/xhp0000404

Colizoli, O., de Gee, J. W., Urai, A. E., & Donner, T. H. (2018). Task-evoked pupil responses reflect internal belief states. *Scientific Reports*, *8*(1), 13702. https://doi.org/10.1038/s41598-018-31985-3

Filipowicz, A., Anderson, B., & Danckert, J. (2016). Adapting to change: The role of the right hemisphere in mental model building and updating. *Canadian Journal of Experimental Psychology/Revue Canadienne de Psychologie Expérimentale*, *70*(3), 201–218. https://doi.org/10.1037/cep0000078

Garrison, J., Erdeniz, B., & Done, J. (2013). Prediction error in reinforcement learning: A meta-analysis of neuroimaging studies. *Neuroscience & Biobehavioral Reviews*, *37*(7), 1297–1310. https://doi.org/10.1016/j.neubiorev.2013.03.023

Hampson, R. E., Opris, I., & Deadwyler, S. A. (2010). Neural correlates of fast pupil dilation in nonhuman primates: Relation to behavioral performance and cognitive workload. *Behavioural Brain Research*, *212*(1), 1–11. https://doi.org/10.1016/j.bbr.2010.03.011

Johnson-Laird, P. N. (2013). Mental models and cognitive change. *Journal of Cognitive Psychology*, *25*(2), 131–138. https://doi.org/10.1080/20445911.2012.759935

Koelewijn, T., de Kluiver, H., Shinn-Cunningham, B. G., Zekveld, A. A., & Kramer, S. E. (2015). The pupil response reveals increased listening effort when it is difficult to focus attention. *Hearing Research*, *323*, 81–90. https://doi.org/10.1016/j.heares.2015.02.004

Lüdecke, D. (2019). *sjstats: Statistical Functions for Regression Models (Version 0.17.7)*. https://doi.org/10.5281/zenodo.1284472

Mazzoni, D. (2019). *Audacity® software is copyright © 1999-2019 Audacity Team. Web site: Https://audacityteam.org/. It is free software distributed under the terms of the GNU General Public License. The name Audacity® is a registered trademark of Dominic Mazzoni.*

Meyniel, F., & Dehaene, S. (2017). Brain networks for confidence weighting and hierarchical inference during probabilistic learning. *Proceedings of the National Academy of Sciences*, *114*(19), E3859–E3868. https://doi.org/10.1073/pnas.1615773114

Meyniel, F., Sigman, M., & Mainen, Z. F. (2015). Confidence as Bayesian Probability: From Neural Origins to Behavior. *Neuron*, *88*(1), 78–92. https://doi.org/10.1016/j.neuron.2015.09.039

Moritz, S., & Bartz-Beielstein, T. (2017). ImputeTS: Time Series Missing Value Imputation in R. *The R Journal*, *9*(1), 207–218. https://doi.org/10.32614/RJ-2017-009

Nash, J. C., & Varadhan, R. (2011). Unifying Optimization Algorithms to Aid Software System Users: Optimx for R. *Journal of Statistical Software*, *43*(9), 1–14.

Nassar, M. R., Wilson, R. C., Heasly, B., & Gold, J. I. (2010). An Approximately Bayesian Delta-Rule Model Explains the Dynamics of Belief Updating in a Changing Environment. *Journal of Neuroscience*, *30*(37), 12366–12378. https://doi.org/10.1523/JNEUROSCI.0822-10.2010

O'Reilly, J. X., Schuffelgen, U., Cuell, S. F., Behrens, T. E. J., Mars, R. B., & Rushworth, M. F. S. (2013). Dissociable effects of surprise and model update in parietal and anterior cingulate cortex. *Proceedings of the National Academy of Sciences*, *110*(38), E3660–E3669. https://doi.org/10.1073/pnas.1305373110

R Core Team. (2019). *R: A Language and Environment for Statistical Computing*. R Foundation for Statistical Computing. https://www.R-project.org/

Reisenzein, R., Horstmann, G., & Schützwohl, A. (2019). The Cognitive-Evolutionary Model of Surprise: A Review of the Evidence. *Topics in Cognitive Science*, *11*(1), 50–74. https://doi.org/10.1111/tops.12292

Schwartenbeck, P., FitzGerald, T. H. B., & Dolan, R. (2016). Neural signals encoding shifts in beliefs. *NeuroImage*, *125*, 578–586. https://doi.org/10.1016/j.neuroimage.2015.10.067

signal developers. (2014). *signal: Signal processing*. http://r-forge.r-project.org/projects/signal/

Simpson, H. M., Molloy, F. M., Hale, S. M., & Climan, M. H. (1968). Latency and magnitude of the pupillary response during an imagery task. *Psychonomic Science*, *13*(6), 293–294. https://doi.org/10.3758/BF03342592

Singmann, H., & Kellen, D. (2017). An introduction to mixed models for experimental psychology. *New Methods in Neuroscience and Cognitive Psychology*.

Sutton, R. S., & Barto, A. G. (2018). *Reinforcement learning: An introduction*.

Urai, A. E., Braun, A., & Donner, T. H. (2017). Pupil-linked arousal is driven by decision uncertainty and alters serial choice bias. *Nature Communications*, *8*(1), 14637. https://doi.org/10.1038/ncomms14637

Van Slooten, J. C., Jahfari, S., Knapen, T., & Theeuwes, J. (2018). How pupil responses track value-based decision-making during and after reinforcement learning. *PLOS Computational Biology*, *14*(11), e1006632. https://doi.org/10.1371/journal.pcbi.1006632

Winn, M. B., Wendt, D., Koelewijn, T., & Kuchinsky, S. E. (2018). Best Practices and Advice for Using Pupillometry to Measure Listening Effort: An Introduction for Those Who Want to Get Started. *Trends in Hearing*, *22*, 233121651880086. https://doi.org/10.1177/2331216518800869

Yu, A. J., & Dayan, P. (2005). Uncertainty, Neuromodulation, and Attention. *Neuron*, *46*(4), 681–692. https://doi.org/10.1016/j.neuron.2005.04.026

Zeileis, A., & Grothendieck, G. (2005). zoo: S3 Infrastructure for Regular and Irregular Time Series. *Journal of Statistical Software*, *14*(6), 1–27. https://doi.org/10.18637/jss.v014.i06