Target vs. Distractor: Does the Role of a Category in Comparisons Influence Learning? Evidence from Skin Cancer Classification

Victoria L. Jacoby (vjacoby@ucla.edu)[1]
Christine M. Massey (cmassey@psych.ucla.edu)[1]
Philip J. Kellman (kellman@cognet.ucla.edu)[1,2]

[1]Department of Psychology, University of California, Los Angeles, Los Angeles, CA 90095, USA
[2]Department of Surgery, David Geffen School of Medicine, University of California, Los Angeles, Los Angeles, CA 90095, USA

Abstract

Recent research indicates that paired comparisons can accelerate perceptual learning of challenging dermatological lesion categories. Here we investigated whether the role of object categories as targets or distractors differentially influences learning outcomes. The frequency with which a given category occupied the target position was manipulated across three learning conditions: Always-Never, where half of 10 categories were always shown as target and the other half never shown as target; Often-Rarely, where half of categories appeared 75% as targets and 25% as distractors, with reversed presentation frequency for the other half; and Equal Split learning, in which all categories appeared as targets or distractors equally often. After learning, transfer results indicated that all conditions yielded equivalent overall learning, but categories prioritized more often as targets exhibited greater learning gains. These findings implicate differential processing of images in comparisons, even when no information regarding target vs. distractor was given prior to feedback.

Keywords: categories; comparison; discriminative contrast; perceptual learning; skin cancer

Introduction

The classification of visual stimuli into meaningful perceptual categories is fundamental in many cognitive domains and a crucial driver of advanced expertise. In many medical imaging domains, the detection of abnormalities often requires the discovery and interpretation of complex visual patterns. Cancer detection and lesion classification in dermatology is one such domain, where perceptual classifications of trained experts guide important medical outcomes. The results of these classifications have real and serious consequences. As skin cancers are currently the most common form of cancer in the United States (CDC, 2023), it is imperative that experts are equipped with the best training possible to increase detection rates of dangerous carcinomas and melanomas.

While increasing declarative knowledge and explicitly learned membership rules can provide valuable assistance in certain instances of perceptual classification, expertise in visual discrimination is largely dependent upon more implicit processes of perceptual learning (PL), experience-induced improvements in perception (Gibson, 1969; Kellman & Garrigan, 2009). In the case of skin cancer detection, a learner may be informed about which features (e.g., size, shape, symmetry, color) are most common in cancerous versus benign lesions; however, because many of these features vary across lesions of the same category, as well as appear very similarly in lesions of different categories, this declarative knowledge alone is insufficient. PL in domains like this grows through classification experiences and feedback with a variety of examples in contrasting categories suitable to allow perceptual selection of relevant discriminative features and patterns, as well as downweighting of irrelevant variations.

Recent work has sought to introduce PL interventions across multiple visually-rich medical domains to help accelerate the development of perceptual expertise (for a review, see Kellman, Jacoby, Massey, & Krasne, 2022). Providing targeted practice with the relevant stimuli in multi-category classification, typically by presenting examples and requiring classification responses, attunes perceptual mechanisms to selectively pick up relevant information and suppress irrelevant information (cf. Petrov, Dosher, & Lu, 2005). Consequently, learners become more accurate and faster in extracting the information most critical for classification (Kellman & Garrigan, 2009; Shiffrin & Schneider, 1977). In dermatology in particular, perceptual learning modules designed to improve the classification of skin lesion categories successfully accelerated category acquisition and transfer for medical students (Rimoin et al., 2015) and novice learners (Kellman, Krasne, Massey, & Mettler, 2023).

Recently, it has been suggested that the learning of these perceptual classifications may be further enhanced by shifting focus from practicing classification of individual displays to simultaneously comparing category exemplars. Jacoby, Massey, and Kellman (2024) tested a *paired comparison* learning approach for teaching the perceptual classification of 10 benign and cancerous skin lesion categories to medically naive participants. In this approach, two skin lesion displays, one from each of two different categories, were presented side by side with the observer instructed to select which one belonged to a given category. For example, an image of a “basal cell carcinoma” lesion might be presented alongside an image of a non-cancerous, “haemangioma” lesion, with the question “Which one is basal cell carcinoma?”. This paired comparison training was compared to two classification-focused approaches in which one or two skin lesions were presented at a time and learners

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

selected a diagnostic label for each item. When tested on a classification-focused assessment consisting of novel lesion examples, those who trained with paired comparisons demonstrated a significant, long-lasting advantage over either classification condition.

The potential for comparisons to enhance learning is not new and has been observed across various areas of learning and cognition (e.g., Gentner & Markman, 1997; Goldstone, Day, & Son, 2010; Medin, Goldstone, & Gentner, 1993), including category learning specifically (e.g., Kurtz, Boukrina, & Gentner, 2013; Spalding & Ross, 1994). Simultaneous comparisons – presentation of exemplars from the same or different categories – allow learners to look between the features presented within each stimulus without needing to rely on memory representations of either category. This may be especially advantageous for complex and/or highly confusable categories, as in dermatology, where the relevant patterns and distinctions are subtle.

However, despite the clear evidence of comparison benefiting learning, the vast majority of categorization studies have involved only single-item classification trials (Markman & Ross, 2003). Further, of the approaches that do utilize simultaneous presentation of multiple items, most learning approaches have used either an entirely passive presentation of items (e.g., Kang & Pashler, 2012; Kok et al., 2013) or required label recall for one or more of the presented items (e.g., Andrews, Livingston, & Kurtz, 2011; Carvalho & Goldstone, 2014; Homa, Powell, & Ferguson, 2014). To our knowledge, outside of the work by Jacoby et al., little to no other research has previously evaluated the use of discrimination-focused, paired comparison trials as a way to learn multiple perceptual classifications. Consequently, little is known about the way learning advances through this method.

Figure 1: An example paired comparison learning trial with feedback: "Which one is Solar Lentigo?" Left image: Solar Lentigo (target); right image: Lentigo Maligna Melanoma (distractor)

The paired comparison format provides learners the opportunity to see two different category examples, while also providing critical feedback regarding each item's category membership. However, given the prompt of each trial, only one category is framed as the *target,* with the other presented category fulfilling the role of a *distractor*. An example of this trial format is shown in Figure 1. Although the opportunity to learn about both presented categories is equal, we were interested in whether the different framing would lead to disparities in the extent to which learning progresses for each presented category.

There are a few reasons for why differences in the processing of targets and distractors in this format may be expected. Learners are often considered to be conservative with regard to their cognitive effort, often only learning and engaging with the minimum amount of information necessary to complete a task (Payne et al., 1993). Early in learning, when the learning strength for all categories is low, items are likely compared closely to complete each trial–which should lead to benefits in the perception of both categories. However, as learning progresses and close comparison of items becomes less necessary, attention may be devoted primarily to the target category. Additionally, given that knowing the category label of the distractor image is not necessary for correctly completing the trial (i.e., one need not know what the distractor category is, just that it is *not* the target category), participants may be less inclined to devote attention to that category during feedback. Finally, while both presented categories are displayed alongside their category label after each trial, the target of a trial gets the benefit of having its category label visually presented twice (once in the instructions and once in feedback). If a category were to regularly appear as a target, rather than a distractor, the increased number of times a learner is exposed to its label may give the impression that that category is more important or possibly more common than the distractor, which in turn could result in an overestimation of likelihood that that category is presented in future encounters.

In the present work, we investigated how learning advances in paired comparison training, and in particular, whether the specific role of a category on the trial (target vs distractor) differentially affects the learning of that category. To evaluate this, participants were tasked with learning the perceptual classification of ten dermatological lesion categories while the frequency with which a specific category showed up in learning as a target or distractor was manipulated across three different learning conditions.

In the *Always-Never* learning condition, half of the to-be learned categories were assigned to appear in learning only as targets, whereas the other half appeared only as distractors. In the *Often-Rarely* learning condition, half of the categories showed up as targets on 75% of trials and as distractors on 25% of trials, with the remaining categories following the opposite scheduling. Finally, in the *Equal Split* condition, all categories showed up in learning equally as often as the target of the learning trial and as the distractor. Within each condition, we aimed to compare the performance on categories prioritized as targets to the performance on categories prioritized as distractors. If the design of these trials provides an asymmetric learning gain that benefits target categories relative to distractor categories, then we would expect to see a positive relationship between the frequency of presentations as a target and final assessment performance. Further, we expected to see the largest disparity in performance among categories in the Always-Never condition, followed by a smaller difference in the Often-Rarely condition, and no difference in the Equal Split condition.

Method

Participants

104 undergraduate psychology students from the University of California, Los Angeles completed this experiment. Participants had no particular medical background or training and received partial course credit for their participation.

Materials

Stimuli consisted of dermoscopic images of 10 different skin lesion categories including four cancerous skin lesion categories and six benign categories. A dermoscope incorporates high magnification and an adjustable illumination system that allows detailed assessment beneath the outer surface of the skin. All images were obtained from the MoleMap Database with the correct image classification determined from combined assessments by expert dermatologists, melanographers, AI systems, and, in many cases, biopsy results. Images were selected on the basis of dermatologic diagnosis, verification via biopsy when appropriate, and good image quality. Due to these criteria and availability, the number of instances varied from 19 to 165 unique images per category. Four exemplars per category were set aside as novel stimuli to be used in assessments.

Categories were divided into two separate lists, each containing five categories. List 1 contained: Actinic Keratosis, Basal Cell Carcinoma, Haemangioma, Nodular Melanoma, and Solar Lentigo. List 2 contained Benign Nevus, Lentigo Maligna Melanoma, Seborrheic Keratosis, Squamous Cell Carcinoma, and Wart. Exemplars from each list are shown in Figure 2. All participants learned both lists of categories in learning.

Design & Procedure

The experiment used a mixed measures design, with three between-participant learning conditions and two within-participant priority lists. Each participant was assigned to one of three learning conditions: Always-Never Learning, Often-Rarely Learning, or Equal Split Learning. Within each learning condition one list of categories was designated as the Target-Priority list and the other was designated as the Distractor-Priority list. The Target-Priority list contained the categories assigned to appear in learning more often as targets than as distractors, whereas the Distractor-Priority list contained the categories that were assigned to appear more often as distractors. This assignment was counterbalanced within each learning condition to control for possible differences in list difficulty.

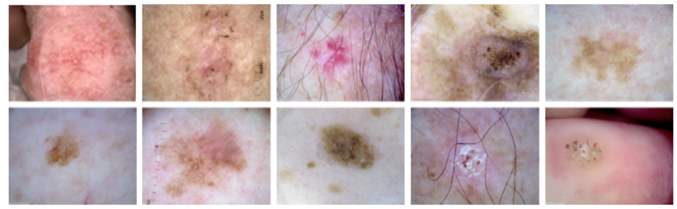

Figure 2: Example category images. The top row depicts List 1 categories in alphabetical order from left to right; the bottom row depicts List 2 categories.

Learning Phase All participants completed a series of paired comparison trials, in which an exemplar from two separate categories would be presented side by side and participants would be asked to identify a given target through a prompt asking "Which is [Category Name]?" An example of this trial can be seen in Figure 1.

In the Always-Never learning condition, when any of the five categories designated as the Target-Priority list appeared on a trial they would always be the target and never be the distractor. Conversely, the five categories designated as the Distractor-Priority list would always appear as the distractor but never the target. As a concrete example, if *solar lentigo* was on the list of categories designated as the Target-Priority list and *lentigo maligna melanoma* was on the list of categories designated as the Distractor-Priority list, a participant could receive a trial that displayed one solar lentigo image and one lentigo maligna melanoma image side by side and asked "Which is Solar Lentigo?" but never a trial that presented the same images and asked "Which is Lentigo Maligna Melanoma?"

In the Often-Rarely learning condition, categories designated as the Target-Priority list would appear as the target of the trial on 75% of presentations and as the distractor on 25%; the reverse was true for categories designated as the Distractor-Priority list.

Finally, in the Equal Split learning condition, all categories appeared equally often as the target and as the distractor. If solar lentigo and lentigo maligna melanoma were presented together in the trial, it was equally likely that the prompt would be "Which is Solar Lentigo?" or "Which is Lentigo Maligna Melanoma?". Although there is no difference between the Target-Priority and Distractor-Priority lists in the Equal Split condition, category lists were still designed as separate priority lists to allow for subsequent analyses.

Participants in all conditions completed 400 learning trials containing a total of 800 images. Importantly, all categories were shown 80 times in learning regardless of the learning condition or target/distractor-priority list. Every trial contained one category from the Target-Priority list and one category from the Distractor-Priority list. Categories were presented in a randomized order. Participants were informed that they would be learning 10 different categories of skin lesions, but they were not made aware of how often a category was to be shown as the target or distractor.

Feedback was provided immediately after each trial and indicated whether the answer was correct/incorrect, as well as labeled both presented images. Participants were given 40 seconds to complete each learning trial and up to 10 seconds to view feedback.

Testing Phase Participants completed an assessment at three different timepoints: before learning, immediately following learning, and after a one-week delay. Assessments required the classification of individually presented skin lesion images, where on each test trial, one image was shown with all ten category labels organized alphabetically below. Participants had 40s to select a category label on each trial; no feedback was given. Each test contained four questions from each category, and only novel (previously unseen) exemplars were used.

Exclusion Criteria

To ensure that participants did not have considerable knowledge of these skin lesion categories prior to starting the experiment, those who scored 30% or greater on the pretest were disqualified and did not participate in the rest of the experiment. Only data from participants who completed all parts of the experiment (pretest, learning phase, immediate posttest, and delayed posttest) were included in the following analyses.

Participants were excluded after data collection if they failed to achieve an average accuracy in the learning phase of 65% (chance accuracy: 50%). In total, 14 participants were excluded for poor learning performance (Always-Never condition: n = 2; Often-Rarely condition: n = 5; Equal Split condition: n = 7). Data from 90 participants were retained.

Dependent Measures and Data Analyses

Accuracy on learning trials and assessments was recorded for all participants and compared across learning conditions. Performance on the posttests was also divided between categories designated as the Target-Priority list and the Distractor-Priority list. Differences between priority lists were compared within each learning condition. To check whether any observed differences in accuracy were due to changes in perception, as opposed to changes in response behavior, false alarm rates and sensitivity (d') were also calculated and compared.

All analyses were conducted using standard parametric measures. Effect sizes are reported for each difference.

Results

Learning Accuracy

All participants completed 400 paired comparison trials during learning. A one-way ANOVA determined that differences in accuracy during learning were not reliable, (Always-Never: M = 0.81, SD = 0.06; Often-Rarely: M = 0.80, SD = .07; Equal Split condition: M = 0.78, SD = .07), $F(2, 87) = 1.78$, $p = .175$, $\eta_p^2 = 0.04$

Assessment Performance

Overall Accuracy Average performance on the pretest assessment was 16.44% (SD = 6.77), and did not differ between conditions, $F(2, 89) = 0.03$, $p = .972$, $\eta_p^2 = .001$.

Posttest accuracy was similar across all three conditions at the immediate posttest (Equal Split: M = .65, SD = .13; Always-Never: M = .62, SD = .13; Often-Rarely: M = .60, SD = .13) and at the delayed posttest (Equal Split: M = .54, SD = .13; Always-Never: M = .53, SD = .17; Often-Rarely: M = .52, SD = .15). A 3 (learning condition) X 2 (posttest phase) mixed measures ANOVA was conducted on the assessment scores. There was a significant main effect of posttest phase, such that participants performed worse on the one-week delayed posttest than on the immediate posttest regardless of learning condition, $F(1, 87) = 54.37$, $p < .001$, $\eta_p^2 = 0.39$. There was no reliable main effect of learning condition, $F(2, 87) = 0.67$, $p = .513$, $\eta_p^2 = 0.02$, nor any reliable learning condition by posttest phase interaction, $F(2, 87) = 0.17$, $p = .844$, $\eta_p^2 = 0.004$.

Target vs. Distractor Accuracy Posttest items were then divided into the categories that had appeared in learning as Target-Priority items and Distractor-Priority items. Figure 3 shows average accuracy on Target-Priority and Distractor-Priority categories for each learning condition at both post-learning assessments.

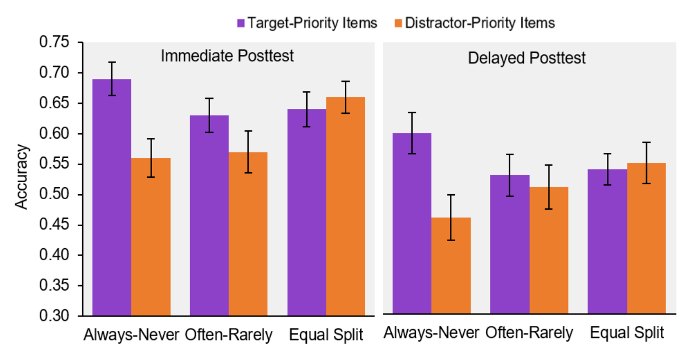

Figure 3: Accuracy (proportion correct) for categories prioritized as targets and for categories prioritized as distractors for each learning condition. Error bars indicate ± 1 standard error of the mean

A 2 (Priority List) by 2 (Posttest Phase) by 3 (Learning Condition) mixed measures ANOVA was conducted on assessment scores. There was a reliable main effect of posttest phase, such that items were classified with lower accuracy at delayed posttest relative to immediate posttest, regardless of learning condition, $F(1,87) = 54.37$, $p < .001$, $\eta_p^2 = 0.39$. Posttest phase did not interact with learning condition, $F(2, 87) = 0.17$, $p = .844$, $\eta_p^2 = 0.004$, or with priority list, $F(2, 87) = 0.001$, $p = .980$, $\eta_p^2 = 0$. Additionally, no three-way interaction was found, $F(2, 87) = .0.52$, $p = .599$, $\eta_p^2 = 0.01$. A significant interaction was found between priority list and learning condition, $F(2, 87) = 4.56$,

p = .013, η_p^2 = 0.10. To further evaluate this relationship, the effect of priority list was assessed separately for each learning condition.

In the Always-Never condition, categories that were assigned to the Target-Priority list were more accurately classified than the categories assigned to the Distractor-Priority list at the immediate posttest (Target-Priority: M = 0.69, SD = 0.15, Distractor-Priority: M = 0.56, SD = 0.17), as well as at the delayed posttest (Target-Priority: M = 0.60, SD = 0.19, Distractor-Priority: M = 0.46, SD = 0.21). Analyses confirmed the average difference between priority lists across both posttests to be significant with a large effect size, $t(29) = 4.40$, $p < .001$, d_z = 0.80.

In the Often-Rarely condition, accuracy on categories prioritized as targets and categories prioritized as distractors did not reliably differ, $t(29) = 0.94$, p = .357, d_z = 0.17, although accuracy was numerically greater for the Target-Priority categories at both the immediate posttest (Target-Priority: M = 0.63, SD = 0.15; Distractor-Priority: M = 0.57, SD = 0.19) and delayed posttest (Target-Priority: M = 0.53, SD = 0.19; Distractor-Priority: M = 0.51, SD = 0.20).

Finally, as was expected, in the Equal Split learning condition, performance on categories designated as the Target-Priority list was similar to those designated as the Distractor-Priority list at both the immediate posttest (Target-Priority: M = 0.64, SD = 0.16; Distractor-Priority: M = 0.66, SD = 0.14) and delayed posttest (Target-Priority: M = 0.54, SD = 0.14; Distractor-Priority: M = 0.55, SD = 0.19). No reliable effect of priority list was found, $t(29) = -0.38$, p = .705, d_z = 0.07.

We also looked at the priority list by learning condition interaction in a different way, to gain insight into the trend across conditions. These results are shown in Figure 4, which displays group means for target-priority categories vs. distractor-priority categories by condition for both immediate and delayed posttest target-accuracy differences. Individual subject data varied considerably, but a single linear fit accounted well for the group mean data ($F(1, 4)$ = 58.56, p = .002) with an R^2 value of .94. The y-intercept for the combined data was -.018, very close to the theoretical value of no difference for target list vs. distractor list in the Equal Split condition. Even with such a small number of data points, the slope of .138 was highly reliably different from zero ($t(4) = 7.65$, p = .0015), indicating that as the target-frequency disparity between target-priority and distractor-priority categories increased, the performance difference for the target and distractor lists increased linearly. Separate linear fits for the immediate and delayed posttest data yielded remarkably similar parameter estimates (slope .142 vs. .135; y intercept -.019 vs. -.016), suggesting that, despite overall accuracy decrements from immediate to delayed tests, the difference between target and distractor list performance showed an approximately invariant effect in the two assessments.

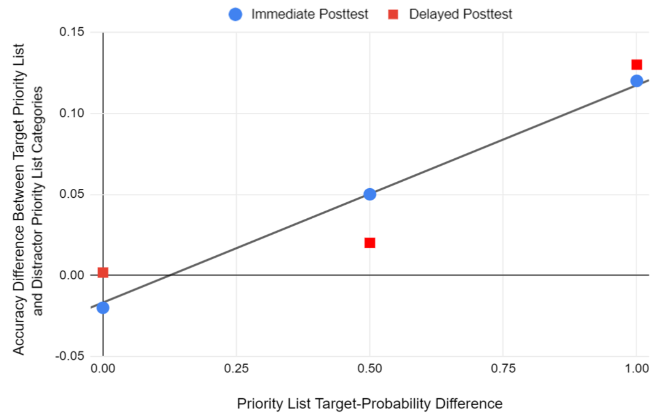

Figure 4: Mean difference in accuracy between Target-Priority and Distractor–Priority lists by Priority List Probability Difference in immediate and delayed posttests.

False Alarm Rates In this multi-category classification paradigm, misclassified trials were recorded as a "miss" for the presented category, a "false alarm" for the category whose label was incorrectly selected, and a "correct rejection" for all remaining 8 categories. A false alarm rate was calculated for each category and, consistent with the posttest accuracy analyses, was averaged across the list of categories prioritized as targets in learning and the list of categories prioritized as distractors. This enabled us to test whether superior performance on Target-Priority categories might have resulted from participants, when unsure, guessing those categories more often than Distractor-Priority categories.

A 2 (Priority List) by 2 (Posttest Phase) by 3 (Learning Condition) mixed measures ANOVA was conducted on false alarm rates. Consistent with the decrease in overall accuracy from immediate to delayed test, there was a reliable main effect of posttest phase, with more false alarms committed at delayed than at immediate posttest, $F(1, 87)$ = 171.22, $p < .001$, η_p^2 = 0.66. There was no main effect of priority list (p = .808) or learning condition (p = .579), and no reliable 2-way or 3-way interactions (all $p > .20$), suggesting that whether a category more frequently occupied a target or distractor position had little or no effect on subsequent classification response biases.

Sensitivity (d') Accuracy (hit rates) and false alarm rates were used to calculate sensitivity, measured as d', for each category before being averaged across priority lists. A 2 (Priority List) by 2 (Posttest Phase) by 3 (Learning Condition) mixed measures ANOVA was conducted on d' scores. There was a main effect of posttest phase, $F(1, 87)$ = 80.66, $p < .001$, η_p^2 = 0.48, indicating greater sensitivity at immediate posttest. There was a main effect of priority list, $F(1, 87)$ = 4.38, p = .039, η_p^2 = 0.05, such that d' was higher for items prioritized as targets rather than distractors. The priority list by learning condition interaction was found to be marginally significant, $F(2, 87)$ = 2.68, p = .074, η_p^2

= 0.06. All other two-way and three-way interactions were not reliable, all $p > .400$.

As expected given the observed differences in the accuracy and the lack of differences in false alarm rates, the general pattern of results for d' paralleled the accuracy results. Within the Always-Never condition, d' was significantly greater for categories prioritized as the target (immediate posttest: $M = 2.29$, $SD = .60$; delayed posttest: $M = 1.84$, $SD = .70$) than categories prioritized as the distractor (immediate: $M = 1.98$, $SD = .67$; delayed: $M = 1.51$, $SD = .73$), $t(29) = 3.31$, $p = .002$, $d_z = 0.60$. In the Often-Rarely condition, d' was numerically greater for Target-Priority categories (immediate: $M = 2.12$, $SD = .57$; delayed: $M = 1.65$, $SD = .66$) than Distractor-Priority categories (immediate: $M = 1.94$, $SD = .73$; delayed: $M = 1.63$, $SD = .72$), but this difference did not reach significance $t(29) = 1.29$, $p = .207$, $d_z = 0.15$. Finally, in the Equal Split condition, differences between Target-Priority list (immediate: $M = 2.21$ $SD = .61$; delayed: $M = 1.71$, $SD = .50$) and Distractor-Priority lists (immediate: $M = 2.26$, $SD = .60$; delayed: 1.72, $SD = .68$) did not differ, $t(29) = -0.30$, $p = .768$, $d_z = 0.05$.

Discussion

The present study investigated whether the role of categories in a paired comparison task influences learning. In all three learning conditions, Always-Never learning, Often-Rarely learning, and Equal Split learning, novice participants improved their ability to classify difficult skin lesion images after a short learning period. Notably, all learning conditions produced roughly equivalent learning gains.

Despite similar overall performance, there were clear differences between conditions in how overall accuracy was achieved. The Always-Never condition showed significantly greater learning gains, with a large effect size, for categories that appeared as targets than for those only shown as distractors, whereas no reliable differences were observed between targets and distractors in the Often-Rarely and Equal Split conditions. The numerical difference, t-value, and effect size in the Often-Rarely condition fell between those of the Always-Never and Equal Split conditions.

Notably, when any image appeared in any comparison trial, there was no information in the display that indicated its role as target or distractor. Only with feedback was the role revealed; even then, however, category feedback was given in the same way for both the target and distractor.

Importantly, in all conditions, categories designated as being part of the Distractor-Priority list produced classification accuracies that far exceeded chance performance, indicating that the inclusion of a category in a comparison, even if never as a target, is sufficient for significant learning gains to occur. This is consistent with prior research that suggests that simultaneous comparisons enhance the discriminability of both presented items (Gibson 1969; Mundy, Honey, & Dwyer, 2007, 2009).

Our assessment results do not suggest a process by which participants became aware of which categories had appeared as targets more often and developed a response bias to guess those category labels more frequently. False alarms in the posttests did not differ by experimental condition. Conversely, increased hit rates for categories prioritized as targets, with the absence of increased false alarms, indicate true improvements in category perception.

What might explain the target-distractor effects we observed? One possibility is that feedback may cause preferential encoding or retention of relevant stimulus information that has just been used to make a classification. When a trial is correctly answered, a process similar to reinforcement learning may strengthen the tendencies to select and weigh more heavily information just used to make that classification. When an error occurs, feedback may initiate signals that downweight the information relied upon. Interestingly, our results suggest that these processes are centered upon the category queried on a given trial, despite the fact that feedback always provided correct labeling for *both* members of the comparison presented. Classic work in PL (Gibson, 1969) suggested that PL can be described as coming to selectively extract distinguishing features -- features and relations that make the difference between one category and another. One might think that learning of distinguishing features would be an inherently symmetrical process: Learning the stimulus properties that make exemplars of Category A different from those of Category B would seem to be the same as learning the properties that distinguish B from A. An intriguing possibility compatible with the present results is that the relation may not be symmetrical. Given the way a comparison task is posed, a learning experience framed as "Choose the image that comes from Category A" may preferentially benefit learning to distinguish A exemplars from others more so than the reverse.

An alternative explanation is that learners may prioritize attention to feedback given for the target category, given its framing as the goal of the trial. Consequently, although learners may come to pick up distinguishing features for the distractor category in each comparison, if they fail to regularly attend to the category label presented for distractors, they may struggle to integrate what they learn across separate trials. In other words, the comparison of target category A and distractor category B provide learning gains for both A and B, and a subsequent comparison between target category C and distractor category B provide similar gains, but if a learner does not come to recognize that both comparisons contained an instance of B, then they may fail to integrate the distinguishing features they have extracted in each comparison.

Acknowledgments

We gratefully acknowledge support for this research from National Institutes of Health grant R01CA236791 to PK and CM.

References

Andrews, J. K., Livingston, K. R., & Kurtz, K. J. (2011). Category learning in the context of co-presented items. *Cognitive Processing, 12,* 161-175.

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.

CDC – Skin Cancer Statistics. (2023, June 8). Retrieved January 30, 2024 from: https://www.cdc.gov/cancer/skin/statistics/index.htm

Gentner, D., & Markman, A. B. (1997). Structure mapping in analogy and similarity. *American Psychologist, 52*(1), 45.

Gibson, E.J. (1969). *Principles of perceptual learning and development.* New York: Appleton-Century-Crofts.

Goldstone, R. L., Day, S., & Son, J. Y. (2010). Comparison. *Towards a theory of thinking: Building blocks for a conceptual framework*, 103-121.

Higgins, E. J., & Ross, B. H. (2011). Comparisons in category learning: How best to compare for what. In L. Carlson, C. Hölscher, & T. Shipley (Eds.), *Proceedings of the 33rd Annual Conference of the Cognitive Science Society* (pp. 1388–1393). Austin, TX: Cognitive Science Society.

Homa, D., Powell, D., & Ferguson, R. (2014). Array training in a categorization task. *Quarterly Journal of Experimental Psychology, 6*7(1), 45-59

Jacoby, V. L., Massey, C.M., & Kellman, P. J. (2024). Comparison training improves perceptual learning of skin cancer diagnoses. Poster presented at the *24th Annual Meeting of the Vision Sciences Society*, St. Petersburg, FL, May, 2024.

Kang, S. H., & Pashler, H. (2012). Learning painting styles: Spacing is advantageous when it promotes discriminative contrast. *Applied Cognitive Psychology, 26*(1), 97-103

Kellman, P. J., & Garrigan, P. (2009). Perceptual learning and human expertise. *Physics of Life Reviews*, *6*(2), 53–84.

Kellman, P. J., Jacoby, V., Massey, C., Krasne, S. (2022). Perceptual learning, adaptive learning, and gamification: Educational technologies for pattern recognition, problem solving, and knowledge retention in medical learning. In: Witchel, H.J., Lee, M.W. (eds) *Technologies in biomedical and life sciences education. Methods in physiology.* Springer, Cham.

Kellman, P. J., Massey, C. M., Krasne, S. & Mettler, E. (2023). Connecting adaptive perceptual learning and signal detection theory in skin cancer screening. In M. Goldwater, F. K. Anggoro, B. K. Hayes, & D. C. Ong (Eds.), *Proceedings of the 45th Annual Conference of the Cognitive Science Society*, 3251-3258.

Kok, E. M., de Bruin, A. B., Robben, S. G., & van Merriënboer, J. J. (2013). Learning radiological appearances of diseases: Does comparison help? *Learning and Instruction, 23*, 90-97.

Kornell, N., & Bjork, R. A. (2008). Learning concepts and categories: Is spacing the "enemy of induction"? *Psychological Science*, *19*(6), 585-592.

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.

Markman, A. B., & Ross, B. H. (2003). Category use and category learning. *Psychological Bulletin, 129*(4), 592.

Medin, D. L., Goldstone, R. L., & Gentner, D. (1993). Respects for similarity. *Psychological Review, 100*(2), 254-278.

Mundy, M. E., Honey, R. C., & Dwyer, D. M. (2007). Simultaneous presentation of similar stimuli produces perceptual learning in human picture processing. *Journal of Experimental Psychology: Animal Behavior Processes, 33*(2), 124-138.

Mundy, M. E., Honey, R. C., & Dwyer, D. M. (2009). Superior discrimination between similar stimuli after simultaneous exposure. *Quarterly Journal of Experimental Psychology, 62*(1), 18-25.

Payne, J. W., Bettman, J. R., & Johnson, E. J. (1993). *The adaptive decision maker*. Cambridge University Press.

Petrov, A. A., Dosher, B. A., & Lu, Z. L. (2005). The dynamics of perceptual learning: an incremental reweighting model. *Psychological Review, 112*(4), 715.

Rimoin, L., Altieri, L., Craft, N., Krasne, S. & Kellman, P. J. (2015). Training pattern recognition of skin lesion morphology, configuration, and distribution. *Journal of the American Academy of Dermatology,* 72(3), 489–495.

Shiffrin, R. M., & Schneider, W. (1977). Controlled and automatic human information processing: II. Perceptual learning, automatic attending and a general theory. *Psychological Review, 84*(2), 127.

Spalding, T. L., & Ross, B. H. (1994). Comparison-based learning: effects of comparing instances during category learning. *Journal of Experimental Psychology: Learning, Memory, and Cognition, 20*(6), 1251-1263.