UC Merced
Proceedings of the Annual Meeting of the Cognitive Science Society

Title
The Role of Comparison and Category Confusability on Concept Acquisition

Permalink
https://escholarship.org/uc/item/3sj0v87q

Journal
Proceedings of the Annual Meeting of the Cognitive Science Society, 48(0)

Authors
Liao, Yongfu
Corral, Daniel

Publication Date
2026

Copyright Information
This work is made available under the terms of a Creative Commons Attribution License, available at https://creativecommons.org/licenses/by/4.0/

Peer reviewed

eScholarship.org Powered by the California Digital Library
University of California

The Role of Comparison and Category Confusability on Concept Acquisition

Yongfu Liao (yliao24@syr.edu) & Daniel Corral (dcorral@syr.edu)
Department of Psychology, Syracuse University

Abstract

Comparison can support category learning by highlighting shared similarities or diagnostic differences between categories. We examined how comparison type (match vs. contrast vs. single-item control) influences featural category learning and whether these effects depend on category confusability (low vs. high). On training trials, the match condition was shown co-presented exemplars from the same category, whereas those in the contrast condition were shown co-presented exemplars from different categories; the control condition was shown individual exemplars. Participants learned artificial categories in a mixed design with comparison type manipulated between subjects and confusability within subjects. Learning was assessed via training performance, embedded endorsement, and posttest endorsement measures. Across all measures, contrast outperformed match, whereas control performed comparably to contrast. Notably, even with fewer exposures, single-item presentation still outperformed match presentation, suggesting that co-presenting exemplars does not always lead to better learning.

Keywords: Concept Learning; Comparison; Confusability; Classification.

Introduction

Improving concept acquisition is a central goal in the cognitive and learning sciences. Comparison has been identified as a particularly powerful learning tool that can support this process (Alfieri et al., 2013). Indeed, comparison has been shown to benefit comprehension, category formation, and concept abstraction (Gentner & Namy 1999; Kotovsky & Gentner, 1996; Namy & Gentner, 2002). Through comparison, learners are provided with the opportunity to align multiple exemplars, enabling them to abstract commonalities and distinguish relevant from irrelevant variation. Comparison can be made between exemplars from the same category (i.e., within-category or *match* comparison) or between exemplars from different categories (i.e., between-category or *contrast* comparison). As an illustrative example, suppose an art instructor wants to help students learn about two painting styles—*impressionist* and *naturalist*. Adopting a within-category comparison strategy would involve presenting examples of impressionist paintings together, with the goal of helping students identify the commonalities among the examples that characterize impressionist paintings. An alternative strategy is to present impressionist paintings together with naturalist paintings, such that distinctions between the two styles may become more apparent through between-category comparison. Together, these forms of comparison provide potentially separate pathways to concept learning by supporting generalization within categories and differentiation between categories (see Corral et al., 2018).

The classification learning paradigm has long been employed to study category learning. In this paradigm, participants are presented a single stimulus and are asked to make a classification judgment on each trial. After responding, participants are given *corrective feedback*, in which they are shown the category label of the stimulus, along with whether their response was correct. Depending on the nature of the stimuli and the category structure, category learning under such a paradigm usually involves training over hundreds or thousands of trials (e.g., Corral & Jones, 2012, 2014; Goldstone, 1994; McKinley & Nosofsky, 1995; Smith & Minda, 1998). Classification learning thus presents learners many exemplars of the to-be-learned categories and offers them the opportunity to utilize comparison between trials.

To further investigate the role of comparison in category learning, recent work has extended the classification learning paradigm to incorporate co-presented exemplars (Corral et al., 2018; Kurtz et al., 2013; Patterson & Kurtz, 2020; Perri et al., 2025, 2026). In this extended paradigm, the co-presented exemplars on each trial are either from the same category (i.e., within-category or *match* comparison) or from different categories (between-category or *contrast* comparison). This paradigm thus naturally leads to the question of whether match or contrast comparison is superior for category learning.

According to theories of feature-based category learning and selective attention (Kruschke, 1992; Mackintosh, 1975; Nosofsky, 1986), between-category comparison draws attention to the diagnostic differences between categories. Continuing with the previous example, although impressionist and naturalist paintings both often depict natural scenes, making them difficult to distinguish, comparing the two side by side should draw attention to differences in brushstroke texture, with impressionist works favoring loose, visible strokes and naturalist works favoring detailed, smooth, and realistic depiction.

Other theoreticians posit that comparing exemplars from the same category (within-category comparison) highlights their shared similarities (Carvalho & Goldstone, 2014), which can help learners abstract the elements that make up the corresponding category (Corral et al., 2018). For example, comparing two impressionist paintings can highlight that both consist of loose, visible strokes.

The question of whether between- versus within-category comparison leads to differential concept acquisition is somewhat understudied. Nevertheless, there are some extant studies that are conceptually related to this question. For instance, Stewart et al. (2002) presented participants with individual exemplars on each trial, where the subsequent

In G. Boleda, I. Dautriche, M. Maldonado, C. Saldana and T. Torrent (Eds.), *Proceedings of the 48th Annual Conference of the Cognitive Science Society* ©2026 the author(s). This work is licensed under a Creative Commons Attribution 4.0 International License (CC BY).

exemplar was either a distant member of the same category as the previous exemplar or a distant member of a different category. Stewart et al. found an effect that they termed the "*category contrast effect*," whereby classification accuracy for a borderline stimulus—an exemplar that is on the border between two categories—was higher when it was preceded by a distant member of a different category than by a distant member of the same category. This finding suggests that when there is a category switch between two trials—conceptually similar to a between-category comparison with two co-presented items—learning is better, compared to when there is no category switch between two trials, which is similar to a within-category comparison. Therefore, the category contrast effect is in line with, if not directly supportive of, the prediction of a contrast advantage in category learning.

In a conceptually related study, Carvalho and Goldstone (2014) directly manipulated the probability of category switches between trials in a classification learning task. In their *blocked condition*, the probability of a category switch between trials was low (25%), while in their *interleaved condition*, this probability was high (75%). Thus, if learners do compare items across trials, both the blocked and interleaved conditions involve between- and within-category comparisons, with the key difference being that the interleaved condition offers more opportunities for between-category comparison, whereas the blocked condition offers more opportunities for within-category comparison. Results from Carvalho and Goldstone (2014) showed that categories were better learned and generalized when the exemplars between the to-be-learned categories were highly confusable (i.e., highly similar). In contrast, for categories that were low in confusability, Carvalho and Goldstone report a blocked advantage. These results might therefore be taken to support a contrast advantage in category learning for highly similar, confusable exemplars, but a match advantage when the exemplars between categories are low in confusability.

Although these findings are suggestive, it is important to note that the studies by Stewart et al. (2002) and Carvalho and Goldstone (2014) involved presenting individual exemplars on each trial and thus enabled between-trial comparison. However, this type of comparison differs from within-trial comparison, which occurs when two or more exemplars are presented on the same trial. Although it is feasible that the same learning processes underlie within- and between-trial comparison, this remains an open question.

Based on the aforementioned theories and findings, one might expect a contrast advantage during classification learning when exemplars are highly confusable and a match advantage when confusability is low between categories. The literature on co-presented exemplars, however, has shown somewhat mixed findings. Using a classification learning task with two co-presented exemplars, Corral et al. (2018) compared learning performance when the exemplars were from the same category (match condition) versus different categories (contrast condition). Their results showed a contrast advantage over match for both featural (Experiment 1) and relational categories. A similar classification paradigm with two co-presented exemplars was used in Patterson and Kurtz (2020) to investigate the role of comparison in learning relational categories. However, they report no difference in learning performance between the contrast and match conditions in their Experiment 2. More recent work used natural rock stimuli to investigate the effects of between- versus within-category comparison on feature-based category learning (Perri et al., 2025, 2026). Although between-category comparison led to better learning than within-category comparison, this advantage did not replicate consistently across experiments. However, it is critical to note that these studies did not directly manipulate confusability, which previous work suggests moderating the effect of comparison (Carvalho & Goldstone, 2014).

Highly confusable categories are perceptually similar and may therefore require greater attention to subtle diagnostic differences to be successfully distinguished. C*ontrast* (i.e., between-category) comparison may facilitate this process by highlighting those differences. On the other hand, low confusable categories are perceptually dissimilar and may require greater attention to subtle commonalities across items to successfully learn their corresponding category structures. *Match* (i.e., within-category) comparison may facilitate this process by highlighting those commonalities.

As such, the effect of comparison might vary as a function of confusability. Specifically, a contrast advantage over match should occur when the to-be-learned categories are highly confusable, whereas a match advantage should occur when these categories are low in confusability. To test these predictions, we report an experiment that incorporates co-presented exemplars that are either from the same (match) category or different (contrast) categories; to assess the extent to which co-presented exemplars impact learning, we also include a control condition where participants were only presented individual exemplars on each trial. We thus examine whether comparison type (match vs. contrast vs. control) affects the learning of artificial, featural categories and whether this effect depends on confusability (low vs. high).

Experiment

The present experiment crossed comparison type (match vs. contrast vs. control, between-subjects) with confusability (low vs. high, within-subjects). During training, an A/B category-learning task was used where participants learned to classify artificial stimuli into two categories. Two sets of stimuli were created (see Figure 1), and for each set, the stimuli were partitioned to create low and high confusability categories.

On each training trial, participants in the match condition were presented with a pair of exemplars side-by-side from the

same category, whereas participants in the contrast condition were presented with two exemplars from different categories. A control condition with single-item presentation was also included to assess baseline learning, in which participants were presented with only one exemplar per training trial. Participants responded by making a classification judgment for the two exemplars and then received *corrective feedback*, with the correct category label(s) for the exemplar(s) shown, along with whether the response was correct.

To compare performance across different training conditions and to assess the progress of learning, an endorsement judgment trial was presented after every 6th training trial for the match and contrast conditions, and after every 12th training trial for the control condition. This was set up so that, at each endorsement judgment, participants in the control, match, and contrast conditions were all exposed to the same number of training exemplars. On an endorsement trial, participants were presented with an exemplar along with a category label and were asked to determine whether the label was correct. After responding, no corrective feedback was given. Endorsement was used instead of single-item classification to eliminate the potential performance benefit of the control group, for which familiarity with single-item classification could carry over from training to the test trials. Endorsement also allowed for a common dependent measure to assess and compare performance across the match, contrast, and control conditions. A posttest endorsement task consisting of repeated and novel exemplars was included after training to assess memory and knowledge generalization. Together, the embedded and the posttest endorsement performance served as our primary dependent measures.

Method

Participants A total of 237 participants was recruited via Prolific (https://www.prolific.com) in December 2025 and was compensated at a rate of $8.24 per hour. Eligibility criteria included being fluent English speakers, residing in the United States, and being at least 18 years old.

Design and Materials This study used a 3 (comparison type: match vs. contrast vs. control; between-subjects) × 2 (confusability: high vs. low; within-subjects) mixed design. Participants were randomly assigned to one of the three between-subject groups: (a) match (n = 81), (b) contrast (n = 82), and (c) control (n = 74).

Four sets of stimuli, each consisting of 64 aliens, were created for the experiment. Two sets were robot-like aliens (see Figure 1A) and two were dog-like aliens (see Figure 1B). For both sets of stimuli, one subset was created for use in the low-confusability conditions and another subset for use in the high-confusability conditions. The stimuli varied along three dimensions—*angle*, *brightness*, and *length*—each of which had four levels, thus resulting in 64 exemplars for each stimulus set. The angle feature corresponded to arm–body angles in robot-like aliens and tail–body angles in dog-like aliens. The brightness feature corresponded to mouth brightness in robot-like aliens and body brightness in dog-like aliens. The length feature corresponded to eye length in robot-like aliens and leg length in dog-like aliens.

For the robot-like aliens in the low-confusability conditions, angle took the values of 5°, 23.75°, 61.25°, and 80°; brightness (on a grayscale ranging from 0 to 1, from white to black) took the values of 0.2, 0.375, 0.725, and 0.9; and length took the values of 183, 239, 351, and 407 pixels. For the robot-like aliens in the high-confusability conditions, angle took the values of 20°, 31.25°, 53.75°, and 65°; brightness took the values of 0.34, 0.445, 0.655, and 0.76; and length took the values of 226, 260, 328, and 362 pixels. For the dog-like aliens in the low-confusability conditions, angle took the values of 5°, 23.75°, 61.25°, and 80°; brightness took the values of 0.2, 0.375, 0.725, and 0.9; and length took the values of 178, 235, 349, and 406 pixels. Finally, for the dog-like aliens in the high-confusability conditions, angle took the values of 20°, 31.25°, 53.75°, and 65°; brightness took the values of 0.34, 0.445, 0.655, and 0.76; and length took the values of 224, 258, 326, and 360 pixels.

The category structure followed a one-dimensional rule. Category membership was defined by a criterial dimension, which was randomly selected from one of the three features independently for each training sessions, such that a participant could have the same or different criterial dimensions across sessions. Stimuli with the criterial dimension having the lower two levels were assigned to Category A, and those having the higher two levels were assigned to Category B (see Figure 1). The level values across the three dimensions were arranged such that the bivariate distributions of any two dimensions were uniform, thereby eliminating any information about category membership beyond that provided by the criterial dimension.

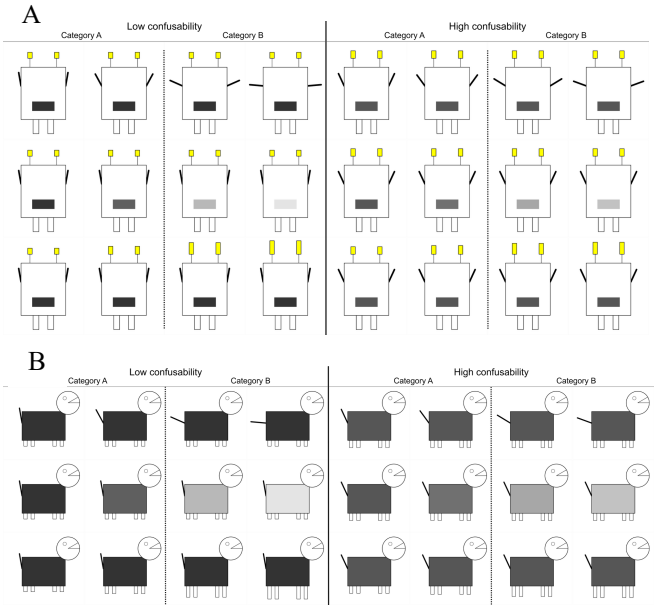

Figure 1: Experiment stimuli for the robot- (A) and dog-like (B) aliens, partitioned by category and confusability.

Procedure All stimuli were presented at the center of the screen on a black background. At the start of each session, participants were shown an instruction telling them that they would be learning about two alien tribes during the training session. Each participant completed two sessions of category training, each using a different set of stimuli. One session used a low-confusability stimulus set, and the other used a high-confusability stimulus set. The order of low- and high-confusability sessions was randomly determined for each participant. Each session either used a robot-like alien stimulus set or a dog-like alien stimulus set. If a robot-like set was used in the first session, then the second session always used a dog-like set, and vice versa. When a robot-like set was used, participants were instructed to learn to categorize the alien stimuli into two tribes, "Nelu" and "Vorka". The same instruction was given when a dog-like set was used, but the two tribe labels used were instead "Toven" and "Sharka". After assigning the stimulus set for a session, the criterial dimension for category membership was randomly selected from the three feature dimensions, independently for each session such that a participant could have the same or different criterial dimensions across the two sessions. Tribe labels were randomly assigned to the categories.

In each session, participants went through a training phase of category learning followed by a posttest. For each participant, in the training phase, 16 Category A aliens and 16 Category B aliens were randomly selected as training exemplars. Of the remaining 32 stimuli within the set, 16 were randomly chosen as candidate stimuli for testing; 8 of these candidates were used in the endorsement trials embedded within the training trials. In the corresponding posttest, the remaining 16 stimuli we used (8 from each category). Additionally, 16 stimuli were also randomly sampled from the training exemplars (8 from each category) and were in the posttest, resulting in a total of 32 posttest endorsement trials (16 new exemplars and 16 old exemplars).

On each training trial, participants in the match and contrast conditions were presented with a pair of alien exemplars side-by-side. In the match condition, the two aliens were randomly drawn from the same category. In the contrast condition, one exemplar was drawn from each of the two categories. For both the match and contrast conditions, the positions of the exemplars (i.e., "left" or "right") were randomized on each trial. In the control condition, on each trial, a single exemplar was drawn from the training stimuli and presented.

All participants completed three training cycles. Each training cycle consisted of 16 training trials for the match and contrast conditions and 32 trials for the control condition. Within each training cycle, the alien stimuli were sampled without replacement. Each training exemplar was thus presented three times during the training phase across all three comparison conditions. Participants were given a self-paced rest break after each training cycle, during which the current progress of the experimental session was shown (e.g., "18 out of 88 trials completed").

When a robot-like alien set was used, participants in the match condition were asked to press "F" if both aliens were "Nelu" and "J" if both were "Vorka". Participants in the contrast condition were asked to press "F" if the alien on the left was a "Nelu" and the one on the right was a "Vorka", and "J" if the left alien was a "Vorka" and the right alien was a "Nelu". Finally, in the control condition, participants were asked to press "F" if the alien was a "Nelu" and "J" if it was a "Vorka". When a dog-like alien set was used, participants in the match condition were asked to press "F" if both aliens were "Toven" and "J" if both were "Sharka". Participants in the contrast condition were asked to press "F" if the alien on the left was a "Toven" and the one on the right was a "Sharka", and "J" if the left alien was a "Sharka" and the right alien was a "Toven". Finally, in the control condition, participants were asked to press "F" if the alien was a "Toven" and "J" if it was a "Sharka".

After each classification response, corrective feedback was presented for 1.5 seconds while the exemplar(s) remained on the screen, with the correct category label(s) shown beneath the alien exemplar(s). If participants responded correctly, "Correct" was displayed in green text at the center of the screen; otherwise, "Wrong" was displayed in red text. On all training trials, the intertrial interval was 400 ms.

After every 6th training trial in the match and contrast conditions, and every 12th training trial in the control condition, participants were presented with an endorsement trial, where an alien exemplar, along with a randomly selected category label, was shown at the center of the screen. Participants were asked to press "C" if the label was correct and "N" otherwise. After responding, the screen was cleared and a "Thank you" prompt was shown for 500 ms.

After each training phase, participants were instructed that they would complete a posttest endorsement task, which was similar to the endorsement trials embedded within the training phase. The posttest consisted of 32 endorsement trials. For each participant, the presentation order of the 32 posttest stimuli was randomized.

Overall, each participant in the control group completed a total of 272 trials (each of the two sessions consisted of 96 training trials, 8 embedded trials, and 32 posttest trials), and each participant in the match and contrast groups completed a total of 176 trials (each of the two sessions consisted of 48 training trials, 8 embedded trials, and 32 posttest trials).

Results

Training Performance Figure 2 shows the mean training performance partitioned by comparison type and confusability. To analyze performance on the training trials, we conducted a 3 × 2 mixed ANOVA, with comparison type (match vs. contrast vs. control) as a between-subjects factor and confusability (low vs. high) as a within-subject factor. A significant main effect was found for comparison type, $F(2,234) = 15.82$, $MSE = .052$, $p < .001$, $\eta_p^2 = .119$. The main effect for confusability, $F(2,234) = 2.94$, $MSE = .027$, $p = .088$, $\eta_p^2 = .012$, and the interaction between comparison and

confusability, $F(2,234) = 2.81$, $MSE = .027$, $p = .062$, $\eta_p^2 = .023$, were not significant.

To further analyze the main effect of comparison type, post hoc LSD tests were conducted, which revealed that participants in the match condition ($M = .68$, $SE = .018$) were outperformed by participants in the contrast condition ($M = .82$, $SE = .018$), $t(234) = 5.50$, $p < .001$, $d = .86$, and by participants in the control condition ($M = .78$, $SE = .019$), $t(234) = 3.73$, $p < .001$, $d = .60$. The difference between the contrast and control conditions was not significant, $t(234) = 1.64$, $p = .232$, $d = .26$.

The analysis above reports results using all training trials such that the total number of exposures to the training exemplars was constant across the match, contrast, and control groups. However, since two, rather than one, exemplars were presented in the match and contrast groups, they only had half the number of training trials as the control group. Therefore, as a supplemental analysis[1] (see Figure 2, top-right panel; also see Figure 3A for the learning curves), we compared the performance of the match and contrast groups to the control group under an equal number of training trials (i.e., by discarding the second half of the trials in the control group). Results showed that participants in the contrast condition ($M = .82$, $SE = .018$) outperformed participants in the control condition ($M = .73$, $SE = .019$), $t(234) = 3.56$, $p = .001$, $d = .57$, and that there were no significant difference between the match ($M = .68$, $SE = .018$) and control condition, $t(234) = 1.88$, $p = .147$, $d = .30$.

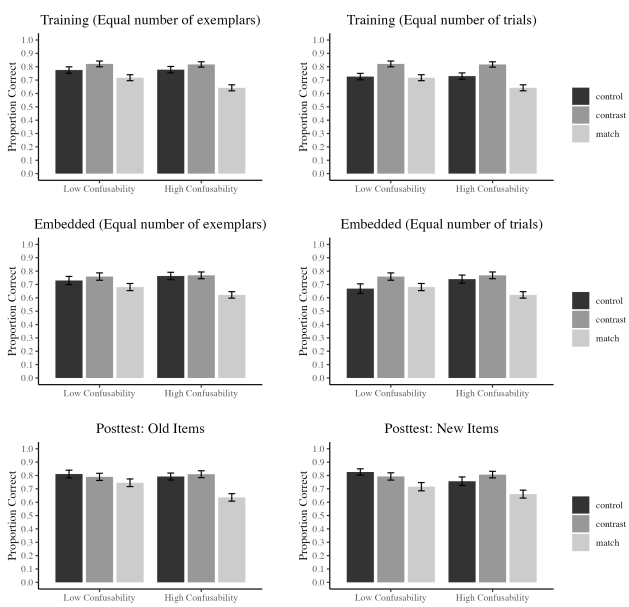

Figure 2: Mean performance by condition. Top row: mean performance on training trials. Middle row: mean performance on embedded endorsement trials. Bottom row: mean performance on posttest endorsement trials.

Embedded Endorsement To analyze performance on the embedded endorsement trials (see Figure 3, middle-left panel), we conducted the same 3 (match vs. contrast vs. control) × 2 (low- vs. high-confusability) mixed ANOVA on the embedded endorsement trials as we did for the training trials. A significant main effect was found for comparison type, $F(2,234) = 8.38$, $MSE = .070$, $p < .001$, $\eta_p^2 = .067$. The main effect for confusability $F(2,234) = 0.07$, $MSE = .044$, $p = .787$, $\eta_p^2 = .0003$, and the interaction between comparison and confusability, $F(2,234) = 2.04$, $MSE = .044$, $p = .132$, $\eta_p^2 = .017$, were not significant.

To analyze the main effect of comparison type, post hoc LSD tests revealed that participants in the match condition ($M = .65$, $SE = .021$) were outperformed by participants in the contrast condition ($M = .76$, $SE = .021$), $t(234) = 3.82$, $p < .001$, $d = .60$, and by participants in the control condition ($M = .75$, $SE = .022$), $t(234) = 3.16$, $p = .005$, $d = .51$. The difference between the contrast and control conditions was not significant, $t(234) = 0.57$, $p = .838$, $d = .09$.

The above analyses include results using all embedded endorsement trials, such that the total number of exposures to the training exemplars was constant across all conditions. As in the training performance section, we report a similar supplemental analysis[2] using a subset of data in which the number of training trials was held constant across the three conditions (i.e., by discarding the second half of the embedded endorsement trials in the control group). Results showed that the differences between the match ($M = .65$, $SE = .021$) and control ($M = .70$, $SE = .022$) conditions, $t(234) = 1.74$, $p = .193$, $d = .28$, and between the contrast ($M = .76$, $SE = .021$) and control conditions, $t(234) = 1.95$, $p = .128$, $d = .31$, were not significant.

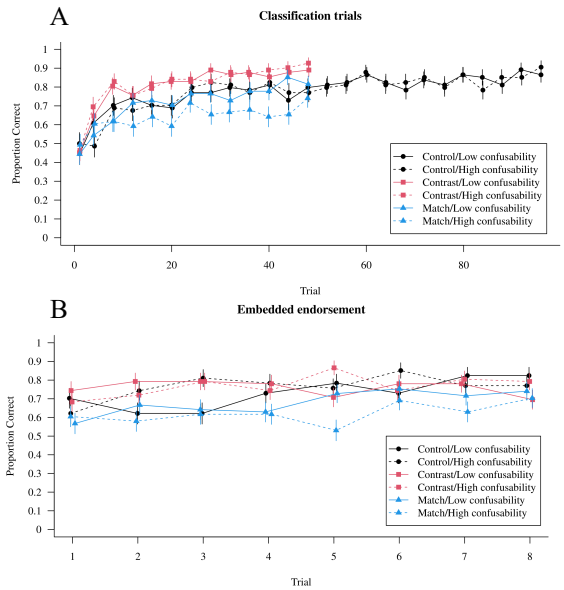

Figure 3: Mean learning curves for the training/classification trials (A) and embedded endorsement trials (B).

[1]In the supplementary analysis, a significant main effect was found for comparison type, $F(2,234) = 15.98$, $MSE = .051$, $p < .001$, $\eta_p^2 = .120$. All other effects were not significant (all $ps > .068$).

[2]In the supplementary analysis, a significant main effect was found for comparison type, $F(2,234) = 7.14$, $MSE = .072$, $p < .001$, $\eta_p^2 = .057$. All other effects were not significant (all $ps > .052$).

Posttest Endorsement To analyze performance on the posttest endorsement task, we conducted a 3 × 2 × 2 mixed ANOVA, with comparison type (match vs. contrast vs. control) as a between-subject factor and confusability (low vs. high) and item type (repeated vs. novel) as within-subject factors. Significant main effects were found for comparison type, $F(2,234) = 9.07$, $MSE = .139$, $p < .001$, $\eta_p^2 = .072$, and confusability, $F(1,234) = 4.44$, $MSE = .071$, $p = .036$, $\eta_p^2 = .019$. There was also a significant three-way interaction between comparison, confusability, and item type, $F(2,234) = 3.10$, $p = .047$, $MSE = .017$, $\eta_p^2 = .026$. The two-way interaction between comparison and confusability was marginally significant, $F(2,234) = 2.87$, $p = .059$, $MSE = .071$, $\eta_p^2 = .024$. All other interactions were not significant (all ps > .623).

Post hoc LSD tests on comparison type revealed that participants in the match condition ($M = .689$, $SE = .021$) were outperformed by participants in the contrast condition ($M = .80$, $SE = .021$), $t(234) = 3.77$, $p < .001$, and by participants in the control condition ($M = .80$, $SE = .020$), $t(234) = 3.57$, $p = .001$.

To interpret the three-way significant interaction, we examined the relationship between comparison and confusability separately for old and new items. For old items, the comparison-by-confusability interaction was statistically significant, $F(2,234) = 4.04$, $MSE = .044$, $p = .019$, $\eta_p^2 = .033$, indicating that the effect of comparison depended on confusability. Further post hoc comparison revealed that this effect was mainly driven by match ($M = .64$, $SE = .026$) being outperformed by contrast ($M = .79$, $SE = .028$), $t(234) = 4.66$, $p < .001$, $d = .73$, and control ($M = .81$, $SE = .026$), $t(234) = 4.09$, $p < .001$, $d = .66$, under the high-confusability condition; no performance differences among the conditions were significant on the low-confusability categories (all ps > .10). In contrast, this comparison-by-confusability interaction was not significant for new items, $F(2,234) = 1.79$, $MSE = .044$, $p = .17$, $\eta_p^2 = .015$, where a reliable advantage of contrast and control over match was not observed.

Discussion

The present study examined how different types of comparison affect feature-based category learning, and whether category confusability moderates this effect. Across all three trial types—training, embedded endorsement, and posttest endorsement—consistent effects of comparison type were found, such that participants in the contrast and control conditions outperformed those in the match condition. These findings—consistent with Corral et al. (2018) and Perri et al. (2025)—add to the literature demonstrating a contrast advantage over match in feature-based category learning. These findings are thus directly in line with theories of feature learning and selective attention, wherein comparing exemplars from different categories is posited to highlight the diagnostic differences between the to-be-learned categories (Nosofsky, 1986).

Marginal comparison-by-confusability interactions suggested that the advantage of contrast over match may be stronger for highly confusable categories than for categories with low confusability. However, because these effects were not reliable, this pattern should be interpreted cautiously and requires further investigation in future work.

This observed pattern is also only partially consistent with the findings from Carvalho and Goldstone (2014), in which a cross-over interaction between presentation type (blocked vs. interleaved) and confusability (low- vs. high-similarity stimuli) was found: interleaved presentation (more opportunities for between-trial, between-category comparison) led to better performance for high-confusability stimuli, whereas blocked presentation (more opportunities for between-trial, within-category comparison) led to better performance for low-confusability stimuli. In the present study, a match advantage was never found under either confusability condition. The present findings thus suggest that the effects observed (by Carvalho and Goldstone) with blocking and interleaving across trials do not extend to within-trial comparisons that involve within- and between-category comparison. Moreover, our results raise the possibility that different learning processes might underlie between- and within-trial comparisons.

Another notable finding is that single-item presentation in the control condition led to comparable, or even better, performance than co-presenting items: participants in the control conditions reliably outperformed those in the match conditions and did not reliably differ from participants in the contrast conditions across the training, embedded endorsement, and posttest trials. Supplemental analyses of the training and embedded endorsement trials—in which the control conditions had only half as many training exemplars as the match and contrast conditions—further reinforce the notion of an advantage for single-item presentation. Under these conditions, the control condition still performed relatively well, performing at least as well as, and numerically better than, match comparison. Although performance did drop and the control condition performed reliably worse than contrast in the training trials, this pattern still suggests an advantage for single-item presentation, given that the control condition had only half as many exposures.

This advantage of single-item presentation does not align with earlier findings showing that co-presenting exemplars leads to better learning than single-item presentation (Kurtz et al., 2013). This finding is also somewhat puzzling, given that between-trial comparison in single-item presentation would presumably place greater load on working memory, as learners must remember stimuli from previous trials to compare them with the stimulus on the current trial. Co-presenting exemplars, by contrast, would presumably reduce this memory load, as it allows within-trial comparison to be made directly by visually comparing two stimuli side by side. However, given this quite robust pattern of an advantage in single-item presentation, especially over match, it may be the case that single-item presentation is more efficient than co-presentation for concept learning. Future work will be needed, however, to further explore and understand the learning mechanisms that might underly this effect.

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

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*(3), 481–495. https://doi.org/10.3758/s13421-013-0371-0

Corral, D. & Jones, M. (2012). Learning of relational categories as a function of higher-order structure. In N. Miyake, D. Peebles & R. P. Cooper (Eds.), *Proceedings of the 34th Annual Meeting of the Cognitive Science Society* (pp. 1434–1439). Austin, TX: Cognitive Science Society. https://escholarship.org/uc/item/3130s373

Corral, D., & Jones, M. (2014). The effects of relational structure on analogical learning. *Cognition, 132*(3), 280–300. https://doi.org/10.1016/j.cognition.2014.04.007

Corral, D., Kurtz, K. J., & Jones, M. (2018). Learning relational concepts from within- versus between-category comparisons. *Journal of Experimental Psychology: General*, *147*(11), 1571–1596. https://doi.org/10.1037/xge0000517

Gentner, D., & Namy, L. L. (1999). Comparison in the development of categories. *Cognitive Development, 14*(4), 487–513. https://doi.org/10.1016/S0885-2014(99)00016-7

Goldstone, R. L. (1994). An efficient method for obtaining similarity data. *Behavior Research Methods, Instruments, & Computers*, *26*, 381-386.

Kotovsky, L., & Gentner, D. (1996). Comparison and categorization in the development of relational similarity. *Child Development, 67*(6), 2797–2822. https://doi.org/10.2307/1131753

Kruschke, J. K. (1992). ALCOVE: An exemplar-based connectionist model of category learning. *Psychological Review*, *99*(1), 22–44. https://doi.org/10.1037/0033-295X.99.1.22

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

Mackintosh, N. J. (1975). A theory of attention: Variations in the associability of stimuli with reinforcement. *Psychological Review*, *82*(4), 276–298. https://doi.org/10.1037/h0076778

McKinley, S. C., & Nosofsky, R. M. (1995). Investigations of exemplar and decision bound models in large, ill-defined category structures. J*ournal of Experimental Psychology: Human Perception and Performance, 21*, 128-148.

Namy, L. L., & Gentner, D. (2002). Making a silk purse out of two sow's ears: Young children's use of comparison in category learning. *Journal of Experimental Psychology: General, 131*(1), 5–15. https://doi.org/10.1037/0096-3445.131.1.5

Nosofsky, R. M. (1986). Attention, similarity, and the identification–categorization relationship. *Journal of Experimental Psychology: General*, *115*(1), 39.

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.

Smith, J. D., & Minda, J. P. (2000). Thirty categorization results in search of a model. *Journal of Experimental Psychology. Learning, Memory, and Cognition*, *26*(1), 3–27. https://doi.org/10.1037//0278-7393.26.1.3

Stewart, N., Brown, G. D. A., & Chater, N. (2002). Sequence effects in categorization of simple perceptual stimuli. *Journal of Experimental Psychology. Learning, Memory, and Cognition*, *28*(1), 3–11. https://doi.org/10.1037//0278-7393.28.1.3