Memory masking vs overwriting in procedural categorization

Intro

In an era defined by ubiquitous digital devices and a seemingly endless stream of media optimized to hijack our attention, the ability to manage, redirect, or erase habitual behavior is more critical than ever. Technological advances have dramatically enhanced the capacity of product developers to induce habitual use, yet our ability to flexibly control or disrupt these behaviors has failed to keep pace. This growing imbalance poses serious challenges—not only to individual well-being but also to the functioning of society—and addressing it is increasingly important.

A major source of habitual behavior likely lies in the neural mechanisms supporting operant stimulus-response (SR) learning. In the animal learning literature, especially in rodent models, behavior is often characterized along a continuum from goal-directed to habitual. Goal-directed actions are sensitive to the agent's current goals and motivational states, enabling flexible adaptation to changing environmental contingencies. In contrast, habitual behaviors are rigid and automatic—triggered directly by environmental cues—and often persist even when the associated reward is devalued. For example, a rat may continue pressing a lever for food despite being sated, much like a human may continue scrolling through social media absent any specific goal or gratification.

In the human cognitive literature, this distinction between goal-directed and habitual behavior closely parallels the division between declarative and procedural learning and memory systems. Declarative memory supports flexible, hypothesis-driven reasoning and explicit knowledge, while procedural memory underlies behaviors learned through direct reinforcement and repeated experience, typically with little cognitive effort or conscious awareness. This division is particularly well studied in the domain of category learning, where different category structures are thought to preferentially engage different memory systems.

In recent work, we reported an intervention that appeared to induce true overwriting of procedural category knowledge. Participants who had previously acquired a category structure through procedural learning no longer showed evidence of this knowledge following the intervention, suggesting that the learned associations had been erased. However, a critical alternative explanation remained untested. Rather than eliminating the procedural memory, the intervention may have merely masked its behavioral expression. On this account, the underlying SR mappings remain intact but lie dormant and may reemerge under appropriate conditions.

In the present study, we directly test this possibility. We provide clear evidence that the intervention does not result in unlearning of procedural category knowledge. Instead, it masks the expression of that knowledge, which remains accessible under suitable retrieval conditions. These findings have important implications for how we understand, measure, and intervene on habit-like behaviors across domains.

Methods

Experiments and conditions

The study consisted of two experiments, each comprising three phases: Learn, Intervention, and Test. In both experiments, participants were assigned to one of two between-subject conditions: Relearn or New Learn. The Learn and Test phases were identical across experiments and conditions. During the Learn phase, participants were trained on a category structure designed to promote procedural learning. In the Test phase, they were assessed on the same or a new category structure, depending on their assigned condition. The critical manipulation occurred during the Intervention phase. In Experiment 1, participants received

fully random feedback during this phase, whereas in Experiment 2, feedback was a mixture of random and veridical. This design allowed us to assess whether the interventions disrupted or merely masked previously acquired procedural knowledge.

Stimuli and categories

The stimuli were circular sine-wave gratings that varied in spatial frequency and orientation. The coordinates of all stimuli were generated by first sampling points in polar coordinates and then converting them into Cartesian coordinates. Specifically, radius values r were sampled from a uniform distribution on the interval $[0, 1]$, and angle values θ were sampled uniformly from the interval $[0, 2\pi]$. These polar coordinates (r, θ) were then transformed into Cartesian coordinates (x, y) using the equations $x = r\cos(\theta)$ and $y = r\sin(\theta)$. This resulted in a set of (x, y) coordinates uniformly distributed within a circle of radius 1 and centered at the origin. Next, (x, y) coordinates were transformed from a circular uniform distribution to an elliptical uniform distribution with horizontal major axis by multiplying the x values by 124.02 and the y values by 28.44. For the Test phase in the New Learning conditions, the resulting coordinates were rotated by 45and translated by $(40, 60)$ for half the stimuli and by $(60, 40)$ for the other half. The resulting stimulus distributions are shown in Figure 1.

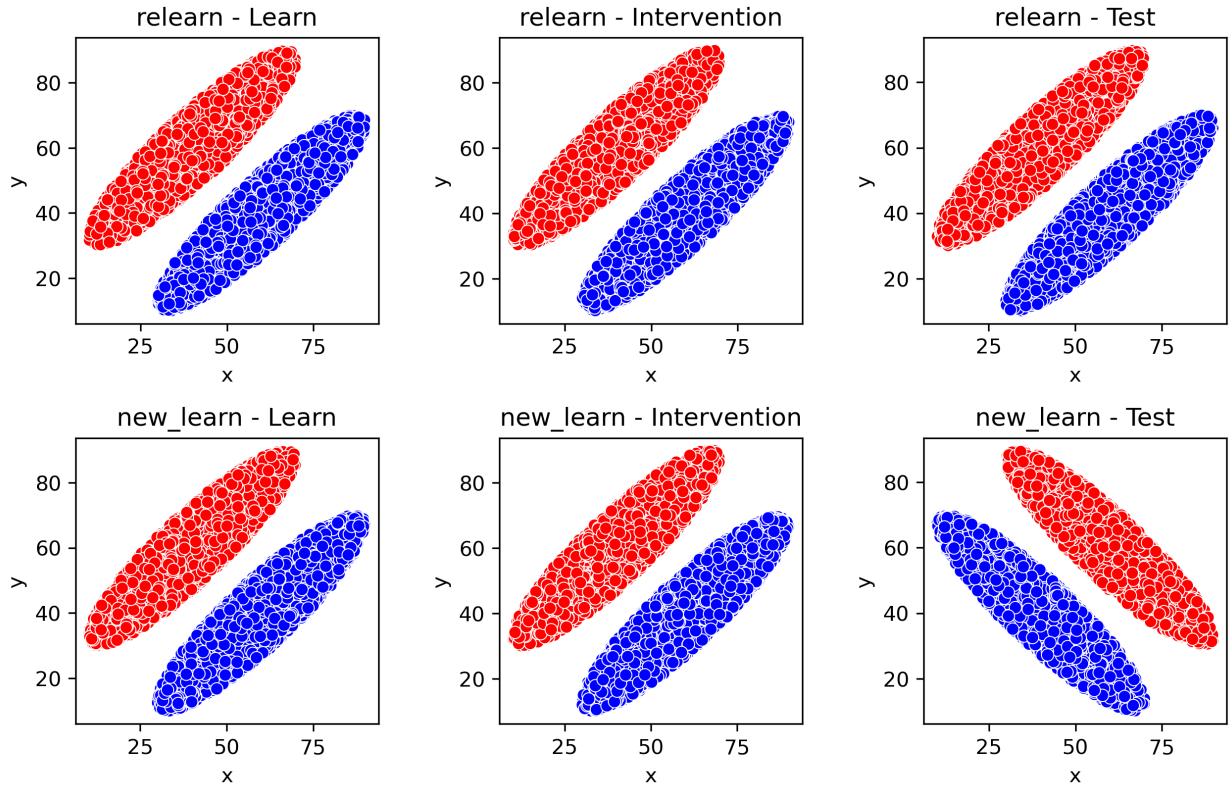

Figure 1

The category structures used in this study were of the information-integration (II) type, in which optimal categorization requires the integration of information across both stimulus dimensions (spatial frequency and orientation). Crucially, in II tasks, no single dimension alone is sufficient for accurate categorization; rather, the decision boundary is diagonal in the stimulus space and does not align with either axis. This stands in contrast to rule-based (RB) category structures, which are typically defined by simple, verbally describable rules along one dimension (e.g., "if orientation ¿ threshold, then category A; else, category B"). RB tasks are generally well-suited to hypothesis-testing and are easily verbalized by participants, making them heavily reliant on explicit, declarative memory systems.

By contrast, II tasks have been shown to favor learning via implicit, procedural systems that rely on gradual, trial-by-trial stimulus-response (SR) association learning. A substantial body of evidence supports this dissociation. For example, ashby1998differentiating demonstrated that performance on II tasks is disrupted by concurrent working memory loads, consistent with their reliance on resource-limited procedural learning. Similarly, maddox2004dissociating found that II category learning is impaired by feedback delays, while RB learning is relatively unaffected—suggesting that II learning depends on temporally contiguous reinforcement signals, a hallmark of procedural learning systems.

Neuroimaging studies provide further support for this distinction. RB learning is associated with activation in the prefrontal cortex and medial temporal lobe structures, which are typically involved in explicit memory and rule formulation nomura2007neural. In contrast, II learning engages the basal ganglia—particularly the striatum—which plays a central role in procedural learning and reinforcement-driven SR mapping poldrack2001interactions, ell2006evidence. Additionally, pharmacological and patient studies indicate that II learning is selectively impaired in populations with striatal dysfunction (e.g., Parkinson's disease), while RB learning remains largely intact n an era defined by ubiquitous digital devices and amaddox2010cognition.

Given these cognitive and neural distinctions, II tasks are widely considered a robust means of isolating procedural learning processes. In the present study, the use of II category structures thus allows us to target and examine the dynamics of procedural memory acquisition, retention, and potential masking or overwriting in the context of experimental manipulations. This distinction is critical, as our primary research question concerns whether procedural SR associations are erased or merely masked by interventions involving feedback disruption.

Procedure

Participants provided informed consent and were given an optional demographic questionnaire to complete. Participants were instructed that their task was to categorize circular sine-wave gratings on the basis of their spatial frequency and orientation, and that each category was equally likely.

Each participant completed a single session consisting of 900 trials. Each phase (Learn, Intervention, Test) consisted of 300 trials. On each trial, participants viewed a fixation cross (1000 ms), followed by a response-terminated stimulus, and then feedback (1000 ms). Responses were given via the "d", and "k" keys. Feedback following correct responses was a green circle that appeared around the stimulus, and feedback following incorrect responses was a red circle. See Figure **??** an illustration of example trials.

Participants

A total of 40 participants were recruited for Experiment 1 (32 female, 8 male), with ages ranging from 18 to 32 years (M = 20.33, SD = 2.61). An additional 40 participants took part in Experiment 2 (X female, X male, 2 nonbinary, and 1 preferred not to disclose), with ages ranging from X to X years (M = X, SD = X). Participants were pseudo-randomly assigned to either the new learning or relearning condition using a blocked allocation method, ensuring equal sample sizes (n = 20) in each condition. To be eligible, participants had to be at least 18 years old with normal or corrected-to-normal vision. All participants were undergraduate students at Macquarie University and received course credit in exchange for participation. Ethics approval for this study was granted by the Macquarie University Human Research Ethics Committee (Ref: 520251317762011).

Decision-Bound Analysis

To identify the decision strategy used by each participant, we fit decision-bound models AshbyValentin2018 to the trial-by-trial response data from the final 100 trials of the Train phase and the first 100 trials of the Test phase separately for each participant. We examined 1-dimensional rule-based models and 2-dimensional procedural models. The rule-based models assumed participants established a criterion on a single stimulus dimension and then categorized stimuli based on whether or not they exceeded this criterion. These models had two free parameters – a response criterion on the attended stimulus dimension, and the variance of perceptual and criterial noise. The 2-dimensional procedural models – i.e., general linear classifier (GLC) –

assumed that participants used a linear decision boundary with an arbitrary slope and intercept to divide the stimulus space into two response regions. The GLC assumes that stimuli are categorized based on their position relative to this boundary. The GLC has three free parameters – a slope and intercept of the decision bound and the variance of perceptual and criterial noise. For details on the models and the model-fitting process, see AshbyValentin2018.

Bayesian Estimation of Procedural Reacquisition

To estimate and compare the probability that participants reacquired a procedural strategy across experiments and conditions, we conducted a series of Bayesian analyses using posterior sampling from Beta distributions.

We defined reacquisition as participants being classified as procedural in both the initial learning phase (Block 2) and the final test phase (Block 6). For each condition (Relearn vs. New Learning) and experiment (Experiment 1 vs. Experiment 2), we computed the posterior distribution over the reacquisition probability, θ, using a Beta prior with parameters $\alpha = 1$ and $\beta = 1$ (i.e., a uniform prior). Given the observed number of procedural reacquisitions (successes) and total participants in each group, we drew 100,000 samples from the resulting Beta posterior distribution for each group:

$$\theta \sim \text{Beta}(\text{successes} + 1,\ \text{failures} + 1) \tag{1}$$

For each condition, we then computed the posterior distribution over the difference in reacquisition probability between Experiment 1 and Experiment 2 by subtracting posterior samples ($\Delta = \theta_1 - \theta_2$). We derived 95% credible intervals from these difference distributions and reported the posterior probability that $\Delta > 0$ (i.e., the probability that reacquisition was more likely in Experiment 1 than in Experiment 2).

Additionally, to assess the effect of condition within each experiment, we compared the posterior distributions of reacquisition rates between the Relearn and New Learning conditions separately for Experiment 1 and Experiment 2.

The resulting posterior distributions and their differences were visualized using histograms, with red dashed lines indicating 95% credible intervals and black dashed lines marking the null value ($\Delta = 0$). This analysis provides direct probabilistic evidence about the likelihood of differences in reacquisition rates across experiments and conditions.

Results

Figure 2 displays mean accuracy per block for each condition and experiment. In Experiment 1 (left panel), participants successfully learned the category structure during the Learn phase but ceased to express this knowledge during the Intervention phase. During the Test phase, participants in the Relearn condition rapidly returned to their previous accuracy levels, indicating that the intervention (random feedback) did not erase the original learning. In contrast, those in the New Learn condition performed significantly worse, consistent with interference from the previously learned, conflicting stimulus-response mappings. A similar pattern was observed in Experiment 2 (right panel), suggesting that a mixed-feedback intervention failed to overwrite the initial learning. These findings support the interpretation that the mixed feedback intervention results we previously reported on masked, rather than erased, procedural category knowledge.

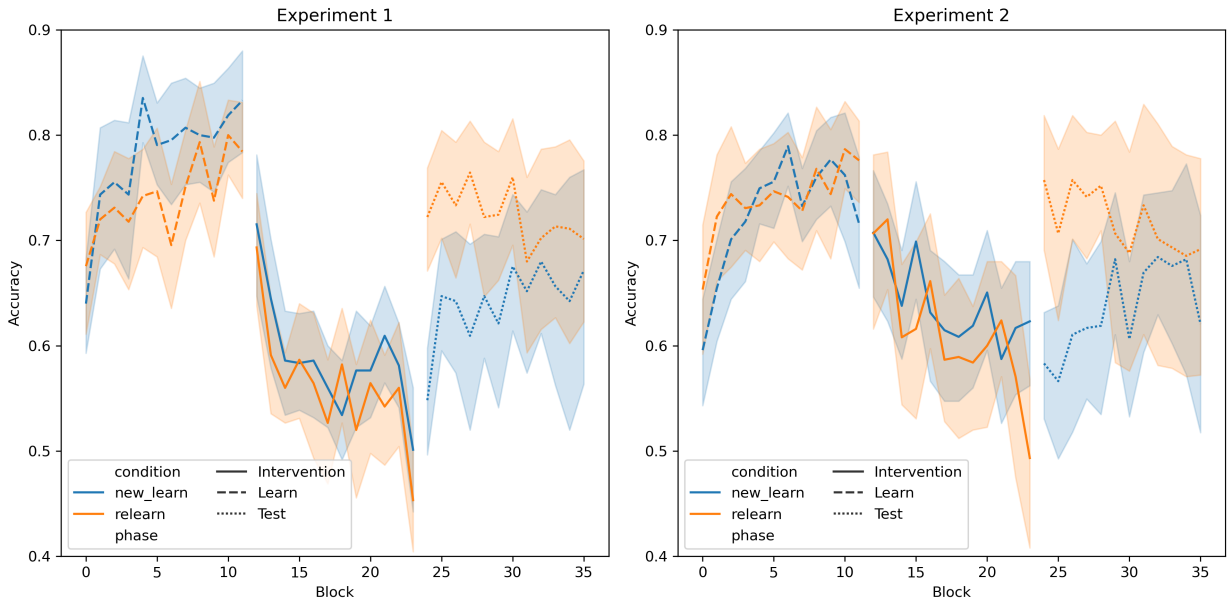

Figure 2: Accuracy over blocks for all participants. Separate lines represent different conditions within each experiment.

Although the results suggest that the mixed-feedback intervention leads to memory masking rather than overwriting, it is essential to rule out potential contamination by the declarative system (e.g., the use of explicit rules). To address this, we fit decision-bound models to each participant's data from the final block of training and the first block of testing. Our primary questions were: (1) did participants initially acquire a procedural strategy, and (2) did they reacquire that strategy during the Test phase? If the intervention caused true unlearning, there should be no difference in the proportion of participants expressing a procedural strategy between the Relearn and New Learn conditions during the Test phase.

Figure 3 shows heatmaps of transitions between best-fit model classes from the end of training to the start of testing, illustrating how participants shifted (or failed to shift) their categorization strategies across phases, separated by experiment and condition. In Experiment 1, participants in the Relearn condition predominantly acquired and reacquired a procedural strategy, indicating that the intervention did not disrupt their underlying category knowledge. In contrast, those in the New Learn condition tended to switch from a procedural to a rule-based strategy during the test phase, suggesting interference rather than erasure. A similar pattern was observed in Experiment 2. This further supports the conclusion that an intervention of mixed feedback masks rather than eliminates procedural category learning.

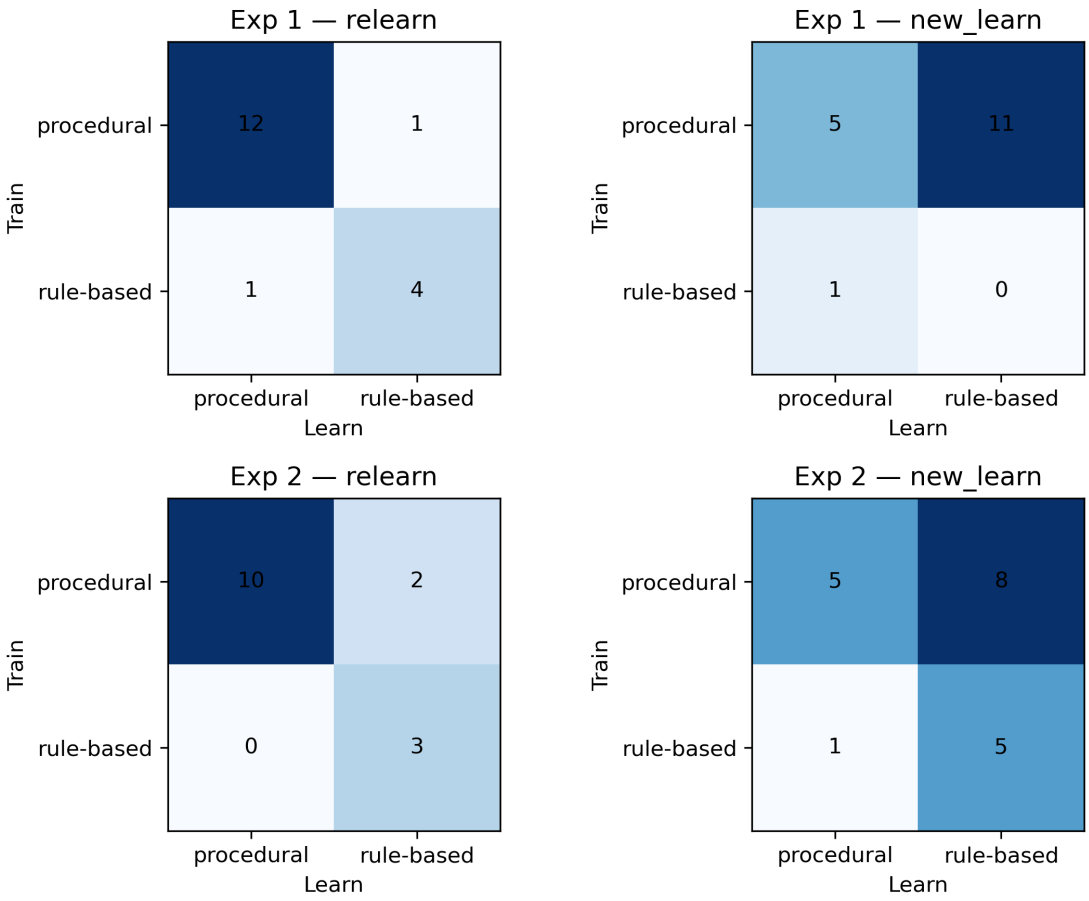

Figure 3: Heatmaps showing transitions between best-fit model classes from the end of training (Y-axis) to the start of testing (X-axis), separated by experiment and condition. Each cell indicates the number of participants who transitioned between model classes.

Figure 4 shows Bayesian posterior estimates of the probability that participants reacquired a procedural strategy during the initial stages of the Test phase. For both the Relearn and New Learn conditions, there was little difference in reacquisition probability between Experiment 1 and Experiment 2, as indicated by the 95% credible intervals that included zero (top and middle rows, rightmost panels). However, participants in the Relearn condition were substantially more likely to reacquire a procedural strategy than those in the New Learn condition. This was true in both experiments, as shown by the 95% credible intervals excluding zero in the bottom row's left and middle panels. Together, these results provide strong evidence that the intervention masked but did not erase procedural category knowledge.

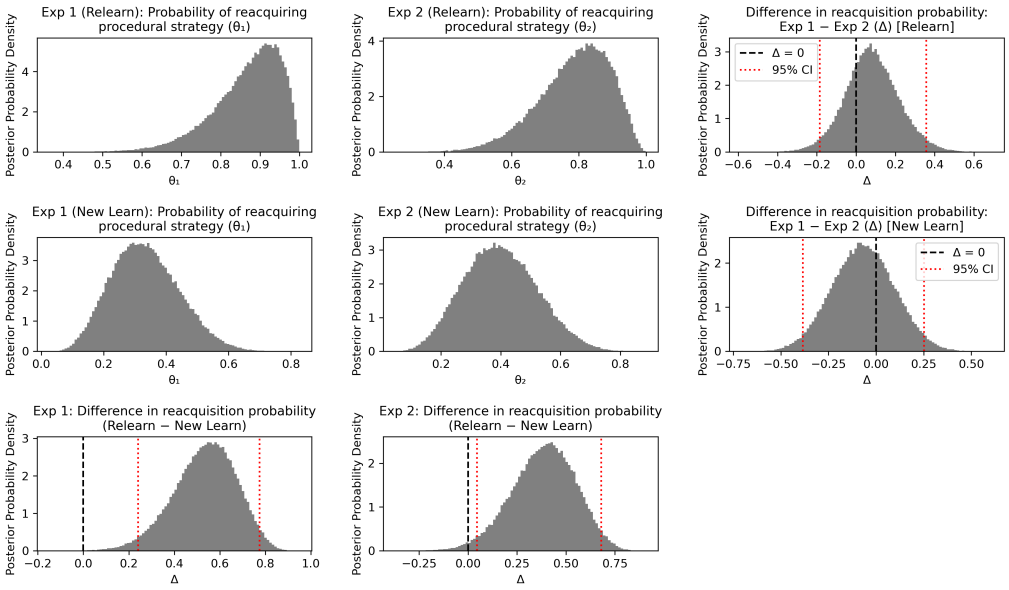

Figure 4: Bayesian posterior distributions over θ, the probability of reacquiring a procedural strategy. Top two rows show experiment comparisons within Relearn and New Learn conditions. Bottom row shows differences between conditions within each experiment. Red lines denote 95% credible intervals; black dashed lines indicate null difference ($\Delta = 0$).

Discussion

...