An EEG investigation of the trial-by-trial updating of complex knowledge structures

Franziska R. Richter[1]

[1]Cognitive Psychology Unit, Institute of Psychology, Leiden University, Leiden, The Netherlands.

Keywords: schema; long-term memory; ERP; P3; prediction error; updating; knowledge;

Abstract

Schemas are higher-level knowledge structures that store an abstraction of multiple previous experiences. They allow us to retain a multitude of information, but without the cost of storing every detail. Schemas are believed to be relatively stable, but occasionally have to be updated to remain useful in the face of changing environmental conditions. Once a schema is consolidated, schema updating has been proposed to be the result of a prediction error (PE) based learning mechanism, similar to the updating of less complex knowledge. However, for schema memory this hypothesis has so far been difficult to test since the tools to track modifications to abstracted memory schemas have not been sensitive enough. Moreover, existing research on electrophysiological correlates of updating have focused on short-term belief updating tasks. Here I am using EEG and continuous memory measures recorded during the encoding of new schema consistent and inconsistent material to test the behavioural and neural correlates of schema updating. Schema updating was assessed in a memory test 24 hours after encountering inconsistent information, demonstrating the long-term effect of such PE-based learning. I observed a stronger relationship between behavioural PE and schema updating measures for inconsistent compared to consistent material, in line with the idea that more updating is required when a schema changes. Moreover, the P3 EEG signal tracked both the PE at the time of learning, as well as the updating of the memory schema one day later in the inconsistent condition. These results demonstrate that schema updating in the face of inconsistent information is indeed driven by PE-based mechanisms, and that similar neural mechanisms underlie the updating of consolidated long-term schemas and short-term belief structures.

Introduction

Memory schemas are complex dynamic knowledge structures that allow us to store vast amounts of information in a concise manner. Schemas evolve slowly over time, and are generally believed to be stable once consolidated (Ghosh & Gilboa, 2014). However, in a dynamically changing world, schemas need to be adjusted or updated from time to time (Piaget, 1952). *Accommodation* is an updating mechanism that leads to small changes to the schemas structure itself.

One popular idea of how memory schemas can be updated (or accommodated) is via means of prediction error (PE) based learning (cf. Henson & Gagnepain, 2010; van Kesteren, Ruiter, Fernández, & Henson, 2012): When a predicted outcome does not match our experience, the resulting PE provides us with a surprise signal that is believed to guide learning. Recent findings suggest that PE-based learning underlies memory updating in many contexts: the updating of schematic rules (Greve, Cooper, Tibon, & Henson, 2019), of episodic memories (Greve, Cooper, Kaula, Anderson, & Henson, 2017; Sinclair & Barense, 2018, 2019), and of single semantic facts (Pine, Sadeh, Ben-Yakov, Dudai, & Mendelsohn, 2018). However, research on the accommodation of complex abstracted knowledge structures such as memory schemas is scarce.

In the context of *short-term* decision making tasks the neural mechanisms of PE-based updating have been studied with fMRI (e.g., Behrens, Woolrich, Walton, & Rushworth, 2007) and EEG (Bennett, Murawski, & Bode, 2015; Jepma, Murphy, Nassar, Rangel-Gomez, Meeter, & Nieuwenhuis, 2016; Kolossa, Kopp, & Fingscheidt, 2015). In these tasks participants adjust their predictions multiple times within an experimental session. Such PE based-updating of short-lived beliefs has been shown to be indexed by the P300 signal (Bennett et al., 2015; Jepma et al., 2018, 2016; Kolossa et al., 2015). The P300 has been suggested to be divisible into two subcomponents, the fronto-centrally distributed P3a and the P3b, which has a

centro-parietal topography (see Fonken, Kam, & Knight, 2019, for a review). Both the P3a (Bennett et al., 2015; Kolossa et al., 2015) and P3b (Jepma et al., 2018, 2016) have been documented to track belief updating. More generally, the P300 has been linked to updating of the (immediate) context in a number of tasks that require the comparison of a current stimulus with a recently preceding stimulus, and with the updating of information in working-memory tasks (Polich, 2012).

However, belief updating in the decision-making experiments described above takes place on a much smaller time scale than what is believed to be necessary for the formation and updating of memory schemas, which are thought to be extracted over hours, days or even months (Ghosh & Gilboa, 2014; O'Reilly, Bhattacharyya, Howard, & Ketz, 2014). Therefore, whether the same neural mechanisms underlie the updating of short-lived beliefs and memory schemas is currently unknown. However, behavioural evidence suggests that PE-based learning also works on longer time-scales (e.g., Greve et al., 2019; Pine et al., 2018). In addition, for some types of memory (e.g., semantic knowledge) updating has been shown to rely on similar mid-brain regions that are also thought to underlie the updating of more short lived beliefs (Garrison, Erdeniz, & Done, 2013; Pine et al., 2018). Moreover, the P3a and the P3b have been linked to better encoding such that larger amplitudes of either potential are associated with better memory in a subsequent test (Azizian & Polich, 2007; Knight & Scabini, 1998; Richter & Yeung, 2016). Therefore, the P300 might also index updating of long-term memory schemas.

A consolidated schema should allow for mismatches (PEs) between schema-based prediction and new inconsistent information. These PEs should consequently induce small modifications to a schema (Ghosh & Gilboa, 2014; cf. Piaget, 1952). With regards to schema memory, however, a major difficulty lies in the tracking of such subtle changes to these complex knowledge structures. Small changes in the memory schema have only recently been made visible in humans using new sensitive continuous memory tools (Richter, Bays, Jeyarathnarajah, & Simons,

2019). Here I am using a similar task to study the neural mechanisms of schema updating using EEG. By using a prediction-based learning paradigm I am able to track the neural correlates of trial-by-trial predictions violations and relate them to schema updating one day later. Based on the findings reviewed above, I predict that the P300 signal should track prediction violations and updating mechanisms in the context of memory schemas.

Materials and Methods

In the current study, I used continuous memory precision measures to track the formation and updating of memory schemas in a three-day experiment, similar to an approach I have used recently in a behavioural paradigm (Richter et al., 2019). A firmly established schema is the foundation for schema-based prediction violations to occur. Therefore, schema updating requires a schema to be at least partially consolidated. Participants in the current study first acquired schemas for different stimulus categories (Animals, Clothes, Furniture, and Food [fruits and vegetables]) in a prediction task on day 1. They then experienced inconsistencies in a subset of the categories on day 2 while their EEG was recorded, and later were tested on their memory for all stimuli (day 3). Specifically, the hypothesis was tested that introducing inconsistent information into a memory schema on day 2 (after a consolidation period) would lead to an increase in PEs (relative to a consistent condition), and that these PEs would be accompanied by increased P300 amplitudes. Moreover, I predicted that the size of these two measures would predict schema updating on day 3.

Procedural overview

Twenty-five participants came into the lab on three consecutive days. The sessions were scheduled 24 hours apart and participants were reimbursed for their

participation with payment or course credits. They all read an information sheet on the experiment and gave informed consent according to the declaration of Helsinki, before they were presented with task instructions. Subsequently, they completed a practice task (34 trials) using instrument-pictures that did not overlap with the image categories used in the main experiment. The 34 practice trials were spread over 4 blocks of length 2, 4, 10 and 18 trials, to give participants a chance to get used to the prediction task and ask additional questions after the blocks. The participants' task (described in more detail below) was to predict where on an invisible circle a picture should be placed. They subsequently had to encode the correct location that was provided as feedback after their prediction. In the practice phase most items (76.47%) correct/feedback locations occurred in a 90-degree segment centered around 65 degrees. Thus, participants initially guessed where a certain picture would go until they learned that stimuli followed a location schema.

In the main task participants learned stimuli from all 4 categories on day 1 and 2. The goal of day 1 was that participants would learn that most of the stimuli in each of the 4 categories were presented in certain category specific areas of the screen. That is, animals would, for example, be clustered in a 90-degree window centred at 110 degrees of the screen (see Figure 1A). The mean location, in degrees, for the 4 schemas was 20, 110, 200, and 290 degrees. The assignment of categories to mean locations was counterbalanced across participants. Within each category most trials (40 out of 50, consistent trials) fell within a category specific 90 degree window. However, a certain number of trials (10 out of 50, inconsistent trials) fell outside of this window to make the location schemas less obvious and ensure gradual learning. On day 1 (initial study phase) and on day 2 (new learning phase) participants completed 10 blocks of 20 trials of schema learning (50 trials each for the four categories). Within the 10 blocks, stimuli of all 4 categories would be presented in equal proportions. (On day 1, data from 1 block was lost from one participant due to a technical error.)

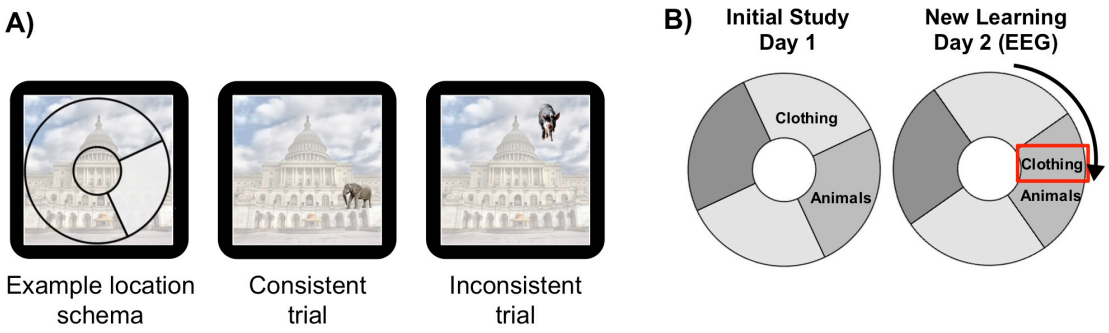

Figure 1. Location schema for the animal category. **(A)** Illustration of the size of the circle segment occupied by a category – here animals. In consistent trials the images of the category are shown within the segment, in inconsistent trials they are presented outside of it. **(B)** On day 2 the category specific circle segment changes for two categories. In this example the clothing category is moved by 90 degrees clockwise.

On day 2 the assignment of categories to circle locations would be retained for 2 of the 4 categories, but would be changed for the other two. For these 'inconsistent' categories the schema mean would shift by 90 degrees clockwise (see Figure 1B). For example, if the schema mean for the animal category was 110 degrees before, it would now be at 200 degrees. The 200 stimuli (again 50 per category) learned on day 2 did not overlap with those learned on day 1 (i.e., they were new exemplars of the same categories studied on day 1). Which category was presented at which location was counterbalanced across participants, and the category shift always occurred in such a manner that all categories (consistent and inconsistent) would be centered either around 20 or around 200 degrees. Thus, the consistent and inconsistent categories where matched for their visual location.

On day 3 participants completed a final precision test on all stimuli learned on the previous 2 days: They were presented with the stimulus at a random location and had to recreate the correct location (that was provided to them in the form of feedback, see details below) from memory.

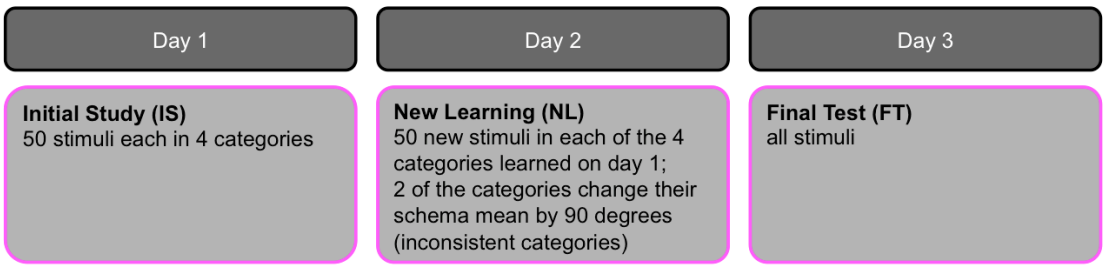

Figure 2. Sessions of the experiment. On day one participants learned stimuli in each category by performing a prediction task. On day 2 new stimuli of each of the categories are learned, but two of the categories become inconsistent. On the final day, the participants complete a precision memory test for all items.

Task

Trials on day 1 and 2 consisted of the following steps: First an objet was presented to participants at a random location on an invisible circle centred on a background image (see Figure 3A) with the word 'Location' presented in the center of the screen in white font. Participants had 5 seconds to move the picture to a position of their choice using the "g" (counter-clockwise) and "h" (clockwise) keys on a standard computer keyboard. They were instructed to use the space bar to lock their response. Initially, participants would guess the location, as no prior schema was established yet. They subsequently had 3 seconds to rate the confidence in their choice on a 100 point scale ranging from certainly incorrect to certainly correct. They again used the "g" (left) and "h" (right) keys to move the slider and locked the response using the space bar. After this, the picture was presented in the correct location for 1 second. The picture than disappeared from the background leaving only a fixation cross and participants had another 4 seconds to memorize the new location before the next trial started.

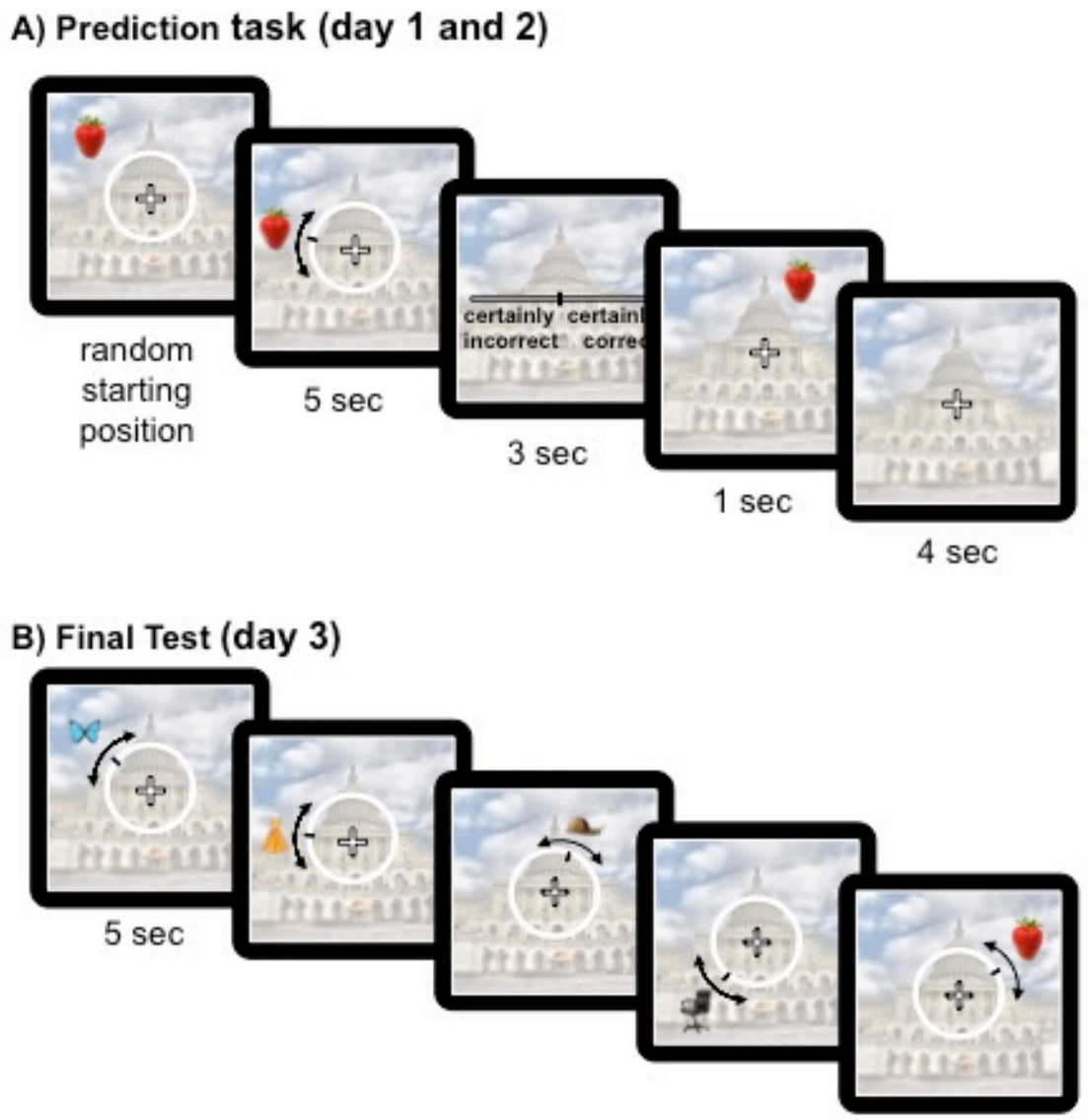

Figure 3. Task structure. (A) Prediction task on day 1 and 2. Participants are presented with an object at a random location and then have 5 seconds to move this object to a location of their choice. They subsequently have 3 seconds to indicate their confidence in their choice. The correct location (feedback) is presented to them for 1 seconds and participants then have another 4 seconds to memorize this location before the next trial starts. (B) In the memory test on day 3 participants are presented with objects of all categories in a random order. They have to recreate the correct location that they have learned on day 1 or 2.

The task on day 3 was slightly different. Participants were presented with all of the stimuli learned on days 1 and 2 in a random order. They were instructed to

move the stimuli to the correct locations (i.e., the locations that they were presented with in the feedback period at the end of each trial on day 1 and day 2). Whenever 1 second was left for participants to make their decision, the word location or the tick mark on the confidence scale would turn red to remind participants to lock their answer. In case participants did not manage to press spacebar in time to confirm their response, the last location the participant moved the cursor to (in the confidence rating) or the location of the picture were recorded. Stimulus presentation was implemented using Matlab (Mathworks, Natick, MA) using Psychtoolbox (www. psychtoolbox.org) software.

EEG data recording and analysis

The EEG was recorded from 32 scalp electrodes on a Biosemi system using AgCl electrodes. Electrode location corresponded to FP1, FPZ, FP2, F7, F3, FZ, F4, F8, FT7, FC3, FCZ, FC4, FT8, T7, C3, CZ, C4, T8, TP7, CP3, CPZ, CP4, TP8, P7, P3, PZ, P4, P8, POZ, O1, OZ, and O2. Two electrodes were attached on the outer canthi of both eyes, and two further electrodes were attached above and below the left eye to record blinks and eye movements. Left and right mastoids were also recorded.

A sampling rate of 512 Hz was used for recording. Before analysis, the EEG data were downsampled to 100 Hz. A common mode sense electrode was used for online referencing and electrodes were re-referenced offline to the average two mastoids to remove any lateral biases.

Data processing was done separately for each participant. Raw EEG data were high pass filtered at 0.5 Hz to remove drift. A low-pass filter of 40 Hz was used for analysis using the “eegfiltnew” method implemented in the EEGLab toolbox. Data were epoched into trials of 4500-ms duration to avoid edge artefacts in further analyses. This time-window included a 1500-ms pre-event period. Preliminary

artefact rejection was done prior to ICA by removing trials with extreme amplitudes (±2000 μV across the entire epoch, as well as ±100 μV in the baseline period). Improbable data (data beyond 5 SD in a single channel or all channels) was furthermore rejected using EEGlab's algorithm. Data were baseline corrected in the time period -100ms to 0ms. Ocular artefact reduction was performed using independent component analysis in EEGlab to identify blink components, which were mathematically subtracted from the data.

In a further artifact rejection step following the removal of the ICA components trials were rejected if at least one electrode showed a difference of more than ±75 μV from the baseline period to 900 ms after the event (i.e., a total of 1 second). All further data analyses were completed in MATLAB using the custom-written routines.

After data preprocessing I inspected the number of surviving trials for each participant to exclude any participant with low trial numbers. The lowest number of trials included for a participant after artifact correction was 111 out of 200 trials. All other participants had more than 122 trials. No participant was excluded. Further data analysis steps are described for the relevant analyses below.

Results

If the Sphericity assumption was violated corrected *p*-values according to Greenhouse-Geisser will be reported alongside *original* degrees of freedom to maintain readability in all analyses below.

Behavioural Results

Schema learning on Day 1 and Day 2

To investigate whether participants learned the location schemas I tracked their absolute error in the prediction task on day 1. (Note that because the pictures were uniformly distributed within a category's 90-degree segment, participants could not reduce their error beyond a certain point, even if they learned the schema perfectly). As evident from Figure 4, participants' performance in the task improved over the course of the initial study period. Performance differed significantly across blocks, $F(9, 207) = 9.726$, $p < .001$, Sphericity not assumed. Follow-up polynomial contrasts indicated a significant linear trend, $F(1, 23) = 35.664$, $p < .001$, indicating a decrease in the absolute error across blocks on Day 1. Thus, participants learned the schema over the course of day 1, evident in a decreasing prediction error.

On day 2 participants should be guided by this schema in the consistent categories, and update their schema about the inconsistent categories. To test this prediction I first calculated an ANOVA with factors consistency (consistent and inconsistent) and block (1 to 10). The ANOVA revealed a significant main effect of consistency, with lower errors in the consistent conditions, $F(1, 24) = 14.095$, $p = .001$, Sphericity assumed. Moreover, a significant main effect of blocks was observed $F(9, 216) = 2.589$, $p = .007$, Sphericity assumed, as well as a marginal interaction between consistency and block $F(9, 216) = 1.737$, $p = .082$, Sphericity assumed.

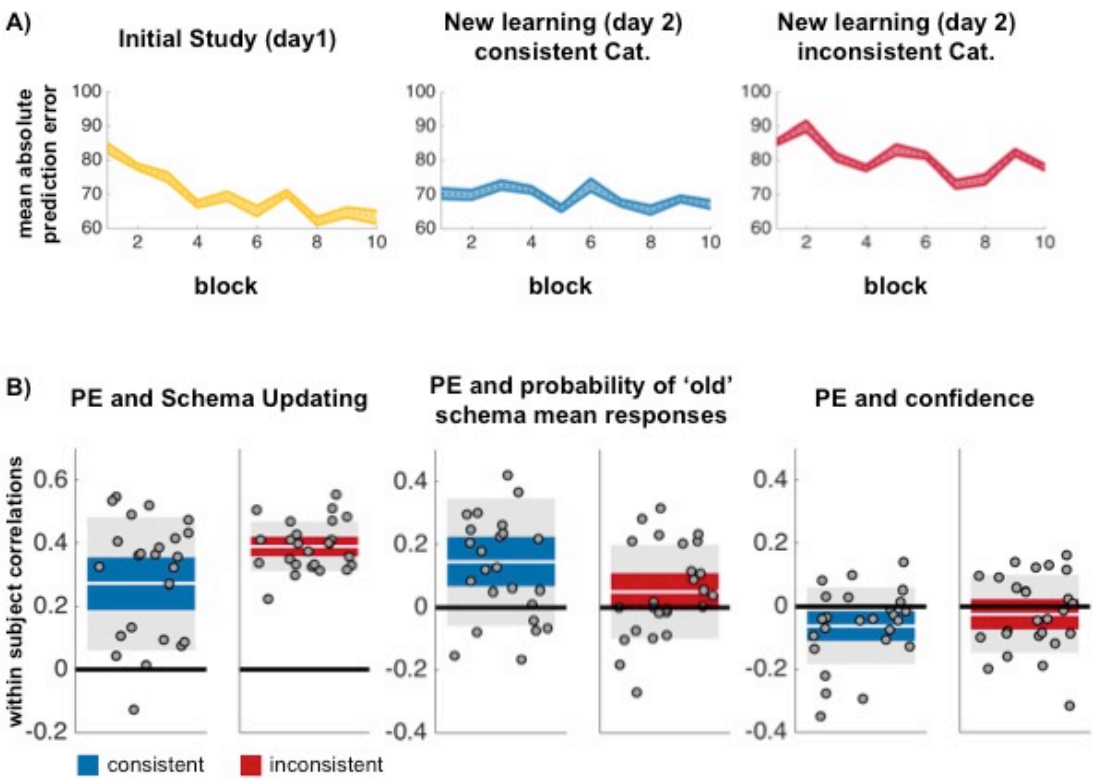

Figure 4. (A) Mean absolute prediction error (dotted line) across individual blocks of initial study (yellow) and new learning. For the new learning phase performance is plotted separately for the consistent categories (blue) and the inconsistent categories (red). Solid lines reflect standard errors of the mean. (B) Within subject correlations between PE and schema updating score (left), prediction error and the probability of responses consistent with the *old* schema mean according to the computational model (center) and PE and confidence (left). In each plot the consistent condition is displayed in blue on the left and the inconsistent condition in red on the right. Each circle represents one subject. The white line represents the mean and the blue/red area the 95% confidence interval. The grey areas indicate *mean* ± 1 *SD.* Illustration of these correlations was created via the notBoxPlot function in Matlab https://nl.mathworks.com/matlabcentral/fileexchange/26508-notboxplot.

Due to the a priori hypothesis that PE should decrease from block 1 to block 10 and that this effect should be stronger in the inconsistent than consistent category, I conducted follow up tests, to see if this was indeed the case. Consistent

with expectations, in the consistent categories PE from the first block on day 2 was not significantly different from that at the end of day 1, $t(24) = 1.01$, $p = .322$, and remained on this level. This result indicates that not much additional learning was taking place on day 2 in the consistent category. There was no significant effect of block, $F(9, 216) = 0.945$, $p = .487$, Sphericity assumed, and no significant linear trend, $F(1, 24) = 1.512$, $p = .231$, in the consistent category on day 2 also in line with this interpretation. PE in the first block of the inconsistent categories, in contrast, was significantly higher than PE in the last block of day 1, $t(23) = 5.305$, $p < .001$. Moreover, in the inconsistent categories there was a significant main effect of block, $F(9, 216) = 3.520$, $p < .001$, Sphericity assumed. PE decreased linearly across blocks, evident in a significant linear trend, $F(1,24) = 8.956$, $p = .006$. However, PE never reached the same level as the last block on day 1 (comparison of Block 10 of day 1 and block 10 of day 2 in the inconsistent categories yielded a significant difference, $t(23) = 3.15$, $p = .0045$. This last finding suggests that participants were still confused (and potentially biased by the initially learned schema mean) at the end of day 2 in the inconsistent condition.

Category consistency affects old and new schema-based responding in the Final Test (day 3)

Modelling and Permutation Testing. Next, I used computational modelling and permutation tests to assess the hypothesis that, in the inconsistent conditions on day 3, participants would show a bias towards the location of the *previously* relevant (day 1) schema mean. Such a bias would be evident in an increased number of responses 90 degrees counter clockwise to the schema mean in the inconsistent condition compared to the consistent condition (where participants had no reason to be biased towards responses counter-clockwise to the current schema).

To test whether there was a bias in responses on day 3, I applied a mixture model comprising (i) a von Mises distribution centred around the target values, (ii) a uniform component and (iii) an additional von Mises distribution centred around -90 degree from the currently relevant schema mean to the data collected during the final test phase (day 3). This second von Mises distribution was included to capture responses that were more consistent with the 'old' schema mean in the inconsistent categories. This approach is similar to that of modelling non-target responses in previous work (Bays, Catalao, & Husain, 2009). The model was run on data pooled across participants to enhance stability of the model, and modelling was done separately for consistent and inconsistent categories.

Based on the model output, I subsequently tested whether participants would show a tendency to bias their responses towards the 'old schema mean' (day 1 schema mean) in the inconsistent categories compared with the consistent categories. For the inconsistent conditions, this day 1 schema mean was always presented -90 degrees from the current schema mean. I used an approach similar to our previous paper (Richter et al., 2019), and described further in Schneegans and Bays (2016). For each trial, I calculated the posterior probability that the response participants gave stemmed from each of the three mixture components [target response, -90 degree (non-target) response, or guessing]. To allow for a valid comparison in the consistent categories (where the 'old' schema mean is identical to the current schema mean) I also modelled the probability of responding -90 degrees from the schema mean. That is, I modelled responses around -90 degrees for both the consistent and inconsistent conditions, but in the inconsistent conditions these responses corresponded to the location of the old schema mean (i.e., the schema mean relevant on day 1). I then used permutation testing to assess the hypothesis that there would be fewer trials assigned to the -90 degree responses in the consistent categories than the inconsistent categories (where participants should be biased towards the old schema mean). The results of this analysis indeed indicated

more responses around -90 degrees ('old schema mean' location) for the inconsistent than the consistent categories ($p < .001$, 1000 iterations). This finding suggests that on day 3 participants are still biased by the day 1 schema in the inconsistent condition.

Prediction error at day 2 predicts schema updating on day 3 more strongly in inconsistent condition

The modeling analysis suggests that, on day 3, participants are more biased towards responses -90 degrees of the current schema in the inconsistent vs. consistent condition (corresponding to the old schema mean in the inconsistent condition). Here I ask whether this bias is diminished for trials in which larger PEs occur. I argued above that schema updating should rely on PEs, and that there should be a strong relationship between PEs and the degree to which participants change their schema, especially in the inconsistent condition. In other words, if prediction errors are stronger on day 2, does more learning occur, evident in smaller errors on day3? In the current prediction design, we have the possibility to *directly* measure the degree to which participants more strongly base their responses on a schema on day 3 than day 2, and we can relate this measure to trial-by trial prediction errors.

To measure "schema updating" I calculated the absolute difference between the responses given by the participants and the schema mean (i.e., an error measure between response and schema mean, rather than the traditional error measure between response and the location learned in the feedback phase) in both the NL phase as well as the FT. Subsequently, I calculated the *schema updating score* for each trial by subtracting the FT error values from the corresponding values during NL. (To make error values comparable between both sessions, errors were z-scored first). Here positive *schema updating scores* indicate that responses are closer to the

current schema mean in the FT (day 3) than in the NL phase (day 2), indicating a stronger orientation towards the current schema. Next I assessed whether PEs predict the degree to which this updating towards current schema occurs. For this purpose, I correlated the PE at day 2 with this schema updating score, and compared the resulting correlation between the consistent and inconsistent categories. I found, as expected, that there was a positive correlation between the updating measure and the absolute PE for both consistent ($r_{consistent}$ = .27, t(1,24) = 6.445, p < .001) and inconsistent categories ($r_{inconsistent}$ = .39, t(1,24) = 24.740, p <.001). This outcome indicates that in either case having a larger PE during NL resulted in more learning (more schema updating). Importantly, and consistent with predictions, this correlation was stronger in the inconsistent than in the consistent categories, where more learning would take place by design (t(1,24) = 3.260, p = .0033, Figure 4A).

A second way to assess the relationship between PE on day 2 and memory on day 3 is to examine the relationship between PE and the probability of responding at -90 degrees. These responses again capture responding with the 'old schema mean' in the inconsistent categories. For the consistent categories, -90 degree responses again provide a baseline of the probability that would be expected if the schema never has been updated. The results of this analysis are presented in Figure 4B. Turning to the consistent categories first, there was a significant relationship between the size of prediction errors and errors around -90 degrees ($r_{consistent}$ = .145, t(1,24) = 3.58, p < .002). In other words, large prediction errors in the consistent category where associated with an increased probability to respond around -90 degrees, indicating that large prediction errors on day 2 lead to relatively large errors during responding at the final test. This effect possibly occurs because large PEs were more likely for trials further away from the schema mean in the consistent condition. In comparison, larger PEs predicted these -90 (or 'old schema mean') responses only marginally in the inconsistent condition ($r_{inconsistent}$ = .05, t(1,24) =

1.684, p = .110), indicating that here larger PEs may have lead to a relative updating of the schema. Thus larger PEs in on day 2 were less strongly associated with erroneous -90 responses in the inconsistent than the consistent category (t(1,24) = -2.802, p = .010). This finding is particularly noteworthy since, in the inconsistent category, participants had been presented with items at the -90 degree position (the old schema mean) on day 1. Still, large PEs on day 2 predicted a relative *decrease* in the probability to respond with the old schema mean in the inconsistent compared to the consistent condition. This finding indicates, again, that following larger PEs more updating took place here, in line with the above analysis.

Confidence is negatively related to PE

I also collected the participants' confidence ratings in order to investigate how confident participants were in their predictions on day 2. A correlation was calculated between the confidence rating a participant gave on each trial and the subsequently observed absolute prediction error. A tendency for a negative relationship (lower PE for high confidence responses) indicating that participants had the metacognitive ability assess their performance was observed in both types of categories. However, this effect was only significant for the consistent categories, mean $r_{consistent}$ = -.0619, t(24) = -2.5564, p = .0173, but not the inconsistent categories, mean $r_{inconsistent}$ = -.0245, t(24) = -0.9998, p = .3274. The fact that the relationship was only significant for the consistent category, could potentially suggest that a relationship between PE and confidence is more pronounced in situations in which the participants can be more certain about their responses. However, the difference of correlations between the consistent and inconsistent categories was not significant, t(24) = 1.1087, p = .2785.

To summarize, the behavioural results indicate that participants successfully learned the schema on day 1. For the inconsistent categories they updated their schemas on day 2, but still remained below day 1 performance (Figure 4A). On day 3 participants showed a bias towards responses corresponding to the old schema mean in the inconsistent categories. The size of the prediction error on day 2 predicted schema updating on day 3, and this effect was stronger in the inconsistent condition (Figure 4B). Lastly, across all trials, confidence and PE were correlated indicating that high confidence was generally associated with lower PEs, but this effect was driven by the consistent condition (Figure 4B). With these behavioural effects established, I next moved on to investigate the neural correlates of prediction errors in the context of schema updating.

EEG Results

For grand-average analyses data was not filtered. For analyses focussing on single trial data, data was first filtered using a 6Hz filter. Moreover, to reduce the effect of outliers, the 2.5% highest and lowest amplitudes were excluded from the analysis (and thus the central 95% of trials were retained). This procedure was done separately for the consistent and inconsistent conditions, to ensure that amplitude differences between the conditions did not affect which trials were excluded. For the trial-wise analysis the mean amplitude in the P300 window was used. Single-trial P3 amplitudes, mean absolute prediction errors, and schema-updating scores were furthermore z-scored, to account for any between subject differences in the across subject analyses.

Larger P3s for the inconsistent categories

The EEG analysis focused on two clusters of interest: the fronto-central and centro-parietal electrodes. These locations had been shown to display belief-updating effects (Bennett et al., 2015; Jepma et al., 2018, 2016; Kolossa et al., 2015), in previous studies, and have displayed differential involvement in subsequent memory effects (Richter & Yeung, 2016). In a first step, I tested the prediction that the P3 should be more pronounced in the inconsistent than the consistent condition during day 2 learning. Consistent with this prediction results showed that in the frontal electrode cluster, the inconsistent categories evoked a larger P3 response than the consistent categories (p = .018). A similar, marginally significant effect was observed in parietal sides (p = .050). The difference in the size of effects on frontal and parietal sides was not significant (p > .05).

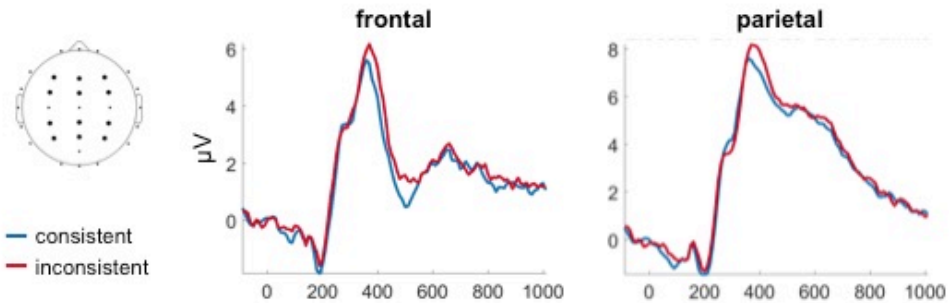

Figure 5. Grand-average feedback-evoked P3 at frontal and parietal sites. The average signal was calculated across clusters of 6 electrodes displayed as black circles in the topography on the left. EEG amplitude is plotted separately for the consistent (blue) and inconsistent (red) categories.

PE predicts trial-by-trial P3 amplitude

Next I examined whether P3 amplitude increased reliably with increasing prediction error. Again, this analysis was performed separately for frontal and posterior electrodes. At the frontal cluster (Figure 6A) increasing PE size was

associated with increasing P3 amplitude. This effect was significant when computed across all subjects and trials t = 2.85, p = .0044, and also when computed within subjects, $t(24)$ = 2.68, p = .013 in the inconsistent categories. A corresponding effect in the consistent categories was not significant, both when computed across all subjects and all trials, t = 0.279, p = .783, as well as when assessed within subjects, $t(24)$ = .187, p = .852. The difference in this within-subject P3-PE correlation between consistent and inconsistent categories was marginally significant, $t(24)$ = 1.85, p = .075.

For the posterior cluster, numerically effects in a similar direction were observed. However, they were not statistically reliable (inconsistent condition: across all subjects and trials, t = 0.707, p = .486; within subjects, $t(24)$ = .495, p = .620; consistent condition: across all subjects and trials, t = 0.108, p = .915; within subjects, $t(24)$ = 0.039, p = .969). The difference between the inconsistent and consistent category within subject effect was not significant at posterior electrodes, t = 0.346, p = .733.

Lastly, to assess location differences, I compared whether the mean within subject P3-PE correlation was stronger at frontal than posterior electrodes. The average within subject correlation was marginally stronger at frontal electrodes in the inconsistent condition, $t(24)$ = 1.790, p = .086. In contrast, an insignificant effect in the consistent condition, $t(24)$ = 0.1383, p = .8912, was potentially to be expected since the relationship between PE and P3 was not significant for this condition at either location.

Thus, consistent with my hypotheses PE predicted P3 size. This effect was only significant at frontal electrodes, and was marginally stronger in the inconsistent than the consistent categories.

Trial-by trial P3 amplitude predicts schema updating

If schema updating is initiated by the P300, P300 should not only correlate with the prediction error, but also with the degree to which participants updated their responses on day 3. Thus, I correlated P300 amplitude in the consistent and inconsistent condition with the *schema updating score* (see calculation above). Analyses revealed that in the inconsistent condition there was a positive correlation between P300 amplitude and later schema updating. This relationship was marginal at frontal electrodes when investigated across all subjects and trials, t = 1.90, p = .057, as well as within subjects, $t(24)$ = 1.88, p = .071. For the consistent category, the relationship was insignificant both across all subjects and trials, t = 1.135, p = .256, and within subjects, $t(24)$ = 982, p = .336. There was no significant difference between the within subject effects of the consistent and inconsistent categories, t = 0.362, p = .721.

At posterior electrodes (see Figure 6B), a significant positive effect was found for the inconsistent category, again across all trials, t = 2.41, p = .016, as well as within subjects, $t(24)$ = 2.13, p = .044. The same analysis for the consistent category revealed a marginal effect across all trials and subjects, t = 1.44, p = .151 and a marginal effect within subjects, $t(24)$ = 1.94, p = .065. There was no significant difference between the within subject effects of the consistent and inconsistent categories, t = 0.651, p = .522.

I also investigated whether larger P3s were associated with a decreased probability of responding with the 'old' schema mean (derived from the model, see calculation above). While the effect was numerically in the predicted direction (a negative correlation), it did not reach significance for either the consistent or inconsistent category both a frontal or at posterior electrodes in both the within and across subject analyses (all *ps* > .4352).

Lastly, to assess location differences, I again compared the mean within-subject correlation for each location-consistency combination separately, here for the relationship between P3 and schema-updating score. Again, no difference between frontal and posterior sites in the average within-subject correlation was observed neither for the inconsistent, $t(24) = -0.4943$, $p = .626$, nor the consistent conditions, $t(24) = -0.316$, $p = .755$.

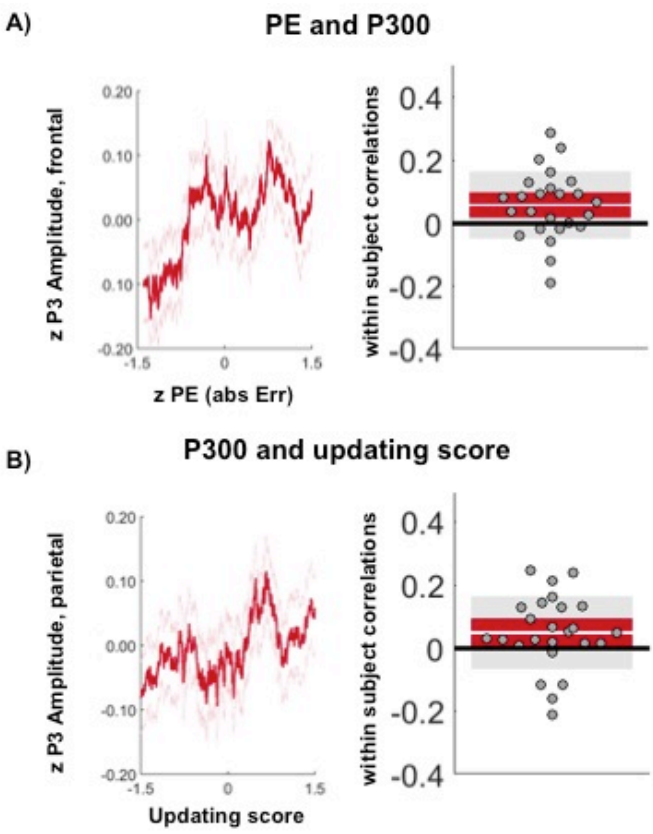

Figure 6. Relationship between PE and P3 amplitude as well as between P3 amplitude and updating scores in the inconsistent condition. (**A**) Relationship between PE and P3 amplitude at frontal electrodes. Left: across subject correlation. Data is pooled across participants. To account for between subject differences single-trial P3 amplitudes and PEs were *z*-scored per participant. The PE is averaged across running bins of 300 trials for display purposes. Right: within subject correlations. Each circle represents one subject. The white line represents the mean and the red area the 95% confidence interval. The grey areas indicated *mean* ± 1 *SD*. (**B**) Relationship between P3 amplitude and updating of the schemas in the final test on day 3

at parietal electrodes. Schema updating was calculated as the difference in absolute error between NL and FT (NL-FT; thus positive values mean updating has taken place). Left: across subject correlation. Data is pooled across participants. To account for between subject differences single-trial P3 amplitudes were also *z*-scored per participant. The updating score is averaged across running bins of 300 trials for display purposes. Note that while the updating scores themselves were not *z*-scored, PE in the NL and FT phase were z-scored within participant prior to calculating the updating scores, in order to account for overall performance differences between the days. Right: within subject correlations. Each circle represents one subject. The white line represents the mean and the red area the 95% confidence interval. The grey areas indicated *mean* ± 1 *SD.* Illustration of the within subject correlation was done via notBoxPlot function in Matlab https://nl.mathworks.com/matlabcentral/fileexchange/26508-notboxplot.

Discussion

The current experiment tested the hypothesis that similar neural mechanisms might underlie the updating of short-lived beliefs in decision-making contexts, as well as the updating of long-term memory schemas. For this purpose, I used a continuous report memory task, in which participants predicted the location of stimuli belonging to different categories in a three-day study. I then tested how they had updated their schema on the third day of testing. By manipulating the consistency of the categories on day 2 (after a consolidation phase), I induced PEs in participants, which were predictive of behavioural updating performance as well as P3 size. The results indicated a close link between PE-based schema updating and the P3a signal. P3b amplitude additionally correlated with PE size revealing potential neural mechanisms underlying the flexible updating of these knowledge structures.

Behaviourally, the described work reports a novel paradigm to study predictions made based on memory schemas, rather than on recently encoded

current belief systems (e.g., Bennett et al., 2015; Jepma et al., 2018, 2016; Kolossa et al., 2015). In contrast to these previous studies, my work does not track updating of *immediate* predictions in the task, but rather updating effects evident in the participants' performance one day later, in a final memory test. This paradigm therefore allows me to draw conclusions about the long-term effects of updating on (relatively) consolidated schemas.

Previous research has speculated that updating of knowledge structures such as schemas might rely on a mismatch between a prediction and the encountered information (Richter et al., 2019; van Kesteren et al., 2012), and similar processes have been observed in other memory tasks (Greve et al., 2019; Henson & Gagnepain, 2010; Krawczyk, Fernández, Pedreira, & Boccia, 2017; Pine et al., 2018). However, due to the absence of sufficiently sensitive methods to measure trial-by-trial differences in schema updating in the laboratory it has been difficult to track changes to complex memory structures such s schemas. The recent development of continuous report memory tasks made it possible to visualise adjustments to pre-existing schemas in the aftermath of encountering schema-inconsistent information (Richter et al., 2019). Here I developed this previous approach further to directly test the hypothesis that PEs are driving memory-schema updating.

My results show that participants are able to learn schemas in a prediction-based task on day 1. Furthermore, they are able to adjust these schemas when they encounter inconsistent information, as evident by the learning curves on day 2. Moreover, I have shown that, on day 3, participants demonstrated a bias towards the old schema information in the inconsistent condition. By using trial-by-trial analysis, I was able to show that this bias was *reduced* when encountering this information on day 2 induced a relatively large PE on day 2, indicating that PEs drive schema updating. Specifically, PEs were prognostic of schema updating, both, when assessing the change in schema-based responding (operationalized as the

difference between prediction and schema mean) on day 2 versus 3, and also when using a modelling approach. Here, larger PEs were associated with a relative decrease in model-identified responses in locations that were consistent with the old schema in the inconsistent condition. This was apparent in reduced evidence of responses -90 degrees from the current schema (i.e., at the position of the old schema for the inconsistent conditions) compared to responses observed in the consistent conditions.

The behavioural results are in line with studies on other types of memory that describe a relationship between PEs and memory updating (Greve et al., 2017, 2019; Pine et al., 2018; Sinclair & Barense, 2018, 2019). My work critically extends these findings to the subtle modification of complex memory constructs like schemas, consistent with what has theoretically been proposed in the literature (van Kesteren et al., 2012). Moreover, the neural mechanisms of schema updating, have so far remained unknown. My work reveals that similar neural mechanisms underlie PE-based updating of long-term memory schemas, as they do for short-term beliefs (Bennett et al., 2015; Jepma et al., 2018, 2016; Kolossa et al., 2015). Specifically, the current study tracks changes to complex mnemonic structures and links these changes to the P300 signal. The P300 potential has been strongly linked to context updating processes, for example in oddball and working memory tasks (Donchin, 1981; Sutton, Braren, Zubin, & John, 1965). The literature on the P300 has distinguished between a more frontally distributed P3a, which has been linked to mismatch and novelty processing, and shown to be dependent on the hippocampus, and the P3b, a parietally centred component, more strongly linked to target detection, and not as strongly linked to hippocampal processing (Fonken et al., 2019). In this study, I investigated both P3a and P3b to assess which of the two ERPs (or both or any) correlated with PE size and schema updating, based on prior findings of both P3a (Bennett et al., 2015; Kolossa et al., 2015) and P3b (Jepma et al., 2018, 2016) involvement in short-term belief updating tasks.

PE and P3

I found that, consistent with predictions, the size of the P3 across trials was larger in the inconsistent than the consistent condition. This effect was significant at the frontal (i.e., P3a), and marginal at the parietal electrode cluster (i.e., P3b, p = .05), and suggests that participants were on average more surprised (larger P3) by the feedback in the inconsistent condition. Similarly, a significant correlation between PE and P3a was observed for the inconsistent categories. In fact, this correlation was stronger for the inconsistent than consistent categories. It is likely that participants accepted a certain degree of noise in the consistent categories (based on their experience of the same level of noise on day 1). In the inconsistent conditions, however, they were able to recognise that the change in locations could not only be explained by noise. Consequently, these new locations elicited an increased surprise signal, evident in a larger P3a. This finding of an increased P3a amplitude in inconsistent conditions is consistent with the idea that frontal areas are recruited in the face of novel, unexpected stimuli (Soltani & Knight, 2000), that attract attention, which might support later memory (Richter & Yeung, 2016). Importantly, in the current study the source of this novelty or surprise signal is solely based on a consolidated memory schema, and not on recently encountered information.

P3 and updating performance

When investigating the relationship between P3 amplitude and the degree to which participants updated their responses in the final test 24 hours after learning, I did not observe any differences between the inconsistent and consistent category. This lack of a difference between conditions indicates that once a surprise signal was elicited, it resulted in an updating of the predicted location, irrespective of the overall consistency of the condition. The correlation between P3 and updating was significant at posterior electrodes (P3b) and marginal at frontal electrodes (P3a). The

P3b signal has been associated with the contextual updating, or more broadly, with the maintenance of representational context (Polich, 2007, 2012). Similarly in the current study, the P3b could index an update to the current context: the new, shifted schema.

Underlying neural mechanisms

Regarding the neurotransmitter system potentially involved in the updating of memory schemas, the P300 is a signal that has been closely linked to catecholaminergic mechanisms in the brain. Here, again, a distinction has been made between the P3a and P3b. The more frontal P3a has been linked to dopaminergic activity (Polich, 2012). Consistent with an involvement of dopaminergic processes, recent research has demonstrated that reward-related PEs (previously strongly linked to the dopamine system (Schultz, Dayan, & Montague, 1997) are positively correlated with incidental memory encoding (Jang, Nassar, Dillon, & Frank, 2019; Rouhani, Norman, & Niv, 2018). Thus, learning based on prediction violations in schema memory might share neural mechanisms with that of reward-mediated episodic memory and external reward may not be necessary in updating long-term beliefs.

The neural processes underlying the P3b might rely on somewhat distinct neural mechanisms: In addition to being affected by dopamine (Polich, 2007; Polich & Criado, 2006) P3b amplitude has also been linked to the locus ceruleus norepinephrine system (De Taeye et al., 2014; Jepma et al., 2018, 2016) and has been suggested to index target-detection, response selection, or rate of learning in the context of decision-making. Moreover, the P3b has been suggested to be generated by temporo-parietal brain areas while the P3a is believed to be generated by frontal areas (Ebmeier et al., 1995; Kirino, Belger, Goldman-Rakic, & McCarthy, 2000).

The involvement of both frontal and posterior ERP components is consistent with the idea proposed in the literature that schema updating might be depended upon two different networks: the initial identification of schema incongruent information may be subserved by fronto-hippocampal interactions (most likely involving ventromedial prefrontal cortex) and subsequent updating may recruit a more posterior hippocampal temporo-parietal network (cf. van Kesteren et al., 2012). Future research, that allows to better track the localisation of these effects is needed to investigate the brain mechanisms further. Moreover it will be of interest to understand how independent the two proposed mechanisms (the top-down attention capture by a stimulus in an inconsistent condition, and the subsequent updating of the newly encoded information and pre-existing memory schema) operate. .

Conclusion

The results provide evidence that memory schemas, similar to more short-lived belief structures, are updated via prediction-error based learning, and that similar neural mechanisms, indexed by the P3, might underlie updating in both cases.

Acknowledgements

I would like to thank Angélica van der Ploeg and Alexandra Felea for help with data collection.

References

Azizian, A., & Polich, J. (2007). Evidence for attentional gradient in the serial position memory curve from event-related potentials. *Journal of Cognitive Neuroscience*, *19*(12), 2071–2081. https://doi.org/10.1162/jocn.2007.19.12.2071

Bays, P. M., Catalao, R. F. G., & Husain, M. (2009). The precision of visual working memory is set by allocation of a shared resource. *Journal of Vision*, *9*(10), 7–7. https://doi.org/10.1167/9.10.7

Behrens, T. E. J., Woolrich, M. W., Walton, M. E., & Rushworth, M. F. S. (2007). Learning the value of information in an uncertain world. *Nature Neuroscience*, *10*(9), 1214–1221. https://doi.org/10.1038/nn1954

Bennett, D., Murawski, C., & Bode, S. (2015). Single-Trial Event-Related Potential Correlates of Belief Updating. *ENeuro*, *2*(5), 1–14. https://doi.org/10.1523/eneuro.0076-15.2015

De Taeye, L., Vonck, K., van Bochove, M., Boon, P., Van Roost, D., Mollet, L., … Raedt, R. (2014). The P3 Event-Related Potential is a Biomarker for the Efficacy of Vagus Nerve Stimulation in Patients with Epilepsy. *Neurotherapeutics*, *11*(3), 612–622. https://doi.org/10.1007/s13311-014-0272-3

Donchin, E. (1981). Surprise!... Surprise? *Psychophysiology*, *18*, 493.

Ebmeier, K. P., Steele, J. D., MacKenzie, D. M., O'Carroll, R. E., Kydd, R. R., Glabus, M. F., … Goodwin, G. M. (1995). Cognitive brain potentials and regional cerebral blood flow equivalents during two- and three-sound auditory "oddball tasks." *Electroencephalography and Clinical Neurophysiology*. https://doi.org/10.1016/0013-4694(95)00173-5

Fonken, Y. M., Kam, J. W. Y., & Knight, R. T. (2019). A differential role for human hippocampus in novelty and contextual processing: Implications for P300. *Psychophysiology*, (May 2018), 1–13. https://doi.org/10.1111/psyp.13400

Garrison, J., Erdeniz, B., & Done, J. (2013). Prediction error in reinforcement learning: A meta-analysis of neuroimaging studies. *Neuroscience and*

Biobehavioral Reviews. https://doi.org/10.1016/j.neubiorev.2013.03.023

Ghosh, V. E., & Gilboa, A. (2014). What is a memory schema? A historical perspective on current neuroscience literature. *Neuropsychologia*, *53*, 104–114. https://doi.org/10.1016/j.neuropsychologia.2013.11.010

Greve, A., Cooper, E., Kaula, A., Anderson, M. C., & Henson, R. (2017). Does prediction error drive one-shot declarative learning? *Journal of Memory and Language*, *94*, 149–165. https://doi.org/10.1016/j.jml.2016.11.001

Greve, A., Cooper, E., Tibon, R., & Henson, R. N. (2019). Knowledge is power: Prior knowledge aids memory for both congruent and incongruent events, but in different ways. *Journal of Experimental Psychology: General*. https://doi.org/10.1037/xge0000498

Henson, R. N., & Gagnepain, P. (2010). Predictive, interactive multiple memory systems. *Hippocampus*, *20*(11), 1315–1326. https://doi.org/10.1002/hipo.20857

Jang, A. I., Nassar, M. R., Dillon, D. G., & Frank, M. J. (2019). Positive reward prediction errors during decision-making strengthen memory encoding. *Nature Human Behaviour*. https://doi.org/10.1038/s41562-019-0597-3

Jepma, M., Brown, S. B. R. E., Murphy, P. R., Koelewijn, S. C., de Vries, B., van den Maagdenberg, A. M., & Nieuwenhuis, S. (2018). Noradrenergic and Cholinergic Modulation of Belief Updating. *Journal of Cognitive Neuroscience*, *30*(12), 1803–1820. https://doi.org/10.1162/jocn_a_01317

Jepma, M., Murphy, P. R., Nassar, M. R., Rangel-Gomez, M., Meeter, M., & Nieuwenhuis, S. (2016). Catecholaminergic Regulation of Learning Rate in a Dynamic Environment. *PLoS Computational Biology*, *12*(10), 1–24. https://doi.org/10.1371/journal.pcbi.1005171

Kirino, E., Belger, A., Goldman-Rakic, P., & McCarthy, G. (2000). Prefrontal activation evoked by infrequent target and novel stimuli in a visual target detection task: An event-related functional magnetic resonance imaging study. *Journal of Neuroscience*.

Knight, R. T., & Scabini, D. (1998). Anatomic bases of event-related potentials and their relationship to novelty detection in humans. *Journal of Clinical Neurophysiology*. https://doi.org/10.1097/00004691-199801000-00003

Kolossa, A., Kopp, B., & Fingscheidt, T. (2015). A computational analysis of the neural bases of Bayesian inference. *NeuroImage*, *106*, 222–237. https://doi.org/10.1016/j.neuroimage.2014.11.007

Krawczyk, M. C., Fernández, R. S., Pedreira, M. E., & Boccia, M. M. (2017). Toward a better understanding on the role of prediction error on memory processes: From bench to clinic. *Neurobiology of Learning and Memory*, *142*, 13–20. https://doi.org/10.1016/j.nlm.2016.12.011

O'Reilly, R. C., Bhattacharyya, R., Howard, M. D., & Ketz, N. (2014). Complementary learning systems. *Cognitive Science*, *38*(6), 1229–1248. https://doi.org/10.1111/j.1551-6709.2011.01214.x

Piaget, J. (1952). The origins of intelligence in children New York. *International Universities Press*.

Pine, A., Sadeh, N., Ben-Yakov, A., Dudai, Y., & Mendelsohn, A. (2018). Knowledge Acquisition is Governed by Striatal Prediction Errors. *Nature Communications*, (2018), 1–14. https://doi.org/10.1038/s41467-018-03992-5

Polich, J. (2007). Updating P300: An integrative theory of P3a and P3b. *Clinical Neurophysiology*, *118*(10), 2128–2148. https://doi.org/10.1016/j.clinph.2007.04.019

Polich, J. (2012). Neuropsychology of P300. In *The Oxford Handbook of Event-Related Potential Components*. https://doi.org/10.1093/oxfordhb/9780195374148.013.0089

Polich, J., & Criado, J. R. (2006). Neuropsychology and neuropharmacology of P3a and P3b. *International Journal of Psychophysiology*, *60*(2), 172–185. https://doi.org/10.1016/j.ijpsycho.2005.12.012

Richter, F. R., Bays, P. M., Jeyarathnarajah, P., & Simons, J. S. (2019). Flexible

updating of dynamic knowledge structures. *Scientific Reports*, *9*(1), 2272. https://doi.org/10.1038/s41598-019-39468-9

Richter, F. R., & Yeung, N. (2016). ERP Correlates of Encoding Success and Encoding Selectivity in Attention Switching. *PLoS One*, *11*(12), e0167396. https://doi.org/10.1371/journal.pone.0167396

Rouhani, N., Norman, K. A., & Niv, Y. (2018). Dissociable Effects of Surprising Rewards on Learning and Memory. *Journal of Experimental Psychology: Learning Memory and Cognition*. https://doi.org/10.1037/xlm0000518

Schneegans, S., & Bays, P. M. (2016). No fixed item limit in visuospatial working memory. *Cortex*, *83*, 181–193. https://doi.org/10.1016/j.cortex.2016.07.021

Schultz, W., Dayan, P., & Montague, P. R. (1997). A Neural Substrate of Prediction and Reward, *275*(MARCH), 1593–1600.

Sinclair, A. H., & Barense, M. D. (2018). Surprise and destabilize: prediction error influences episodic memory reconsolidation. *Learning & Memory*, *25*(8), 369–381. https://doi.org/10.1101/lm.046912.117

Sinclair, A. H., & Barense, M. D. (2019). Prediction Error and Memory Reactivation: How Incomplete Reminders Drive Reconsolidation. *Trends in Neurosciences*. https://doi.org/10.1016/j.tins.2019.08.007

Soltani, M., & Knight, R. T. (2000). Neural origins of the P300. *Critical Reviews in Neurobiology*.

Sutton, S., Braren, M., Zubin, J., & John, E. R. (1965). Evoked-Potential Correlates of Stimulus Uncertainty. *Science*, *150*(3700), 1187 LP – 1188. https://doi.org/10.1126/science.150.3700.1187

van Kesteren, M. T. R., Ruiter, D. J., Fernández, G., & Henson, R. N. (2012). How schema and novelty augment memory formation. *Trends Neurosci*, *35*(4), 211–219. https://doi.org/10.1016/j.tins.2012.02.001