# The Role of Category Confusability in Classification Versus Observation Learning

**Rachel L. Perri, Yu-Wei Chang, Michael L. Kalish & Daniel Corral (dcorral@syr.edu)**
Department of Psychology, Syracuse University

**Abstract**

Concept learning can be studied through training that involves either classification or observation. In classification, learners actively make a judgment about category membership of presented stimuli and receive feedback; in observation, learners study a labeled stimulus. Extant work shows contradictory findings, as some studies show a benefit of classification over observation, whereas others show the opposite pattern. Recent work posits that these findings might be explained by category confusability. We present an experiment that investigates this question by directly manipulating category confusability (within-subjects), wherein subjects either learned through classification or observation. All subjects learned three different pairs of natural categories that shared low, moderate, or high feature similarity. Subjects were presented a single stimulus on each trial; every fourth trial, they made an endorsement judgment. Our findings suggest that training both with (classification) and without (observation) error are equally beneficial for learning natural categories, challenging several prominent category learning models.

**Keywords:** concept learning; categorization; classification training; observation training; similarity

## Introduction

A shared goal to both the cognitive and learning sciences is finding ways to improve concept acquisition. While various tasks can encourage concept formation, researchers usually study this topic using classification paradigms. Classification tasks present learners with an *unlabeled* stimulus and ask them to select the correct category label. After making a classification judgment, learners receive some type of feedback. But another, though less commonly studied, paradigm for concept learning is the observation task. Unlike classification, observation tasks present a *labeled* stimulus and ask learners to study that stimulus carefully.

Despite both tasks presenting identical information about the stimuli, research investigating the differences between these two learning modes shows distinctly varied conclusions in their effectiveness. For instance, several studies show that learners using classification are advantaged over learners using observation (Ashby et al., 2002; Carvalho & Goldstone, 2015; Edmunds et al., 2015; Estes, 1994; Love, 2002; Markant & Gureckis, 2014). Accordingly, researchers have proposed different mechanisms suggested to aid concept learning through classification, such as *hypothesis testing by selection* (Markant & Gureckis, 2014); a retention advantage associated with *generation* (Chechile & Soraci, 1999) or *retrieval practice* (Carpenter, 2009, 2011); and *selective attention* processes that draw attention to the diagnostic properties of a category (Carvalho & Goldstone, 2015; Markman & Ross, 2003).

Conversely, other studies suggest that observation leads to better learning than classification (Levering & Kurtz, 2015; Patterson & Kurtz, 2020; Perri et al., 2025, 2026). Ergo, there are suggested mechanisms to explain instances in which observation is preferable to classification. For example, some work raises the possibility that, during training, observation may be less cognitively demanding than classification (Perri et al., 2025, 2026); and subsequently, less prone to proactive interference (Corral & Carpenter, 2024).

Furthermore, other studies show evidence for similar levels of learning between classification and observation (Corral & Carpenter, 2024; Lee & Ahn, 2018).

Recent work by Perri et al. (2026) has shown an observation advantage over classification using natural rock categories. Critically, this work also shows that *confusability* was highly predictive of performance, such that highly confusable categories were better learned through observation than classification. These findings raise the possibility that the differential benefits of classification or observation might depend on the degree of exemplar similarity between or among the to-be-learned categories. While the aforementioned work controls for confusability statistically, it does not directly manipulate it. Therefore, it remains directly unexamined whether training-mode advantages in memory and generalization depend upon category confusability.

To this end, the present study tests whether the effect of learning mode varies as a function of category confusability by systematically varying (within-subjects) the confusability (*low*, *moderate*, or *high)* of the to-be-learned rock categories.

Manipulating confusability effectively varies the within- and between-category similarities. In our study, categories that are low in confusability have a greater intervening distance between them in psychological space, and thus, are easier to distinguish. Categories that are highly confusable have a smaller intervening distance, and thus, are harder to distinguish. Finally, categories that are moderately confusable sit at the midpoint between the distances for low and high confusability.

There are several predictions that can be derived from prior research. A classification task requires learners to switch their attention between hypotheses and feedback, whereas the observation task does not require this type of switching and might therefore be less cognitively demanding (Corral & Carpenter, 2024). As a result, learners using observation may be better equipped to spread their attention across dimensions of the category structure (Levering & Kurtz, 2015). With attention on multiple dimensions of a category, observation learners may be more attuned to within-category correlations (Corral & Carpenter, 2024), in turn developing a broader understanding of the category as a whole (Levering & Kurtz,



In G. Boleda, I. Dautriche, M. Maldonado, C. Saldana and T. Torrent (Eds.), *Proceedings of the 48th Annual Conference of the Cognitive Science Society* 

2015). Thus, highly confusable categories (i.e., categories with low between-category distances) might be easier to learn with observation than classification. Additionally, classification, which draws attention to diagnostic dimensions (Markman & Ross, 2003), may be better for learning categories that are low in confusability. Such categories have high between-category distances, and unlike highly confusable categories, can be separated along a single dimension (Nosofsky et al., 2017). In this case, classification may act to highlight critical dimensions for category membership.

## Experiment

We tested these predictions by crossing learning mode (classification vs. observation) with confusability (low vs. moderate vs. high). During training, subjects were presented with three sequential A/B category-learning tasks, each of which consisted of learning about two different natural rock categories.

On each training trial, subjects in the classification condition were asked to classify the presented exemplar, 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 condition, the exemplar on each trial was shown with the corresponding category label and they were asked to study the exemplar carefully.

To assess the rate at which learning progressed during training, category knowledge was assessed on every fourth trial through an endorsement judgment. Specifically, on every fourth 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 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 assessment trials to reduce the likelihood that subjects in the classification condition would gain an unintended advantage by being tested in the same format as their training (for a similar approach, see Levering & Kurtz, 2015; Patterson & Kurtz, 2020; Perri et al., 2025, 2026). 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.

### Method

**Subjects** A total of 156 participants were recruited from Prolific (https://www.prolific.com) in return for monetary payment. Participants were paid at a rate of $9 an hour.

**Design** This study used a 2 (learning mode: classification vs. observation) × 3 (confusability: low vs. moderate vs. high) mixed-subjects design. Learning mode was a between-subjects factor, to which subjects were randomly assigned to one of two conditions: classification ($n = 70$), and observation ($n = 86$). Confusability was a within-subjects factor, to which all subjects completed a low, moderate, and highly confusable categorization task.

**Stimuli** All stimuli were taken from Nosofsky et al. (2017), accessible via the OSF repository (https://osf.io/w64fv/). The stimuli used are naturally occurring categories of rocks. There were three superordinate rock types: (a) *Igneous*, (b) *Metamorphic*, and (c) *Sedimentary*. Within each superordinate type, there were 10 subordinate rock categories (e.g., (a) *Obsidian*, (b) *Marble*, (c) *Conglomerate*) with 12 exemplars each. Each exemplar can be represented along a combination of eight featural dimensions (Nosofsky et al., 2017). We selected a subset of subordinate categories from each superordinate type. Specifically, for each superordinate category, one anchor category and three additional discrimination categories that varied in their confusability to the anchor, were chosen (see Table 1 for each possible subordinate category pairing within each superordinate category).

First, we calculate the Euclidean distance among all pairs of exemplars, across all eight dimensions, within each subordinate category (i.e., *within-category similarity*). After averaging the pairwise distances, we get one mean within-category distance value for each subordinate category. The formula for confusability handles pairs of categories. Thus, to define average within-category distance for a pair of categories, we take the mean of their respective within-category distances. Then, we calculate the Euclidean distance between every exemplar from one category to every exemplar from another category, across all eight dimensions, and average across all pairs (i.e., *between-category similarity*). Finally, we use the average within-category distance and between-category distance to calculate a confusability ratio for each pair of subordinate categories.

That is, to determine the confusability of a pair of to-be-learned categories, we divided how similar one category was to another (i.e., between-category similarity) by how similar all items from one category were (i.e., within-category similarity). Therefore, we define confusability as an inverse similarity function—that is, in terms of the ratio of between-category distance to average within-category distance—such that higher ratios reflect less confusable categories.

In the present study, *Igneous*, *Metamorphic*, and *Sedimentary* were randomly assigned to the three confusability levels for each participant. In each category-learning task, they learn to distinguish two subordinate rock categories from the same superordinate type. As a result, we compute the confusability of rock pairs only from the same superordinate type. We plotted these ratios with the subordinate category on the X axis, and ratio value on the Y axis (see Figure 1). From these ratios, we selected one

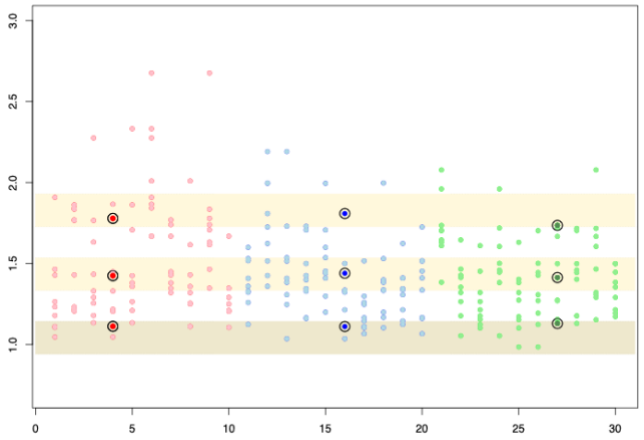


Figure 1: Plot of confusability ratios ordered within superordinate type. The X-axis has the subordinate rock category numbers ($x < 11$ = *Igneous* rocks; $10 < x < 21$ = *Metamorphic* rocks; $x > 20$ = *Sedimentary* rocks), and the Y-axis has the confusability ratio. Lower ratios indicate more confusable pairs. The outlined data points represent the low, moderate, and highly confusable categories for the subset.

subordinate rock category as the anchor [1] in each superordinate rock type. A rock category qualifies as an anchor if it has at least one category pairing within each of the defined threshold for low, moderate, and high confusability ratios. Ratios from all pairs were fit to a normal distribution (outliers > 3 SD removed). Using the fitted mean (μ) and standard deviation (σ), the following thresholds were defined based on ±1.2 SD and bin widths of 0.5 SD: (Low Confusability) μ + 0.95σ, μ + 1.45σ, (Moderate Confusability) μ - 0.25σ, μ + 0.25σ, (High Confusability) μ - 1.45σ, μ + 0.95σ. Among possible candidates, we picked the best set of three anchors that has the most similar ratios in all three thresholds, as shown in Figure 1. Low confusability is reflected by higher ratios and indicates that a given subordinate rock category has a larger psychological distance between itself and the anchor rock type than other pairings within that superordinate type. Effectively, this distance makes such categories more distinct in psychological space, therefore more difficult to confuse. Moderate confusability is reflected by ratios at the midpoint and indicates that a subordinate rock category is an average distance to the anchor than other pairings within that superordinate type. Finally, high confusability is reflected by lower ratios and indicates that a subordinate rock category has a smaller psychological distance between itself and the anchor rock type—these pairings are less distinct in psychological space.

**Procedure** All instructions and stimuli were presented on a computer monitor, and all responses were entered using a computer keyboard. All stimuli were presented sequentially 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 three different pairs of rocks. Each participant learned about one confusability pairing from each superordinate type (e.g., *anchor Metamorphic & discrimination Metamorphic)*.

The confusability pairings were randomly sampled, subject to the constraint that one of the pairings was low in confusability, one pairing was moderately confusable, and one pairing was highly confusable. For example, if a low confusability pairing is selected from the Igneous rocks, that level of confusability is subsequently disqualified from further selections. For each subject, the pairings were randomly ordered.

Table 1: Subordinate Categories and Confusability Levels.

| **Rock Category** | **Anchor** | **Low** | **Moderate** | **High** |
|---|---|---|---|---|
| Igneous | Gabbro | Pumice | Granite | Peridotite |
| Metamorphic | Migmatite | Anthracite | Marble | Phyllite |
| Sedimentary | Rock Gypsum | Bituminous Coal | Breccia | Rock Salt |

With 12 exemplars per category, we randomly selected eight of these from each category to use for training. The other four exemplars were set aside to use for the test endorsement task. For the selected training exemplars, we also ensured that the subset's *within-category similarity* was not far from the whole set's within-category similarity. Specifically, exemplars were resampled until the subset's within-category similarity fell within ±10% of the within-category similarity of the complete category.

The training exemplars were randomly presented one at a time in the center of the screen, blocked by confusability. For each subject, the name of the six rock categories that they were required to learn was randomly assigned *Category A* through *Category F*. The rock names were made to be arbitrary to reduce the effect of prior knowledge on category learning and were randomly selected for pairing with a given category for each subject.

On each training trial, subjects were presented with one rock from the corresponding superordinate category (e.g., an *anchor rock* from the *Sedimentary type*). In the classification condition, subjects were asked to select from two options, which always corresponded to the labels for the to-be-learned categories in a given block of trials. Subjects were asked to type a certain letter for each category. The keypress for each

[1] Although one rock category was designated as an anchor within each of the three superordinate types, this assignment served only to control for variation across participants. Within any given participant, only two rock categories were distinguished within each superordinate type; therefore, the specific choice of anchor category should not affect task performance.

category always corresponded with its category label (e.g., *Press A if this is Category A, Press C if this is Category C)*

After each response, subjects received corrective feedback for two seconds, which showed the correct category label in green text directly below the exemplar. If the subject responded correctly, the word 'Correct' was also presented in green text, otherwise 'Wrong' was presented in red text.

In the observation condition, each rock was presented with the corresponding category label in white text below the presented exemplar. These subjects were instructed to study the exemplar carefully. After two seconds, subjects were shown a prompt, centered at the top of the screen, that instructed them to type a given letter corresponding to the rock category to move on (with the same key options as in the classification condition). For all subjects, after each response on the training trials, the screen was cleared for 400ms.

A training phase was 48 trials, and all subjects completed two phases for a total of 96 training trials. To be clear, during each phase, subjects learned about all six categories. Within each phase, the stimuli were presented quasi-randomly, specifically, blocked by confusability. Furthermore, all subjects were given a self-paced rest break after every 16 training trials, in which they were shown (a) the number of trials that they completed and (b) the number of trials that were left in the experiment.

During training, an endorsement task was given to all subjects on every fourth trial (the *embedded endorsement* task). During endorsement, a training stimulus was presented at the center of the screen, along with a randomly selected category label. Subjects were asked to type 'Y' if the label was correct or 'N' if it was incorrect. After, the screen was cleared and a 'Thank you' prompt was presented at the center of the screen for 500 ms.

After the training phase, all subjects received a test phase endorsement task, which consisted of 48 trials and was nearly identical to the endorsement task from the training phase. The test phase endorsement task consisted of rock exemplars from each category that were presented during training (repeated items) and novel rocks from the same category. There were thus 24 repeated exemplars and 24 novel exemplars. For each subject, the repeated items were randomly selected from the exemplars that were used during training, and as noted earlier, novel items were randomly selected for each subject at the beginning of the experiment. Subjects were asked to type 'S' if the label was correct or 'K' if it was incorrect. For each subject, the order in which the stimuli were presented in the test phase endorsement task was randomized.

## Results

### Embedded Endorsement

To analyze performance on the embedded endorsement task, we conducted a 2 (learning mode: classification vs. observation) × 3 (confusability: low vs. moderate vs. high) mixed ANOVA. The results revealed a significant main effect of confusability, such that endorsement accuracy varied across the levels of confusability. Specifically, subjects performed better on low confusability categories ($M$ = 0.843, $SE$ = 0.015) than moderately confusable categories ($M$ = 0.775, $SE$ = 0.015), $p$ = .002, and highly confusable categories ($M$ = 0.627, $SE$ = 0.017), all $p$s < 0.001. Additionally, subjects performed better on moderately confusable categories than highly confusable categories, $p$ < 0.001. There was no other significant main effect or interaction (all $p$s > 0.05).

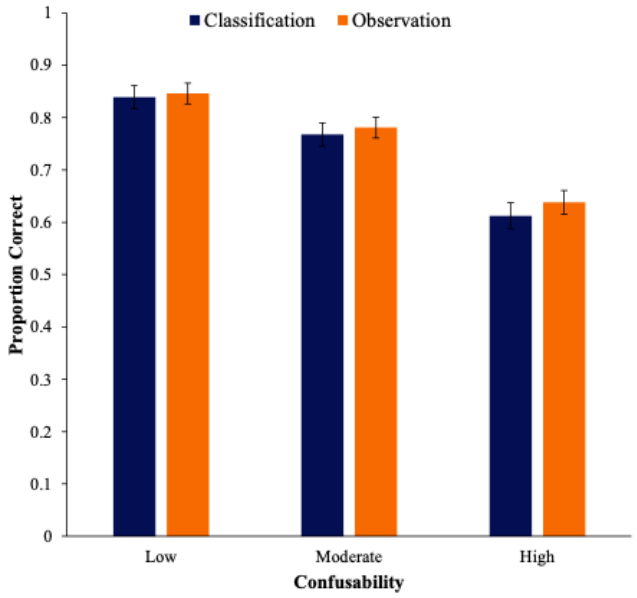


Figure 2: Proportion correct on the embedded endorsement task by learning mode within each level of confusability.

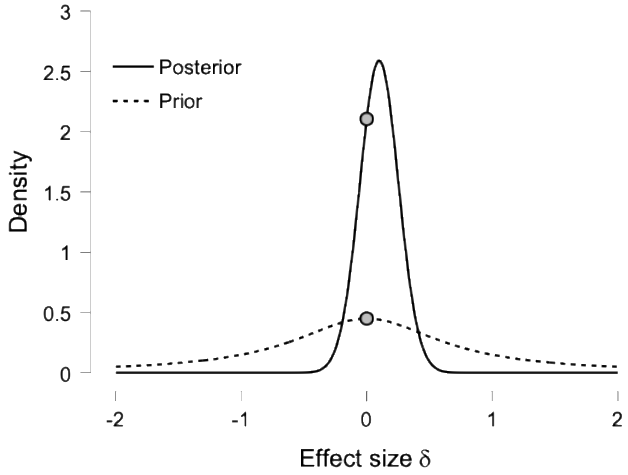


Figure 3: The Bayes Factor results on the embedded endorsement task with prior distribution (dashed line) and posterior distribution (solid line) for effect size (δ). The evidence supports the null hypothesis.

The following tests showed that participants in all conditions performed well above chance on the embedded endorsement task (all *p*s < .001). Furthermore, a one-tailed repeated measures *t*-test indicated that performance was significantly better on the second training block ($M = 0.765$, $SE = 0.014$) than on the first training block ($M = 0.731$, $SE = 0.012$), $t(310) = 1.820$, $p = .035$, which shows evidence of learning between training blocks.

**Benefit of Learning Modes** Taken together, these findings suggest that both classification and observation led to similar levels of concept learning. To test this conclusion more directly, we ran a two-tailed Bayesian Independent *t*-test on JASP version 0.95.3 (JASP Team, 2025) with a noninformative, unbiased Cauchy prior $r = .707$. We observed a Bayes Factor ($BF_{0a}$) of 4.68 indicating that the alternative hypothesis is 4.68 times less likely to the null hypothesis (see Figure 3). The evidence from the present study is against the advantage of observation during training.

## Test Endorsement

Next, to analyze performance on the test phase endorsement task, we conducted a mixed ANOVA with learning mode (classification vs. observation) as a between-subject factor and confusability (low vs. moderate vs. high) as a within-subject factor. Due to a coding error, we could not reliably identify novel items on the test phase, so we collapsed across these items.

Again, the results revealed a significant main effect of confusability, such that endorsement accuracy varied across each level of confusability. Specifically, subjects performed better on low confusability categories ($M = 0.752$, $SE = 0.015$) than moderately confusing categories ($M = 0.674$, $SE = 0.015$), and highly confusable categories ($M = 0.596$, $SE = 0.012$), both *p*s < .001. Additionally, subjects performed better on moderately confusable categories than highly confusable categories, $p < .001$. There was no other significant main effect or interaction (all *p*s > 0.05). Tests again showed that subjects in all conditions performed well above chance in the test phase, all *p*s < .001.

**Benefit of Learning Modes** To further test the effect of learning modes on performance, we ran a two-tailed Bayesian Independent *t*-test. With the same noninformative prior, the result showed a $BF_{01}$ of 5.63 indicating that the Bayes Factor is 5.63 times in favor of the null hypothesis (see Figure 5). This provides suggestive evidence for the claim that there is no meaningful difference in the performance between classification and observation. Taken together, our data showed that participants in both learning modes learned all the categories but did so equally well.

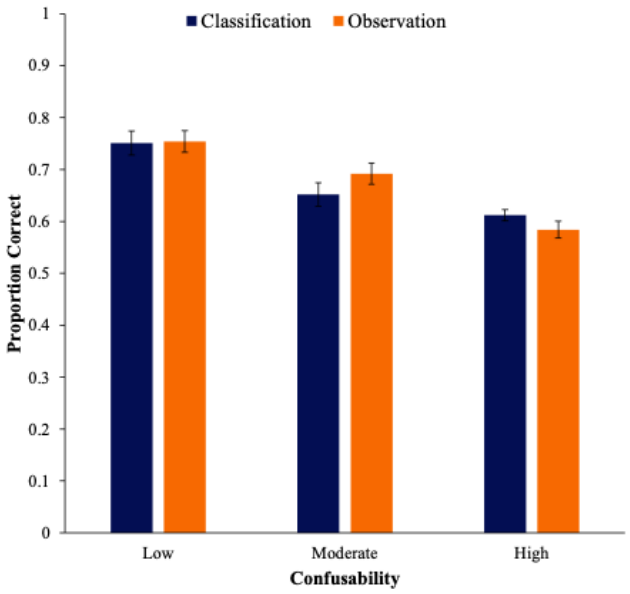


Figure 4: Proportion correct on the test endorsement task by learning mode within each level of confusability.

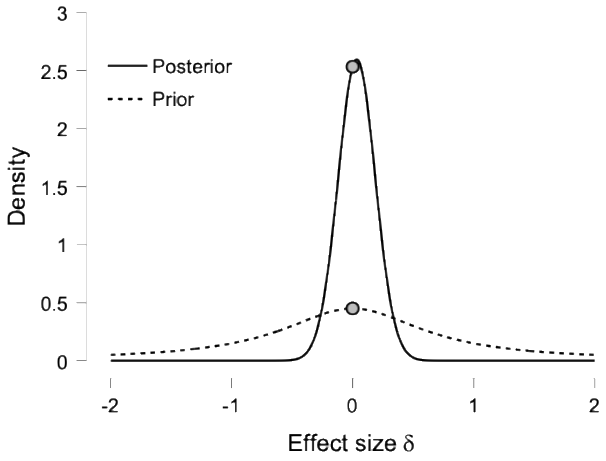


Figure 5: The Bayes Factor results on the test endorsement task with prior distribution (dashed line) and posterior distribution (solid line) for effect size (δ). The evidence supports the null hypothesis.

## Discussion

The results of this experiment suggest that both observation and classification learners experience similar levels of learning. While these findings do not support the predictions that follow from recent work (Perri et al., 2026), they suggest that both observation and classification may be equally beneficial for natural category learning. That is, both groups, across all levels of confusability perform well above chance on both the embedded and posttest endorsement tasks, which is an indicator that both groups learned all of the categories.

Likewise, there was a main effect of confusability such that the least confusing categories were learned better than moderately confusing categories, which in turn were both learned better than the most confusable categories. Such an effect is indicative of learning, as this difference between groups could not arise otherwise. Still, performance is above chance at each level of confusability. Subjects also performed better on the second training phase than on the first, which shows direct evidence of learning. When taken together, the present findings show that classification and observation led to equal levels of learning (for similar findings, see Corral & Carpenter, 2024).

As shown in the present study, participants showed very similar levels of performance when trained with (classification) or without (observation) an observable error. This is challenging for theories of category learning which rely on error signals to drive changes in the learner (e.g., ALCOVE; Kruschke, 1992, SUSTAIN; Love et al., 2004). Such theories are generally mute about the way covert responses differ from overt ones during learning; and other than ad-hoc modifications to learning rates, such models have no theoretical way to account for the presence or absence of a classification advantage.

A possible explanation for the indistinguishable performance between learning modes may be related to how task difficulty was operationalized. In the present study, task difficulty was defined in terms of confusability, which is calculated as the ratio of the within-category similarity to between-category similarity. By representing rocks in a psychological space, confusability provides an integrated measure across all feature dimensions. However, because the stimuli were in natural categories, the features of the rocks may not have been equally salient, and therefore, may not have been equally weighted during learning. Under these circumstances, confusability may be more accurately estimated using weighted similarities rather than assuming equal contribution across dimensions.

One possible resolution follows from the results of Edmunds, et al. (2015), who showed a consistent classification advantage. They speculated that the dimensionality of the discrimination determined the value of classification learning. Our stimuli differ from theirs by being both natural and high dimensional. If our manipulation of discriminability was also a manipulation of dimensionality, our null result would be problematic for their theory, but work is needed to determine if our learners attended to different numbers of dimensions during learning before such a conclusion can be established.

Future studies could begin by identifying optimal dimension weights, examine whether the number of relevant dimensions varies as a function of confusability, and compare these effects between natural and artificial categories.

## Conclusions

The results of this study add to a library of work showing mixed effects of training mode on concept learning, with some studies reporting a classification advantage (e.g., Ashby et al., 2002; Carvalho & Goldstone, 2015), others reporting an observation advantage (Levering & Kurtz, 2015; Perri et al., 2026), and still others finding no reliable differences between the two (Corral & Carpenter, 2024). Importantly, the current findings demonstrate that both training modes are equally beneficial for concept learning. Despite direct manipulation of category confusability, classification and observation produced comparable learning outcomes, well above chance performance. This suggest that neither training with nor without error is uniformly superior. Thus, identifying the conditions under which a classification or observation advantage emerges remains a central open question for theories of concept learning, and warrant further empirical and computational investigation

## References


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. (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., & 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

Edmunds, C. E. R., Milton, F., & Wills, A. J. (2015). Feedback can be superior to observational training for both rule-based and information-integration category structures. *The Quarterly Journal of Experimental Psychology*, *68*(6), 1203–1222. https://doi.org/10.1080/17470218.2014.978875

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

JASP Team (2025). JASP (Version 0.95.4) [Computer software].

Kruschke, J. K. (2020). ALCOVE: An exemplar-based connectionist model of category learning. In *Connectionist Psychology* (pp. 107–138). Psychology Press.

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

Love, B. C., Medin, D. L., & Gureckis, T. M. (2004). SUSTAIN: A network model of category learning. *Psychological Review*, *111*(2), 309. https://doi.org/10.1037/0033-295X.111.2.309

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., 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

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

Perri, R. L., Kalish, M., & Corral, D. (2025). Classification versus observation through within- and between-category comparison. *Proceedings of the Annual Meeting of the Cognitive Science Society*, *47*. https://escholarship.org/uc/item/04j2p27n.

Perri, R. L., Kalish, M., & Corral, D. (2026). *The role of comparison in observation versus classification learning.* Manuscript in preparation.