Classification Versus Observation through Within- and Between-Category Comparison

Rachel L. Perri (rlperri@syr.edu)
Department of Psychology, Marley Education Center
Syracuse, NY 13210 USA

Michael L. Kalish (mlkalish@syr.edu)
Department of Psychology, 342 Marley Education Center
Syracuse, NY 13210 USA

Daniel Corral (dcorral@syr.edu)
Department of Psychology, 342 Marley Education Center
Syracuse, NY 13210 USA

Abstract

Inductive concept learning requires making inferences about target categories based on specific examples. Two factors which influence this process are type of learning task and the nature of the items available for comparison. However, the literature remains inconsistent on which combination of factors best facilitates concept learning. Moreover, much of the present literature focuses on artificial categories with arbitrary boundaries, leaving open the question of how best to improve learning for natural categories. We report two experiments on natural category learning, which cross learning mode (classification vs observation) with comparison type (match vs. contrast vs. control). Across both experiments, we find evidence of an observation advantage and some evidence for a contrast advantage (Experiment 1). These findings offer evidence against a classification advantage during natural category learning, which some studies have shown, and highlight the critical need for investigating the factors that impact the efficacy of classification and observation learning.

Keywords: Concept Learning; Comparison; Observation; Classification.

Introduction

Finding ways to improve concept learning is a central goal in both the cognitive and learning sciences. Traditionally, concept learning has been studied through classification. In a typical classification task, learners are presented a given stimulus and are asked to classify it. After making a classification judgment, learners are typically provided *corrective feedback*, wherein they are shown the correct answer along with whether their response was correct. This process continues iteratively until a given number of trials have been completed

Classification is posited to aid concept learning through various learning mechanisms, as it incorporates principles of (a) *hypothesis testing* (Markant & Gureckis, 2014), (b) the *generation effect* (Chechile & Soraci, 1999), (c) *retrieval practice* (Carpenter, 2009, 2011), and *comparison* (Alfieri et al., 2013). Furthermore, engaging in classification has been posited to draw attention to the diagnostic properties of the categories that are being learned (Markman & Ross, 2003; Yamuchi & Markman, 1998, 2000).

Although less commonly studied, concept learning can also be examined through *observation*. Observation paradigms are similar to classification tasks but differ in that learners are typically presented a stimulus along with the corresponding category label. As such, learners are not required to make an overt classification judgment and are instead asked to study the stimulus carefully (see Corral & Carpenter, 2024; Levering & Kurtz, 2015; Patterson & Kurtz, 2020). Previous work has hypothesized that observation aids concept learning by focusing attention on the internal structure of the category (Levering & Kurtz, 2015).

One question to consider is whether classification and observation lead to differences in concept learning. Given that both types of learning procedures provide learners with the same information, it is sensible that both would produce similar levels of learning. However, classification might involve more active learning processes (e.g., hypothesis testing, generation) than observation, and the former might thus lead to better concept learning than the latter.

Alternatively, recent work has raised the possibility that classification is more cognitively demanding than observation (Corral & Carpenter, 2024). To elaborate, for any given classification trial, learners do not know the category of the stimulus and must therefore consider hypotheses from each of the to-be-learned categories. In contrast, during observation learning, each stimulus is presented with the corresponding category label. As a result, for any observation trial, learners can restrict the hypotheses that they consider to the corresponding category. Observation might thus lead to better concept learning than classification.

Critically, the research comparing classification to observation has produced somewhat conflicting results. Some studies have found a classification advantage over observation (Ashby et al., 2002; Carvalho & Goldstone, 2015; Estes, 1994; Jacoby et al., 2010; Love, 2002; Markant & Gureckis, 2014; Ramscar et al., 2010; Steininger et al., 2022; Yang & Shanks, 2018), whereas others have found the

In D. Barner, N.R. Bramley, A. Ruggeri and C.M. Walker (Eds.), *Proceedings of the 47th Annual Conference of the Cognitive Science Society*

opposite (Levering & Kurtz, 2015; Patterson & Kurtz, 2020). Furthermore, some studies have failed to show learning differences between classification and observation (Lee & Ahn, 2018) or have shown that both produce similar levels of learning (Corral & Carpenter, 2024).

These inconsistent findings are difficult to interpret, because the corresponding studies vary on a wide array of dimensions from one another (e.g., artificial vs. natural categories (Levering & Kurtz, 2015; Yang & Shanks, 2018), featural vs. relational categories (Ashby et al., 2002; Patterson & Kurtz, 2020), abstract vs. perceptual categories (Jacoby et al., 2010 vs. Corral & Carpenter, 2024)). Nevertheless, perhaps one of the most consistent findings in this literature is that studies with naturalistic or real-world stimuli often show that classification leads to better learning than observation (e.g., Jacoby et al., 2010; Steininger et al., 2022; Yang & Shanks, 2018). Using natural categories might thus be a reasonable starting point to explore whether a consistent result can be replicated across multiple experiments.

To this end, the present paper compares classification versus observation learning with well-normed, naturalistic rocks that have been carefully calibrated (Nosofsky et al., 2017; Nosofsky et al., 2018). A core benefit of using these stimuli is that they have been extensively studied and thus offer a high degree of stimulus control.

As a secondary question, we also examine how comparison might impact classification and observation learning. Comparison is a powerful learning tool (Alfieri et al., 2013), but its impact on category learning has been somewhat understudied. Two relatively straightforward comparisons that can be made during category learning is to compare two exemplars from the same category (i.e., within-category or *match* comparison) or to compare two exemplars from different categories (i.e., between-category or *contrast* comparison).

Theorists have posited that within-category comparison can highlight the similarities between exemplars, which can facilitate abstraction of the elements that define the corresponding category (Corral et al., 2018; also see Carvalho & Goldstone, 2014). In contrast, theories on feature learning and selective attention lead to the prediction that between-category comparison aids learning by guiding attention to the diagnostic features of the to-be-learned categories (Nosofsky, 1986). Although these predictions have not been tested extensively, a series of studies used artificial categories and showed that between-category comparison led to better learning than within-category comparison (Corral et el., 2018). However, whether this effect extends to natural categories and whether it varies as a function of classification or observation are open questions.

In the present paper, we report two experiments that investigate whether learning mode (classification vs. observation) affects the learning of natural categories, and whether this outcome varies as a function of comparison type (match vs. contrast vs. control).

One prediction that follows from previous studies with natural or real-world stimuli (e.g., Jacoby et al., 2010; Steininger et al., 2022; Yang & Shanks, 2018) is that classification should lead to better learning than observation. Another prediction that follows from previous work on comparison (Corral et al., 2018) is that learning should be better for between-category comparison (contrast) than within-category comparison (match).

Another set of predictions can be derived based on the learning processes that are posited to arise from classification and observation and within- and between-category comparison. To remind the reader, both classification (Markman & Ross, 2003; Yamuchi & Markman, 1998, 2000) and between-category comparison (Corral et al., 2018) are posited to highlight the diagnostic features of the to-be-learned categories, whereas observation and within-category comparison draw attention to within-category similarity (Levering & Kurtz, 2015). One possibility is that combining classification with within-category comparison or observation with between-category comparison will better enable subjects to acquire both the diagnostic features of the to-be-learned categories, along with the feature correlations within each category. Thus, both classification with within-category comparison and observation with between-category comparison might lead to better learning than both classification with between-category comparison and observation with within-category comparison.

Experiment 1

We tested these predictions by crossing learning mode (classification vs. observation) with comparison type (match vs. contrast vs. control). During training, subjects were presented an A/B category-learning task, which consisted of learning about two natural rock categories. These stimuli were taken from Nosofosky et al. (2017).

On each training trial, subjects in the match conditions (i.e., within-category comparison; see Figure 1A) were presented two side-by-side exemplars from the same category, whereas subjects in the contrast conditions (i.e., between-category comparisons; see Figure 1B) were presented two exemplars from different categories. To test the relative learning efficacy of these comparison conditions, we also included a control condition, in which subjects were only presented one exemplar per training trial.

On each training trial, subjects in the classification conditions were asked to classify the presented exemplar(s), after which they were shown *corrective feedback*, in which the correct answer was shown along with whether the subject's response was correct. For subjects in the observation conditions, the exemplar(s) on each trial were shown with the corresponding category label and they were asked to study the exemplar(s) carefully.

To assess the rate at which learning progressed during training, category knowledge was assessed on every fifth trial through an endorsement judgment. Specifically, on every fifth training trial subjects were presented an exemplar along with a category label and were required to determine whether

this label was correct. Subjects were provided with no feedback after their endorsement was made. Performance on these endorsement trials was our main dependent measure of interest. Endorsement was used instead of classification on the test trials so that subjects in the classification conditions completed a different task at test than they did at training, thereby reducing the likelihood that they would gain an unintended benefit from classification training (for a similar approach, see Levering & Kurtz, 2015; Patterson & Kurtz, 2020). Lastly, all subjects completed a test phase endorsement, which consisted of another endorsement task. To assess memory and knowledge generalization, this task consisted of items that were repeated from training, as well as novel items.

A.

B.

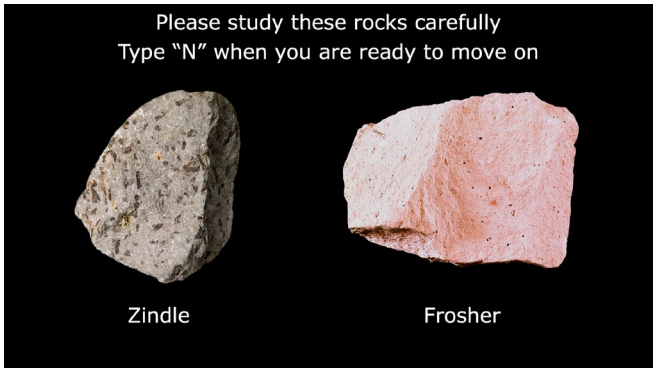

Figure 1: An example of a training trial from the match- (Panel A) and contrast- (Panel B) observation conditions.

Method

Subjects A total of 234 Syracuse University undergraduate students from an introductory psychology course participated in this study for partial fulfillment of a course requirement.

Design This study used a 2 (learning mode: classification vs. observation) × 3 (comparison: match vs. contrast vs. control) between-subjects design. Subjects were randomly assigned to one of the six conditions: classification/match (n = 37), classification/contrast (n = 40), classification/control (n = 39), observation/match (n = 39), observation/contrast (n = 41), and observation/control (n = 38).

Stimuli All stimuli were taken from Nosofsky et al. (2017) and involved naturally occurring categories of rocks. These stimuli are accessible via the OSF repository (https://osf.io/w64fv/). There were three superordinate rock categories: (a) *Igneous*, (b) *Metamorphic*, and (c) *Sedimentary*. Within each superordinate category, there were 10 subordinate rock types (e.g., *Obsidian*, *Marble*, *Conglomerate*). Each subordinate rock category contained 12 exemplars.

Procedure All instructions and stimuli were presented on a 24-inch LCD computer monitor, and all responses were entered using a computer keyboard. All stimuli were presented at the center of the screen on a black background.

At the beginning of the experiment, subjects were instructed that their job was to learn to distinguish between two types of rocks: *Froshers* and *Zindles*; made up category names were used in the place of the actual rock names, so as to reduce the chance that subjects' prior knowledge about the categories might impact their learning. For each subject, one superordinate rock category was randomly selected; and within that category, two of its subordinate categories were randomly selected. For each subject, the name of the two rock categories (i.e., Froshers and Zindles) that they were required to learn was randomly assigned.

For each subject, 10 exemplars from each of the to-be-learned categories were randomly selected for the training phase of the experiment. Two exemplars from each of these categories were thus randomly selected to not be included in the training and served as the novel items in the test phase endorsement task.

On each training trial, subjects in the match and contrast conditions were presented two side-by-side rock exemplars, which were either in the same category (match) or in different categories (contrast); the side of the screen on which each rock exemplar was presented was randomly selected on each trial. In the control conditions, only one rock exemplar was presented per trial.

On each training trail, subjects in the classification/match condition were asked to type 'X' if both rocks were Froshers or 'N' if both were Zindles; subjects in the classification/contrast condition were asked to type 'X' if the rock on the left was a Frosher and the rock on the right was a Zindle or 'N' if the rock on the left was a Zindle and the rock on the right was a Frosher; subjects in the classification/control condition were asked to type 'X' if the rock was a Frosher or 'N' if it was a Zindle.

After each classification response, corrective feedback was presented for two seconds, wherein the correct category label was shown in green text directly beneath the corresponding rock exemplar(s). If the subject responded correctly, the word 'Correct' was also presented in green text, otherwise the word 'Wrong' was presented in red text. After two seconds, the screen was cleared.

For the training trials for subjects in the observation conditions, each rock was presented with the corresponding category label, which was presented in white text directly below the corresponding rock exemplar(s). These subjects were asked to study the rock exemplar(s) carefully and were required to wait for two seconds, after which they were shown a prompt that instructed them to type a given key to move on. For subjects in the observation/match condition, if both rocks were Froshers, they were asked to type 'X'; if both rocks were Zindles they were asked to type 'N'. For subjects in the observation/contrast condition, if the rock on the left was a Frosher and the rock on the right was a Zindle, they were asked to type 'X'; if the rock on the left was a Zindle and the rock on the right was a Frosher they were asked to type 'N'. Subjects in the observation/control condition were asked to type 'X' if the rock was a Frosher or 'N' if it was a Zindle. After each response, the screen was cleared.

Every fifth trial, all subjects were given an endorsement task (the 'embedded' task), in which a stimulus was presented at the center of the screen, along with a category label that was randomly selected. Subjects were asked to type 'L' if the label was correct or 'K' if it was incorrect. The screen was then cleared and a 'Thank you' prompt was presented at the center of the screen for 500 ms.

A training cycle was 10 trials for subjects in the match and contrast conditions and 20 trials for subjects in the control conditions. All subjects completed a total of 12 training cycles, which amounted to 120 trials for subjects in the match and contrast conditions and 240 trials for subjects in the control conditions. This design thus requires *either* including a comparison manipulation where subjects view a different number of stimuli *or* complete a different number of trials. Because presenting control subjects a different number of stimuli might lead to a considerable disadvantage, we decided to hold the number of presented stimuli constant across all conditions, and instead provided control subjects a greater number of trials than subjects in the match and contrast conditions. As such, all subjects were shown the same number of stimuli during the training phase (240). However, subjects in the control condition completed twice as many training trials as subjects in the match and contrast conditions, because the former were presented one exemplar per training trial, whereas the latter were presented two exemplars on each training trial. For each subject, on every training cycle, the order in which the stimuli were presented was randomized. Furthermore, all subjects were given a self-paced rest break after each training cycle, in which they were shown (a) the number of trials that they had completed and (b) the number of trials that were left in the experiment.

After the training phase, all subjects were given a test phase endorsement task (the 'test phase'), which consisted of eight trials and was nearly identical to the endorsement task from the training phase. The test phase endorsement task consisted of two rock exemplars from each category, which were presented during training; we refer to these exemplars as *repeated* items. For each subject, the repeated items were randomly selected from the exemplars that were used during the training phase, subject to the constraint that two exemplars were selected from each category. The test phase endorsement task also consisted of two novel rock exemplars from each category, which as noted earlier, were randomly selected for each subject at the beginning of the experiment. For each subject, the order in which the stimuli were presented in the test phase endorsement task was randomized. On all trials (i.e., training, endorsement, and test phase endorsement), the intertrial interval was 400 ms.

Results and Discussion

To analyze performance on the embedded endorsement task (see Figure 2), we conducted a 2 (learning mode: classification vs. observation) × 3 (comparison type: match vs. contrast vs. control) ANOVA. The results revealed a marginal, non-significant main effect of learning mode, $F(1,228) = 2.649$, $p = .105$, $MSE = .020$, $\eta^2 = .011$, such that subjects in the observation condition ($M = .861$, $SE = .013$) performed numerically better than subjects in the classification condition ($M = .831$, $SE = .014$). The results also showed a main effect of comparison, $F(2,228) = 4.300$, $p = .015$, $MSE = .020$, $\eta^2 = .036$. To examine this main effect further, we conducted post hoc least significant difference tests, which revealed that subjects in both the contrast ($M = .867$, $SE = .016$) and control ($M = .864$, $SE = .014$) conditions outperformed subjects in the match conditions ($M = .808$, $SE = .019$; both ps $< .016$).

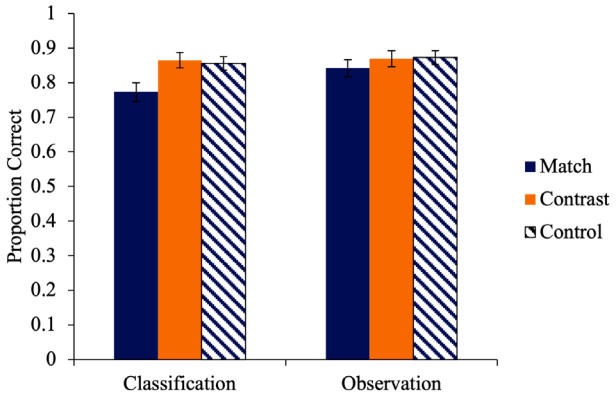

Figure 2: Average endorsement accuracy and standard errors of the mean during training for each condition in Experiment 1.

To analyze performance on the test phase endorsement task, we conducted a mixed ANOVA with learning mode (classification vs. observation) and comparison type (match vs. contrast vs. control) as between-subject factors and item (repeat vs. novel) as a within-subject factor. However, these results showed no reliable main effects and no interactions, all ps $> .422$.

Although Experiment 1 did not reveal a reliable main effect of learning mode on the embedded endorsement task, we do note that a non-significant marginal advantage of observation over classification was observed. Critically, this outcome is

in the opposite direction of what prior research that has used natural or real-world stimuli has shown (e.g., Jacoby et al., 2010; Steininger et al., 2022; Yang & Shanks, 2018), which has demonstrated a classification advantage over observation.

Furthermore, a contrast advantage was observed over the match condition on the embedded endorsement task. This finding replicates previous findings (Corral et al., 2018). However, it is important to note that the control condition also outperformed the match condition on the embedded task. This finding was somewhat unexpected, as previous work suggests that co-presented exemplars lead to better learning than presenting learners individual exemplars (Kurtz et al., 2013; Patterson & Kurtz, 2020). Moreover, no performance differences were found between the contrast and control group on the embedded task (p = .888).

One possible reason that no interaction was found between learning mode and comparison type and why no reliable performance differences were observed on the test phase endorsement task is that subjects completed too many training trials. As a result, ceiling effects might have emerged and obscured stronger learning differences from being observed.

Experiment 2

To address this possibility, we conducted a second experiment, in which we cut the training trials down from 120 to 70. Besides this difference, the design, stimuli, and procedure were identical to Experiment 1.

Method

A sample of 196 Syracuse University undergraduate students from an introductory psychology course participated in exchange for partial fulfillment of a course requirement. Subjects were randomly assigned to six conditions: classification/match (n = 33), classification/contrast (n = 33), classification/control (n = 31), observation/match (n = 32), observation/contrast (n = 35), and observation/control (n = 32).

Results and Discussion

To examine performance on embedded endorsement task (see Figure 3), we ran a 2 (learning mode: classification vs. observation) × 3 (comparison type: match vs. contrast vs. control) ANOVA. The results revealed a main effect of learning mode, $F(1,190) = 4.985$, $p = .027$, $MSE = .026$, $\eta^2 = .025$, as subjects in the observation condition ($M = .858$, $SE = .015$) outperformed subjects in the classification condition ($M = .806$, $SE = .017$). No main effect of comparison type and no interaction was observed between learning mode and classification type (both ps >.156).

Next, to analyze performance on the test phase endorsement task, we conducted a mixed ANOVA with learning mode (classification vs. observation) and comparison type (match vs. contrast vs. control) as between-subject factors and item (repeat vs. novel) as a within-subject factor. Although no main effects or interactions were statistically reliable (all ps > .061), we do note that a marginal, non-significant main effect of learning mode occurred, $F(1,190) = 3.582$, $p = .060$, $MSE = .070$, $\eta^2 = .013$. Specifically, subjects in the observation condition ($M = .865$, $SE = .017$) numerically outperformed subjects in the classification condition ($M = .814$, $SE = .020$) on the test phase endorsement task.

Taken together, the present results replicate the primary findings from Experiment 1 and offer stronger evidence for the existence of observation learning advantage over classification. However, the present findings do not replicate the contrast advantage over the match condition that was observed in Experiment 1, nor do they replicate the advantage of the control condition over the match condition. As in Experiment 1, no interaction between learning mode and comparison type was observed on either endorsement task.

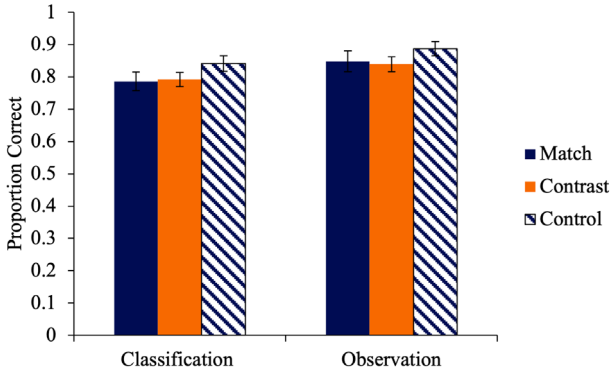

Figure 3: Average endorsement accuracy and standard errors of the mean during training for each condition in Experiment 2.

General Discussion

Across two experiments, the primary finding is that observation led to better category learning than classification, which occurred on the embedded task. Although this advantage was only marginal in Experiment 1, it was statistically reliable in Experiment 2. The results from Experiment 2 thus replicate the pattern of an observation advantage that was found in Experiment 1.

We do note, however, that only a small trace of this benefit appears to have extended to the test phase endorsement task, as no performance differences were observed between classification and observation in Experiment 1 and only a marginal advantage of observation over classification was observed in Experiment 2. One possible reason that the observation advantage appears to become weaker on the test phase endorsement task is that subjects across all conditions are able to learn the categories relatively well during training. Thus, the observation advantage that arises during training might become notably weaker by the time subjects get to the test phase endorsement task. In line with this idea, Experiment 2 consisted of 50 fewer training trials than

Experiment 1, and revealed a stronger effect on both endorsement tasks.

Critically, the embedded task allows us to directly assess learning as it occurs, whereas the test phase endorsement task occurs after subjects have learned the to-be-learned categories. For this reason, the embedded endorsement task offers a more direct assessment of concept learning than the test phase endorsement task.

We also note that a contrast advantage was observed over the match condition on the embedded task in Experiment 1. Although this effect is in line with previous work on between- versus within-category comparison (Corral et al., 2018), this effect was not observed in Experiment 2, which revealed no performance differences among the different comparison types. Moreover, in Experiment 1, subjects in the control condition outperformed subjects in the match condition on the embedded task. This latter finding is notable, as previous work has shown that co-presented exemplars generally lead to better concept learning than presenting one exemplar at a time (Kurtz et al., 2013; Patterson & Kurtz, 2020).

Given these inconsistencies, our findings on comparison type are somewhat difficult to interpret. One possibility is that with the rock stimuli used in the present set of experiments, the contrast advantage is more variable than with the stimuli that have been used in previous studies (see Corral et al., 2018). However, because the question on comparison type is secondary to our primary question about learning mode (i.e., classification vs. observation), we leave this issue for future research to investigate. We also note that we did not find any evidence of an interaction between learning and type of comparison, which suggests that although observation seemed to lead to better learning than classification, this benefit does not appear to depend on the type of comparison that learners engage in.

Critically, our primary findings that observation leads to better learning than classification of the natural rock categories used in this study contrast with previous work that has shown a classification advantage with natural or real-world stimuli (e.g., Jacoby et al., 2010; Steininger et al., 2022; Yang & Shanks, 2018). A primary motivation for the two experiments reported in this paper was that various studies have used natural or real-world stimuli to show a classification advantage. We thus aimed to use natural categories as a starting point to demonstrate findings that were consistent with previous studies that examined classification versus observation with natural or real-world stimuli. However, not only were we not able to replicate the classification advantage that previous studies have shown with natural or real-world stimuli, but we in fact found evidence against a classification advantage.

One possible explanation for these results is that natural categories are in fact better learned using observation, rather than classification. However, this possibility seems somewhat unlikely, as it suggests that prior research showing a classification advantage is incorrect (Jacoby et al., 2010; Steininger et al., 2022; Yang & Shanks, 2018). Therefore, a more likely explanation might correspond to the category structures used in the present experiments (see Nosofsky et al., 2017; Nosofsky et al., 2018). To fully investigate the implications of this idea, future research should specifically examine whether the benefits of learning through classification versus observation depend on differences in the to-be-learned category structures.

Conclusions

The present results add to a growing body of literature, wherein some studies show a classification advantage (e.g., Ashby et al., 2002; Carvalho & Goldstone, 2015; Estes, 1994), others show an observation advantage (Levering & Kurtz, 2015; Patterson & Kurtz, 2020), and others show no differences in learning (Corral & Carpenter, 2024). Critically, the present paper demonstrates that these variable findings extend to natural or real-world stimuli. What drives the variability in whether a classification or observation advantage arises thus remains an important open question. We hope that the present paper draws attention to this question and spurs further investigation into this issue.

References

Alfieri, L., Nokes-Malach, T. J., & Schunn, C. D. (2013). Learning through case comparisons: A meta-analytic review. *Educational Psychologist*, *48*(2), 87–113. https://doi.org/10.1080/00461520.2013.775712

Ashby, F. G., Maddox, W. T., & Bohil, C. J. (2002). Observational versus feedback training in rule-based and information-integration category learning. *Memory & Cognition*, *30*, 666–677. https://doi.org/10.3758/BF03196423

Carpenter, S. K. (2009). Cue strength as a moderator of the testing effect: The benefits of elaborative retrieval. *Journal of Experimental Psychology: Learning, Memory, and Cognition*, *35*(6), 1563–1569. https://doi.org/10.1037/a0017021

Carpenter, S. K. (2011). Semantic information activated during retrieval contributes to later retention: Support for the mediator effectiveness hypothesis of the testing effect. *Journal of Experimental Psychology: Learning, Memory, and Cognition*, *37*(6), 1547–1552. https://doi.org/10.1037/a0024140

Carvalho, P. F., & Goldstone, R. L. (2014). Putting category learning in order: Category structure and temporal arrangement affect the benefit of interleaved over blocked study. *Memory & Cognition*, *42*, 481–495. https://doi.org/10.3758/s13421-013-0371-0

Carvalho, P. F., Goldstone, R. L. (2015). The benefits of interleaved and blocked study: Different tasks benefit from different schedules of study. Psychonomic Bulletin & Review, *22*, 281–288. https://doi.org/10.3758/s13423-014-0676-4

Chechile, R. A., & Soraci, S. A. (1999). Evidence for a multiple-process account of the generation effect. *Memory*, *7*(4), 483–508. https://doi.org/10.1080/741944921

Corral, D., Kurtz, K. J., & Jones, M. (2018). Learning relational concepts from within-versus between-category

comparisons. *Journal of Experimental Psychology: General*, *147*(11), 1571. https://doi.org/10.1037/xge0000517

Corral, D., & Carpenter, S. K. (2024). Acquiring complex concepts through classification versus observation. *Cognitive Research: Principles and Implications*, *9*(81), 1–22. https://doi.org/10.1186/s41235-024-00608-z

Estes, W. K. (1994). Toward a statistical theory of learning. *Psychological Review*, *101*(2), 282–289. https://doi.org/10.1037/h0058559

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

Kurtz, K. J., Boukrina, O., & Gentner, D. (2013). Comparison promotes learning and transfer of relational categories. *Journal of Experimental Psychology: Learning, Memory, and Cognition*, *39*(4), 1303–1310. https://doi.org/10.1037/a0031847

Lee, H. S., & Ahn, D. (2018). Testing prepares students to learn better: The forward effect of testing in category learning. *Journal of Educational Psychology*, *110*(2), 203–217. https://doi.org/10.1037/edu0000211

Levering, K. R., & Kurtz, K. J. (2015). Observation versus classification in supervised category learning. *Memory & Cognition*, *43*, 266–282. https://doi.org/10.3758/s13421-014-0458-2

Love, B. C. (2002). Comparing supervised and unsupervised category learning. *Psychonomic Bulletin & Review*, *9*, 829–835. https://doi.org/10.3758/BF03196342

Markant, D. B., & Gureckis, T. M. (2014). Is it better to select or to receive? Learning via active and passive hypothesis testing. *Journal of Experimental Psychology: General*, *143*(1), 94–122. https://doi.org/10.1037/a0032108

Markman, A. B., & Ross, B. H. (2003). Category use and category learning. *Psychological Bulletin*, *129*(4), 592–613. https://doi.org/10.1037/0033-2909.129.4.592

Nosofsky, R. M. (1986). Attention, similarity, and the identification–categorization relationship. *Journal of Experimental Psychology: General*, *115*, 39–57. http://dx.doi.org/10.1037/0096-3445.115.1.39

Nosofsky, R. M., Sanders, C. A., Gerdom, A., Douglas, B. J., & McDaniel, M. A. (2017). On learning natural-science categories that violate the family-resemblance principle. *Psychological Science*, *28*(1), 104–114. https://doi.org/10.1177/0956797616675636

Nosofsky, R. M., Sanders, C. A., Meagher, B. J., & Douglas, B. J. (2018). Toward the development of a feature-space representation for a complex natural category domain. *Behavior Research Methods*, *50*, 530–556. https://doi.org/10.3758/s13428-017-0884-8

Patterson, J. D., & Kurtz, K. J. (2020). Comparison-based learning of relational categories (you'll never guess). *Journal of Experimental Psychology: Learning, Memory, and Cognition*, *46*(5), 851–871. https://doi.org/10.1037/xlm0000758

Ramscar, M., Yarlett, D., Dye, M., Denny, K., & Thorpe, K. (2010). The effects of feature-label-order and their implications for symbolic learning. *Cognitive Science*, *34*(6), 909–957. https://doi.org/10.1111/j.15516709.2009.01092.x

Steininger, T., Wittwer, J. & Voss, T. (2022). Classifying examples is more effective for learning relational categories than reading or generating examples. *Instructional Science*, *50*, 771–788. https://doi.org/10.1007/s11251-022-09584-7

Yamauchi, T., & Markman, A. B. (1998). Category learning by inference and classification. *Journal of Memory and Language*, *39*(1), 124–148. https://doi.org/10.1006/jmla.1998.2566

Yamauchi, T., & Markman, A. B. (2000). Inference using categories. *Journal of Experimental Psychology: Learning, Memory, and Cognition*, *26*(3), 776–795. https://doi.org/10.1037/0278-7393.26.3.776

Yang, C., & Shanks, D. R. (2018). The forward testing effect: Interim testing enhances inductive learning. *Journal of Experimental Psychology: Learning, Memory, and Cognition*, *44*(3), 485–492. https://doi.org/10.1037/xlm0000449