Reviewed Preprint
v1 • December 18, 2024
Not revised

Reviewed Preprint
v2 • July 21, 2025
Revised by authors

Reviewed Preprint
v3 • February 4, 2026
Revised by authors

**For correspondence:**
h.mcdermott@fu-berlin.de

**Competing interests:** No competing interests declared

**Funding:** See page 14

**Reviewing editor:** Clare Press, University College London, United Kingdom



# Dissociable dynamic effects of expectation during statistical learning

**Hannah H McDermott[1,2,3] ✉, Federico De Martino[3], Caspar M Schwiedrzik[4,5,6], Ryszard Auksztulewicz[1,3]**

[1]Department of Education and Psychology, Freie Universität Berlin, Berlin, Germany • [2]Berlin School of Mind and Brain, Berlin, Germany • [3]Faculty of Psychology and Neuroscience, Maastricht University, Maastricht, Netherlands • [4]Neural Circuits and Cognition Lab, European Neuroscience Institute Göttingen - A Joint Initiative of the University Medical Center Göttingen and the Max Planck Institute for Multidisciplinary Sciences, Göttingen, Germany • [5]Perception and Plasticity Group, German Primate Center, Leibniz Institute for Primate Research, Göttingen, Germany • [6]Cognitive Neurobiology, Research Center One Health Ruhr, University Alliance Ruhr, Faculty of Biology and Biotechnology, Ruhr-University Bochum, Bochum, Germany

## eLife Assessment

This **important** study is of relevance for the fields of predictive processing, perception and learning, with a well-designed paradigm allowing the authors to avoid several common confounds in investigating predictions, such as adaptation. Using a state-of-the-art multivariate EEG approach, the authors test the opposing process theory and find evidence in support of it. Overall, the empirical evidence is **solid**, however, some conclusions rest on limited evidence and need further work to reconcile the present results with previous studies.

https://doi.org/10.7554/eLife.103689.3.sa2

## Abstract

The brain is thought to generate internal predictions, based on previous statistical regularities in the environment, to optimise behaviour. Predictive processing has been repeatedly demonstrated and seemingly explains expectation suppression (ES), or the attenuation of neural activity in response to expected stimuli. However, the mechanisms behind ES are unclear and various models of the mechanisms supporting ES have been suggested with conflicting evidence. Sharpening models propose that expectations suppress neurons that are not tuned to the expected stimulus, increasing the signal-to-noise ratio for expected stimuli. In contrast, dampening models posit that expectations suppress neurons that are tuned to the expected stimuli, increasing the relative response amplitude for unexpected stimuli. Previous studies have used decoding analyses to examine these effects, with increases in decoding accuracy interpreted in terms of sharpening and decreases related to dampening. The opposing process theory (OPT) has suggested that both processes may occur at different time points, namely that initial sharpening is followed by later dampening of the neural representations of the expected stimulus as learning progresses. In this study we aim to test this theory and shed light on the dynamics of expectation effects, both at single trial level and over time. Thirty-one participants completed a statistical learning task consisting of paired scene categories whereby a “leading” image from one category is quickly followed by a “trailing” image from a different category. Multivariate EEG analyses focussed on decoding stimulus information related to the trailing image category. Within-trial, decoding analyses showed that stimulus expectation increased decoding accuracy at early latencies and decreased decoding accuracy at later latencies, in line with OPT. However, across trials, stimulus expectation decreased decoding accuracy in initial trials and increased decoding accuracy in later trials. We theorise that these dissociable dynamics of expectation effects within and across trials can be explained in the context of hierarchical learning mechanisms. Our single trial results provide evidence for the OPT, while our results over time suggest that sharpening and dampening effects emerge at different stages of learning in the visual domain.

## Introduction

Influential theories of perception and learning suggest a key role of prediction in both processes (1–3). However, it remains unclear whether the influence of prediction on neural activity across sources of prediction, time scales and learning stages can be explained by a single mechanism (4–7). Repetition suppression (RS), or the attenuation of neural activity in response to repeated (and therefore predictable) stimuli, has been well documented across a range of sensory modalities, species, and experimental paradigms (8,9). Predictive processing theories explain RS as a manifestation of minimising prediction errors (PEs) through adaptive changes in predictions about the content and precision of sensory inputs (8). Expectation suppression (ES), or the attenuation of neural activity in response to expected stimuli (regardless of its frequency) appears to be a related phenomenon, but ES is more controversial in terms of its underlying mechanisms. In an earlier study attempting at dissociating the two phenomena, it has been shown that RS and ES affect neural activity in separable time windows (10). More recently, however, Feuerriegel et al. (11) pointed out major challenges to the evidence underlying ES in the visual system, such as the conflation of ES with RS, surprise effects, attention effects, and stimulus novelty. While empirical evidence for ES is more fragmentary (11), it is in principle a more convincing neural signature of predictive processing, as it is less readily explained than RS by mere neural adaptation (12), which does not necessarily reflect cognitive processes.

At least two types of neural activity modulation have been linked to the effects of expectation. Sharpening models propose that expectations suppress neurons that are not tuned to the expected stimulus, thus inhibiting information inconsistent with the expectation (13,14). Depending on the level of analysis, this may result in sharper tuning curves (i.e., increased signal-to-noise ratio for expected stimuli) for a particular neuron or small population, amounting to relative expectation enhancement; and/or in sparse responses (i.e., net activity decrease) in a larger population (15). In contrast, dampening models posit that expectations suppress neurons that are tuned to the expected stimuli, effectively “explaining away” the prediction error responses (16). By cancelling information in line with expectations, uninformative activity in the sensory stream is prevented while favouring novel information (17,18). Studies attempting to settle the debate have produced mixed results on both accounts. Kok et al. (13) and Garlichs & Blank (19) both found evidence for neural sharpening in the visual domain using fMRI and multivariate pattern analysis (MVPA). Another study using 7T fMRI found both suppression of irrelevant neural responses in visual areas and selective enhancement of relevant neural responses in higher-order frontoparietal cortices (20). In contrast, Richter et al. (18,21) demonstrated evidence in favour of dampening of neural responses in the visual domain. In the latter study, after training participants on a statistical learning task, stimulus features were decoded from fMRI signals separately in expected vs unexpected trials. The rationale of the analysis was that, if expectation leads to dampening, the response amplitudes will be lower, and as a result stimulus decoding should be lower in expected trials; conversely, if expectation leads to sharpening, the signal-to-noise ratio will increase, and as a result stimulus decoding should increase in expected trials. Both studies found lower decoding for expected trials, consistent with suppressed activity of neuronal populations tuned towards expected stimuli. However, electrophysiological research in the area is scarce (17,22), despite its potential for parsing out temporal differences which may go unnoticed in fMRI studies with poor temporal resolution.

Taking temporal differences into account, the recently suggested Opposing Process Theory (OPT) has suggested that sharpening and dampening may both occur, but at different time points during the predictive process (23). This model seeks to reconcile contrasting findings as some studies show that expected events are perceived with greater intensity 50 ms after presentation but this bias reverses around 200 ms (24,25), while classically reported ES effects most often take place ~150 ms after presentation (26). OPT suggests that within each trial, initial processing relies on prior knowledge to sharpen sensory representations towards the expected stimuli, and a later processing stage follows, dampening the neural representations of the expected stimulus. This is based in part on trends in prior research. Studies have found that while expected events are perceived with greater intensity 50 ms after presentation, this bias reverses by 200 ms such that

unexpected events are perceived with greater intensity (24,25). While this theory is largely untested thus far, it will prove a highly useful benchmark for future research. Namely, by implementing decoding analyses in a time-resolved manner we can better parse out the intricacies of this model. Sharpening and dampening models have previously been investigated using time-resolved methods, albeit with different study designs. Han et al. (27) found that stimulus-specific event-related activity is dampened for expected tones in the auditory cortex, while studies examining the inferior temporal (IT) cortex in monkeys have found that IT neurons represent less accurately predictable compared with unpredictable stimuli (16,17). It is worth noting that while the above studies did change the pairings of stimuli between trials, they used small pools of stimuli (<30) which were repeated many times and implemented extensive training and exposure to the stimuli prior to collecting data.

Theoretical work has proposed that, similar to the effects of prediction within a trial, learning should also be linked to neural activity suppression, albeit at hierarchically higher levels of neural processing (28). However, empirical work yielded more nuanced results. Previous research has focussed on the mechanisms of ES after learning has taken place, while the effects of expectation across trials (i.e., during learning) have been largely overlooked. In an earlier fMRI study (29), it has been shown that repetition effects are qualitatively different for the initial vs. later stimulus repetitions, gradually enhancing vs. suppressing BOLD activity, respectively. In contrast, EEG work has shown that neural responses are more strongly modulated by initially formed repetition-based predictions than by their subsequent revisions (30). Furthermore, in an MEG study differentiating the effects of expectation and familiarity, both factors were linked to suppressed neural activity but expectation (previously seen stimuli with expected vs. unexpected structure) attenuated activity in the lateral occipital cortex, whereas familiarity (previously seen vs. unseen stimuli with no predictable structure) additionally led to temporal sharpening of evoked responses in early visual regions (31). However, the study only characterised these effects after learning, averaging across many trials. While theoretical accounts such as the reverse hierarchy theory of perceptual learning (32) and hierarchical Bayesian framework of statistical learning (33) suggest that perceptual learning effects progress from higher to lower levels of the visual system, in classical predictive coding hierarchies the time scales increase along the processing hierarchy in an ascending manner (28). Thus, it is unclear whether dampening vs. sharpening effects (previously found to occur in anterior vs. posterior regions) (20) should be identified at earlier vs. later learning stages.

Here, we assessed the effects of expectation on EEG-based stimulus decoding during statistical learning while controlling for repetition effects, which may render the expectation effects impossible to parse out given the similarity between the two phenomena. We focused on empirically testing the Opposing Process Theory on visually evoked responses. We expected to find univariate evidence of expectation suppression, by way of reduced ERP amplitude to trailing images in valid trials over occipital channels. Furthermore, in line with the OPT, we hypothesised that sharpening effects will lead to higher decoding accuracy for valid trailing images earlier within trials, and dampening effects will lead to higher decoding for invalid images later within trials. We also explored the dynamics of expectation effects across trials. In brief, our results show support for OPT, however with different dynamics within trials than across learning stages.

# Results

Participants (N=31) were exposed to a sequence of visual scenes (Fig. 1A) while their neural activity was recorded using EEG (see Methods for more details). The sequence was manipulated with respect to the expectation of image categories. In each trial, participants were presented with two images from nine different categories in quick succession, with five 'Leading' categories, and four 'Trailing' categories. Over 3500 categorised images were used to avoid the repetition of individual stimuli and control for RS. Category pairs and the transitional probabilities between them were determined by the transitional probability matrix depicted in Fig. 1B. Each participant viewed 1728 trials without prior training and each image was only presented once.

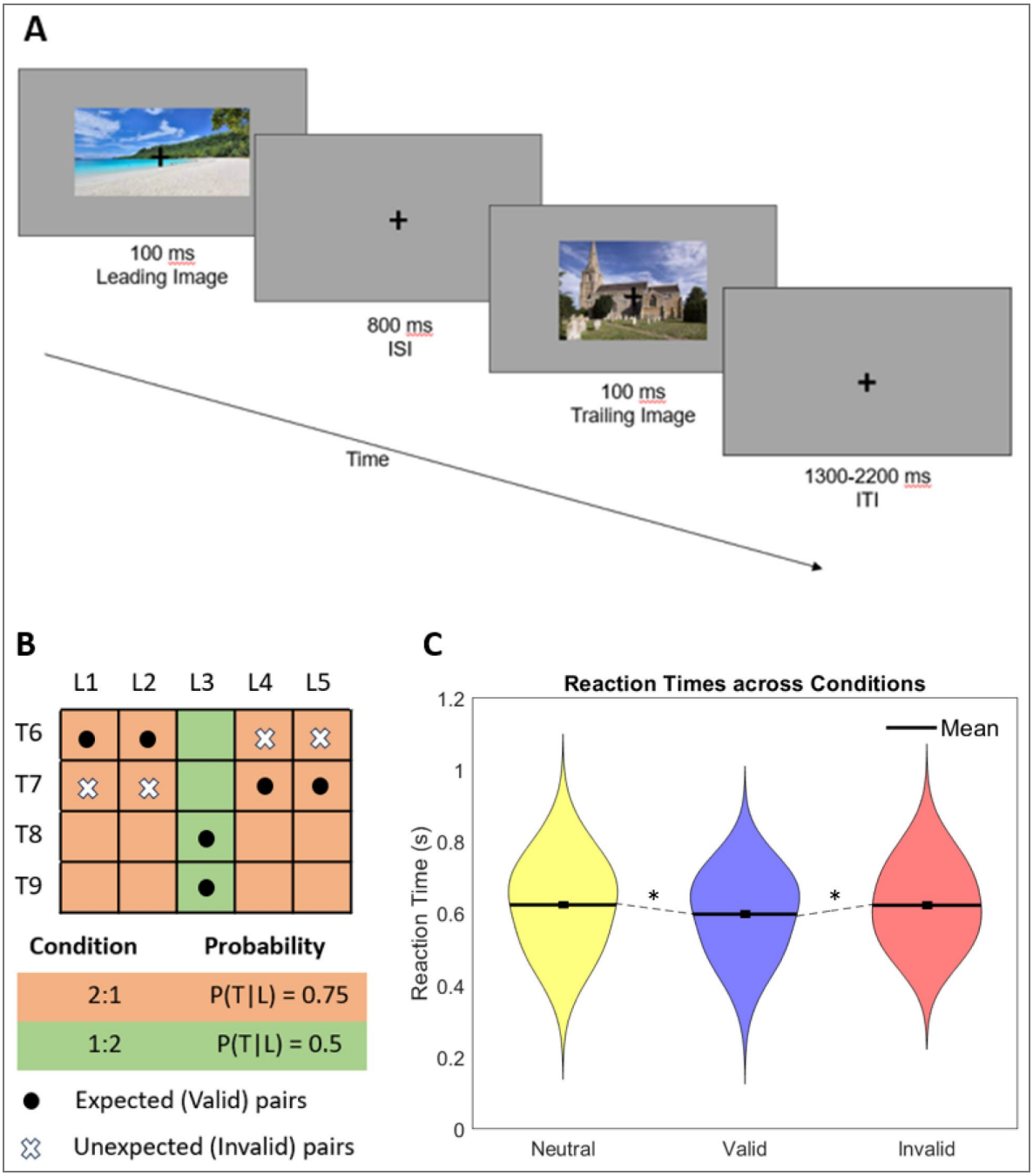


**Figure 1.** **Paradigm overview.**

***A.*** A single trial, with two example images. Images were presented for 100 ms each with 800 ms interstimulus interval, and an intertrial interval of 1300-2200 ms. Participants were required to respond to upside down images with button press; all upside down images were trailing. ***B.*** Transitional probability matrix determining category pairs. Five leading categories and four trailing categories were used. Orange represents the 2:1 condition, and green represents the 1:2 (control) condition. Cells with dots represent the valid pairs, cells with Xs represent the invalid pairs, and empty cells represent non-existent pairs. ***C.*** Reaction times across Valid, Invalid, and Neutral conditions. Asterisks indicate significant results at $p<0.05$ after correction for multiple comparisons.

## Behavioural results

After the EEG recording, a majority of participants (N=21) were asked to perform a categorization task on images drawn from the same categories. Transitional probabilities in this task were kept the same as in the EEG session, and participants performed a speeded indoor/outdoor categorization task on the 'Trailing' images. A repeated measures ANOVA was conducted to compare reaction times (RTs) across conditions in order to assess implicit learning, which showed significant differences between conditions ($F(_{2,40})$ = 4.231, $p$ = 0.022, $\eta^2$ =0.175). Paired t-tests were conducted to further elucidate RT differences in between conditions. Participants reacted significantly faster in valid trials (mean = 598.9 ms, SEM = 7.2 ms) than in invalid trials (mean = 623.6 ms, SEM = 7.2 ms; $t_{20}$ = -2.18, $p$ = 0.041, $d$ = 0.476, two-tailed). Participants also reacted significantly faster in valid trials than in neutral trials (mean = 624.7 ms, SEM = 5.7 ms; $t_{20}$ = 2.88, $p$ = 0.00, $d$ = 0.629, two-tailed). There was no significant difference in reaction times between invalid and neutral trials ($t_{20}$ = 0.120, $p$ = 0.906, $d$ = 0.026, two-tailed).Significant differences between both valid and invalid, and valid and neutral conditions suggest that associations between pairs were learned implicitly (Fig 1C).

## Univariate EEG analyses

In order to assess the impact of expected vs unexpected trials on visually evoked ERPs, a paired t-test was conducted on the epoched amplitudes in a cluster of occipital channels (Oz, O1, O2) at channel Oz. Since previous research has questioned the robustness of univariate effects of ES (11), we conducted a target region-of-interest analysis on the channel showing peak amplitude of the visually evoked response (Fig. 2B). Unexpected trials resulted in a significantly more pronounced negative deflection 150 ms after image onset ($t_{2926}$=3.3789, $p$ <= 0.00123) after FWE correction. (Fig. 2A). This is in line with the idea that neural responses to expected stimuli are attenuated in ES. A more conservative and exploratory whole-brain analysis, correcting for all channels and time points failed to reach significance, consistent with the issues raised by Feuerriegel et al. (11). The group level topography of the amplitudes differences between valid and invalid trials are represented in Fig. 2B.

## Decoding analyses

To test the hypothesis derived from the OPT that expectation should have opposing effects at earlier (sharpening) vs. later (dampening) response latencies, we used multivariate decoding analysis. Following previous work (21,27), we reasoned that sharpening should lead to increased decoding accuracy for expected stimuli, while dampening should lead to increased decoding accuracy for unexpected stimuli. To this end, we calculated a linear SVM-based category decoding accuracy at each time point based on EEG amplitudes, separately for valid vs. invalid trials (27). Three separate decoding analyses were conducted: "sensory decoding" (decoding the visual category of the presented image, done separately for the leading and trailing images), and "memory decoding" (decoding the preceding visual category based on the trailing images). The above analyses were all subject to cluster-based family-wise error (FWE) correction for multiple comparisons across time points. FWE is a standard method of correcting for multiple comparisons and decreasing the cluster-based false positive ratio implemented in SPM. It uses random field theory assumptions (34) to account for the spatiotemporal correlation between neighbouring data points. Clusters were thresholded to 0.05 uncorrected, then reported clusters which were significant at cluster-level $p_{FWE} < 0.05$.

In the sensory decoding analysis based on leading images, decoding accuracy was above chance for both valid ($T_{max}$= 4.79, $p_{FWE} < 0.001$) and invalid trials ($T_{max}$= 4.02, $p_{FWE} < 0.001$) from 100 ms, with no significant difference between them ($T_{max}$= 2.47, $p_{FWE}$ = 0.987) (Fig. 2C). This initial analysis largely served as a sanity check, since no difference between decoding the category of valid and invalid images should be observed as the leading image cannot be invalid.

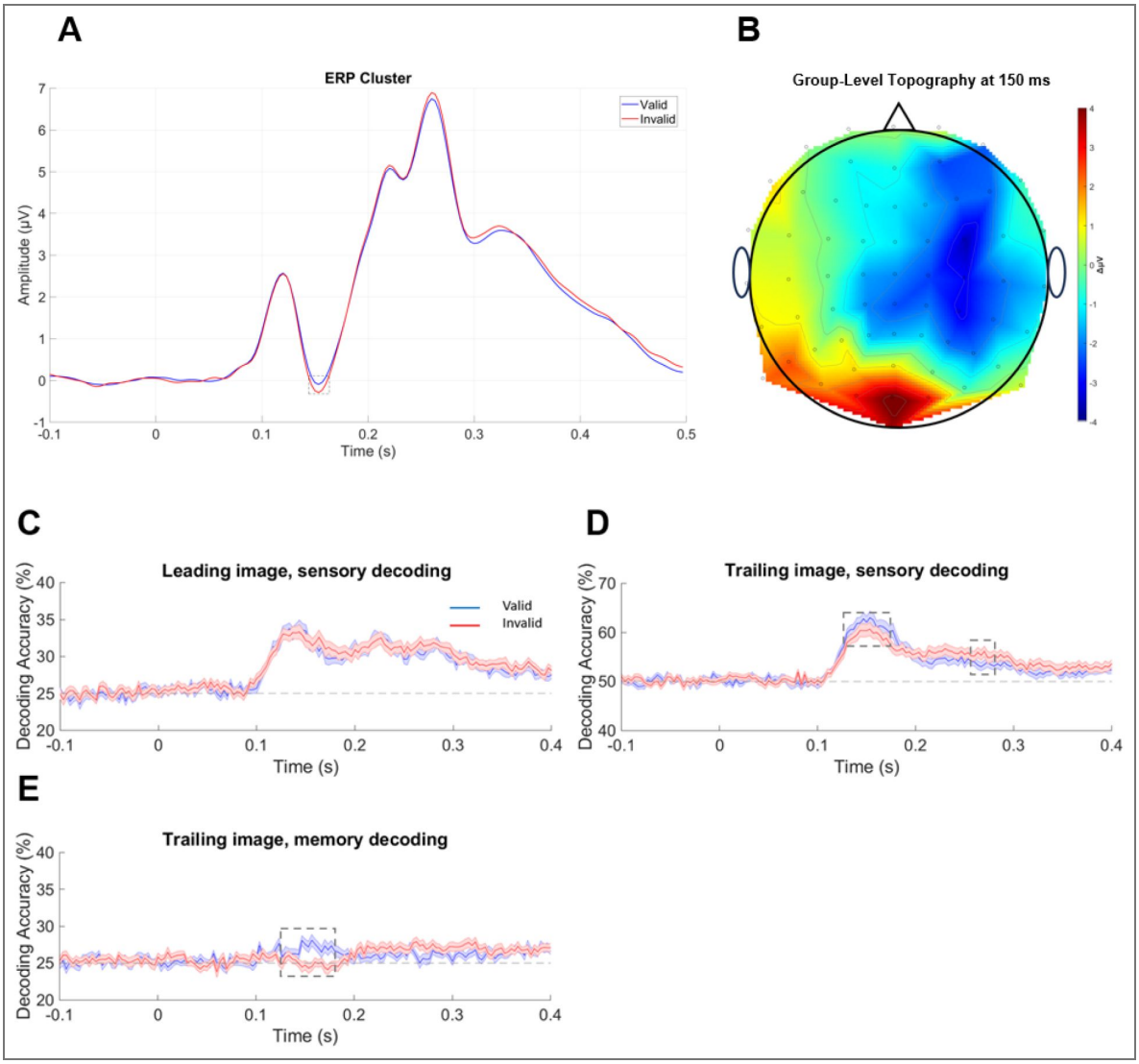


**Figure 2. Results of data analysis.**

**A.** Grand average ERP statistical analyses in a cluster of occipital channels at Channel Oz. Significant difference between valid and invalid trial amplitudes. **B.** Topography plot of the t-values resulting from paired t-tests comparing Valid and Invalid EEG data at 150 ms. **C.** SVM decoding results for Leading image sensory decoding. Dashed vertical line at 0 s indicates stimulus onset. **D.** Decoding results for Trailing image sensory decoding. Dashed rectangles denote latencies of significant effects ($p_{FWE} < 0.05$). Dashed vertical line at 0 s indicates stimulus onset. **E.** Decoding results for Trailing image memory decoding. Dashed rectangle denotes latencies of significant effects ($p_{FWE} < 0.05$). Dashed vertical line at 0 s indicates stimulus onset.

On the contrary, in sensory decoding of trailing images, decoding accuracy was significantly different between valid and invalid images 123-180 ms after image onset, where decoding the category of valid images was significantly higher than invalid images (cluster-level statistics: $T_{max}$ = 3.62, $p_{FWE}$ < 0.001), but decoding the category of invalid images was significantly higher than valid images 280-296 ms after image onset ($T_{max}$ = 3.89, $p_{FWE}$ < 0.001) (Fig. 2D). Therefore, stimulus category expectation had opposing effects on trailing category decoding based on EEG signals early vs. late within a trial. The decoding accuracy of both valid ($T_{max}$ = 5.10, $p_{FWE}$ < 0.001) and invalid trials ($T_{max}$ = 3.83, $p_{FWE}$ < 0.001) was also above chance in this condition from 100 ms.

Finally, as with sensory decoding of trailing images, memory decoding accuracy based on valid trailing images was significantly higher than invalid trailing images 123-180 ms after image onset ($T_{max}$ = 3.96, $p_{FWE}$ < 0.001) (Fig. 2E). Both valid ($T_{max}$ = 4.67, $p_{FWE}$ = 0.047; from 100 ms) and invalid ($T_{max}$ = 3.40, $p_{FWE}$ < 0.001; from 200 ms) trailing images yielded above-chance decoding.

In order to better assess how the difference in prediction and memory decoding between valid and invalid pairs changes during learning, we separated each data set into four trial bins. In order to assess the utility of examining each bin in more detail, we calculated the trial-by-trial time-series of the decoding accuracy benefit for valid vs. invalid for each participant and averaged this benefit across time points for each of the two significant time windows. Based on this, we fitted a logarithmic model to quantify the change of this benefit over trials, then found the trial index for which the change of the logarithmic fit was < 0.1%, indicating stabilized decoding accuracy. This demonstrates that expectation effects developed rapidly early in learning and stabilised approximately after 50% of trials, after which almost zero learning took place. Given the results of this analysis and to ensure a sufficient number of trials, we focussed our further analyses on bins 1-2 to directly assess the effects of learning.

We separately analysed the early (123-180 ms) and late (280-296 ms) within-trial time window identified in the main analysis of sensory decoding based on the trailing images (Fig. 2D) as well as the early (123-180 ms) time window identified in the memory decoding analysis (Fig. 2E). We fit each of these effects to a linear mixed effects model with Holm-Bonferroni correction for multiple comparisons across bins. Outliers larger than 3 standard deviations based on Cook's distance were removed.

In the early time window of sensory decoding, the main effects of validity ($T_{79}$ = 0.84141, $p$ = 0.403) and bin ($T_{79}$ = 1.9242, $p$ = 0.058) were not significant. The interaction between validity and bin was significant, however ($T_{79}$ = -2.1679, $p$ = 0.033). Cohen's f for the full model was $f$ = 1.688. Post-hoc analyses of the effects of each bin showed no significant differences between valid and invalid trials in the first bin ($T_{28}$ = 1.2036, $p$ = 0.239, $f$ = 0.052) but significant differences in the second bin ($T_{23}$ = 3.8704, $p$ < 0.001, $f$ = 0.651) (Fig. 3A). In other words, the decoding benefit of valid predictions - visible at early post-stimulus latencies when averaging over all trials - took approximately 15 minutes (or 50 expected trials per image pairing) to become significant. Additional post-hoc analyses comparing the first bin and the second bin found significant differences in invalid trials ($T_{24}$ = -2.208, $p$ = 0.037, $f$ = 0.203), but not in valid trials ($T_{26}$ = 1.449, $p$ = 0.159, $f$ = 0.081). Outliers larger than 3 standard deviations based on Cook's distance were removed prior to this analysis.

Analyses of the later peak in sensory decoding found the opposite effect. The main effect of validity ($T_{77}$ = 3.5402, $p$ < 0.001), and the main effect of bin ($T_{77}$ = 2.2399, $p$ = 0.028) both contributed significantly. The interaction between them ($T_{77}$ = -2.6854, $p$ = 0.009) also showed significant differences. Cohen's f for the full model was $f$ = 1.5943. In contrast to the early peak, post-hoc analyses showed that the first bin had significant differences between valid and invalid trials ($T_{24}$ = 4.293, $p$ < 0.001, $f$ = 0.768) but the second bin did not ($T_{24}$ = 0.2673, $p$ = 0.791, $f$ = 0.003) (Fig. 3B). In other words, the decoding benefit of invalid predictions - visible at late post-stimulus latencies when averaging over all trials - disappeared after approximately 15 minutes (or 50 expected trials per image pairing). As with the early peak, additional post-hoc analyses comparing the first bin and the second bin found significant differences in invalid trials ($T_{26}$ = -2.607, $p$ = 0.015, $f$ = 0.261), but not in valid trials ($T_{22}$ = 1.273, $p$ = 0.216, $f$ = 0.074). Outliers larger than 3 standard deviations based on Cook's distance were removed prior to this analysis.

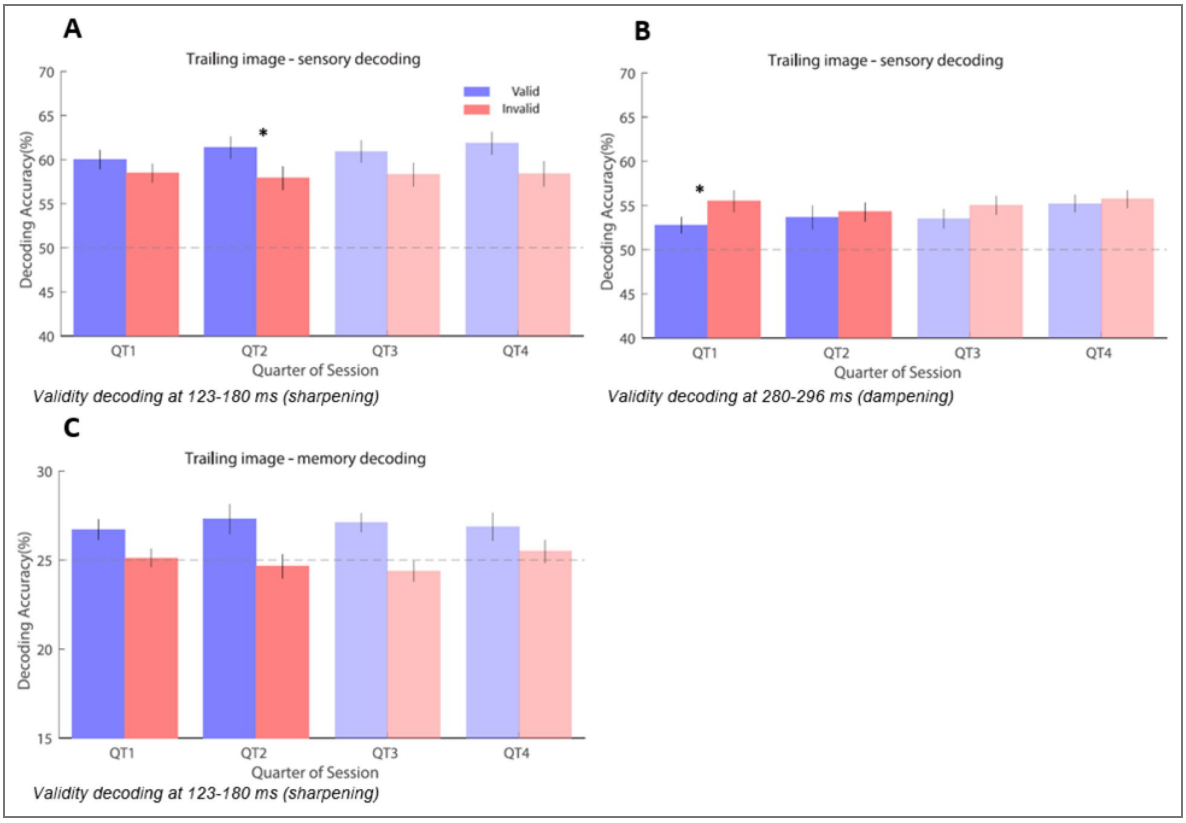


**Figure 3.** **Analysis of validity effects on decoding over trial bins.**

Error bars indicate SEM. Dashed horizontal lines indicate chance level. Asterisks indicate significant results at $p<0.05$ after correction for multiple comparisons. ***A.*** Learning over time at 123-180 ms in sensory decoding. ***B.*** Learning over time at 280-196 ms in sensory decoding. ***C.*** Learning over time at 280-296 ms in memory decoding.

Finally, in the memory decoding, each of the main effects for validity ($T_{1,77}$ = -0.407, $p$ = 0.685), bin ($T_{1,77}$ = 0.751, $p$ = 0.455), and the interaction between them ($T_{1,77}$ = -0.794, $p$ = 0.429) failed to find significant effects after removal of outliers (Fig. 3C). Cohen's f for the full model was $f$ = 0.997.

In order to examine the extent to which participants' neural responses were linked to their behavioural responses, we calculated Pearson correlation coefficients for behavioural responses (difference in RTs), neural responses (difference in decoding accuracy) during the sharpening effect, and neural responses during the dampening effect based on the difference in mean responses for each participant. Correlations between the behavioural responses and the sharpening effect ($r_{19}$ = 0.121, $p$ = 0.602) and between the behavioural responses and the dampening effect ($r_{19}$ = 0.127, $p$ = 0.584) were not significant. It should be noted however that EEG and behavioural data were not recorded simultaneously, therefore this null finding should be treated with caution.

# Discussion

We set out to investigate the effects of expectation on stimulus decoding during statistical learning and demonstrated dynamic differences in decoding expected vs. unexpected stimulus categories, both within and across trials. EEG-based decoding of expected vs. unexpected visual scene categories showed that within trials, a relative decoding boost for expected stimuli was followed by a relative decoding boost for unexpected stimuli, in line with the recent Opposing Process Theory (OPT) (20). However, across trials, the decoding boost for unexpected stimuli was observed earlier than the decoding boost for expected stimuli, possibly due to a dynamic relationship between cortical processing hierarchies (32,35,36).

The OPT (23) combines both sharpening and dampening models of stimulus expectation, which have previously been mapped onto a relative decoding boost for expected vs. unexpected stimuli respectively (27). In this model, hypothesis units (units representing the brain's 'best guess' about the outside world) are initially strongly weighted towards the expected, however when agents encounter events that elicit surprise, increased gain on surprising inputs leads to high fidelity representations of unexpected events across hypothesis units. In simple terms, early sharpening (which favours processing of expected information) is followed by later dampening (which in turn allows novel information to be processed) within trials. This is supported by our within-trial results, whereby we found that decoding of expected images was significantly higher than that of unexpected images early during the trial, while the opposite effect was found later during the trial. Differences in early vs late prediction error detection have previously been observed using fMRI (5). This study used a visual detection task to demonstrate that immediate implicit (early) detection of PEs enables fast but partial comparison of bottom-up sensory input with top-down predictive information, but effortful explicit (late) processing permits more comprehensive assessment of the type of mismatch between expected and actual input. While the authors did not interpret their results in terms of dampening or sharpening models, it is possible that the differences observed are a result of changing temporal dynamics in the neural response which are difficult to demonstrate using fMRI. In this context, higher-level processes would modulate detection of prediction error through top-down processes and regulate adaptation of predictive models as a result.

Interestingly, the early (~150 ms) decoding accuracy boost found for expected trials coincided with the univariate ERP amplitude increase found for unexpected trials. Such seemingly opposing effects are commonly explained based on the notion that sharpening leads to fewer units responding to the expected stimulus, and the population response becoming sparser and more selective (15). According to Friston's original theoretical model assessment of predictive coding (1), sharpening and dampening occur in parallel but in separate prediction and error neurons which would reside in deep and superficial layers of the brain respectively. As such, studies that have argued for sharpening models (13,14), have focussed primarily on the response to expectations in the occipital cortex and V1, while studies that have favoured dampening models (18,21,27) have centred more on the neural response to violated expectations in the ventral visual stream. Using fMRI, Thomas et al. (36) found that expected and unexpected events are represented across the

cortical column in V1. However, while expected events were represented similarly across layers, unexpected events were only well represented in superficial layers. One recent EEG study has attempted to directly elucidate the mechanisms of the OPT. Xu et al. (37) used auditory cues followed by a predictable series of flashes to investigate the mechanisms of ES. In line with OPT, they found that the N1 was augmented, while the N2 was attenuated. However, this study focussed entirely on ERP analyses (rather than multivariate decoding or gain/tuning models), which limits the interpretability of the results in terms of sharpening vs. dampening. Other studies have also found evidence for dampening after learning only. For example, two studies (16,17) examining the inferior temporal (IT) cortex in monkeys using single- and multi-unit recordings have found that IT neurons represent less accurately predictable vs. unpredictable stimuli. Similarly, MEG studies have shown that representations of expected stimuli in the neural signal appear shortly before stimulus onset (38) and that prior expectations lead to overall dampened activity in the auditory cortex (27). An fMRI study also showed overall dampened activity in the ventral visual stream (18). It is worth noting that several of the studies which showed sharpening found sharpening for predictions which were implemented as task-relevant (13,14), while some of the studies which showed dampening found effects in response to task-irrelevant predictions (18,21). However, it is unlikely that effects of task-relevance are being parsed out in the present study, where no instructions regarding relevance were given to participants, and behavioral tests were only administered after implicit learning. None of the aforementioned studies showed a reversal of sharpening and dampening effects, but it is worth noting that their experimental design did not allow for testing these effects over the course of learning as they implemented extensive training and/or exposure paradigms. Some studies did use categorised stimuli to reduce this impact, but exemplars were still repeated multiple times over the course of the experiment (17, 39), or ensured that stimuli were equally familiar across expectation conditions (16,27). However, these studies given that failing to fully account for the confounding effects of RS as adaptation effects can depend on stimulus history beyond the most recently encountered stimulus and long-lag repetition effects can occur when the first and second presentation of a stimulus is separated by hundreds of intervening stimuli (11), which may mask the ES effects being examined. The present study controls for confounding RS effects by using stimuli that are novel and unfamiliar, foregoes a training session to examine ES effects both during learning and after learning has occurred, and utilises the high temporal resolution of EEG, providing a unique insight into the neural dynamics of these phenomena, albeit without a correlation between neural and perceptual variables.

In contrast to our within-trial results, when we examined how this process developed over multiple trials, we found that dampening occurred within the first ~15 minutes of recording and was only later followed by sharpening. These effects seem to be driven by changes in responses to unexpected trials, while expected trials did not change reliably over time. This apparent stability suggests that expectation-related representations are less sensitive to the ongoing learning-related changes that affect unexpected trials. We hypothesise that this is a result of a dynamic relationship between hierarchical levels of cortical processing. Previous work has shown that in MEG (40), early-latency responses typically reflect early cortical stages, while late-latency responses typically reflect higher cortical stages. Our results are in line with this in that across trials, late-latency decoding differences arise within the initial trials, suggesting that higher-order regions quickly process statistical regularities during learning. Similarly, early-latency decoding differences occur later during the experiment, consistent with the theory that lower-order regions need more time to accumulate evidence, while the distinction between higher-order dampening and lower-order sharpening can also be observed in rapid forms of learning, as seen in rapid one-shot perceptual learning (20). In short, 'invalid' trials are more salient early in the experiment as priors are still being formed, which leads to predictions occurring earlier in each trial, later in the experiment. Earlier studies have pointed to a dissociation between early and late ventral stream effects, which may also fit into this framework (18,21). Other studies using statistical learning have indicated slower dynamics, in that some regions, such as the hippocampus, begin preferentially representing PEs early during learning, and predictions are preferentially represented later during learning (41). This is consistent with our finding that early dampening is followed by later sharpening over time, given that sharpening and dampening are defined by the relative difference

in decoding accuracy between expected and unexpected trials. Another recent fMRI study found that in the presence of predictions, early stages of the processing hierarchy exhibit well separable and high-dimensional neural geometries resembling those at the top of the hierarchy, which is accompanied by a systematic shift in tuning properties from higher to lower areas, enriching lower areas with higher-order, invariant representations in place of their feedforward tuning properties (42). In the context of these studies, our results could be explained by higher-order regions quickly detecting stimulus statistics and show evidence of predictive dampening to expected inputs, and then gradually delegate those predictions to lower-order regions which start showing evidence of predictive sharpening. This suggestion is somewhat speculative however, and is thus intended to stimulate further research. It should also be noted that we demonstrate these across-trial dynamics in the visual domain only, and other domains (such as action and audition) may elicit different results and still require additional research.

In addition to the dynamic effects of expectation on sensory decoding of the trailing image, we also found that valid expectation increases decoding accuracy of the memory of the leading image. Within trials, this effect was found at similar latencies as the positive effect of expectation on sensory decoding; however, across trials, the effect on memory decoding started earlier and persisted throughout most of the experiment. A memory trace linking the trailing and leading image is a key component of associative learning (43). Furthermore, the finding that within a trial, sensory and memory decoding occur at similar latencies and are subject to similar effects of expectation, is broadly consistent with work showing simultaneous sensory, mnemonic, and predictive representations within the same neocortical regions (44,45). However, the differential dynamics of these effects across trials suggest that memory and prediction encoding is also partly dissociable, not only at the level of the hypothesised circuit mechanisms, likely involving the neocortex and the hippocampus (2), but also over time. This suggestion is consistent with a neuroimaging study showing that during associative learning, the hippocampus gradually switches from signalling PEs (i.e., mismatch between memory trace and actual stimulus) to signalling predictions (i.e., stimulus expectation independent of the actual stimulus) (41).

In summary, our study is the first to show dynamic effects of expectation on stimulus processing during learning, both within and across trials. We have shown using EEG and decoding analyses that early dampening is followed by later sharpening over time, but within trials early sharpening is followed by later dampening. This provides direct evidence for the OPT, while also indicating that sharpening and dampening effects emerge at different learning stages.

# Methods

## Participants

Thirty-one healthy participants (25 female, 1 non-binary, 5 male) aged 18-35 (mean = 22.7) were recruited from the student population at Freie Universität Berlin. One participant was left-handed. Participants had no history of neurological illness, were not currently taking psychotropic medication, and had normal or corrected-to-normal vision. The study was approved by the Ethics Committee of the Department of Education and Psychology at Freie Universität Berlin. Participants were compensated with either €30, or research participation credits. Data from one participant was excluded due to a technical issue which resulted in an absence of triggers in the EEG data. The final sample size (N=30) corresponded to previous research (46).

## Experimental Design and Statistical Analysis

### Stimuli and experimental paradigm

#### *Main task*

Approximately 3500 colour images across nine categories were generated using artificial intelligence software Craiyon (Craiyon LLC, 2022). Images were visually inspected to ensure they appeared realistic. Some physically implausible examples were discarded. Stimuli were presented on a 38 cm LCD monitor with 1280 x 1024 resolution and 60 Hz refresh rate, the viewing distance

was 62 cm, and stimuli were presented at 2.55°. The experimental paradigm was adapted from Richter et al. (18) with several modifications, as described below. Participants were exposed to pairs of visual stimuli, comprising artificially generated natural scenes drawn from different scene categories. In each trial, participants were presented with two images from different categories in quick succession, each presented for 100 ms with 800 ms interstimulus interval and 1300-2200 ms intertrial interval (Fig. 1A). A fixation cross was presented throughout the experiment. Each image belonged to one of nine scene categories, with five 'Leading' categories ('barn', 'beach', 'library', 'restaurant', 'cave') and four 'Trailing' categories ('church', 'conference room', 'castle', 'forest'). Category pairs and the transitional probabilities between them were determined by the transitional probability matrix depicted in Fig. 1B. As previously established, many prior studies have overlooked the confounding effects of RS in their experimental paradigms. To control for RS, categorised images were used to avoid the repetition of stimuli. As such, each individual image was only presented once. In the experimental condition, categories were paired in a 2:1 manner, where two different 'Leading' categories could result in one 'Trailing' category with 75% validity. For example, both 'beach' and 'barn' as 'Leading' categories would result in 'church' as a 'Trailing' category with 75% validity. In the control condition, one 'Leading' category led to two 'Trailing' categories with 50%/50% accuracy. Participants were not informed of the associations between categories, and were instead instructed to respond by button press when images appeared upside down, which occurred in ~5% of trials. This acted as catch trials to ensure participants maintained attention on the images. The occurrence of these catch trials was randomised, as was trial order. Participants did not undergo a training session, such that the neural signatures of early learning processes may be investigated using EEG. Each participant completed eight blocks consisting of 216 trials per block, totalling ~ 90 mins of testing.

#### *Categorisation task*

EEG data was collected from the first 10 participants as a pilot version of the experiment, after which the categorisation task was added to the experiment, resulting in a lower number of participants with behavioral results. After completion of the main task, the majority of participants (n=21) performed a categorisation task to ensure learning had taken place. In this task, participants were presented with the same images as before and instructed to indicate via button press whether the 'Trailing' image takes place indoors or outdoors. This task aimed to assess implicit reaction time (RT), as the statistical regularities learned in the main task could be used to predict what category the 'Trailing' image would belong to. Participants were not informed of the intent behind this task or instructed to make use of what they had learned in the main task.

## Data Analysis

#### *Categorisation task analysis*

Behavioural data from the categorisation task was analysed in terms of RTs. All RTs exceeding 2 SD above or below the mean across subjects were excluded as outliers. RTs for valid,invalid, and neutral trials underwent a log transformation before being averaged separately per participant. A repeated measures ANOVA was performed to compare RTs in each condition, then three separate paired t-tests were conducted comparing valid vs. invalid, valid vs. neutral, and invalid vs. neutral. Removal of RTs exceeding 2 SDs above or below the mean resulted in two participants having no neutral trials, thus these empty values were replaced with the mean of the remaining subjects. In order to examine the extent to which participants' neural responses were linked to their behavioural responses, we calculated Pearson correlation coefficients for behavioural responses, neural responses during the sharpening effect (123-180 ms), and neural responses during the dampening effect (280-296 ms) based on the difference in mean decoding across time within both time windows for each participant.

#### *EEG data preprocessing*

EEG data were acquired at 2048 Hz from 64 electrodes using a BioSemi ActiveTwo system. Electrodes were arranged according to the international 10/20 system. Preprocessing was conducted using custom SPM12 (Wellcome Trust Centre for Neuroimaging, University College London; *RRID:SCR_007037*) for Matlab (The MathWorks; *RRID:SCR_001622*) as well as custom

Matlab scripts. Continuous EEG data were high-pass filtered above 0.1 Hz and notch-filtered between 48-52 Hz using 5th-order zero-phase Butterworth filters. Data were then downsampled to 300 Hz. Blinks were detected using two horizontal and two vertical electro-oculogram (EOG) electrodes placed above and below the left eye, and to the left and right of the eyes. An epoch of -200 ms to 400 ms relative to artefact onset was applied to detected eye blinks and spatiotemporal confounds were calculated using Singular Value Decomposition (SVD). Top two principal components were removed from the segments of data associated with each eye blink (47). Data were then further denoised using the Dynamic Separation of Sources (48) which maximises the reproducibility of stimulus-evoked responses across trials in order to increase the signal-to-noise ratio of ERPs. During this process, data were re-referenced offline to the average of all channels. Continuous data were epoched between 200 ms before leading stimulus onset to 2600 ms after stimulus onset. Each epoch was baseline-corrected to the mean of the pre-stimulus onset (from -200 ms to 0 ms relative to stimulus onset). Epochs with an average root mean square (RMS) amplitude exceeding the median by two standard deviations (SDs) were excluded from analysis in order to remove contamination by transient artefacts. This resulted in rejecting 11.2% (± 1.03% the standard error of the mean (SEM) across participants) trials on average. Data were then averaged across trials and a low-pass filter of 48 Hz was applied. Catch trials were removed prior to analysis.

### *Decoding analyses*

Single-trial data underwent decoding based on linear support vector machines (SVM) implemented using custom MATLAB code. As part of feature selection, Principal Component Analysis (PCA) was conducted to reduce dimensionality of sensor-level EEG data, and subsequent components were selected based on signal-to-noise (SNR) with a cutoff threshold of 8 dB. As a result, an average of 7.6 components, SD = 0.498, were used across participants. Given the poor spatial resolution of EEG, considering all channels for feature extraction increases the computational cost, risk of overfitting, and required number of trials. As such, PCA can be useful for dimensionality reduction prior to SVM-based decoding (49,50). Three separate decoding analyses were conducted, decoding the visual category of both leading and trailing images, decoding the predictable trailing images based on the leading images, and memory decoding for trailing images. Valid and invalid were analysed separately at each step. In both analyses (of valid vs invalid trials), the decoder was trained and tested using leave-8-out cross-validation, such that in each test set, single trials corresponding to all 8 unique combinations of 4 leading and 2 trailing images were included (see Fig. 1B). Visual category decoding was used to differentiate neural responses to leading vs. trailing categories, while memory decoding aimed to assess memory of the leading category based on the trailing category.

The initial analysis aimed at quantifying decoding as a function of peristimulus time. To this end, we calculated the decoding accuracy per time point, using all trials. The resulting decoding time series were converted to one-dimensional images, entered into a GLM, outliers larger than 3 standard deviations based on Cook's distance were removed, and subjected to (1) a one-sample t-test against chance-level, equivalent to a fixed-effects analysis (51), and (2) a paired t-test between valid and invalid categories. While the fixed effects approach (one-sample t-test) enables us to establish whether some consistent informative patterns are detectable in these particular subjects, the results from our paired t-tests support inference to the wider population. Given that EEG time-series are autocorrelated over time, statistical tests were corrected for multiple comparisons over time using cluster-based FWE correction. A follow-up analysis aimed at quantifying changes in decoding over learning. To this end, based on the time windows in which we found significant decoding differences between valid and invalid trials (see below), we averaged the decoding accuracy estimates within each time window, separately for four bins of trials. In order to assess the utility of examining each bin in more detail, we calculated the trial-by-trial time-series of the decoding accuracy benefit for valid vs. invalid for each participant and averaged this benefit across time points for each of the two significant time windows, resulting in 200 trials per participant. Based on this, we fitted a logarithmic model to quantify the change of this benefit over trials, then found the trial index for which the change of the logarithmic fit was < 0.1%.

Logarithmic fits have been shown to accurately describe learning curves in statistical learning (52–54), informing our choice of model. In the early time window (123-180 ms) the median proportion of trials after which the decoding benefit stabilises is 11%, with the 99th percentile being reached after 38% of trials. In the later time window (280-196 ms), the median proportion of trials after which the decoding benefit stabilises is 18%, with the 99th percentile being reached after 47% of trials. Given that decoding stabilises after approximately 50% of trials (indicating that virtually zero learning took place after this point), and in order to balance this with having a sufficient number of trials for analysis, we focussed our further analyses on bins 1-2 to directly assess the effects of learning. Each bin encompassed approximately 15 minutes of testing or 432 trials on average.

# Data availability

Scripts and stimuli are available at *https://github.com/hannahmcderm/Expectation_EEG_eLife* while data is available at *https://osf.io/x7ydf*.

# Additional information

## Funding


| Funder | Grant reference number | Author |
|---|---|---|
| Deutsche Forschungsgemeinschaft (DFG) | AU423/2-1 | Ryszard Auksztulewicz |
| EC \| European Research Council (ERC) | https://doi.org/10.3030/101001270 | Federico De Martino |
| Deutsche Forschungsgemeinschaft (DFG) | SCHW1683/2-1 | Caspar M Schwiedrzik |


## Author ORCID iDs

**Hannah H McDermott**: https://orcid.org/0009-0005-5399-2917
**Federico De Martino**: https://orcid.org/0000-0002-0352-0648
**Caspar M Schwiedrzik**: https://orcid.org/0000-0003-0661-8859
**Ryszard Auksztulewicz**: https://orcid.org/0000-0001-9078-3667

# References

1. **Friston K** (2005) A theory of cortical responses. *Philos Trans R Soc B Biol Sci* **360**:815-36
2. **Jiang LP,** Rao RPN (2024) Dynamic predictive coding: A model of hierarchical sequence learning and prediction in the neocortex. *PLOS Comput Biol* **20**:e1011801
3. **Walsh KS,** McGovern DP, Clark A, O'Connell RG (2020) Evaluating the neurophysiological evidence for predictive processing as a model of perception. *Ann N Y Acad Sci* **1464**:242-68
4. **Blom T,** Bode S, Hogendoorn H (2021) The time-course of prediction formation and revision in human visual motion processing. *Cortex* **138**:191-202
5. **Malekshahi R,** Seth A, Papanikolaou A, Mathews Z, Birbaumer N, Verschure PFMJ, et al. (2016) Differential neural mechanisms for early and late prediction error detection. *Sci Rep* **6**:24350
6. **Wang R,** Shen Y, Tino P, Welchman AE, Kourtzi Z (2017) Learning Predictive Statistics: Strategies and Brain Mechanisms. *J Neurosci* **37**:8412-27
7. **Wischnewski M,** Schutter DJLG (2019) Electrophysiological correlates of prediction formation in anticipation of reward- and punishment-related feedback signals. *Psychophysiology* **56**:e13379

8. **Auksztulewicz R,** Friston K (2016) Repetition suppression and its contextual determinants in predictive coding. *Cortex* **80**:125-40
9. **Grotheer M,** Kovács G (2016) Can predictive coding explain repetition suppression?. *Cortex* **80**:113-24
10. **Todorovic A,** Lange FP de. (2012) Repetition suppression and expectation suppression are dissociable in time in early auditory evoked fields. *J Neurosci* **32**:13389-95
11. **Feuerriegel D,** Vogels R, Kovács G (2021) Evaluating the evidence for expectation suppression in the visual system. *Neurosci Biobehav Rev* **126**:368-81
12. **Barron HC,** Garvert MM, Behrens TEJ (2016) Repetition suppression: a means to index neural representations using BOLD?. *Philos Trans R Soc B Biol Sci* **371**:20150355
13. **Kok P,** Jehee JFM, de Lange FP (2012) Less is more: Expectation sharpens representations in the primary visual cortex. *Neuron* **75**:265-70
14. **Yon D,** Gilbert SJ, de Lange FP, Press C. (2018) Action sharpens sensory representations of expected outcomes. *Nat Commun* **9**:4288
15. **Kremláček J,** Kreegipuu K, Tales A, Astikainen P, Põldver N, Näätänen R, et al. (2016) Visual mismatch negativity (vMMN): A review and meta-analysis of studies in psychiatric and neurological disorders. *Cortex* **80**:76-112
16. **Kumar S,** Kaposvari P, Vogels R (2017) Encoding of predictable and unpredictable stimuli by inferior temporal cortical neurons. *J Cogn Neurosci* **29**:1445-54
17. **Meyer T,** Olson CR (2011) Statistical learning of visual transitions in monkey inferotemporal cortex. *Proc Natl Acad Sci* **108**:19401-6
18. **Richter D,** Ekman M, Lange FP de (2018) Suppressed sensory response to predictable object stimuli throughout the ventral visual stream. *J Neurosci* **38**:7452-61
19. **Garlichs A,** Blank H (2024) Prediction error processing and sharpening of expected information across the face-processing hierarchy. *Nat Commun* **15**:3407
20. **González-García C,** He BJ (2021) A Gradient of Sharpening Effects by Perceptual Prior across the Human Cortical Hierarchy. *J Neurosci* **41**:167-78
21. **Richter D,** Heilbron M, de Lange FP (2022) Dampened sensory representations for expected input across the ventral visual stream. *Oxf Open Neurosci* **1**:kvac013
22. **Schwiedrzik CM,** Freiwald WA (2017) High-Level Prediction Signals in a Low-Level Area of the Macaque Face-Processing Hierarchy. *Neuron* **96**:89-97.e4
23. **Press C,** Kok P, Yon D. (2020) The Perceptual Prediction Paradox. *Trends Cogn Sci* **24**:13-24
24. **Press C,** Gherri E, Heyes C, Eimer M. (2010) Action Preparation Helps and Hinders Perception of Action. *J Cogn Neurosci* **22**:2198-211
25. **Yon D,** Press C. (2017) Predicted action consequences are perceptually facilitated before cancellation. *J Exp Psychol Hum Percept Perform* **43**:1073-83
26. **Hughes G,** Desantis A, Waszak F (2013) Mechanisms of intentional binding and sensory attenuation: the role of temporal prediction, temporal control, identity prediction, and motor prediction. *Psychol Bull* **139**:133-51
27. **Han B,** Mostert P, de Lange FP (2019) Predictable tones elicit stimulus-specific suppression of evoked activity in auditory cortex. *NeuroImage* **200**:242-9
28. **Kiebel SJ,** Daunizeau J, Friston KJ (2008) A Hierarchy of Time-Scales and the Brain. *PLOS Comput Biol* **4**:e1000209
29. **Müller NG,** Strumpf H, Scholz M, Baier B, Melloni L (2013) Repetition Suppression versus Enhancement—It's Quantity That Matters. *Cereb Cortex* **23**:315-22
30. **Fitzgerald K,** Auksztulewicz R, Provost A, Paton B, Howard Z, Todd J (2021) Hierarchical Learning of Statistical Regularities over Multiple Timescales of Sound Sequence Processing: A Dynamic Causal Modeling Study. *J Cogn Neurosci* **33**:1549-62

31. **Manahova ME,** Mostert P, Kok P, Schoffelen JM, de Lange FP (2018) Stimulus Familiarity and Expectation Jointly Modulate Neural Activity in the Visual Ventral Stream. *J Cogn Neurosci* **30**:1366-77
32. **Ahissar M,** Hochstein S (2004) The reverse hierarchy theory of visual perceptual learning. *Trends Cogn Sci* **8**:457-64
33. **Fiser J,** Lengyel G (2019) A common probabilistic framework for perceptual and statistical learning. *Curr Opin Neurobiol* **58**:218-28
34. **Kilner JM,** Kiebel SJ, Friston KJ (2005) Applications of random field theory to electrophysiology. *Neurosci Lett* **374**:174-8
35. **Nigam T,** Schwiedrzik CM (2024) Predictions enable top-down pattern separation in the macaque face-processing hierarchy. *Nat Commun* **15**
36. **Thomas ER,** Haarsma J, Nicholson J, Yon D, Kok P, Press C. (2024) Predictions and errors are distinctly represented across V1 layers. *Curr Biol* **4**:p2265-2271.e4
37. **Xu M,** Meng J, Yu H, Jung TP, Ming D (2020) Dynamic Brain Responses Modulated by Precise Timing Prediction in an Opposing Process. *Neurosci Bull* **37**:70-80
38. **Kok P,** Mostert P, de Lange FP (2017) Prior expectations induce prestimulus sensory templates. *Proc Natl Acad Sci* **114**:10473-8
39. **Sherman BE,** Graves KN, Huberdeau DM, Quraishi IH, Damisah EC, Turk-Browne NB (2022) Temporal dynamics of competition between statistical learning and episodic memory in intracranial recordings of human visual cortex. *J Neurosci* **42**:9053-68
40. **Cichy RM,** Pantazis D, Oliva A (2014) Resolving human object recognition in space and time. *Nat Neurosci* **17**:455-62
41. **Aitken F,** Kok P (2022) Hippocampal representations switch from errors to predictions during acquisition of predictive associations. *Nat Commun* **13**:3294
42. **Riels K,** Ramos Campagnoli R, Thigpen N, Keil A (2022) Oscillatory brain activity links experience to expectancy during associative learning. *Psychophysiology* **59**:e13946
43. **Libby A,** Buschman TJ (2021) Rotational dynamics reduce interference between sensory and memory representations. *Nat Neurosci* **24**:715-26
44. **Cappotto D,** Kang H, Li K, Melloni L, Schnupp J, Auksztulewicz R (2022) Simultaneous mnemonic and predictive representations in the auditory cortex. *Curr Biol* **32**:2548-2555.e5
45. **Barron HC,** Auksztulewicz R, Friston K (2020) Prediction and memory: A predictive coding account. *Prog Neurobiol* **192**:101821
46. **Huang YT,** Wu CT, Fang YXM, Fu CK, Koike S, Chao ZC (2024) Crossmodal hierarchical predictive coding for audiovisual sequences in the human brain. *Commun Biol* **7**:965
47. **Ille N,** Berg P, Scherg M (2002) Artifact Correction of the Ongoing EEG Using Spatial Filters Based on Artifact and Brain Signal Topographies. *J Clin Neurophysiol* **19**:113
48. **de Cheveigné A,** Simon JZ (2008) Denoising based on spatial filtering. *J Neurosci Methods* **171**:331-9
49. **Zhang T,** Chen W, Li M (2019) Classification of inter-ictal and ictal EEGs using multi-basis MODWPT, dimensionality reduction algorithms and LS-SVM: A comparative study. *Biomed Signal Process Control* **47**:240-51
50. **Rahman MA,** Hossain MF, Hossain M, Ahmmed R. (2020) Employing PCA and *t*-statistical approach for feature extraction and classification of emotion from multichannel EEG signal. *Egypt Inform J* **21**:23-35
51. **Allefeld C,** Görgen K, Haynes JD (2016) Valid population inference for information-based imaging: From the second-level t-test to prevalence inference. *NeuroImage* **141**:378-392
52. **Kang H,** Auksztulewicz R, Chan CH, Cappotto D, Rajendran VG, Schnupp JWH (2023) Cross-modal implicit learning of random time patterns. *Hear Res* **438**:108857

53. **Siegelman N,** Bogaerts L, Kronenfeld O, Frost R (2018) Redefining "learning" in statistical learning: What does an online measure reveal about the assimilation of visual regularities?. *Cogn Sci* **42**:692-727

54. **Choi D,** Batterink LJ, Black AK, Paller KA, Werker JF (2020) Preverbal infants discover statistical word patterns at similar rates as adults: Evidence from neural entrainment. *Psychol Sci* **31**:1161-73

# Peer reviews

## Reviewer #3 (Public review):

Summary:

In their study McDermott et al. investigate the neurocomputational mechanism underlying sensory prediction errors. They contrast two accounts: representational sharpening and dampening. Representational sharpening suggests that predictions increase the fidelity of the neural representations of expected inputs, while representational dampening suggests the opposite (decreased fidelity for expected stimuli). The authors performed decoding analyses on EEG data, showing that first expected stimuli could be better decoded (sharpening), followed by a reversal during later response windows where unexpected inputs could be better decoded (dampening). These results are interpreted in the context of opposing process theory (OPT), which suggests that such a reversal would support perception to be both veridical (i.e., initial sharpening to increase the accuracy of perception) and informative (i.e., later dampening to highlight surprising, but informative inputs).

Strengths:

The topic of the present study is of significant relevance for the field of predictive processing. The experimental paradigm used by McDermott et al. is well designed, allowing the authors to avoid common confounds in investigating predictions, such as stimulus familiarity and adaptation. The introduction provides a well written summary of the main arguments for the two accounts of interest (sharpening and dampening), as well as OPT. Overall, the manuscript serves as a good overview of the current state of the field.

Weaknesses:

In my opinion the study has a few weaknesses. Some method choices appear arbitrary (e.g., binning). Additionally, not all results are necessarily predicted by OPT. Finally, results are challenging to reconcile with previous studies. For example, while I agree with the authors that stimulus familiarity is a clear difference compared to previous designs, without a convincing explanation why this would produce the observed pattern of results, I find the account somewhat unsatisfying.

https://doi.org/10.7554/eLife.103689.3.sa1

## Author response:

The following is the authors' response to the previous reviews

> ***Reviewer 1***
>
> *Minor*
>
> *The main substance of my previous comment I suppose targeted a deeper issue - namely whether such a result is reflecting a resolution to a 'neural prediction' puzzle or a 'perceptual prediction' puzzle. Of course, these results tell us a great deal about a potential resolution for how dampening and sharpening might co-exist in the brain - but*

> *in the absence of corresponding perceptual effects (or a lack of correlation between neural and perceptual variables - as outlined in this revision) I do wonder if any claims about implications for perception might need moderation or caveating. To be honest, I don't think the authors *need* to make any more changes along these lines for this paper to be acceptable - it is more an issue they might wish to consider themselves when contextualizing their findings.*

Thank you for the thoughtful comment. We have now added a caveat to the relevant section of the discussion to make it clearer that we are discussing neural results, not perceptual results (p.20, lines 378-379).

> *I am also happy with the changes that the authors have made justifying which claims can and cannot made based on a statistical decoding test against 'chance' in a single condition using t-tests. I was perhaps a little unclear when I spoke about 'comparisons against 0' in my original review, when the key issue (as the authors have intuited!) is about comparisons against 'chance' (where e.g., 0% decoding above chance is the same thing as 'chance'!). The authors are of course correct in the amendment they have made on p.29 to make clear this is a 'fixed effects analysis' - though I still worry this could be a little cryptic for the average reader. I am not suggesting that the authors run more analyses, or revise any conclusions, but I think it would be more transparent if a note was added along the lines of "while the fixed effects approach (one-sample t-test) enables us to establish whether some consistent informative patterns are detectable in these particular subjects, the results from our paired t-tests support inference to the wider population".*

This sentence has been added for increased transparency (p. 27, lines 544-547).

> ***Reviewer 3***
>
> *Major*
>
> *(1) In the previous round of comments, I noted that: "I am not fully convinced that Figures 3A/B and the associated results support the idea that early learning stages result in dampening and later stages in sharpening. The inference made requires, in my opinion, not only a significant effect in one-time bin and the absence of an effect in other bins. Instead to reliably make this inference one would need a contrast showing a difference in decoding accuracy between bins, or ideally an analysis not contingent on seemingly arbitrary binning of data, but a decrease (or increase) in the slope of the decoding accuracy across trials. Moreover, the decoding analyses seem to be at the edge of SNR, hence making any interpretation that depends on the absence of an effect in some bins yet more problematic and implausible". The authors responded: "we fitted a logarithmic model to quantify the change of the decoding benefit over trials, then found the trial index for which the change of the logarithmic fit was < 0.1%. Given the results of this analysis and to ensure a sufficient number of trials, we focused our further analyses on bins 1-2". However, I do not see how this new analysis addresses the concern that the conclusion highlights differences in decoding performance between bins 1 and 2, yet no contrast between these bins are performed. While I appreciate the addition of the new model, in my current understanding it does not solve the problem I raised. I still believe that if the authors wish to conclude that an effect differs between two bins they must contrast these directly and/or use a different appropriate analysis approach.*
>
> *Relatedly, the logarithmic model fitting and how it justifies the focus on analysis bin 1-2 needs to be explained better, especially the rationale of the analysis, the choice of parameters (e.g., why logarithmic, why change of logarithmic fit < 0.1% as criterion, etc), and why certain inferences follow from this analysis. Also, the reporting of the associated results seems rather sparse in the current iteration of the manuscript.*

We thank the reviewer for this important point. Following your suggestion, we conducted additional post-hoc tests directly comparing the first and second bins. We found significant differences between bins in the invalid trials, but not the valid trials, suggesting that sharpening/dampening effects are condition specific. This is discussed in the manuscript on p.14, lines 268-271; p.15, 280-284; p.20, lines 382-386.

A logarithmic analysis was chosen as learning is usually found to be a nonlinear process; learning effects occur rapidly before stabilising relatively early, as seen in Fig. 2D. This is consistent with other research which found that logarithmic fits efficiently describe learning curves in statistical learning (Kang et al., 2023; Siegelman et al., 2018; Choi et al., 2020). By utilising a change of logarithmic fit at <0.1% as a criterion, it is ensured that virtually zero learning took place after that point, allowing us to focus our analysis on learning effects as they developed and providing a more accurate model of representational change. This is explained in the manuscript on p.13, lines 250-251; p.27-28, lines 557-563.

> *(2) A critical point the authors raise is that they investigate the buildup of expectations during training. They go on to show that the dampening effect disappears quickly, concluding: "the decoding benefit of invalid predictions [...] disappeared after approximately 15 minutes (or 50 trials per condition)". Maybe the authors can correct me, but my best understanding is as follows: Each bin has 50 trials per condition. The 2:1 condition has 4 leading images, this would mean ~12 trials per leading stimulus, 25% of which are unexpected, so ~9 expected trials per pair. Bin 1 represents the first time the participants see the associations. Therefore, the conclusion is that participants learn the associations so rapidly that ~9 expected trials per pair suffice to not only learn the expectations (in a probabilistic context) but learn them sufficiently well such that they result in a significant decoding difference in that same bin. If so, this would seem surprisingly fast, given that participants learn by means of incidental statistical learning (i.e. they were not informed about the statistical regularities). I acknowledge that we do not know how quickly the dampening/sharpening effects develop, however surprising results should be accompanied with a critical evaluation and exceptionally strong evidence (see point 1). Consider for example the following alternative account to explain these results. Category pairs were fixed across and within participants,i.e. the same leading image categories always predicted the same trailing image categories for all participants. Some category pairings will necessarily result in a larger representational overlap (i.e., visual similarity, etc.) and hence differences in decoding accuracy due to adaptation and related effects. For example, house barn will result in a different decoding performance compared to coffee cup barn, simply due to the larger visual and semantic similarity between house and barn compared to coffee cup and barn. These effects should occur upon first stimulus presentation, independent of statistical learning, and may attenuate over time e.g., due to increasing familiarity with the categories (i.e., an overall attenuation leading to smaller between condition differences) or pairs.*

We apologise for the confusion, there are 50 expected trials per bin per condition. The trial breakdown is as follows. Each participant completed 1728 trials, split equally across 3 mappings (two 2:1 maps and one 1:2 map), giving 1152 trials in the 2:1 mapping. Stimuli were expected in 75% of trials (864), leaving 216 per bin, and 54 per leading image in each bin. We have clarified this in the script (p.14, line 267; p.15, line 280). This is in line with similar studies in the field (e.g. Han et al., 2019).

> *(3) In response to my previous comment, why the authors think their study may have found different results compared to multiple previous studies (e.g. Han et al., 2019; Kumar et al., 2017; Meyer and Olson, 2011), particularly the sharpening to dampening switch, the authors emphasize the use of non-repeated stimuli (no repetition suppression and no familiarity confound) in their design. However, I fail to see how familiarity or RS could account for the absence of*

*sharpening/dampening inversion in previous studies.*

*First, if the authors argument is about stimulus novelty and familiarity as described by Feuerriegel et al., 2021, I believe this point does not apply to the cited studies. Feuerriegel et al., 2021 note: "Relative stimulus novelty can be an important confound in situations where expected stimulus identities are presented often within an experiment, but neutral or surprising stimuli are presented only rarely", which indeed is a critical confound. However, none of the studies (Han et al., 2019; Richter et al., 2018; Kumar et al., 2017; Meyer and Olson, 2011) contained this confound, because all stimuli served as expected and unexpected stimuli, with the expectation status solely determined by the preceding cue. Thus, participants were equally familiar with the images across expectation conditions.*

*Second, for a similar reason the authors argument for RS accounting for the different results does not hold either in my opinion. Again, as Feuerriegel et al. 2021 correctly point out: "Adaptation-related effects can mimic ES when the expected stimuli are a repetition of the last-seen stimulus or have been encountered more recently than stimuli in neutral expectation conditions." However, it is critical to consider the precise design of previous studies. Taking again the example of Han et al., 2019; Kumar et al., 2017; Meyer and Olson, 2011. To my knowledge none of these studies contained manipulations that would result in a more frequent or recent repetition of any specific stimulus in the expected compared to unexpected condition. The crucial manipulation in all these previous studies is not that a single stimulus or stimulus feature (which could be subject to familiarity or RS) determines the expectation status, but rather the transitional probability (i.e. cue-stimulus pairing) of a particular stimulus given the cue. Therefore, unless I am missing something critical, simple RS seems unlikely to differ between expectation condition in the previous studies and hence seems implausible to account for differences in results compared to the current study.*

*Moreover, studies cited by the authors (e.g. Todorovic & de Lange, 2012) showed that RS and ES are separable in time, again making me wonder how avoiding stimulus repetition should account for the difference in the present study compared to previous ones. I am happy to be corrected in my understanding, but with the currently provided arguments by the authors I do not see how RS and familiarity can account for the discrepancy in results.*

The reviewer is correct in that the studies cited (Han et al., 2019; Kumar et al., 2017; Meyer and Olson, 2011) ensure that participants are equally familiar with the images across expectation conditions. Where the present study differs is that participants are not familiar with individual exemplars at all. Han et al., 2019 used a pool of 30 individual images, and subjects underwent exposure sessions lasting two hours each daily for 34 days prior to testing. Kumar et al., 2017 used a pool of 12 images with subjects being exposed to each sequential pair 816 times over the course of the training period. Meyer & Olsen, 2011 used pure tones at five different pitch levels. While familiarity of stimuli across conditions was controlled for in these studies in the sense that familiarity was constant across conditions, novelty was not controlled for. The present study uses a pool of ~3500 images, which are unrepeated across trials.

Feuerriegel et al., 2021 also points out: “There are also effects of adaptation that are dependent on the recent stimulation history extending beyond the last encountered stimulus and long-lag repetition effects that occur when the first and second presentation of a stimulus is separated by tens or even hundreds of intervening images”. Bearing this in mind, and given the very small pool of stimuli being used by Han et al., 2019; Kumar et al., 2017; Meyer and Olson, 2011, it stands to reason that these studies may still have built-in but unaccounted for effects relating to the repetition of exemplars. Thus, our avoidance of those possible

confounds, in addition to foregoing any prior training, may elicit differing results. Furthermore, as pointed out by Walsh et al. 2020, methodological heterogeneity (such as subject training) can produce contrasting results as PP makes divergent predictions regarding the properties of prediction error given different permutations of variables such as training, transitional probabilities, and conditional probabilities. In our case, the use of differing methodology was intentional. These issues have been discussed in more detail on p.5, lines 112-115; p.19, lines 368-377; p.20, lines 378-379).

> *Minor*
>
> *(1) The authors note in their reply to my previous questions that: "As mentioned above, we opted to target our ERP analyses on Oz due to controversies in the literature regarding univariate effects of ES (Feuerriegel et al., 2021)". This might be a lack of understanding on my side, but how are concerns about the reliability of ES, as outlined by Feuerriegel et al. (2021), an argument for restricting analyses to 1 EEG channel (Oz)? Could one not argue equally well that precisely because of these concerns we should be less selective and instead average across multiple (occipital) channels to improve the reliability of results?*

The reviewer is correct in suggesting that a cluster of occipital electrodes may be more reliable than reporting one single electrode. We have amended the analysis to examine electrodes Oz, O1, and O2 (p.9, lines 187-188; p.11, lines 197-201).

> *(2) The authors provide a github link for the dataset and code. However, I doubt that github is a suitable location to share EEG data (which at present I also cannot find linked in the github repo). Do the authors plan to share the EEG data and if so where?*

Thank you for bringing this to my attention. EEG data has now been uploaded at osf.io/x7ydf and linked to the github repository (p.28, lines 569-570).

> *(3) The figure text could benefit from additional information; e.g. Fig.1C and Fig.3 do not clarify what the asterisk indicates; p < ? with or without multiple comparison correction?*

Thank you for pointing out this oversight, the figure texts have been amended (p. 9, line 168; p.16, line 289).

https://doi.org/10.7554/eLife.103689.3.sa0