**Category Learning Effects on Memory**

Kevin O’Neill, Audrey Liu, Siyuan Yin, Timothy Brady, and Felipe De Brigard

**Author Note**

Correspondence may be directed to Kevin O’Neill, Center for Cognitive Neuroscience, Duke University, .... Email: [kevin.oneill@duke.edu](mailto:kevin.oneill@duke.edu).

**Abstract**

Recent views on categorization suggest that people may invoke one of several cognitive processes for category learning under different circumstances, each relying on a distinct memory system. It has been suggested that the exact system used depends on the nature of the stimuli, with rule-based categories recruiting activity in the hippocampus and prefrontal cortex, and information-integration categories relying on the striatum. However, little research has been devoted to studying exactly how strong this association between category structures and memory systems is, and what effects differences in learning may have on subsequent cognition. In this paper, we address both of these questions by asking participants to learn rule-based categories under a variety of learning conditions (practice only, instruction only, both, neither), and by looking at resulting differences in recognition memory. Replicating previous work, we show an increase in both hits and false alarms for in-category items compared with out-of-category items. Interestingly, we find no differences in recognition or RT between groups under different conditions of learning, suggesting that learning in all conditions produced schema-like representations. In accordance with the multiple-memory systems perspective of category learning, our data suggest that rule-based categories may be learned using the same network under a variety of different conditions. \*add citations

*Keywords*: Category Learning; Recognition Memory

**Introduction**

Research in category learning has been converging on the idea that category learning is supported by distinct memory systems. For example, while simple rule-based categories are learned episodically via the hippocampus and vmPFC, categories without verbally expressible rules (i.e., information-integration categories) have been shown to be learned procedurally through reinforcement learning in the striatum.

**Methods**

**Participants**

867 participants were recruited via Amazon Mechanical Turk (<https://www.mturk.com>). All participants were from the United States, had at least 100 approved HITs, had an overall HIT approval rate of at least 95%, and received $2.00 in compensation for their participation. To ensure that participants learned the category and remained on-task during the experiment, data from 153 participants were excluded because of failure to learn the category below 85% accuracy during the last 20 trials of learning, as in De Brigard et al. (2017), or excessive mean response time during learning or test (greater than the grand mean + three standard deviations, 5.95 seconds and 5.46 seconds respectively), leaving 714 participants for data analysis. All participants were provided informed consent in accordance with the Duke University IRB.

**Materials**

Stimuli consisted of MATLAB (2018b)-generated flowers, used previously in De Brigard et al. (2017). These flowers vary over five features, with each feature having three possible values: number of petals (four, six, or eight), petal color (blue, green or yellow), center shape (circle, triangle, or square), center color (orange, purple, or turquoise), and number of sepals (one, two, or three). Figure 1 depicts three flowers that illustrate all possible values of the five features. All flowers were displayed on the center of the screen with a white background.

**Procedure**

Procedure followed Experiment 4 of De Brigard et al. (2017), with a few modifications. This paradigm includes three phases: learning, study, and test. Participants began the experiment by reading an instruction screen for a minimum of 30 seconds which detailed the five stimulus features and the possible values those features could take. This instruction screen also displayed two example stimuli for illustration. Participants were instructed that they would see flowers on the screen, one at a time, and would be asked to determine whether each flower belonged to the species *avlonia*. Participants were told that *avlonias* differed from other flowers in one simple way (e.g., only *avlonias* have four petals), but that they must discover what makes *avlonia* flowers unique. Participants were told that they must initially guess, but that they would eventually learn what makes a flower an *avlonia*. The feature and value that constituted this Learned category (*avlonia*) was counterbalanced across participants. In order to isolate effects of the Learned category from possible effects of stimulus features not associated with conceptual information, participants were also assigned a Not-Learned category, defined by a value of a different feature, of which they were unaware. The Not-Learned category was never mentioned to the participants, statistically independent of the Learned category, and counterbalanced across participants. During each of the three phases of the experiment, each value of each feature was displayed in 1/3 of the trials for that phase, so that the co-occurrence of all feature/value combinations was uniform. Accordingly, 1/3 of all flowers presented were members of the Learned category.

In addition to the above category conditions, we introduced two further between-subjects manipulations on learning: whether or not the participant was explicitly instructed of the the learned category’s discriminating feature and value (Instruction: Instructed vs. Not-Instructed), and whether the participant actively categorized flowers during learning, or merely watched as flowers were categorized as *avlonias* or not-*avlonias* on the screen (Practice: Practiced vs. Not-Practiced). These manipulations were also fully counterbalanced between participants.

Participants first learned to categorize flowers into the species *avlonia*. In the Instructed condition, participants were told explicitly how to identify *avlonias*, (e.g. “*Avlonias* are flowers that have six petals”). In the Not-Instructed condition, however, participants were told only that they would have to learn what feature and value defined the species *avlonia*. In the Practiced condition, participants completed 72 self-paced trials in which they pressed the “y” key if the flower was an *avlonia*, or the “n” key otherwise. Immediate feedback (“Correct” or “Incorrect”) was presented after each key-press for 1s. In the Not-Practiced condition, participants instead passively viewed 72 trials in which a flower was shown for 3s, and a categorization (“Avlonia” or “Not Avlonia”) was presented immediately after for 1s. Of the 72 flowers presented, 16 flowers were in the Learned category but not the Not-Learned category, 16 flowers were in the Not-Learned category but not the Learned category, 8 flowers were in both categories, and 32 flowers were in neither category.

In the study phase, participants were asked to memorize 18 flowers. Participants read instructions for this phase for a minimum of 30s. Each flower was shown for 5s following a 1s inter-trial interval, and none of these flowers were shown previously in the learning phase. Of these 18 flowers, 4 flowers were in the Learned category but not the Not-Learned category, 4 flowers were in the Not-Learned category but not the Learned category, 2 flowers were in Both categories, and 8 flowers were in Neither category. Participants were told that they would receive a bonus if they could remember a high number of flowers (XX participants were in fact given a $X.XX bonus for a hit rate exceeding 85%).

Finally, in the test phase, participants were told that they would see 54 flowers, one by one, and asked to press the “y” key if the flower was old (presented during the study phase), or to press the “n” key otherwise. Participants also read instructions for this phase for a minimum of 30s. Each trial was self-paced with a 1s inter-trial interval. Of these 54 flowers, 18 were presented during study. Of the remaining 36 flowers (lures), 8 flowers were in the Learned category but not the Not-Learned category, 8 flowers were in the Not-Learned category but not the Learned category, 4 flowers were in Both categories, and 16 flowers were in Neither category. None of the lures appeared in the learning or study phases.

**Results**

**Learning Phase**

Because the participants in the Not-Practiced condition did not make any responses during learning, we limit analysis of this phase to those in the Practiced condition. As found in De Brigard et al. (2017), participants in the Not-Instructed condition started at near chance (*M* = 65.8%) categorization accuracy in the first 10 trials, and gradually rose to near ceiling (*M* = 98.7%) accuracy in the last 10 trials. In contrast, participants in the Instructed condition began with high (*M* = 89.3%) accuracy, and gradually rose to near ceiling (*M* = 99.1%) accuracy. This confirms that explicit instruction allowed participants to successfully learn the *avlonia* category before practice, although participants could learn the category without explicit instruction.

**Test Phase**

As outcome variables during the testing phase, we measured reaction time (RT), hits, and false alarms (FAs), as well as hierarchical estimates of sensitivity (*d'*) and bias (*C*) under signal detection theory. For all outcome variables, we fit Bayesian mixed-effects models in R using the **brms** package with 5 chains, 1000 iterations of burn-in, and 1500 iterations of sampling. Means and standard deviations for all variables are reported in Table 1.

To examine the effects of our manipulations on RT, we employed 2 (Learned: yes vs. no) x 2 (Not-Learned: yes vs. no) x 2 (Practice: yes vs. no) x 2 (Instruction: yes vs. no) x 2 (Old: yes vs. no) Bayesian mixed-effect model with random intercepts for each subject, random slopes for Learned, Not-Learned, and Old, and an ex-Gaussian distribution. There was...

For hits and FAs, we used similar 2 (Learned) x 2 (Not-Learned) x 2 (Practice) x 2 (Instruction) Bayesian mixed-effect models with random intercepts for each subject, uncorrelated random slopes for Learned and Not-Learned, and Bernoulli distributions with a logit link function. We used weakly informative Gaussian priors on all coefficients, with means of 0 and standard deviations of 4. Looking at hits, we found weak evidence that flowers in the Learned category were recollected more frequently than flowers not in the Learned category, *β* = .34, *CI* = [.12, .56], *BF* = 2.08. In contrast, there was strong evidence for against an effect of Not-Learned, *β* = .19, *CI* = [-.02, .40], *BF* = 0.13, Practice, *β* = -.10, *CI* = [-.32, .11], *BF* = 0.04, and Instruction, *β* = -.06, *CI* = [-.25, .14], *BF* = 0.03. Similarly, there was moderate to strong evidence against all 2, 3, and 4-way interactions between these variables (all BFs < 0.15). For FAs, we again found weak evidence that lures in the Learned category were also more likely to be falsely recollected, *β* = .27, *CI* = [.10, .45], *BF* = 1.82. In contrast, there was strong evidence against an effect of Not-Learned, *β* = .09, *CI* = [-.06, .24], *BF* = .04, Practice, *β* = -.03, *CI* = [-.22, .16], *BF* = .03, and Instruction, *β* = .10, *CI* = [-.07, .27], *BF* = .04. There was also moderate to strong evidence against all 2, 3, and 4-way interactions between these variables (all BFs < .25).

Finally, we made hierarchical estimates of SDT parameters (namely *d’* and *C*) using a 2 (Learned) x 2 (Not-Learned) x 2 (Practice) x 2 (Instruction) x 2 (Old) Bayesian mixed-effect model with random intercepts for each subject, uncorrelated random slopes for Learned, Not-Learned, and Old, and a Bernoulli distribution with a probit link function. We used weakly informative Gaussian priors on all coefficients, with means of 0 and standard deviations of 2. In this model, the intercept represents estimates of -*C*, the effect of Old represents estimates of *d’*, effects of our factors represent effects on *C*, and interactions with Old represent effects on *d’*. There was strong evidence for a main effect of Old, suggesting that subjects had above-zero estimates of *d’*, *β* = .23, *CI* = [.14, .32], *BF* > 1000. There was also weak evidence for an effect of Learned, *β* = .17, *CI* = [.07, .27], *BF* = 3.57, suggesting that estimates of *C* were higher for flowers in the Learned category () compared to flowers not in the Learned category (). In contrast, there was strong evidence against effects of Practice, *β* = -.02, *CI* = [-.13, .10], *BF* = .03, Instruction, *β* = 06., *CI* = [-.04, .16], *BF* = .05, and Not Learned, *β* = .05, *CI* = [-.04, .14], *BF* = .05, on estimates of *C.* There was also moderate to strong evidence against effects of Practice, *β* = -.05, *CI* = [-.18, .09], *BF* = .04, Instruction, *β* = -.10, *CI* = [-.22, .02], *BF* = .09, Learned, *β* = .04, *CI* = [-.11, .18], *BF* = .04, and Not Learned, *β* = .07, *CI* = [-.08, .21], *BF* = .05, on estimates of *d’*. Finally, there was weak to strong evidence against all other interactions between Learned, Not Learned, Practice, Instruction, and Old (all BFs < 0.3).

**Discussion**

In this study we aimed to discover how different forms of learning impact schema formation and deployment. To do so, we tested participant's memory after learning simple rule-based categories under different learning constraints (practice only, instruction only, both, neither). We found little behavioral difference between these groups, suggesting that they may be equally effective with regard to forming lasting schemas. However, we did find that the Learned category had a significant impact on hits, FAs, and response bias (*C*)*,* supporting the idea that all of these methods of category learning produce representations that are synonymous with schemas (**CITE**).

Future work in this area could reveal conditions in which the representations produced under these different learning constraints are dissociable. Because this paradigm allows for testing the memory effects of newly learned categories independently of the perceptual or statistical features of the stimuli, we believe that it would be fertile ground for testing this hypothesis in fMRI.