Was That My Cue? Reactivity to Category-Level Judgments of Learning

Anthony Cruz (acruz27@uwo.ca)
Department of Psychology, University of Western Ontario
London, ON N6A 5C2 Canada

John Paul Minda (jpminda@uwo.ca)
Department of Psychology, University of Western Ontario
London, ON N6A 5C2 Canada

Abstract

Making a judgment of learning (JOL) during study can improve later test performance, a phenomenon called JOL reactivity. In paired-associates learning, JOLs improve memory for strongly (not weakly) related word pairs. JOLs appear to strengthen cue-target associations, enhancing future performance on tests sensitive to those associations. We investigated whether JOL reactivity would emerge in feedback-based category learning, wherein participants learn novel stimulus-response associations. We investigated whether this effect would be present for novel test items and if it would depend upon stimulus-category relatedness. Participants completed a category learning task; some performed JOLs throughout learning. At test, participants categorized novel and previously studied stimuli of varying degrees of stimulus-category relatedness. We found JOL reactivity for both novel and previously studied stimuli, and no effect of relatedness. Our experiment provides preliminary evidence that JOL reactivity can be produced in feedback-based category learning. COVIS theory provides an excellent framework for future investigations.

Keywords: Concepts and categories; Learning

Introduction

A learner's ability to monitor their learning is essential to self-guided study. Before an assessment, a learner may consider how well they remember the material in deciding how to study, what material to study, and the extent to which further study is necessary (Metcalfe, 2009). To probe learners' self-monitoring of performance, researchers might ask learners to predict how likely they are to correctly remember material on a future test. This metacognitive task, the judgment of learning (JOL), sees widespread use in a variety of learning and memory paradigms. A rapidly developing body of evidence suggests that making a JOL can have an impact on memory, a finding typically referred to as JOL reactivity (see Double et al., 2018 for a meta-analysis). JOL reactivity has been observed in word list memorization (Hourihan & Tullis, 2015; Kubik et al., 2022; Yang et al., 2015) and inferential learning (Lee & Ha, 2019), but is predominantly examined in paired-associates learning (e.g., Chang & Brainerd, 2023; Halamish & Undorf, 2023; Myers et al., 2020; Rivers et al., 2023; Witherby et al., 2023).

JOL reactivity was first experimentally demonstrated by Soderstrom et al. (2015). In their paired-associates learning task, participants studied a list of 60 cue-target word pairs, of which half were strongly related (e.g., *blunt-sharp*) and half were weakly related (e.g., *boxer-terrible*). Each pair was studied for 8s. Halfway through the presentation of a given word pair, some participants were asked to make a JOL, judging their likelihood of successfully recalling the target word on a future cued recall test. After a 3-minute retention interval, participants took a cued recall test in which they were presented with the cue word and asked to produce the associated target word. For strongly related word pairs, participants who performed JOLs had higher cued recall performance than those who did not. In contrast, the presence of JOLs had no impact on cued recall performance for the weakly related word pairs. This pattern of results, wherein JOLs are beneficial for strongly related but not weakly related word pairs, has since been demonstrated repeatedly in the paired-associates literature (Chang & Brainerd, 2023; Halamish & Undorf, 2023; Mitchum et al., 2016; Myers et al., 2020; Rivers et al., 2021, 2023; Witherby et al., 2023).

The leading hypothesis explaining JOL reactivity is the cue-strengthening hypothesis (Soderstrom et al., 2015). This hypothesis assumes that when participants make JOLs, they call upon an existing cue-target relationship (Halamish & Undorf, 2023). This cue-target relationship is strengthened by the act of making a JOL, improving later memory performance on a test that is sensitive to that relationship (Myers et al., 2020). It has been suggested that the cue-target relationship must be semantic to benefit from JOLs, as the typical pattern of JOL reactivity is not produced when participants connect otherwise unrelated cue-target pairs via mental imagery (Witherby et al., 2023) or morphological similarities (Rivers et al., 2023). It has also been suggested that this effect is item-specific and does not lead to general changes in learning strategy (Rivers et al., 2021). The cue-strengthening hypothesis is strongly defined by its relationship to the paired-associates learning task. In order to develop a comprehensive theory of JOL reactivity, it is imperative that the effects of JOLs on learning are assessed in a wider variety of learning tasks (Halamish & Undorf, 2023; Myers et al., 2020; Rivers et al., 2021). Our primary research goal is to determine whether feedback-based category learning also benefits from JOL reactivity. If so, we are interested in whether the cue-strengthening hypothesis could be adapted to explain JOL reactivity in this learning paradigm.

In L. K. Samuelson, S. L. Frank, M. Toneva, A. Mackey, & E. Hazeltine (Eds.), *Proceedings of the 46th Annual Conference of the Cognitive Science Society.*

Categorization is a cognitive process by which stimuli are sorted into functional equivalence classes using many-to-one stimulus-response mappings (Kéri, 2003). Categorization is essential to inference making; when the features that define a category are learned, it is possible to categorize never-before-seen (novel) stimuli, and subsequently make inferences about them. For example, an animal's physical features might make it possible to classify it as a "dog," and one can infer various non-physical traits of dogs, such as disposition and diet. Category learning is often studied using an observational learning paradigm, a paradigm in which participants are presented with category members and labels for a fixed duration of study time, in a similar vein to paired-associates learning (e.g., Do & Thomas, 2023; Kang & Pashler, 2012; Kornell & Bjork, 2008).

Lee & Ha (2019) explored JOL reactivity in observational category learning. Participants made JOLs at one of three levels of granularity: item-by-item (*Rate your likelihood of correctly identifying new paintings by the given artist at a later test*), category-level (*Rate your likelihood of correctly classifying new paintings by Artist X at a later test*), and global-level (general questions about performance, such as *Rate your likelihood of correctly identifying who created new paintings by the artists you have just studied*). At test, participants only classified novel stimuli. Participants correctly classified stimuli at higher rates following category- and global-level JOLs, but not item-by-item JOLs. Critically, this study provides evidence that JOL reactivity can occur in category learning. Lee & Ha's (2019) distinction between the different levels of JOL granularity falls in line with the cue-strengthening hypothesis; item-by-item JOLs do not benefit categorization test performance because item-by-item information is not probed on this type of test (Myers et al., 2020). However, this study refutes the cue-strengthening hypothesis' prediction that JOL reactivity requires a pre-existing semantic relationship, as the stimuli presented during learning were novel to participants. Moreover, it refutes the idea that the effect is item-specific, as it was observed in novel test items.

In this experiment, we explore JOL reactivity in a feedback-based, procedural category learning task. During this type of category learning, it is supposed that participants gradually learn stimulus-response associations (Ashby et al., 2003). If JOL reactivity is necessarily item-specific, as proposed in the paired-associates literature (Rivers et al., 2021), then we should only expect JOL reactivity for previously studied items at test. However, models of categorization generally suppose that it is regions of perceptual space, rather than specific stimuli, that become associated with novel category labels (e.g., Ashby et al., 1998; Smith & Minda, 1998). Considering this, as well as Lee & Ha's (2019) results, we expect JOL reactivity for both novel and previously studied test items. This is our primary prediction in this experiment.

One model describing categorization comes from general recognition theory (Ashby & Perrin, 1988). In this multivariate generalization of signal detection theory, it is assumed that while learning novel categories, learners draw boundaries separating the different regions of perceptual space associated with each category. This idea of separating stimuli into categories using a boundary in perceptual space is a staple of the category learning literature (e.g., Epping & Busemeyer, 2023; Roark et al., 2022; Shamloo & Hélie, 2020). After learning the boundary, items further from the boundary may be perceived as more strongly associated with their respective category than items close to the boundary (Seger et al., 2015). In the paired-associates literature, it has been found repeatedly that JOLs enhance memory for strongly (but not weakly) related word pairs (Chang & Brainerd, 2023; Halamish & Undorf, 2023; Mitchum et al., 2016; Myers et al., 2020; Rivers et al., 2021, 2023; Soderstrom et al., 2015; Witherby et al., 2023). In this experiment, we compare stimuli that are strongly and weakly associated with their assigned categories by using distance from the category boundary as an index of relatedness. We are interested in determining whether JOL reactivity depends on this index of relatedness. This would provide a potential generalization of the cue-strengthening hypothesis to account for JOL reactivity in category learning.

We expect to find reactivity to category-level JOLs in feedback-based, procedural category learning. In feedback-based category learning, participants begin learning by randomly guessing the category to which each stimulus belongs, gradually learning via immediate corrective feedback. This paradigm makes it possible to assess participants' degree of learning throughout the learning phase. In observational and paired-associates learning, this is not possible. Participants in this experiment learned to categorize artificial visual stimuli (sinusoidal gratings) varying along continuous dimensions into two categories separated by a linear boundary. Throughout learning, some participants made category-level JOLs while others did not. Unlike Lee & Ha (2019), we included a mix of novel and previously studied items at test, allowing us to determine if the effect of reactivity would differ between retention and generalization. Since participants are presumed to learn to associate distinct regions of perceptual space with the appropriate category labels, we expected this benefit to be present in both novel and previously studied items during the test. Our results are in line with this prediction. In an attempt to use the cue-strengthening hypothesis to account for these results, we assumed that items further from the category boundary would be more strongly associated with their assigned category, predicting that these items would benefit more from JOL reactivity than items close to the boundary. Our results did not support this prediction. We discuss future directions and implications of our results.

Methods

Data were collected between 27 September 2023 and 31 April 2024. Participants completed this online experiment on their smartphones. There were three between-participant conditions: No Pause (NP), Pause (P), and Category-Level JOL (CLJ). The CLJ condition served as the experimental condition; we expected that participants in this condition would outperform those in the remaining conditions. During learning, participants learned to categorize 128 stimuli by performing six blocks of feedback-based learning. After each learning block, some participants took a took a pause (P) or performed category-level JOLs (CLJs). The No Pause and Pause conditions were both experimental controls, with the latter being time-matched to the CLJ condition. At test, participants classified a mix of previously studied (retention) and novel (generalization) stimuli.

Participants

An a priori power analysis was not conducted. Participants were $N = 192$ ($n_{NP} = 60$, $n_P = 64$, $n_{CLJ} = 68$) undergraduate students compensated with course credit. Participants spoke fluent English, owned a smartphone with internet access, and had normal or corrected-to-normal vision. All experimental procedures and materials were approved by Western's Research Ethics Board. Participants were randomly assigned to a learning condition after beginning the online experiment.

Materials

Stimuli were generated using the GRT package (Matsuki, 2017) in R version 4.1.1 (R Core Team, 2021). Stimuli were grayscale sinusoidal gratings varying in spatial frequency (f) and orientation angle (θ). Parameters for learning phase stimuli were sampled from two multivariate Gaussian distributions with identical covariance matrices, such that the Pearson correlation between f and θ was $r = .78$ in each category. 64 stimuli were generated for each category, resulting in 128 unique stimuli for the learning phase. Categories were labelled "A" or "B." To minimize the potential effects of the labels or button locations, the stimuli that corresponded to each label were counterbalanced across participants. Mean f (measured in cycles per image) for each group was $(\mu_A, \mu_B) = (11,17)$. Mean θ (in degrees relative to vertical) was $(\mu_A, \mu_B) = (81,64)$. Categories were separated by a linear boundary that perfectly accounted for category membership. The deterministic nature of this boundary made it possible (albeit unlikely) for participants to derive a general, 100% accurate categorization strategy This boundary was defined by a linear combination of the stimulus dimensions, $\theta = 30.89 + 2.54f$. Along this boundary, every unit increase in spatial frequency is associated with a 2.54 degree increase in orientation angle. Figure 1 shows the entire stimulus space and boundary used in this study. Boundaries of this nature, which integrate information from multiple stimulus dimensions, are thought to be learned via stimulus-response association (Ashby et al., 1998; Ashby & Valentin, 2017). During learning, the maximum possible accuracy attainable using single-dimensional rule-based strategies was 73.44% using an f-based strategy or 69.53% using a θ-based strategy. We simulated 20000 random responders in this task and defined "above chance" performance as any accuracy at or above the 99th percentile of these simulated random responders' accuracies, 60.16%.

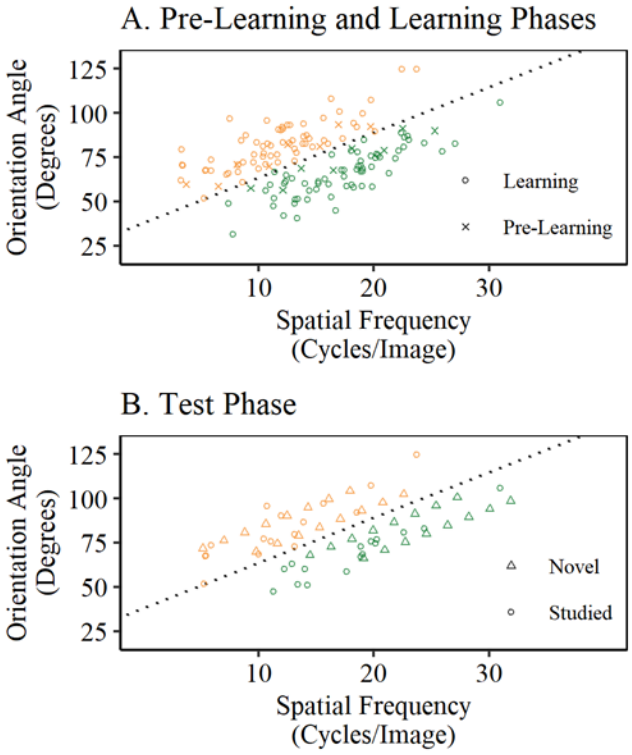

Figure 1: Stimulus space used for pre-learning (A), learning (A), and test (B) phases.

Sixteen unique stimuli were generated for the pre-learning phase. Stimulus parameters were generated sampled from a 4x4 grid rotated about the category boundary. The grid was centered at the point $(f, \theta) = (14.5,74.9)$, a point which lies on the category boundary. In a similar manner, parameters for novel test phase stimuli were generated by rotating an 8x4 grid about the category boundary. This grid was centered at the point $(f, \theta) = (16.8,80.5)$. 32 stimuli generated for the learning phase were presented again at test, resulting in a total of 64 test phase stimuli. The shortest Euclidean distance between each test stimulus and the optimal categorization boundary was measured. A cutoff of 5.4 units separated "close" and "far" test items. For both the pre-learning and test phase, novel stimulus parameters were chosen such that they occupied similar regions of perceptual space to the stimuli used during the learning phase, and such that they ran parallel to the category boundary. Stimuli were 352px x 352px in size

at a resolution of 96 DPI and were scaled to fill 45% of the participant's smartphone screen width during the experiment. See Figure 2 for examples of stimuli used in the experiment.

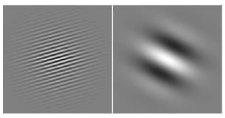

Figure 2: Example stimuli.

Procedure

Participants completed a Qualtrics form indicating their consent to participate and verifying they were using their smartphones. Then, participants were redirected to the experiment URL, where they were randomly assigned to a learning condition. The experimental procedure included a pre-learning phase, a learning phase, and a test phase. Pre-learning consisted of a similarity judgment task, learning consisted of six blocks of feedback-based category learning, and test consisted of one block of categorization without feedback. Participants completed the experiment in one 90-minute session. The experiment was programmed in jsPsych 7.1.2 (de Leeuw, 2015) and hosted on Pavlovia.org.

Pre-Learning Phase. The pre-learning phase consisted of a similarity judgment task. In one trial, participants saw two stimuli side-by-side (in a randomized order) in the center of their phone screen and were asked to evaluate how similar they appeared to be. Images were programmed to each take 45% of the width of each participant's phone screen. Participants used a 1-8 Likert scale to provide their responses and a 10-second time limit was imposed for each trial. There were constant on-screen instructions stating that 8 should refer to pairs that are "identical or nearly identical" and 1 to pairs that are "extremely dissimilar." Participants were shown all 136 possible pairs of similarity judgment stimuli (16 identical, 64 between-category, and 56 within-category). Stimuli from this task were not used during learning or test but were designed to occupy similar regions of perceptual space. This pre-learning task is included to be consistent with previous uses of this category learning paradigm (Cruz & Minda, 2024). This pre-exposure to the category structure may facilitate later learning (Folstein et al., 2009, 2010).

Learning Phase. During the learning phase, participants completed 6 blocks of category learning with 128 trials per block, resulting in 768 total learning trials. Participants were instructed to sort stimuli into Category A or Category B by pressing the A and B buttons at the button of the screen with their thumb. They were instructed to use only one finger for category judgments for the duration of the experiment. To minimize the potential effects of the A/B labels or button locations, the stimuli that corresponded to each label were counterbalanced across participants. Each learning trial began with 500ms of fixation. The stimulus would then appear until the participant made a judgment or until 10s had passed. This was followed by 700ms of corrective feedback ("CORRECT!" or "INCORRECT!"). Figure 3 depicts one learning trial. If participants failed to provide a response in time, they were asked to "Please respond more quickly." There were 128 unique stimuli presented during learning, and each was presented once per block in a randomized order.

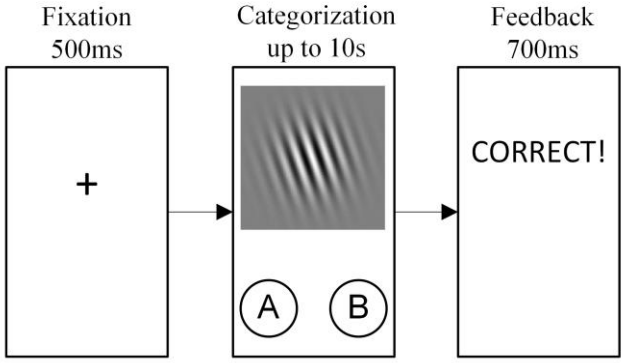

Figure 3: Example of one category learning trial.

The experimental manipulations took place during the learning phase. Participants in the No Pause condition did not receive breaks between adjacent learning blocks; participants in this condition transitioned from one learning block to the next with no explicit indication that one block of learning had ended. The Pause condition involved a 4.2-second pause after each learning block in which participants were told to "Please take a moment to pause." The duration of this pause was chosen as the median time it took participants to complete category-level JOLs in a pilot study. In the Category-Level JOL condition, each block of learning was followed by two category-level JOLs (one for each category) presented in a random order. Participants were given the following prompt: "How likely are you to correctly identify members of **Category X** in the future? Please use the full range of the scale." Responses were made using a 0-100 sliding scale. Numerical labels were omitted. Instead, the left- and right-hand sides of the scale were labelled "Wild Guess" and "Certain Correct," respectively (Jacoby et al., 2010). The slider began in the center of the scale (50) but was required to be adjusted before participants submitted a response.

Test Phase. The test phase involved a categorization test with 64 unique trials. Test trials were structured very similarly to learning trials. Each trial started with 500ms of fixation. Then, a stimulus was shown, and participants had to decide whether it belonged to Category A or Category B. In lieu of feedback, each trial was followed by 700ms of a blank screen. The test immediately followed the final block of category learning, with instructions appearing that indicated to participants that they were entering the test phase. Of the 64 test stimuli, 32 were novel (generalization) items and 32 were previously studied (retention) items.

Results

Statistical values are reported to three significant digits. Analyses were conducted in R version 4.3.1 (R Core Team, 2023). All data and stimuli from this experiment are available upon request. Only participants who reached accuracy significantly above chance levels (at least 60.18%) during the final learning block were included in analyses, leaving a final sample size $N = 120$ ($n_{NP} = 35$, $n_P = 39$, $n_{CLJ} = 46$).

Learning Phase

A 3 (learning condition) x 6 (learning block) ANOVA was conducted with accuracy as the dependent variable. The main effect of learning condition was not significant, $F(2,117) = 0.557, p = 0.574, \eta_p^2 = 0.009$. The main effect of learning block was significant, $F(5,585) = 20.6, p = 5.59 \times 10^{-19}, \eta_p^2 = 0.15$. See Table 1 for accuracy across each block of learning. The interaction between learning block and learning condition was not significant, $F(10,585) = 1.65, p = 0.09, \eta_p^2 = 0.027$.

Table 1: Accuracy by learning block and condition.

	Learning Block					
	1	2	3	4	5	6
No Pause						
Mean	0.6828	0.7064	0.7018	0.721	0.7185	0.7271
SD	0.1138	0.1126	0.1074	0.101	0.0882	0.0861
Pause						
Mean	0.6855	0.7332	0.7370	0.732	0.7394	0.7476
SD	0.0836	0.0887	0.0855	0.105	0.0934	0.0744
Category-Level JOL						
Mean	0.6596	0.7120	0.7356	0.735	0.7437	0.7568
SD	0.0922	0.0979	0.0990	0.114	0.1040	0.0828

Planned contrasts were conducted in the first and final blocks of learning. The first set of planned contrasts compared accuracy in the Category-Level JOL condition to the remaining learning conditions. This contrast was not significant in block 1, $t(117) = -1.35, p = 0.179$, or block 6, $t(117) = 1.27, p = 0.205$. The second set of planned contrasts compared accuracy in the Category-Level JOL condition to accuracy in the Pause condition. This contrast was not significant in block 1, $t(117) = -1.23, p = 0.221$, or block 6, $t(117) = 0.521, p = 0.604$.

Test Phase

A 3 (learning condition: NP vs. P vs. CLJ) x 2 (stimulus type: Retention vs. generalization) x 2 (distance from boundary: Close vs. far) ANOVA was conducted with test accuracy as the dependent variable. The main effect of novelty was significant, $F(1,69) = 17.9, p = 6.91 \times 10^{-5}, \eta_p^2 = 0.206$. Participants performed better on retention items ($M = 0.755, SD = 0.127$) than generalization items ($M = 0.702, SD = 0.116$). The main effect of distance from boundary was significant, $F(1,69) = 122, p = 7.09 \times 10^{-17}, \eta_p^2 = 0.638$. Participants classified items close to the boundary ($M = 0.654, SD = 0.114$) less accurately than items far from the boundary ($M = 0.804, SD = 0.125$). The interaction between stimulus type and distance from boundary was significant, $F(1,69) = 14.8, p = 2.69 \times 10^{-4}, \eta_p^2 = 0.176$. Among items close to the boundary, retention items ($M = 0.703, SD = 0.142$) were classified more accurately than generalization items ($M = 0.605, SD = 0.124$). Among items far from the boundary, performance did not differ by stimulus type (See Figure 4A). All other interactions were not significant, $ps > .4$.

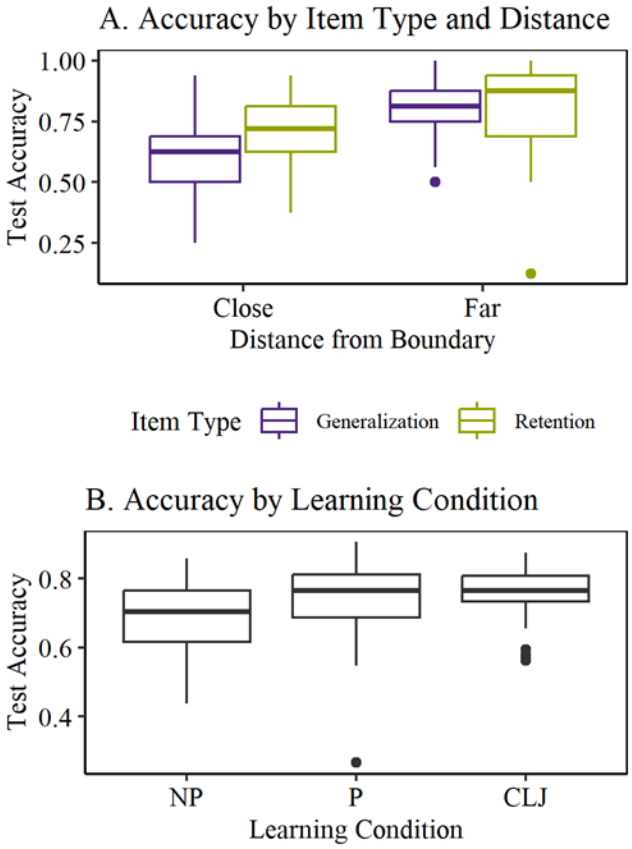

Figure 4: Test accuracy by condition.

The main effect of learning condition was significant, $F(2,69) = 5.83, p = 0.005, \eta_p^2 = 0.145$ (See Figure 4B). We followed this significant omnibus test with planned contrasts. The first compared accuracy in the Category-Level JOL condition to the remaining learning conditions. The result was significant, $t(117) = 2.32, p = 0.0218$. Participants had significantly higher accuracy in the Category-Level JOL condition ($M = 0.757, SD = 0.429$) than in the remaining conditions ($M = 0.713, SD = 0.452$). The second planned contrast compared the Pause condition

to the Category-Level JOL conditions. The result was not significant, $t(117) = 0.916, p = 0.362$. We still note that participants in the Category-Level JOL condition ($M = 0.757, SD = 0.429$) reached higher levels of accuracy than those in the Pause condition ($M = 0.736, SD = 0.441$).

Discussion

Our results provide preliminary evidence that feedback-based, procedural category learning benefits from JOL reactivity. As predicted, we found JOL reactivity for both novel and previously studied test items, building upon Lee & Ha's (2019) results which also suggested that JOLs could improve performance on novel test items. This refutes the idea that JOL reactivity is item-specific (Rivers et al., 2021). Participants in our experiment classified sinusoidal gratings, stimuli without any pre-existing semantic associations. Thus, our results refute the idea that a pre-existing, semantic cue-target relationship is a pre-requisite for JOL reactivity (Rivers et al., 2023; Witherby et al., 2023). Our predicted interaction between relatedness and JOL did not emerge. This interaction is central to the literature on JOL reactivity in paired-associates learning (Chang & Brainerd, 2023; Halamish & Undorf, 2023; Mitchum et al., 2016; Myers et al., 2020; Rivers et al., 2021, 2023; Soderstrom et al., 2015; Witherby et al., 2023). Our results suggest that the cue-strengthening hypothesis, in its current form, is not able to account for JOL reactivity in category learning.

As an alternative to semantic relatedness, we suggest that cues and targets must be connected by some form of associative learning. Such a definition would encompass semantic associations as well as the results we demonstrate here, wherein participants seem to be associating regions of perceptual space with distinct category labels. We believe the COVIS model provides a strong framework for guiding future inquiries on JOL reactivity in category learning. According to the COVIS model, procedural category boundaries similar to the one we used in this experiment are learned via stimulus-response association (Ashby et al., 1998; Ashby & Valentin, 2017). In contrast, COVIS would assume that simple single-dimensional category boundaries are learned in a declarative manner. Unlike procedural categories, a participant's ability to learn declarative categories is resilient to manipulations of motor responses (Ashby et al., 2003) and feedback (Maddox et al., 2003; Smith et al., 2014), indicating that stimulus-response association is not critical to these categories. Future work might attempt to contrast JOL reactivity in declarative and procedural category learning. If learned stimulus-response associations are strengthened by JOLs, we should expect JOL reactivity for procedural, but not declarative categories. Additionally, we believe computational models and self-report measures could provide insight into how JOLs impact participants' categorization strategies.

Although we demonstrated JOL reactivity at test, this effect does not appear at all during learning, despite JOLs being performed throughout. Perhaps the knowledge that learning has ended and testing has begun is necessary to elicit reactivity. We note that the feedback-based category learning paradigm is uniquely positioned to assess learning throughout the learning phase. This is not possible in paired-associates or observational category learning, as participants in these tasks do not perform cued recall or categorizations during learning. It is thus unclear if JOL reactivity has an impact during learning in these tasks, which is a large limitation given that paired-associates learning comprises a great deal of the JOL reactivity literature. This is a further reason to continue studying JOL reactivity using feedback-based category learning.

We note some limitations of our experiment. This category learning task was quite challenging, with a substantial number of participants failing to exceed chance-level performance. Our results also suggest a substantial role of memory in participants' learning, as they demonstrated much higher test accuracy on previously studied items compared to novel items. It is unclear how our results would look if the task were easier to learn, or more conducive to generalization. The data in this experiment were collected on smartphones. This approach to data collection is convenient and more ecologically valid than traditional lab-based learning. However, it means that there is a large amount of variance that we are unable to account for. Potential sources include viewing distances, viewing angles, screen brightness, and external distractions. We may have observed larger effect sizes had these sources of variance been controlled. It is also possible that participants' poor performance is, in part, accounted for by their low commitment to a smartphone-based task. Future work will address these limitations.

Our findings contribute to the existing JOL reactivity literature by demonstrating JOL reactivity in a feedback-based category learning task. This reactivity is seen for both retention and generalization items, indicating that the JOL benefit is not item-specific. We believe that COVIS theory provides an excellent framework for testing further predictions related to JOL reactivity. We expect that JOL reactivity may be present for procedural, but not declarative categories. We also consider the potential for other categorization models to play a role, such as prototype (Smith & Minda, 1998) and exemplar (Nosofsky, 1986) models. Though we do not discuss these models in-depth here, they may also prove to be promising subjects in future work. Regardless, a stronger theory of JOL reactivity could have substantial implications in education. Intermittent testing is beneficial to learning (Halamish & Bjork, 2011; Roediger & Karpicke, 2006) but is a source of anxiety for many students. JOLs could serve as a strong alternative to intermittent testing. Despite clear limitations, our results represent a first step toward developing a more comprehensive theory of JOL reactivity.

References

Ashby, F. G., Alfonso-Reese, L. A., Turken, A. U., & Waldron, E. M. (1998). A neuropsychological theory of multiple systems in category learning. *Psychological Review*, *105*(3), 442–481. https://doi.org/10.1037/0033-295x.105.3.442

Ashby, F. G., Ell, S. W., & Waldron, E. M. (2003). Procedural learning in perceptual categorization. *Memory & Cognition*, *31*(7), 1114–1125. https://doi.org/10.3758/bf03196132

Ashby, F. G., & Perrin, N. A. (1988). Toward a unified theory of similarity and recognition. *Psychological Review*, *95*(1), 124–150. https://doi.org/10.1037/0033-295X.95.1.124

Ashby, F. G., & Valentin, V. V. (2017). Chapter 7 - multiple systems of perceptual category learning: Theory and cognitive tests. In H. Cohen & C. Lefebvre (Eds.), *Handbook of categorization in cognitive science* (second, pp. 157–188). Elsevier. https://doi.org/10.1016/B978-0-08-101107-2.00007-5

Chang, M., & Brainerd, C. J. (2023). Changed-goal or cue-strengthening? Examining the reactivity of judgments of learning with the dual-retrieval model. *Metacognition and Learning*, *18*(1), 183–217. https://doi.org/10.1007/s11409-022-09321-y

Cruz, A., & Minda, J. P. (2024). The spacing effect in remote information-integration category learning. *Memory & Cognition*. https://doi.org/10.3758/s13421-024-01569-w

de Leeuw, J. R. (2015). jsPsych: A JavaScript library for creating behavioral experiments in a web browser. *Behavior Research Methods*, *47*(1), 1–12. https://doi.org/10.3758/s13428-014-0458-y

Do, L. A., & Thomas, A. K. (2023). The underappreciated benefits of interleaving for category learning. *Journal of Intelligence*, *11*(8). https://doi.org/10.3390/jintelligence11080153

Double, K. S., Birney, D. P., & Walker, S. A. (2018). A meta-analysis and systematic review of reactivity to judgements of learning. *Memory*, *26*(6), 741–750. https://doi.org/10.1080/09658211.2017.1404111

Epping, G. P., & Busemeyer, J. R. (2023). Using diverging predictions from classical and quantum models to dissociate between categorization systems. *Journal of Mathematical Psychology*, *112*, 102738. https://doi.org/10.1016/j.jmp.2022.102738

Folstein, J. R., Gauthier, I., Green, J. L., & Palmeri, T. J. (2009). An effect of mere exposure on visual category learning. *Journal of Vision*, *9*(8), 874–874. https://doi.org/10.1167/9.8.874

Folstein, J. R., Gauthier, I., & Palmeri, T. J. (2010). Mere exposure alters category learning of novel objects. *Frontiers in Psychology*, *1*, 40. https://doi.org/10.3389/fpsyg.2010.00040

Halamish, V., & Bjork, R. A. (2011). When does testing enhance retention? A distribution-based interpretation of retrieval as a memory modifier. *Journal of Experimental Psychology: Learning, Memory, and Cognition*, *37*(4), 801–812. https://doi.org/10.1037/a0023219

Halamish, V., & Undorf, M. (2023). Why do judgments of learning modify memory? Evidence from identical pairs and relatedness judgments. *Journal of Experimental Psychology: Learning, Memory, and Cognition*, *49*(4), 547–556. https://doi.org/10.1037/xlm0001174

Hourihan, K. L., & Tullis, J. G. (2015). When will bigger be (recalled) better? The influence of category size on JOLs depends on test format. *Memory & Cognition*, *43*(6), 910–921. https://doi.org/10.3758/s13421-015-0516-4

Jacoby, L. L., Wahlheim, C. N., & Coane, J. H. (2010). Test-enhanced learning of natural concepts: Effects on recognition memory, classification, and metacognition. *Journal of Experimental Psychology: Learning, Memory, and Cognition*, *36*(6), 1441–1451. https://doi.org/10.1037/a0020636

Kang, S. H. K., & Pashler, H. (2012). Learning painting styles: Spacing is advantageous when it promotes discriminative contrast. *Applied Cognitive Psychology*, *26*(1), 97–103. https://doi.org/10.1002/acp.1801

Kéri, S. (2003). The cognitive neuroscience of category learning. *Brain Research. Brain Research Reviews*, *43*(1), 85–109. https://doi.org/10.1016/s0165-0173(03)00204-2

Kornell, N., & Bjork, R. A. (2008). Learning concepts and categories: Is spacing the "enemy of induction"? *Psychological Science*, *19*(6), 585–592. https://doi.org/10.1111/j.1467-9280.2008.02127.x

Kubik, V., Koslowski, K., Schubert, T., & Aslan, A. (2022). Metacognitive judgments can potentiate new learning: The role of covert retrieval. *Metacognition and Learning*, *17*(3), 1057–1077. https://doi.org/10.1007/s11409-022-09307-w

Lee, H. S., & Ha, H. (2019). Metacognitive judgments of prior material facilitate the learning of new material: The forward effect of metacognitive judgments in inductive learning. *Journal of Educational Psychology*, *111*(7), 1189–1201. https://doi.org/10.1037/edu0000339

Maddox, W. T., Ashby, F. G., & Bohil, C. J. (2003). Delayed feedback effects on rule-based and information-integration category learning. *Journal of Experimental Psychology: Learning, Memory, and Cognition*, *29*(4), 650–662. https://doi.org/10.1037/0278-7393.29.4.650

Matsuki, K. (2017). *Grt: General recognition theory*. https://CRAN.R-project.org/package=grt

Metcalfe, J. (2009). Metacognitive judgments and control of study. *Current Directions in Psychological Science*, *18*(3), 159–163. https://doi.org/10.1111/j.1467-8721.2009.01628.x

Mitchum, A. L., Kelley, C. M., & Fox, M. C. (2016). When asking the question changes the ultimate answer: Metamemory judgments change memory. *Journal of Experimental Psychology: General*, *145*(2), 200–219. https://doi.org/10.1037/a0039923

Myers, S. J., Rhodes, M. G., & Hausman, H. E. (2020). Judgments of learning (JOLs) selectively improve memory depending on the type of test. *Memory & Cognition*, *48*(5), 745–758. https://doi.org/10.3758/s13421-020-01025-5

Nosofsky, R. M. (1986). Attention, similarity, and the identification-categorization relationship. *Journal of Experimental Psychology: General*, *115*(1), 39–61. https://doi.org/10.1037//0096-3445.115.1.39

R Core Team. (2021). *R: A language and environment for statistical computing*. R Foundation for Statistical Computing. https://www.R-project.org/

R Core Team. (2023). *R: A language and environment for statistical computing*. R Foundation for Statistical Computing. https://www.R-project.org/

Rivers, M. L., Dunlosky, J., Janes, J. L., Witherby, A. E., & Tauber, S. K. (2023). Judgments of learning enhance recall for category-cued but not letter-cued items. *Memory & Cognition*, *51*(7), 1547–1561. https://doi.org/10.3758/s13421-023-01417-3

Rivers, M. L., Janes, J. L., & Dunlosky, J. (2021). Investigating memory reactivity with a within-participant manipulation of judgments of learning: Support for the cue-strengthening hypothesis. *Memory*, *29*(10), 1342–1353. https://doi.org/10.1080/09658211.2021.1985143

Roark, C. L., Lescht, E., Hampton Wray, A., & Chandrasekaran, B. (2022). Auditory and visual category learning in children. *Proceedings of the Annual Meeting of the Cognitive Science Society*, *44*(44). https://escholarship.org/uc/item/3ts4j331

Roediger, H. L., & Karpicke, J. D. (2006). Test-enhanced learning: taking memory tests improves long-term retention. *Psychological Science*, *17*(3), 249–255. https://doi.org/10.1111/j.1467-9280.2006.01693.x

Seger, C. A., Braunlich, K., Wehe, H. S., & Liu, Z. (2015). Generalization in category learning: The roles of representational and decisional uncertainty. *The Journal of Neuroscience: The Official Journal of the Society for Neuroscience*, *35*(23), 8802–8812. https://doi.org/10.1523/JNEUROSCI.0654-15.2015

Shamloo, F., & Hélie, S. (2020). A study of individual differences in categorization with redundancy. *Journal of Mathematical Psychology*, *99*. https://doi.org/10.1016/j.jmp.2020.102467

Smith, J. D., Boomer, J., Zakrzewski, A. C., Roeder, J. L., Church, B. A., & Ashby, F. G. (2014). Deferred feedback sharply dissociates implicit and explicit category learning. *Psychological Science*, *25*(2), 447–457. https://doi.org/10.1177/0956797613509112

Smith, J. D., & Minda, J. P. (1998). Prototypes in the mist: The early epochs of category learning. *Journal of Experimental Psychology: Learning, Memory, and Cognition*, *24*(6), 1411–1436. https://doi.org/10.1037/0278-7393.24.6.1411

Soderstrom, N. C., Clark, C. T., Halamish, V., & Bjork, E. L. (2015). Judgments of learning as memory modifiers. *Journal of Experimental Psychology. Learning, Memory, and Cognition*, *41*(2), 553–558. https://doi.org/10.1037/a0038388

Witherby, A. E., Babineau, A. L., & Tauber, S. K. (2023). Does interactive imagery influence the reactive effect of judgments of learning on memory? *Journal of Intelligence*, *11*(7). https://doi.org/10.3390/jintelligence11070139

Yang, H., Cai, Y., Liu, Q., Zhao, X., Wang, Q., Chen, C., & Xue, G. (2015). Differential neural correlates underlie judgment of learning and subsequent memory performance. *Frontiers in Psychology*, *6*, 1699. https://doi.org/10.3389/fpsyg.2015.01699