From generating to violating predictions: The effects of prediction error on episodic memory

Gözem Turan[1,2], Isabelle Ehrlich[1,2], Yee Lee Shing[1,2], and Sophie Nolden[1,2]

[1] Department of Psychology, Goethe University Frankfurt, Frankfurt Am Main, Germany

[2] IDeA Center for Individual Development and Adaptive Education, Frankfurt am Main, Germany

Author Note

Gözem Turan https://orcid.org/0000-0002-2847-2144

Data, scripts, and additional online materials are openly available at the project's Open Science Framework page (https://osf.io/pfgyb/). We have no conflicts of interest to disclose. The study was funded by a Starting Grant from the European Union for YLS (ERC-2018-StG-PIVOTAL-758898). The work of YLS was also supported by the German Research Foundation (Project-ID 327654276, SFB 1315, "Mechanisms and Disturbances in Memory Consolidation: From Synapses to Systems"), and the Hessisches Ministerium für Wissenschaft und Kunst (HMWK; project "The Adaptive Mind"). The work of SN was also supported by a Research grant Focus A/B, Goethe-University Frankfurt am Main, ("Dynamics of auditory and visual memory representations in the aging brain").

Correspondence concerning this article should be addressed to Gözem Turan or Sophie

Nolden, Department of Psychology, Goethe University Frankfurt, PEG-Gebäude, Theodor-W.-Adorno-Platz 6, D-60629 Frankfurt am Main, Germany. Email: turan@psych.uni-frankfurt.de or nolden@psych.uni-frankfurt.de

Abstract

Generating predictions about environmental regularities, relying on these predictions, and updating these predictions when there is a violation from incoming sensory evidence are considered crucial functions of our cognitive system for being adaptive in the future. The violation of a prediction can result in a prediction error (PE) which affects subsequent memory processing. In our preregistered studies, we examined the effects of different levels of PE on episodic memory. Participants were asked to generate predictions about the associations between sequentially presented cue-target pairs, which were violated later with individual items in three PE levels, namely low, medium, and high PE. Hereafter, participants were asked to provide old/new judgments on the items with confidence ratings, and to retrieve the paired cues. Our results indicated a better recognition memory for low PE than medium and high PE levels, suggesting a memory congruency effect. On the other hand, there was no evidence of memory benefit for high PE level. Together, these novel and coherent findings strongly suggest that high PE does not guarantee better memory.

Keywords: prediction error, episodic memory, predictive processing

Introduction

According to the predictive processing framework (Bar, 2007; Friston, 2010; Henson & Gagnepain, 2010), key functions of the cognitive system are to generate predictions about environmental regularities and to update these predictions when there is a violation from incoming sensory evidence, giving rise to a prediction error (PE). Building up abstracted knowledge can be achieved by extracting the statistical regularities of the environment. The violation of the abstracted knowledge can result in a PE. It may be important to remember these events giving rise to PEs to ensure better predictions in the future.

The memory benefit of events giving rise to PE has been demonstrated in a series of recent studies (Bein et al., 2021; Brod et al., 2018; Greve et al., 2017; Kafkas & Montaldi, 2018; Quent et al., 2022). When we encounter an event that gives rise to PE, we tend to remember it better, possibly because it can be important for improving our predictions in the future. For example, to investigate the memory benefit for such events, Kafkas and Montaldi (2018) used a rule learning task in which participants learned a contingency relationship between six different symbols and two stimulus categories, i.e., natural or human-made. Then, they violated the previously experimentally-induced relationships either in memory encoding or retrieval phases. Their results showed that the presentation of unpredicted stimuli enhanced the subsequent recollection performance regardless of the position of the violation (i.e., at memory encoding or retrieval).

On the contrary, another body of research has shown that events that are in line with our predictions are remembered better, i.e., the memory congruency effect (Alba & Hasher, 1983; Anderson, 1981; Craik & Tulving, 1975). For example, congruent associations (e.g., wood-chair) would be easier to remember than incongruent associations (e.g., wood-cookie). This

account has been largely corroborated by recent behavioral evidence, e.g., in studies with item-scene pairs (Brod & Shing, 2019; Liu et al., 2018; Ortiz-Tudela et al., 2017; van Kesteren et al., 2013), item-location pairs (Atienza et al., 2011) and non-preexisting relations (Ostreicher et al., 2010).

To bridge these diverging findings in the literature, a recent model called Schema-Linked Interactions between Medial Prefrontal and Medial Temporal Lobe (SLIMM, van Kesteren et al., 2012) suggests a U-shape relationship between prediction and memory, in which different brain systems are involved. The model postulates that memory benefit is proportional to the degree of PE, where the degree is calculated via the difference between the prediction and the actual outcome. That is, events that are correctly predicted (i.e., low PE) would lead to better memory, as the memory congruency effect suggests (Brod et al., 2013), and this process is supported by the medial prefrontal cortex (mPFC). By means of the mPFC, already existing connections between representations upon which predictions are built become strengthened, facilitating the retrieval of the respective information. Similarly, events giving rise to PE (i.e., high PE) would also improve learning and memory (Henson & Gagnepain, 2010). The model postulates that the medial temporal lobe (MTL) creates “snapshots” for those events, resulting in a memory advantage. Lastly, events that are neither strongly predicted nor unpredicted would lead to medium PE. Since the activation of mPFC and MTL is weak for those events with medium PE, they would not benefit memory. Taken together, the varying differences between predictions and actual outcomes would result in low, medium, to high PE levels, which in turn is postulated to exhibit a U-shape relationship with episodic memory: Events of two ends of PE, namely low and high PEs, are assumed to be remembered better compared to medium levels.

A recent study by (Greve et al., 2018) showed evidence for this U-shape function. In

their series of experiments, the authors first led their participants to learn a rule about the pairing of object exemplars, which was then manipulated in three levels based on the strength of matching level with previous learned associations, namely, congruent, incongruent, and unrelated. While the rule remains unchanged for the congruent level, the incongruent level has a reversed rule. On the other hand, for the unrelated level, the rule reversed after the first trial. The authors aimed to establish a rule about the paired objects and subsequently violate or confirm them on the critical trial just before testing memory performance. Importantly, for the unrelated level, there was no rule to establish. Even though their results were in line with the U-shape function, it should be noted that the medium level was unrelated to the previous learned associations. One can argue that the poor memory performance for the medium level might be related to the requirement to create new associations instead of representing the medium level in the spectrum.

Another study from the same group of researchers (Quent et al., 2022) addressed this issue via a continuous function of prediction. The authors conducted a virtual reality study in which participants had to explore a virtual kitchen with kitchen objects positioned in different locations which had varying degrees of congruency based on semantic predictions. For example, kettle placed at the counter would be predicted (low PE), kettle placed at table would be neither strongly predicted nor unpredicted (medium PE), and kettle placed at trash can would be unpredicted (high PE). The authors used both recall and an alternative forced-choice task to test subsequent memory for the object-location pairs. Their results followed the U-shape function of PE, suggesting better memory for predicted and unpredicted events when compared to the medium level. On the other hand, the authors also pointed out that it remains unknown to investigate the U-shape function when predictions are driven from an episodic context rather

than pre-experimental knowledge (Ortiz-Tudela et al., 2021).

We aimed to address the issue of the two conflicting ends of the U-shape, i.e., the memory congruency effect and the boosting effect of PE, in an episodic memory context. Importantly, previous studies that showed the benefits of PE on subsequent memory were either missing a medium level (Kafkas & Montaldi, 2018) or their medium level was unrelated to experimentally-induced associations (Greve et al., 2018). In our two preregistered studies, we attempted to address this issue by creating a medium level that was being related to the previously induced prediction learning. We asked our participants to learn associations between cue-target pairs (i.e., Experiment 1: musical instrument sound and object categories; Experiment 2: environment and item categories) and to generate predictions based on these associations. Hereafter, we violated their predictions with individual items in three PE levels, low, medium, and high, respectively. The subsequent memory for the individual items and their associations were assessed. We expected that this paradigm would help us to test the U-shape function of PE as a continuum, with the medium level related to experimentally induced prediction learning. Testing the SLIMM model with these two preregistered studies that were meant to be conceptual replicates of each other, we hypothesized that there is a U-shape relationship between PE level and recognition memory performance. We further hypothesized that low and high PE levels would have better recognition memory in comparison to medium PE level. Moreover, we expected that performance on the association test varies as a function of PE and there is a significant difference between confidence ratings across different PE levels.

Experiment 1

Experiment 1 sought to test the U-shape function of different PE levels on recognition memory using associations between auditory and visual stimulus categories. During the first day,

participants were presented with sounds of musical instruments and asked to predict the upcoming object category which can be either natural or human-made. More importantly, the musical instrument categories predicted the object categories in varying degrees (please see Figure 1D). In line with the study from (Schapiro et al., 2012), repeated pairing of associations would enable one to generate better predictions over time, through statistical learning. On the second day, they were again presented with sounds of musical instruments and asked to predict the upcoming object category in their mind based on what they have learned on the previous day. After their prediction, individual objects were presented to create three levels of PE, low, medium, and high PE. It should be noted that three levels of PE were not only based on the varying levels of contingency, but also on categorical differentiation which comes from subcategories of natural and human-made objects (see Procedure section below for details). A surprise memory task followed the encoding phase in which item memory for the objects and associative memory for the object-sound pairs were tested.

Method

Participants

60 participants (46 females, aged 18–29 years, mean age = 21.92 (SD = 2.84)) were recruited for the study. They were recruited through advertisements across the campuses of Goethe University of Frankfurt, student social media groups, and Prolific (https://www.prolific.co/). For their compensation, participants received either course credits or eight € per hour. All participants had normal or corrected-to-normal vision and hearing. Participants who reported a history of neurologic or psychiatric disorder were excluded from participation. They all signed an informed consent approved by the local ethics committee of the

Goethe University Frankfurt before their participation. The study design and analyses were preregistered on the Open Science Framework (https://osf.io/wybtn) before data collection.

Since the main aim of the study was to understand the effect of PE on recognition memory, an absence of generating accurate predictions on the associations between object and sound categories would make the memory results difficult to interpret. Therefore, in line with our preregistration, we decided to exclude participants with poor learning performances of less than a 65 % accuracy rate. Ten participants who could not reach the criterion were excluded from the further analysis steps.

Material

The stimuli consisted of sounds of musical instruments and object pictures. The sounds of musical instruments were selected from four categories: guitar, trumpet, violin, and piano with eight distinct sounds per category. A total number of 196 object pictures was selected from the BOSS database (Brodeur et al., 2014). The objects were equally divided into two main object categories, natural and human-made, with two sub-categories each. For natural objects the sub-categories consisted of animals and fruits/vegetables/nuts, for human-made objects the sub-categories consisted of household and toys/school/sports objects.

Procedure

The study was conducted over two sessions taking place on two consecutive days and lasting one hour each (see Figure 1). On the first day, participants were trained to learn the sound-object category associations to build up predictions. On the next day, they were asked to complete encoding and recognition phases. Since the study took place online due to the pandemic, we implemented additional steps into the data collection procedure to gain traction on data quality (Newman et al., 2021). Each session started with a video call with the participant to

check their overall well-being and physical environment. All participants were informed that they must be in a quiet room, sitting in a comfortable chair, using a computer with a stable internet connection, and to minimize distractions to be able to focus on the task. Also, we subdivided each task into several blocks and suggested our participants to take short breaks in between. The information about online testing was followed by the task instructions. The instructions were given in both spoken (during the video call) and written form (during the task). Once the participant completed the task, they were asked to video call the experimenter again to give feedback about their participation and talk about any unforeseen problems that might resulted in the incompletion of the task. In addition to the video call, they were strongly encouraged to watch the video prepared by our team to get familiarized with online testing prior to the study (https://www.psychologie.uni-frankfurt.de/102061001/Instructions_for_Online_Testing___English_Version). The presentation of stimulus and response collection were programmed in PsychoPy v2021.1.4 and https://pavlovia.org was used to run the task (Peirce, 2007).

Prediction Learning Phase. The prediction learning phase served to build up predictions about the sound-object category associations (see Figure 1A). For each of four sound categories, there were one strongly (70 %), one mildly (20 %), and two weakly (10 %) associated object categories. Most importantly, strongly and mildly associated categories were derived from one of the main categories, namely natural or human-made. Figure 1D demonstrates an example of the structure of the association between sound and object categories. In the figure, guitar sounds were strongly associated with animals. Consequently, whereas the mild association was fruits/vegetables/nuts, the weak associations were human-made categories.

During the task, participants were asked, after hearing the sound, to predict the upcoming object category. The contingency structure of the task was unknown to participants, and they had to learn the associations across trials. Each trial started with a fixation cross at the center of the screen for 1000 ms and was followed by the sound for 3000 ms. Participants were then asked to predict the category of the upcoming object based on two levels, natural or human-made. After their response, an object from the sixteen exemplars was presented for 3000 ms following the designated contingencies. Although the category prediction task was self-paced with no time limit, participants were encouraged to be as fast and accurate as possible. All participants completed 200 trials equally spread over five blocks. The association between sound and object categories was counterbalanced across participants to keep the sub-category structure stable.

Encoding Phase. The second session of the study took place approximately 24 hours after the first session and started with the encoding phase (Figure 1B). In this phase, we aimed to violate predictions with individual object pictures which have not been presented before. The following contingencies were used: 50 %, 30 %, and 20 % for strongly, mildly, and weakly associated pairs, respectively, to maintain the original contingencies as close as possible to the prediction learning phase while increasing the number of trials possible for the weakly associated category. As in the prediction learning phase, the strongly and mildly associated categories were derived from one of two main object categories.

Each trial started with a fixation cross at the center of the screen for 1000 ms and the musical instrument sound was presented for 3000 ms. After hearing the sound, participants were presented with a blank screen and asked to predict the most likely object category in their mind for 2000 ms. More importantly, they were previously informed about the four different object categories, namely, animals, fruits/vegetables/nuts, household, and toys/school/sports objects,

and were instructed to predict one of those categories. Participants were informed about the four object categories on Day 2 because it was aimed to create medium PE level via varying levels of contingencies and semantic subcategories. The paired object picture was presented for 3000 ms. Then, participants were asked to indicate whether the presented object belongs to the same category which they predicted based on a 4-level Likert scale (from 1: Strongly yes to 4: Strongly no). The encoding phase consisted of three blocks of 40 trials, for a total of 120 trials.

Recognition Phase. Immediately after the encoding phase, a surprise recognition phase started (Figure 1C). In addition to the 120 object pictures from the previous phase, 60 new objects were included. The new objects were equally selected from the four object categories. Each trial started with a fixation cross for 1000 ms and was followed by the object picture presentation for 3000 ms. Participants were required to give an old/new judgment and to provide their confidence based on a 4-level Likert scale (from 1: very sure to 4: very unsure). Hereafter they were asked to select the sound (between two options) that was associated with the presented object. They listened to two sounds one after another and indicated which one was paired with the object. To avoid guessing biases based on the previously learned associations, the alternatives were selected from the same musical instrument category. All responses were self-paced with no time limit. A total of 180 trials was distributed across three blocks.

Statistical Analyses

To test the effect of PE levels on recognition memory performances, we assessed the participants' hit responses based on correct answers for old items. Before testing the main hypothesis, cumulative accuracy scores were computed to exclude participants with poor learning performances (below 65 %). Also, mean prediction rates from the encoding phase were

assessed as a sanity check for our PE manipulation. For all phases of the study, trials with reaction time shorter than 100 ms or longer than 1500 ms were excluded.

To determine whether PE level was a significant predictor of recognition memory performance, we conducted a linear mixed effect model with participant as random intercept to account for between-participant variability in hit responses. The model included PE level (low, medium, and high PE) and confidence ratings (from 1: very sure to 4: not unsure) as fixed within-participant factor. Model estimations were determined with maximum likelihood ratio and the statistical significance of the fixed effects was determined using $\chi 2$ (chi-squared) tests. In addition to our primary analyses on the effect of PE level on hit responses, we also analyzed the responses for the associated sound pair with the same model specification. All analyses were performed using custom-made R scripts with the lmer function in the lme4 package (Bates et al., 2015) that can be found on the OSF page (https://osf.io/pfgyb/).

In addition to our primary hypothesis of mean differences across PE levels, we also exploratorily analyzed the relationship between learning, PE, and recognition performance. Assumingly, participants whose learning performance was better would benefit more from high PE compared to participants with lower learning performance. We assessed learning performance via cumulative accuracy from the last third of trials from the prediction learning phase. A linear mixed effect model was calculated with participant as random intercept. We included learning performance and PE level as fixed factors into the model.

Deviations from the registered protocol

The current study was preregistered prior to the data collection (https://osf.io/wybtn) and there are some deviations from it which are important to mention (Claesen et al., 2021). Firstly, even though we planned to test 45 participants assuming an effect size of .25 to observe .80

power, additional participants had to be recruited due to problems related to exclusion criteria, online testing and failed data transmission. The second main difference from the preregistered plan was in the encoding phase. We had planned to ask our participants to report the category of the presented object (e.g., animal or fruit). However, during the pilot studies, it revealed that participants were not attentive to the associations between sounds and objects, since attending the object only would be sufficient to solve the task which was selecting its category. Therefore, generating different PE levels for associations between sound and object categories was not possible although it was crucial for the study. We correspondingly decided to update the encoding phase in which participants indicated if the presented object belongs to the same category which they predicted. Lastly, the preregistered plan was to test hypotheses with a within factors repeated measures ANOVA. We changed our statistical model to a linear mixed model because it allows us to control for the variance attributed to random factors (e.g., participants) and it was more suitable for an unbalanced number of observations since in the present study the number of trials differed among PE levels due to the nature of the experimental design.

Results

First, to check if participants generate predictions about the sounds of musical instruments and object categories, we examined the accuracy results from the prediction learning phase. Mean accuracy across participants was .87 (SD = .33) during the prediction learning phase that was significantly above chance levels of performance ($t(49) = 28.75$, $p < .001$, $d = 4.07$). Figure 3A shows that the accuracy of predicting the upcoming object category increased with trials. This shows that participants could generate predictions about the associations between musical instruments and object categories.

Then, to test our manipulation on the PE levels, participants' category judgments during the encoding phase were investigated. Ratings about the distance between the presented object category and the category they previously predicted were collected. While the highest ratings (i.e., 4- Strongly no) mean that the presented object did not belong to the same category that participants predicted earlier, the lowest ratings (i.e., 1- Strongly yes) indicate that the presented object was from the same category that they predicted. The results can be seen in Figure 3B. The model for the encoding phase with participant as random intercept and PE level as fixed factor showed that there was a significant main effect of PE ($\chi^2(2) = 292.3$, $p < .001$), indicating that higher ratings were obtained for high PE trials ($\beta = 1.98$, $t = 16.88$, $p < .001$) and medium PE trials ($\beta = .56$, $t = 5.76$, $p < .001$) compared to low PE trials. This result indicates a strong verification for our manipulation on the PE levels, such that our participants reported that objects from high and medium PE levels are not from the category they predicted.

The model to predict recognition memory performance for PE levels with participant as random intercept indicated a significant main effect of PE ($\chi^2(2) = 8.27$, $p = .02$). Post-hoc contrasts showed that hit responses for high PE levels ($\beta = -.24$, $z = -2.79$, $p = .01$) were lower than low PE level (see Figure 3C). In addition to PE level, we added confidence ratings as fixed effect for random intercept and random slope into the model and compared both models. The model fit was significantly improved, $\chi 2(81) = 871.58$, $p < .001$; AIC first: 6293.6, second: 5584.1. The main effect of confidence ($\chi^2(3) = 169.08$, $p < .001$) was significant, indicating higher hit responses obtained at the highest confidence rating (1- Very sure) compared to the other confidence ratings (2- Sure: $\beta = -1.77$, $z = -6.89$, $p < .001$; 3- Unsure: $\beta = -2.78$, $z = -9.18$, $p < .001$; 4- Very unsure: $\beta = -3.42$, $z = -8.99$, $p < .001$). The interaction effect between PE and confidence was significant ($\chi^2(6) = 12.91$, $p = .04$), which indicated that participants' hit

responses were lower for high PE level compared to low PE level at lower confidence ratings (3- Unsure: $\beta = .81$, $z = 2.57$, $p = .01$; 4- Very unsure: $\beta = 1.08$, $z = 2.37$, $p = .02$). The results can be seen in Figure 3D. Lastly, we applied a similar model to test the association memory accuracy. There was neither significant main nor interaction effect of PE level on association memory accuracy, $\chi 2(2) = 1.90$, $p = .45$.

We further tested our exploratory question about the relationship between learning, PE, and recognition memory. The model was run to predict recognition memory performance with PE levels and cumulative accuracy scores with fixed effects and participant as random intercept. Results showed significant main effect of PE ($\chi^2(2) = 8.40$, $p = .02$). Post hoc comparisons indicated that highest hit responses were measured for low PE level ($\beta = .75$, $t = 4.87$, $p < .001$). Neither the effect of learning nor the interaction was significant.

Discussion

Experiment 1 showed that while three levels of PE were successfully constructed during the encoding phase by generating predictions about the associations between sound and object categories in the prediction learning phase, the hypothesized U-shape function of PE on recognition memory was not observed. In contrast, our results indicated that recognition memory was better for low PE than high PE, suggesting a memory congruency effect. As one can argue that the U-shape function of PE is observable mostly for associative memory because of its role in creating “snapshots” for high PEs (Henson & Gagnepain, 2010), we also assessed associations between individual items and their sound pairs, but we did not observe a memory benefit for high PE. Even though these findings provide contradicting evidence for the U-shape relationship between PE and memory, it is crucial to rule out that the absence of this PE effect might be linked to the task-related differences. For example, we argue that it should be revisited how

predictions are experimentally built up. It is very likely to make a difference if pre-experimental knowledge (i.e., semantic categories) is used or certain contingency structures are established during the prediction learning phase. In other words, in Experiment 1, the contingency structure was built up during the very first phase of the study with varying degrees of PE which was also based on semantic categories of the objects. The issue with this task structure might be that participants were not sensitive to the individual items shown during the encoding phase because participants had been acquainted with being presented with objects which were from different semantic categories than their predictions. Experiment 2 sought to address these issues.

Experiment 2

Experiment 2 followed the same rationale as Experiment 1: Testing the U-shape function of PE on memory with all levels based on experimentally-induced prediction learning. However, there are three main differences between our two experiments worthwhile highlighting. First, due to the nature of Experiment 1, the PE levels were based on semantic sub-categorization (e.g., animals and fruits/vegetables/nuts for the natural object category). We aimed to rule out any potential effects of previous semantic knowledge by using associations between artificial creatures called “Wubbels” and their environments in Experiment 2 (Watson et al., 2019). Secondly, in Experiment 1, varying contingencies started from the very first phase (i.e., the prediction learning phase). During the prediction learning phase, the musical instrument categories predicted the object categories to varying degrees, thus this setup might have resulted in participants having the understanding that their predictions can be sometimes incorrect. As a consequence, participants might not have been sensitive to medium and high PE trials during the encoding phase, because they already knew from the first phase that sometimes the presented object does not match with their predictions. Therefore, in Experiment 2, we decided to have a

deterministic prediction learning phase in which the contingency was set to 100 %. Lastly, in Experiment 2 we measured two additional aspects of associative memory, i.e., Wubbel-scene pair and Wubbel-location pair (see Procedure section for details).

As in Experiment 1, the study consisted of three phases: Prediction learning, encoding, and recognition (see Figure 2). In the prediction learning phase, participants were asked to learn Wubbel-scene associations via feedback across trials. During the encoding phase, they were presented with individual, unique Wubbels that varied to certain degrees to create three PE levels. In the end, recognition and associative memory were assessed to test the U-shape function of PE on memory.

Method

Participants

51 participants (28 females, aged 18-35 years, mean age = 23.14 (SD = 4.49)) were recruited in this study. They were recruited through advertisements across the campus, student social media groups, and Prolific (https://www.prolific.co/). The inclusion criteria were having normal or corrected-to-normal vision and hearing. Participants with a history of neurologic or psychiatric disease were not recruited for the study. All participants signed an informed consent approved by the local ethics committee prior to their participation and they were given either course credits or honorarium for their compensation. The preregistered study design and analyses can be found in the Open Science Framework (https://osf.io/bwujz). With the same rationale as in Experiment 1, we excluded one participant from the analysis due to poor learning performance with a less than 40% accuracy rate.

Material

We used associations between the artificial creatures called "Wubbels" and certain environments to prevent the effect of prior knowledge (Watson et al., 2019). The scenes for environments were selected from the ECOS database (https://sites.google.com/view/ecosdatabase/) with four distinct categories: beach, snowy mountains, desert, and savannah. There were four different exemplars for each environment category.

The Wubbels were created with the Autodesk 3DS Max software using the provided script by Watson et al. (2019). The creation of Wubbels followed a schema according to which the main feature defining the affiliation with one of the two species is the body shape, i.e., longish or roundish. The body shapes of the four Wubbel families were consistent with their primary species. That is, families one and two, which are affiliated with the *long-shaped* species, have *concave* and *oblong* body shapes, whereas families three and four, which are affiliated with the *round-shaped* species, have compressed *oblong* and *spherical* body shapes (see Figure 2D). We structured Wubbels into two main groups, the so-called Wubbel parents and Wubbel children to use them during the prediction learning and the remaining phases (i.e., encoding and recognition), respectively.

The main difference between Wubbel parents and Wubbel children was that the children have varying features. Except for the Wubbel parents, i.e., the prototype pairs used for the prediction learning phase, the additional features of the Wubbel children, such as hat shape (6 instances), arm shape (4 instances), skin type (6 instances), body color (13 instances), pattern color (13 instances), and pattern (50 instances) were variable. They were attached to the body strategically to ensure that the Wubbels are as dissimilar from another as possible. To this end, a

matrix with all possible feature combinations was built from which 80 combinations with a distance of at least three features were randomly drawn. The drawn combinations provided the building instructions for the Wubbels. That is, we aimed to guarantee that at least three features do not overlap when one compares all Wubbels to one another. This procedure resulted in a set of 80 unique Wubbel children which were used in the encoding and recognition phases. In contrast to the Wubbel children, the parents differ as the assigned color patterns are for one prototype vertically striped and for the other prototype horizontally striped rainbow colors, respectively. Similarly, as for the Wubbel children, each prototype had a unique combination of the remaining features complementing its body. All color and pattern patches were created with Python 3.7.4 using the OpenCV library.

Procedure

There were three study phases on two consecutive days and each lasted for one hour. On day 1 in the prediction learning phase, participants learned associations between Wubbels and environments. The second session on day 2 started with the encoding phase and was followed by the recognition phase. This study was also run online. Therefore, we followed the same structure as in Experiment 1 including the video call with participants in order to increase data quality. The presentation of stimulus and response collection were programmed in PsychoPy v2021.1.4 and https://pavlovia.org was used to run the task.

Prediction Learning Phase. The purpose of the prediction learning phase was to let participants learn the Wubbel-environment associations to be able to generate predictions. The phase started with a cover story in which the Wubbels, their species, and their families were introduced. Participants were told that each Wubbel family lives in a different environment and they were asked to learn these combinations as quickly as possible (Figure 2A). There were two prototypes for each four families and four scene pictures for each four scene categories. The associations between Wubbels and environments were predetermined, and the contingency was 100 %.

Each trial started with a scene and a fixation cross in the center for 500 ms. Then, two Wubbels from different families were presented in two out of four possible screen locations (i.e., top-right, top-left, bottom-right, and bottom-left). Here, participants were asked to decide which Wubbel matches with the presented scene by giving a response. After their response, feedback was presented via a green (correct) or red (incorrect) frame around the chosen Wubbel according to the pre-determined associations based on 100 % contingency for 3000 ms. The prediction learning task was self-paced with no time limit, but we encouraged our participants to be as fast and accurate as possible. The total number of trials was 96 for one block. The associations between Wubbels and environments were counterbalanced across participants.

Encoding Phase. The second session started with the encoding phase approximately 24 hours after the first session (Figure 2B). To create three different levels of PE, individual, unique Wubbel pictures, which we told the participants as Wubbel children, were used. We used twelve Wubbel-children for each family and divided them into three conditions, low, medium, and high PE levels. In detail, a Wubbel child shown in an environment of its own family would elicit a low PE, a Wubbel child from a different family but within the same species would lead to a medium PE in that same environment, and a Wubbel child from the other species would lead a high PE. As in the example presented in Figure 2, participants learned in the prediction learning phase that one family prototype of the long-shaped Wubbel species lives on the beach. When a child from this family was presented with a beach during encoding, it would lead to low PE because participants would have expected to find a family member with such a longish body shape. When the same child was presented with an environment category associated with the other long-shaped family, i.e., a desert, it would result in medium PE level. Lastly, presenting the same child with any environment associated with a round-shaped family, i.e., a savannah, would create high PE.

Prior to the task, participants were first instructed about the structure of Wubbel families and their children. It was explained that the children have the same body shapes as their parents and that they are similar to their relatives as they have similar body shapes (please see Figure 2D). It was also described that they are very different from the other two unrelated families who have very different body shapes. Moreover, they were told that they can see them in different environments since the children of Wubbels always visit each other because they enjoy meeting other Wubbels. The contingencies for PE levels were equal, meaning twelve Wubbels for each

PE level. The participants' task was to indicate the matching level between Wubbels and the environments.

Each trial started with a fixation cross at the center of the screen for 500 ms and a scene was presented for 3000 ms. During the scene presentation, participants were asked to predict the most likely Wubbel family in their minds based on what they have learned the day before. Then, a Wubbel child was presented on one out of four possible screen locations for 3000 ms. Participants were asked to indicate whether the presented Wubbel child belongs to the same family which they predicted or not, using a 4-level Likert scale (from 1: Strongly yes to 4: Strongly no). This was followed by a 2500 ms blank screen to wait for the next Wubbel child. The total number of trials was 48 and the association structure from the previous day was the same. Different from Experiment 1, we told participants that there will be a memory test for the Wubbel children, their features, the environment as well as the location on the screen they were shown at. Thus, whereas Experiment 1 was incidental, this study was based on intentional learning.

Recognition Phase. The recognition phase followed the encoding phase (Figure 2C). The Wubbels from the previous phase were presented with 32 new Wubbel children - eight from each family. Each trial started with a fixation cross for 500 ms. Then, participants had to indicate whether they have seen the presented Wubbel before or not and rate their confidence based on a 4-level Likert scale (from 1: very sure to 4: very unsure). Regardless of their response, scene association and scene location tasks were employed for old Wubbel children. Participants had to choose the correct scene from four alternatives. Importantly, the alternative scenes were from the same scene category to prevent guessing biases. Lastly, participants were asked in which location on the screen the Wubbel was presented. All responses were self-paced with no time limit. All participants completed 80 trials.

Statistical Analyses

The steps for statistical analyses were identical to Experiment 1.

Deviations from the registered protocol

Our preregistered study and analysis plan can be found here: https://osf.io/bwujz. Due to unforeseen reasons, we had to deviate in several aspects during the study, with the reasons being summarized here. As in Experiment 1, due to problems related to online testing, we had to test more participants than we originally reported. The initial sample size was 40 to obtain .80 power with an effect size of .40 at the standard .05 alpha error probability. During the pilot task, Wubbels had two main features, namely body shape, and color. However, it had not been anticipated that the color information overshadowed body shape information, as a consequence, participants considered solely the body shape information to accomplish the task. Unfortunately, we could not create the different levels of PE even though it was our main manipulation of the task. Thus, we decided to only have the body shape as the main characteristics to define species

but not color information. The other important deviation was to have a different structure for the encoding phase. The pilot task with the like/dislike task did not show significant difference in PE levels. Therefore, we changed the structure and asked participants to evaluate if the presented Wubbel matched with the scene. The last deviation concerned the analysis plan. Similar to Experiment 1, we decided to run linear mixed models instead of repeated measure ANOVA due to the aforementioned reasons.

Results

We first checked prediction learning performance during the first phase. Mean accuracy across participants was .88 (*SD* =.32) which was significantly above chance levels of performance ($t(50) = 29.34$, $p < .001$, $d = 4.11$). As in Experiment 1, learning performance for the associations between Wubbels and scene categories increased with trials indicating that participants were able to generate accurate predictions (Figure 3A). Then, we tested the effect of PE level on the category judgments in the encoding phase (Figure 3B). The results showed a main effect of PE, $\chi^2(2) = 690.66$, $p < .001$. The ratings were higher for high PE trials ($\beta = 2.39$, $t = 25.83$, $p < .001$) and medium PE ($\beta = .96$, $t = 10.58$, $p < .001$) compared to low PE. Together with these results, we further supported our PE manipulation.

As in Experiment 1, the model to predict hit responses for PE levels with participants as random intercept showed a significant main effect of PE ($\chi^2(2) = 13.63$, $p = .01$). Post-hoc comparisons indicated that hit responses for high PE level were lower than low PE level, $\beta = -.36$, $z = -3.23$, $p = .01$. Results can be seen in Figure 3C. Next, we continued by adding confidence ratings to the model. The model fit was significantly improved, $\chi 2(81) = 153.36$, $p < .001$; AIC first: 3173.2, second: 3181.8. The main effects of PE level ($\chi^2(2) = 9.30$, $p = .01$) and confidence ratings ($\chi^2(3) = 8.24$, $p = .05$) were significant but there was no interaction effect,

$\chi 2(6) = 5.28$, $p = .51$. Post hoc comparisons only showed that higher hit responses were recorded at rating level 2- Sure, $\beta = .53$, $z = 1.77$, $p = .07$ (Figure 3D). Neither the results for association memory for scene pair ($\chi 2(2) = 2.94$, $p = .23$) nor the results for association memory for location ($\chi 2(2) = .02$, $p = .99$) indicated a main effect of PE level.

For our exploratory analysis, the mixed-effect logistic regression to predict recognition memory performance for different PE levels and cumulative accuracy scores from the prediction learning phase indicated that there was a significant main effect of PE ($\chi^2(2) = 12.14$, $p < .01$). Post hoc comparisons indicated that hit responses for high PE levels ($\beta = -.01$, $t = -.02$, $p = .02$) were lower than for low PE levels. Neither the effect of learning nor the interaction was significant. These findings demonstrate that beyond the proxy of recognition memory, better performances were obtained at low PE compared to high PE level, which was in line with Experiment 1.

Discussion

Experiment 2 replicated the better recognition memory for low PE, i.e., memory congruency effect, and the absence of memory benefit for high PE which were found in Experiment 1. This stands in contrary to some previous studies (Brod et al., 2018; Greve et al., 2017; 2019; Kafkas & Montaldi, 2018; Quent et al., 2022) that documented better memory for events that elicit high PE. Despite not using pre-experimental knowledge to rule out the semantic memory processes and despite having a fully deterministic contingency in the prediction learning phase, better recognition memory results were obtained only for low PE. Notably, Experiments 1 and 2 both indicate coherently the lack of evidence for memory advantages for high PE.

General Discussion

Within two preregistered studies, we examined the effects of different PE levels on recognition memory performance. To test the hypothesized U-shape function of PE on episodic memory, we first asked our participants to learn novel contingency structures and generate predictions, then violated these predictions on three levels (i.e., low, medium, and high PE) with all levels being related to experimentally-induced prediction learning. We showed that participants were able to learn from the provided regularities and successfully formed predictions. Even though our findings indicated a strong verification for our manipulation of PE, there was no memory advantage for high PE level neither in the recognition nor in the association tasks. Rather, we consistently found a memory advantage for low PE trials, in line with memory congruency effect (Alba & Hasher, 1983; Anderson, 1981; Craik & Tulving, 1975). In addition to our primary findings, we exploratorily investigated the relationship between learning performance and recognition memory. We hypothesized that a better learning performance would lead to better recognition performance for high PE trials. Contrary to our hypothesis, the results were in line with our main findings suggesting a memory congruency effect but no memory advantage for high PE. Thus, the current studies clearly show that high PE does not guarantee subsequent memory benefit.

A body of research has shown that events giving rise to PE are remembered better, such that they facilitate new learning for better predictions in the future (Bein et al., 2021; Brod et al., 2018; Greve et al., 2017; Kafkas & Montaldi, 2018; Quent et al., 2022). Although PE is sometimes taken for granted as a driver of new learning (Greve et al., 2017), its direct behavioral effect on memory may depend on several factors, such as the task sensitivity concerning how PEs are experimentally generated and tested. One explanation why there was no memory benefit for high PE trials in the current studies can be the differences in the experimental paradigms, for

example, how the encoding and the recognition phases were structured. In the following sections, we will discuss these points.

The first of these differences in the experimental designs can be examined via the differences in encoding of PEs. The studies which reported high PE benefits on memory had their PE manipulation and memory test based on the item level (Bein et al., 2021; Kim et al., 2017). In the current studies, we let our participants to generate predictions on the category level, because the task was to learn the associations between cue-target categories and predict the upcoming target category based on the cue. Nevertheless, participants were tested on their memory for the target (i.e., item level) at retrieval. Our participants might have only focused on the category-level information during encoding rather than on individual items which were tested later. Evidence for the effect of PE on association memory has also been demonstrated in previous studies (Greve et al., 2017; Quent et al., 2022). Unfortunately, the obtained association memory performance was below chance in our experiments. The difficulty level of study materials can be an issue both for musical instruments and the Wubbels. For example, participants might have found it challenging to discriminate guitar sounds from each other. On top of that, since processing the guitar sound category was already informative enough to accomplish the task, participants might have had a shallow encoding for the sounds.

Yet another explanation for why the current studies did not show a memory advantage for high PE levels would stem from our experimental approach to assessing memory performance. Previous studies that provided a PE benefit tested recognition memory via alternative forced-choice tests (Greve et al., 2017; Quent et al., 2022) or mixed lists with similar lures (Bein et al., 2021; Frank et al., 2020), unlike our study which used an old/new paradigm. One can argue that PE creates distinct memory traces (i.e., snapshots) which leads to better recognition memory in

return. However, it may not be possible to evaluate its memory traces via old/new paradigms since only a level of familiarity would be sufficient to dissociate the old items from the new ones. Even though we assessed the memory performance via confidence ratings to deal with the issue of familiarity, the results still were not suggesting a memory benefit for high PE with high confidence.

On the contrary, our results were in line with the rich literature on the memory congruency effect (Alba & Hasher, 1983; Anderson, 1981; Craik & Tulving, 1975). One possible reason could be that high PE items were ignored because the task was making "better" predictions to correctly respond what will be presented. During the encoding phase, participants were asked to predict the upcoming object category in their mind, and they were presented with objects with varying degrees of PE. In order to make correct predictions for the future, participants might have tended to leave out the items which were not in line with their predictions. This semantic encoding has been considered to have a great adaptive value to enhance the functionality of the memory system even though it may result in distortions (Schacter et al., 2011). Moreover, a recent framework (i.e., PACE, Gruber & Ranganath, 2019) postulated that enhanced memory encoding for PE is also based on the evaluation of information that can be valuable in the future. PEs might not be sufficient to trigger new learning because its appraisal is not important. As stated previously, high PE trials in our studies were not informative for the task at hand, potentially leading to these trials not being encoded better. One can thus infer that PE does not necessarily benefit memory. If events that give rise to PE are not evaluated as informative for future functioning, our memory system may tend to ignore them and rely more on the existing predictions.

Besides preceding explanations regarding why high PE does not benefit subsequent memory, additional evidence might derive from the research on cognitive conflict. A recent study (Ptok et al., 2021) showed that where the manipulation for conflict took place might have a crucial impact on the memory benefit. The authors run a series of experiments to investigate the effect of locus of processing conflict on memory benefit. They found memory benefit for incongruent items when the conflict is on the to-be-tested item. On the other hand, changing the attentional focus from the to-be-tested-item to the response does not lead to better memory. For example, Lisa (female name) with an incongruent distractor, male, would lead to a better recognition memory, whereas "Lisa – press right button" as an incongruent response information does not show a memory benefit. The authors concluded that having a violation and attentional focus on the to-be-tested item predicts the subsequent memory benefit. On the other hand, in our studies, although we had our PE (cf. conflict) on the individual level, namely the to-be-tested items, the task was to decide whether the presented object matched with participants' prediction. Therefore, participants might need to revisit their previous knowledge about the associations in order to do the task. As a consequence, they might have had an attentional switch from the item to the category level. This attentional switch could explain the better recognition memory for low PE trials than for high PE trials.

To conclude, our two preregistered studies provided novel paradigms to generate and violate PEs in varying degrees that indicated memory advantage for the events in line with predictions but not for the ones giving rise to PE. These findings suggest that it remains elusive to illustrate the U-shape relationship between prediction and memory. We conclude that it is important to investigate the specific condition in which a U-shape relationship could be reliably found. Relatedly, we showed in another study (Ortiz-Tudela et al., 2022) an inverted U-shape

function instead of the U-shape function as suggested in the SLIMM Model (van Kesteren et al., 2012) in which we used a different experimental manipulation of continuous PE through prior strength. This indicated that the uncertainty level of generated prediction can modulate how PE affects memory. Our convergent results underscore that the effects of PE on episodic memory are complex, and there are potentially other modulating factors that may offer a better roadmap for further exploring PE as a driver of new learning and a possible reason for better memory.

References

Alba, J. W., & Hasher, L. (1983). Is memory schematic? *Psychological Bulletin*, *93*(2), 203–231. https://doi.org/10.1037/0033-2909.93.2.203

Allen, L., Brand, A., Scott, J., Altman, M., & Hlava, M. (2014). Credit where credit is due. *Nature*, *508*(7496), 312–313. https://doi.org/10.1038/508312a

Anderson, J. R. (1981). Effects of prior knowledge on memory for new information. In *Memory & Cognition* (Vol. 9, Issue 3).

Atienza, M., Crespo-Garcia, M., & Cantero, J. L. (2011). Semantic Congruence Enhances Memory of Episodic Associations: Role of Theta Oscillations. *Journal of Cognitive Neuroscience*, *23*(1), 75–90. https://doi.org/10.1162/JOCN.2009.21358

Bar, M. (2007). The proactive brain: using analogies and associations to generate predictions. *Trends in Cognitive Sciences*, *11*(7), 280–289. https://doi.org/10.1016/j.tics.2007.05.005

Bates, D., Mächler, M., Bolker, B. M., & Walker, S. C. (2015). Fitting Linear Mixed-Effects Models Using lme4. *Journal of Statistical Software*, *67*(1), 1–48. https://doi.org/10.18637/JSS.V067.I01

Bein, O., Plotkin, N. A., & Davachi, L. (2021). Mnemonic prediction errors promote detailed memories. *Learning and Memory*, *28*(11), 422–434. https://doi.org/10.1101/LM.053410.121

Bradski, G. (2000). The OpenCV Library. *Dr. Dobb's Journal of Software Tools.*

Brod, G., Hasselhorn, M., & Bunge, S. A. (2018). When generating a prediction boosts learning: The element of surprise. *Learning and Instruction*, *55*(2018), 22–31. https://doi.org/10.1016/j.learninstruc.2018.01.013

Brod, G., & Shing, Y. L. (2019). A boon and a bane: Comparing the effects of prior knowledge on memory across the Lifespan. *Developmental Psychology*, *55*(6), 1326–1337. https://doi.org/10.1037/dev0000712

Brod, G., Werkle-Bergner, M., & Lee Shing, Y. (2013). The influence of prior knowledge on memory: A developmental cognitive neuroscience perspective. *Frontiers in Behavioral Neuroscience*, *7*, 1–13. https://doi.org/10.3389/fnbeh.2013.00139

Brodeur, M. B., Guérard, K., & Bouras, M. (2014). Bank of Standardized Stimuli (BOSS) phase ii: 930 new normative photos. *PLoS ONE*, *9*(9). https://doi.org/10.1371/journal.pone.0106953

Claesen, A., Gomes, S., Tuerlinckx, F., & Vanpaemel, W. (2021). Comparing dream to reality: an assessment of adherence of the first generation of preregistered studies. *Royal Society Open Science*, *8*(10). https://doi.org/10.1098/RSOS.211037

Craik, F. I., & Tulving, E. (1975). Depth of processing and the retention of words in episodic memory. *Journal of Experimental Psychology: General*, *104*(3), 268–294. https://doi.org/10.1037/0096-3445.104.3.268

Frank, D., Montemurro, M. A., & Montaldi, D. (2020). Pattern Separation Underpins Expectation-Modulated Memory. *Journal of Neuroscience*, *40*(17), 3455–3464. https://doi.org/10.1523/JNEUROSCI.2047-19.2020

Friston, K. (2010). The free-energy principle: A unified brain theory? *Nature Reviews Neuroscience*, *11*(2), 127–138. https://doi.org/10.1038/NRN2787

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. (2018). Knowledge is power: Prior knowledge aids memory for both congruent and incongruent events, but in different ways. *Journal of Experimental Psychology: General*, *148*(2), 325. https://doi.org/10.1037/XGE0000498

Gruber, M. J., & Ranganath, C. (2019). How Curiosity Enhances Hippocampus-Dependent Memory: The Prediction, Appraisal, Curiosity, and Exploration (PACE) Framework. *Trends in Cognitive Sciences 23*(12), 1–12. https://doi.org/10.1016/j.tics.2019.10.003

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

Kafkas, A., & Montaldi, D. (2018). Expectation affects learning and modulates memory experience at retrieval. *Cognition*, *180*, 123–134. https://doi.org/10.1016/J.COGNITION.2018.07.010

Kim, G., Norman, K. A., & Turk-Browne, N. B. (2017). Neural Differentiation of Incorrectly Predicted Memories. *Journal of Neuroscience*, *37*(8), 2022–2031. https://doi.org/10.1523/JNEUROSCI.3272-16.2017

Liu, Z. X., Grady, C., & Moscovitch, M. (2018). The effect of prior knowledge on post-encoding brain connectivity and its relation to subsequent memory. *NeuroImage*, *167*, 211–223. https://doi.org/10.1016/J.NEUROIMAGE.2017.11.032

Newman, A., Bavik, Y. L., Mount, M., & Shao, B. (2021). Data Collection via Online Platforms: Challenges and Recommendations for Future Research. *Applied Psychology*, *70*(3), 1380–1402. https://doi.org/10.1111/APPS.12302

Ortiz-Tudela, J., Milliken, B., Botta, F., LaPointe, M., & Lupiañez, J. (2017). A cow on the prairie vs. a cow on the street: long-term consequences of semantic conflict on episodic encoding. *Psychological Research*, *81*(6), 1264–1275. https://doi.org/10.1007/S00426-016-0805-y

Ortiz-Tudela, J., Nolden, S., Pupillo, F., Ehrlich, I., Schommartz, I., Turan, G., & Shing, Y. L. (2021). *Not what U expect: Effects of Prediction Errors on Episodic Memory.* PsyArXiv. https://doi.org/10.31234/OSF.IO/8DWB3

Ostreicher, M. L., Moses, S. N., Rosenbaum, R. S., & Ryan, J. D. (2010). Prior experience supports new learning of relations in aging. *The Journals of Gerontology. Series B, Psychological Sciences and Social Sciences*, *65B*(1), 32–41. https://doi.org/10.1093/GERONB/GBP081

Peirce, J. W. (2007). PsychoPy-Psychophysics software in Python. *Journal of Neuroscience Methods*, *162*(1–2), 8–13. https://doi.org/10.1016/j.jneumeth.2006.11.017

Ptok, M. J., Hannah, K. E., & Watter, S. (2021). Memory effects of conflict and cognitive control are processing stage-specific: evidence from pupillometry. *Psychological Research*, *85*(3), 1029–1046. https://doi.org/10.1007/s00426-020-01295-3

Quent, J. A., Greve, A., & Henson, R. N. (2022). Shape of U: The Nonmonotonic Relationship Between Object–Location Memory and Expectedness. *Psychological Science*, *33*(12), 2084–2097. https://doi.org/10.1177/09567976221109134

Schacter, D. L., Guerin, S. A., & st. Jacques, P. L. (2011). Memory distortion: an adaptive perspective. *Trends in Cognitive Sciences*, *15*(10), 467–474. https://doi.org/10.1016/J.TICS.2011.08.004

Schapiro, A. C., Kustner, L. v., & Turk-Browne, N. B. (2012). Shaping of object representations in the human medial temporal lobe based on temporal regularities. *Current Biology*, *22*(17), 1622–1627. https://doi.org/10.1016/j.cub.2012.06.056

Van Rossum, G., & Drake Jr, F. L. (1995). *Python reference manual.* Centrum voor Wiskunde en Informatica Amsterdam.

van Kesteren, M. T. R., Beul, S. F., Takashima, A., Henson, R. N., Ruiter, D. J., & Fernández, G. (2013). Differential roles for medial prefrontal and medial temporal cortices in schema-dependent encoding: from congruent to incongruent. *Neuropsychologia*, *51*(12), 2352–2359. https://doi.org/10.1016/J.NEUROPSYCHOLOGIA.2013.05.027

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

Watson, M. R., Voloh, B., Naghizadeh, M., & Womelsdorf, T. (2019). Quaddles: A multidimensional 3-D object set with parametrically controlled and customizable features. *Behavior Research Methods*, *51*(6), 2522–2532. https://doi.org/10.3758/S13428-018-1097-5

Author contributions: Author contributions are codded according to the CRediT taxonomy (Allen et al., 2014).

Gözem Turan: Conception, Methodology, Computation, Formal Analysis, Investigation – performed the experiments, Investigation – data collection, Data curation, Writing – writing the initial draft, Writing – review & editing, Writing – visualization, Project Administration

Isabelle Ehrlich: Conception, Methodology, Computation, Investigation – performed the experiments, Investigation – data collection, Resources, Data curation, Writing – review & editing, Project Administration

Yee Lee Shing: Conception, Methodology, Writing – review & editing, Supervision, Project Administration, Funding Acquisition

Sophie Nolden: Conception, Methodology, Writing – review & editing, Supervision, Project Administration, Funding Acquisition

Author notes: Data, scripts, and additional online materials are openly available at the project's Open Science Framework page (https://osf.io/pfgyb/). We have no conflicts of interest to disclose. The study was funded by a Starting Grant from the European Union for YLS (ERC-2018-StG-PIVOTAL-758898). The work of YLS was also supported by the German Research Foundation (Project-ID 327654276, SFB 1315, "Mechanisms and Disturbances in Memory Consolidation: From Synapses to Systems"), and the Hessisches Ministerium für Wissenschaft und Kunst (HMWK; project "The Adaptive Mind"). The work of SN was also supported by a Research grant Focus A/B, Goethe-University Frankfurt am Main, ("Dynamics of auditory and visual memory representations in the aging brain"). All authors approved the final version of the manuscript for submission. We would like to thank our

student assistants Yi You Tan and Daniel Urban for their help with data collection.

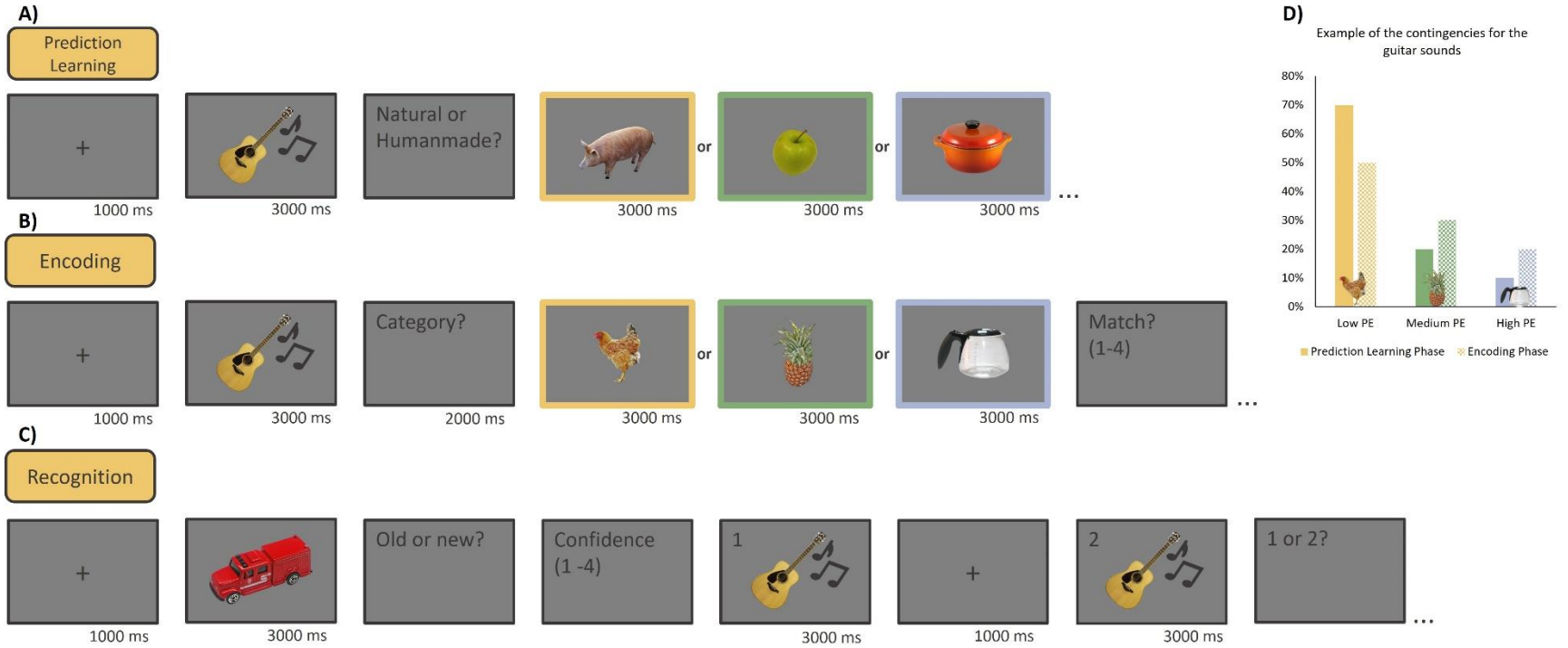

Figure 1. Study design for Experiment 1. The study was conducted on two consecutive days. A) On day 1, participants were asked to build up predictions about the sound-object category associations. In each trial of prediction learning phase, participants were presented with a musical instrument sound and asked to predict and indicate the upcoming object category based on two levels (i.e., natural and human-made). After their response, an object exemplar was shown. The contingency structure for the associations between sounds of musical instruments and object categories was unknown to participants. As also seen in panel D, the guitar sounds were followed by exemplar objects from animal categories 70% of the times (Low PE, yellow). Thus, the contingencies for exemplar objects from fruit categories (Medium PE, green) and human-made categories (High PE, blue) were 20% and 10%, respectively. B) On Day 2, during the encoding phase, a musical instrument sound was firstly presented. Participants were asked to predict the most likely object category in their mind among four different sub-categories. Participants were then presented with object pictures and asked to indicate if the presented object belongs to the same category they predicted. The contingency structure as follows, 50%, 30%, and 20% for low, medium, and high PE levels, respectively. C) During the recognition phase, participants were asked to make old/new

judgments on the test pictures with their confidence ratings, and they were asked to indicate the paired sound as well. D) Example of the contingency structure for the guitar sound.

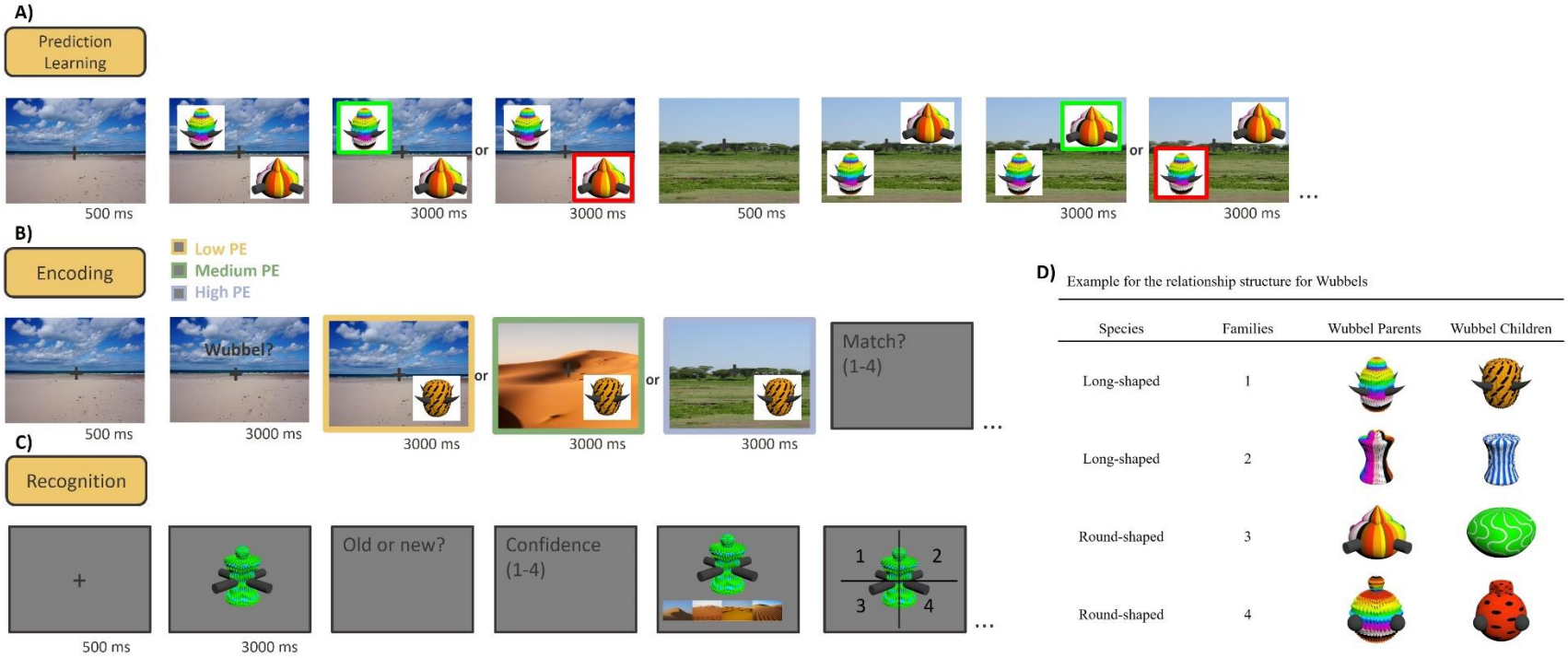

Figure 2. Study design for Experiment 2. The study was conducted on two consecutive days. A) On day 1, participants were asked to learn Wubbel family-environment category associations. In each trial of prediction learning phase, participants were presented with two Wubbel parents from different families on different screen locations and asked to indicate which Wubbel matches with the presented scene. Participants were then presented with feedback based on their response. As in the example, participants were expected to learn that one of the long-shaped Wubbel family lives on the beach. B) On day 2, the encoding phase was first run. Participants were presented with a scene and asked to predict the most likely Wubbel family in their minds. Then, a Wubbel child was presented on one out of four possible screen locations. Participants were further asked to indicate whether the presented Wubbel child belongs to the family which they predicted. As in the example, a long-shaped Wubbel child presented in a beach scene would elicit a low PE (yellow). Presenting the same Wubbel on a desert which was an environment for the other long-shaped family would lead a medium PE (green). Lastly, presenting the same Wubbel with the environments associated with the round-shaped families would elicit a high PE (blue). C) During the recognition phase, participants were asked to make old/new judgments and report their confidence.

Participants were then asked to indicate the paired scene and paired location. D) Example for the relationship structure for Wubbel families and children.

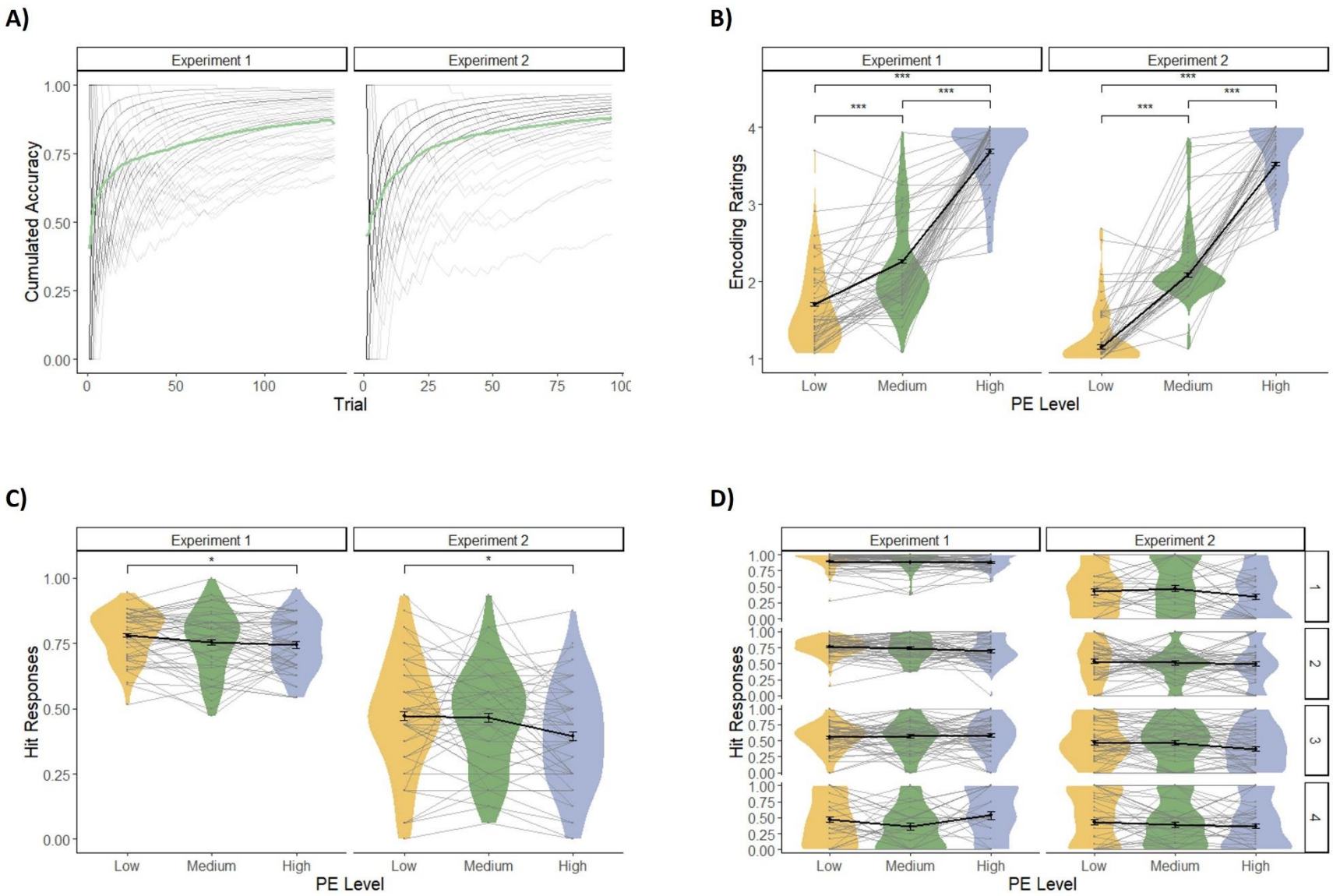
A)
Experiment 1
Experiment 2
Cumulated Accuracy
1.00
0.75
0.50
0.25
0.00
0
50
100
0
25
50
75
100
Trial
B)
Experiment 1
Experiment 2
***
Encoding Ratings
4
3
2
1
Low
Medium
High
PE Level
C)
Experiment 1
Experiment 2
*
Hit Responses
1.00
0.75
0.50
0.25
0.00
Low
Medium
High
PE Level
D)
Experiment 1
Experiment 2
Hit Responses
1.00
0.75
0.50
0.25
0.00
1
2
3
4
Low
Medium
High
PE Level

Figure 3. Results for Experiment 1 and 2. A) Cumulated accuracy for prediction learning. Grey lines indicate the performance of single participants. Green lines indicate the group mean. B) Encoding ratings for low, medium, and high PE levels. C) Hit responses for low, medium, and high PE levels. D) Hit responses for low, medium, and high PE levels separated by confidence ratings (1- Very sure, 2- Sure, 3- Unsure, 4- Very Unsure). Grey lines indicate the performance of single participants. Black lines indicate the group mean with error bars reflecting ± SEM. Asterisks denote statistically significant differences, * $p < .05$, *** $p < .001$.