Title
Comparisons in Adaptive Perceptual Category Learning

Journal
Proceedings of the Annual Meeting of the Cognitive Science Society, 44(44)

Authors
Jacoby, Victoria L
Massey, Christine M.
Mettler, Everett
et al.

Publication Date
2022

Peer reviewed

Comparisons in Adaptive Perceptual Category Learning

Victoria L. Jacoby (vjacoby@ucla.edu)[1]
Christine M. Massey (cmassey@psych.ucla.edu)[1]
Everett Mettler (mettler@ucla.edu)[1]
Philip J. Kellman (kellman@cognet.ucla.edu)[1]
[1]Department of Psychology, University of California, Los Angeles
405 Hilgard Ave, Los Angeles, CA 90095 USA

Abstract

Recent work suggests that learning perceptual classifications can be enhanced by combining single item classifications with adaptive comparisons triggered by each learner's confusions. Here, we asked whether learning might work equally well using *all* comparison trials. In a face identification paradigm, we tested single item classifications, paired comparisons, and dual instance classifications that resembled comparisons but required two identification responses. In initial results, the comparisons condition showed evidence of greater efficiency (learning gain divided by trials or time invested). We suspected that this effect may have been driven by easier attainment of mastery criteria in the comparisons condition, and a negatively accelerated learning curve. To test this idea, we fit learning curves and found data consistent with the same underlying learning rate in all conditions. These results suggest that paired comparison trials may be as effective in driving learning of multiple perceptual classifications as more demanding single item classifications.

Keywords: perceptual learning; categories; comparison; adaptive learning; face perception

Introduction

Many real-world learning tasks involve classification of objects, displays, or situations into perceptual categories, such as a person recognizing a face as belonging to a particular person or a dermatologist classifying a skin lesion into one of several diagnostic categories. Improvement in such tasks rests heavily on processes of perceptual learning – experience-driven improvements in the pickup of information (Gibson, 1969). There are a number of different ways in which perception improves with experience in some domain (Gibson, 1969; Goldstone, 1998; Kellman, 2002). Of these, learning to classify exemplars into perceptual categories requires the discovery and selective pickup of commonalities amongst members within a category, as well as the discovery of features and relations that distinguish different categories (Gibson, 1969; Homa & Chambliss, 1975). These improvements in the pickup of information are often accompanied by more rapid classification that occurs with reduced effort and attentional load.

Because of the importance of multi-category perceptual classification to many educational and training situations, there has been considerable interest in how to optimize this kind of learning in instruction and learning technology.

Comparisons

Considerable research has identified comparison as important for facilitating the learning of categories (Medin, Goldstone, & Genter, 1993; Spalding & Ross, 1994). Comparison has been studied in various ways including as a learning strategy by which multiple stimuli are assessed in terms of their common structural features (Gentner & Namy, 1999; Kurtz & Gentner, 2013; Lowenstein, Thompson & Gentner, 1999). One form of comparison consists of simultaneous presentation of information from the same or alternative categories. These simultaneous comparisons may be especially helpful in perceptual learning tasks involving categories with variable instances that share common structural characteristics. Concurrent presentation of items from different categories may allow learners to more easily discover distinguishing characteristics. Evidence suggests that such concurrent exposure can lead to improved differentiation for those items later on (e.g., Mundy, Honey, & Dwyer, 2007, 2009), as well as enhanced transfer performance in perceptual and category learning paradigms (Andrews, Livingston, & Kurtz, 2011; Carvalho & Goldstone, 2014; Higgins & Ross, 2011; Homa, Powell, & Ferguson, 2014; Kurtz, Boukrina & Gentner, 2013).

Recent research has begun to study the effectiveness of simultaneous comparisons in the context of adaptive learning. In studies conducted by Jacoby, Massey, and Kellman (2021), participants received a combination of single item classification and simultaneous comparison trials. Most trials involved single classification trials, but when two errors involving the same pair of categories were made, an *adaptively-triggered comparison* trial was generated. On these trials, participants were presented with a category label and instructed to choose between two images from the confused categories before resuming standard trials. Ultimately, when compared to a condition containing only single classification trials, the inclusion of comparison trials resulted in faster and more efficient learning. A separate study showed that adaptively-triggered comparisons were more effective in enhancing learning than an equal number of non-adaptive comparison trials inserted into the learning phase. The benefit of adaptively-triggered comparisons was attributed to their providing targeted opportunities to identify distinguishing features and resolve learner-specific confusions between categories.

Whereas comparison trials that are presented when learners struggle with certain category discriminations are helpful, an interesting question is whether using all comparison trials throughout learning might be as good or even better than presenting individual classifications with selective comparisons. Unlike most active learning

In J. Culbertson, A. Perfors, H. Rabagliati & V. Ramenzoni (Eds.), *Proceedings of the 44th Annual Conference of the Cognitive Science Society*.

approaches which require the classification of each presented item, comparison trials restrict participants to choosing between a limited set of options (often two). Without the cognitive load of considering all possible categories, this format may enable participants to devote more attention to extracting perceptual invariants that will advance learning. However, there is a concern that learning in this format alone may be too easy to produce meaningful, long-lasting learning. In particular, it may be difficult to transition from differentiating between only two items to making future classification judgments across all learned categories. Additionally, with guessing rates at 50% for paired-comparison trials, there is the potential to induce a misleading sense of fluency, as participants may be able to progress quickly through learning based partially on chance responding. In the present work, we evaluated the utility of paired-comparisons for learning relative to single item classification trials in which learners must identify an item with one out of a large set of possible response categories.

Adaptive Spacing

In the present study, all conditions employed an adaptive learning system that spaced categories as a function of the individual's accuracy and response times. Detailed descriptions of this adaptive response-time based scheduling (ARTS) system can be found elsewhere (Mettler, Massey & Kellman, 2011, 2016).

Prior research has demonstrated a benefit of spaced item presentation on the learning and transfer of perceptual categories, (Kang & Pashler, 2012; Kornell & Bjork, 2008), and the application of adaptive methods have been evidenced to amplify this benefit (Mettler & Kellman, 2014). While adaptive systems have been successfully tested in a variety of perceptual learning domains, they have not been systematically examined with respect to type of classifications that learners make on each trial. Assessing the present learning formats within an adaptive context could contribute to understanding of learning processes and also yield useful information about presentation modes in learning technology systems.

We assessed learning and transfer performances across three types of learning conditions: an All Comparisons condition, a Single Classification condition, and a Dual Classification condition, in which participants chose a name to go with each of two faces presented side by side. Dual classification was included as a hybrid format that might leverage possible learning benefits of both single item classifications and paired comparisons. Following learning, participants were tested on their ability to classify previously seen as well as novel images of the learned face categories.

Method

Participants

75 total participants were recruited from the University of California, Los Angeles subject pool to participate in this experiment.

Stimuli

The stimuli used were five distinct pictures of 22 different human male faces for a total of 110 unique images taken from a larger database (Min, Kose, & Dugelay, 2014). Four images of each of the 22 categories were used in the learning phase. Non–face details such as hairstyle or visible clothing could vary across images within the same category; however, the distance from camera, background, and final image size (256 X 256 pixels) remained identical across all exemplars. Variations in pose and non-face details encouraged participants to focus on learning perceptual structure of faces rather than specific image details. The fifth image in each category was set aside for use as a novel stimulus in immediate and delayed posttests.

Each learning category consisted of face images from the same person, and each category was identified with a name. The names were chosen to be unremarkable, and were taken from the Social Security list of most common names given in the United States in 2000-2009.

Design & Procedure

Each participant was assigned to one of three learning conditions. All participants completed a learning phase followed by an immediate posttest. Delayed posttests were completed one week later.

In the Single Classification condition, on each trial, one face was presented with all 22 possible name options organized alphabetically below. Participants selected the name corresponding to the face presented. In the All Comparisons condition, two faces, each from a separate category, were presented side-by-side under the prompt "Which one is [Category Name]." Participants were instructed to click on the image of the person named. Immediate feedback was given, with the correct label being presented for both images. In the Dual Classification condition, two faces were presented side by side, as in the All Comparisons condition, but participants chose the correct name for each face, as in the Single Classification condition. Participants could compare the two faces in the Dual Classification condition, but they were not directly instructed to do so. Figure 1 shows the layout for the 3 trial types.

In all conditions, participants were given 40 seconds to complete each learning trial and up to 10 seconds to view feedback. Feedback was given in the form of a green checkmark appearing alongside their answer choice when correct and a red 'X' appearing when incorrect. Additionally, when an incorrect answer was given, the correct name label appeared below the face.

Categories were adaptively scheduled and interleaved through the Adaptive Response Time-Based Sequencing (ARTS) system. During learning, each learning item is assigned a priority score indicating the relative benefit of that item appearing on the next learning trial. The priority score for each item, updated after every trial, was a function of learner accuracy, response times, trials elapsed since last presentation, and progress toward meeting mastery criteria.

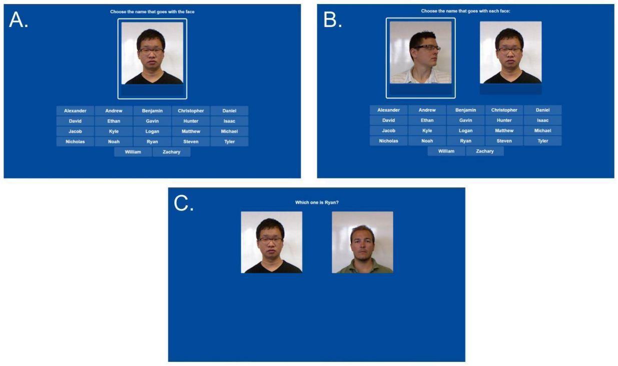

Figure 1. Trial layout for the three learning conditions. A. Single Classification ("Choose the name that goes with the face") B. Dual Classification ("Choose the name that goes with each face") C. All Comparisons ("Which one is Ryan?")

(See Mettler, Massey & Kellman, 2016 for computational details.) The sequencing algorithm presented the highest priority item on each trial. Incorrect responses indicated low learning strength for an item and generated a high priority for recurrence. However, ARTS also utilizes an enforced delay, such that even a high priority item cannot recur while feedback from a recent instance still resides in working memory. In the present study, we used an enforced delay of 3 classifications with some "jitter," such that the delay was sometimes 2 or 4 (each 25% of the time). As an individual's learning strength for a given learning item increased (indicated by accuracy and lower RTs), the ARTS algorithm automatically generated lower priority, and longer recurrence intervals, as an inverse function of the logarithm of reaction time. For the All Comparisons condition, where two faces were presented but only one response was made, one face category was considered the target on each trial, and accuracy and RT from that trial influenced only the scheduling of that category.

In all conditions, a category was considered mastered and subsequently retired from the learning set after four consecutive correct classifications were made. Retired categories only re-emerged when necessary to achieve correct spacing intervals for the remaining, unmastered categories. Immediately following learning, participants were administered a posttest. One previously seen exemplar per category, as well as one novel exemplar per category, were randomized and presented sequentially for classification. The layout of each test trial was identical to the learning trials in the Single Classification condition; however, no feedback was given during testing. A delayed posttest, administered one week later, was identical in content and structure to that of the immediate posttest.

Dependent Measures and Data Analysis

For each participant, we recorded the number of classifications invested in learning to achieve mastery for each category, the total time invested in learning, and accuracies on the immediate and delayed posttests. Learning efficiencies were also calculated. Efficiency measures provide a way of measuring learning that incorporates variations in both learning and posttest performances. Because mastery criteria were used, participants differed in the number of trials and time invested in learning. Efficiency combines accuracy gain and the investment in learning by dividing accuracy gains by classifications or time invested. A classification-based efficiency measure divided posttest accuracies by the number of classifications made for each participant. For this measure, we counted each trial in the Dual Classification condition as two classifications. A second approach looked at the total amount of *time* invested in the learning phase. This time-based efficiency measure was calculated for each participant by dividing posttest accuracy by time, measured in minutes. As an additional method for comparing conditions, we modeled learning rates during training by fitting exponential functions for each participant (Dosher & Lu, 2007; Cochrane & Green, 2021). Initial levels of knowledge were assumed to be zero, asymptotic learning was assumed to be 100%, and the learning time was used along with posttest percent correct to obtain the single free parameter of learning rate for each participant. Condition comparisons for all measures were compared using ANOVA and other standard parametric statistical methods. Since we sought to compare differences among the conditions, we conducted planned comparisons between pairs of conditions for each measure. Previous research on adaptive comparisons (Jacoby, et al., 2021) revealed condition differences with medium to large effect sizes; we anticipated similar effect sizes in the present study.

Results

We first examine the differences among conditions in the number of learning classifications and time required to reach mastery criteria during learning and in accuracy scores on posttests. We then turn to our primary measure of learning efficiency which relates investment during learning to posttest performance as a rate and is a particularly informative measure when all participants learn to criterion. Finally, we report the results of modeling the average learning rates for each condition to assess the degree to which the learning curves were similar or divergent.

Learning Measures

Learning classifications to criterion differed across the three conditions and favored the All Comparisons condition. A 3-way ANOVA on classifications invested by learning condition revealed a reliable main effect of condition, $F(2, 74) = 11.92$, $p < .001$, $\eta_p^2 = .244$. Contrasts between conditions revealed a reliable difference between the mean number of classifications made in the the All Comparisons condition ($M = 136.88$, $SD = 38.23$) and Single Classification condition ($M = 235.28$, $SD = 97.78$), $t(48) = 4.69$, $p < .001$, $d = 1.32$, as well as between the All Comparisons condition and the Dual Classification condition (M = 220.44, SD = 81.77), $t(48) = 4.63$, $p < .001$, $d = .83$.

Total time invested in learning was assessed in a separate ANOVA. We observed a reliable main effect of condition, $F(2, 74) = 18.57$, $p < .001$, $\eta_p^2 = .334$, with reliable differences between the mean time (min) invested in the All Comparisons condition (M = 16.36, SD = 8.51) and the Single Classification condition (M = 34.56, SD = 12.94, $t(48) = 5.876$, $p < .001$, $d = 1.66$, as well as between the All Comparisons condition and Dual Classification condition (M = 29.92, SD = 11.02), $t(48) = 4.87$, $p < .001$, $d = 1.34$.

Accuracy Measures

Accuracy in both the immediate and delayed posttests was highest for the Single Classification condition, followed by the Dual Classification condition, and then the All Comparisons condition. These patterns were confirmed by the analyses. A 3 X 2 mixed factor ANOVA with a between-subjects factor of learning condition and a within-subjects factor of posttest phase showed a main effect of posttest phase, $F(1, 72) = 182.85$, $p < .001$, $\eta_p^2 = .717$ as well as a main effect of learning condition, $F(2, 72) = 17.87$, $p < .001$, $\eta_p^2 = .332$, and no reliable interaction. At immediate posttest, accuracy was highest for the Single Classification condition (M = .88, SD = .10), followed by the Dual Classification condition (M = .84, SD = .14), and the All Comparisons condition (M = .62, SD = .20). There was a reliable difference between the Single Classification and All Comparisons conditions, $t(48) = 6.02$, $p < .001$, $d = 1.70$, as well as between the Dual Classification and All Comparisons conditions, $t(48) = 4.50$, $p < .001$, $d = 1.27$.

At the delayed posttest, accuracy was again highest for the Single Classification condition (M = .64, SD = .19), followed by the Dual Classification condition (M = .63, SD = .19), and then the All Comparisons condition (M = .41, SD = .20). The difference between the Dual Classification condition and All Comparisons condition was significant, $t(48) = 3.92$, $p < .001$, $d = 1.11$, as was the difference between the Single Classification and All Comparisons condition $t(48) = 4.09$, $p < .001$, $d = 1.15$.

Subsequent analyses examined posttest performance broken down into old versus novel items. At immediate posttest, there was a reliable main effect of item type, such that all conditions performed better on old items than novel items, $F(1, 72) = 40.125$, $p < .001$, $\eta_p^2 = .358$. This pattern was also present in the delayed posttest, with old items being classified correctly more often than novel items, $F(1, 72) = 10.84$, $p = .002$, $\eta_p^2 = .131$. There was no interaction found between condition and item type at either phase of the posttest, both $p > .130$.

Efficiency Measures

Efficiency comprised the primary measure in this work, because it combines the classifications or time invested in learning and the learning outcome of accurate identification performance. It may be considered a rate measure of learning.

Figure 2 shows the efficiency results based on time by condition on the immediate and delayed posttests. For convenience, we express this measure in terms of accuracy gain per 10 min of learning time. At the immediate posttest, efficiency based on time favored the All Comparisons condition such that for every 10 minutes invested there was a 47.45% mean increase in posttest accuracy (SD = 27.33), whereas those in the Single Classification condition saw a 28.90% (SD = 10.50) increase, and those in the Dual Classification condition saw a 32.36% (SD = 13.51) increase. A 3 x 2 ANOVA on condition and posttest version showed a main effect of posttest phase, $F(1, 72) = 119.62$, $p < .001$, $\eta_p^2 = .624$, and a main effect of condition, $F(2, 72) = 5.33$, $p = .007$, $\eta_p^2 = .129$, as well as a significant condition by posttest interaction, $F(2, 72) = 6.55$, $p = .002$, $\eta_p^2 = .154$. A test of simple main effects revealed significant differences in efficiency at both the immediate posttest, $F(2, 72) = 7.02$, $p = .002$, $\eta_p^2 = .159$, and delayed posttest, $F(2, 72) = 3.25$, $p = .045$, $\eta_p^2 = .129$. Planned contrasts revealed reliable differences between the All Comparisons condition and the Single Classification condition, $t(48) = 3.17$, $p = .003$, $d = .84$, as well as between the All Comparisons and Dual Classification conditions, $t(48) = 2.48$, $p = .017$, $d = .70$.

At the delayed posttest, time-based efficiency again favored the All Comparisons condition such that for every 10 minutes invested in learning, we observed a 32.65% (SD = 23.37) mean increase in accuracy. Those in the Single Classification condition saw a 21.33% (SD = 93.75) accuracy increase, and those in the Dual Classification condition saw a 24.48% (SD = 12.17) increase. The All Comparison condition reliably exceeded the Single Classification condition, $t(48) = 2.23$, $p = .030$, $d = .63$. The difference between the All Comparisons and the Dual Classification conditions was not statistically reliable, $p = .128$, nor was the difference between the Single Classification and Dual Classification conditions, $p = .317$.

Based on classifications completed in learning rather than time, the mean efficiency rate at immediate posttest was highest for the All Comparison condition (M = .0051, SD = .003), followed by the Dual Classification condition (M = .0043 SD = .0016), and the Single Classification condition (M = .0043 SD = .002. At the delayed posttest, efficiency was again highest in the All Comparisons condition (M = .0034, SD = .002), followed by the Dual Classification condition (M = .0033, SD = .002), and then the Single Classification condition (M = .0032, SD = .002). However, these differences at both the immediate and delayed posttests were not reliable. A 3 X 2 mixed factor ANOVA on learning condition and posttest phase revealed a significant main effect of posttest phase, $F(1, 72) = 128.73$, $p < .001$, $\eta_p^2 = .641$, as well as a significant condition by posttest interaction, $F(2, 72) = 3.43$, $p = .038$, $\eta_p^2 = .087$. However, tests of simple main effects did not reveal a significant difference among conditions at either phase of the posttest, both $p > .280$.

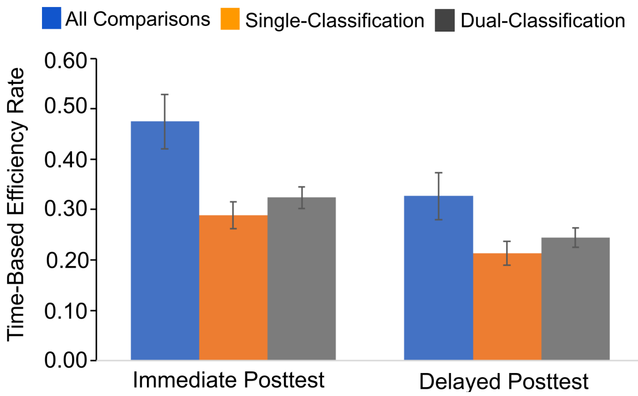

Figure 2. Time-based efficiency in immediate and delayed posttest by learning condition.

Learning Rate

The efficiency results suggest that the All Comparisons condition actually produced the best learning. If so, this would be a remarkably useful finding, in that individual trials requiring the learner only to make a two-choice discrimination are relatively easy, and the time to criterion was shortest. Individual trials took about 8.82 sec in the Single Classification condition but only 7.17 sec in the All Comparisons condition. (These times include the time learners spent examining feedback.)

Despite the appeal of the possible finding that using all comparisons might be a superior approach to learning, both in ease and efficiency, we were concerned that this result might be somewhat illusory. All conditions used the same mastery criterion of 4 consecutive accurate responses (for a category), but this criterion was easier to achieve in the All Comparisons condition, due to a chance accuracy rate of .50 (as compared to 1/22 in the classification conditions). Note that this difference in chance accuracy was not inadvertent; it was a consequence of the research goal to test paired comparisons vs. multi-category item classification. However, it allowed the All Comparisons condition to achieve apparently higher efficiencies, despite lower accuracy at posttest, due to a much lower number of learning classifications or time.

These considerations are especially important given that learning curves tend to be negatively accelerated. A condition that is actually proceeding along the same learning curve might actually appear to give "more bang for the buck" if it stops at an earlier point in the course of learning than another condition that continues longer. Yet, these might be equivalent as learning conditions in that they are sampling from the same underlying rate of learning.

To investigate this possibility, we modeled learning curves in the three conditions based on their learning investments and posttest outcomes. Rates of learning were modeled using a simple exponential function relating the time spent in learning to the performance level at immediate posttest (c.f., Dosher & Lu, 2007; Heathcote, Brown & Mewhort, 2000). Posttest accuracy is given as:

$$posttest\ accuracy = (1 - i) - e^{(-at)} \qquad (1)$$

where i is the initial level of knowledge (set to zero for this experiment), a is the learning rate and t is time. This analysis allowed us to compare conditions because participants in all conditions took a common posttest and all were measured for time in training.

Equation 1 was fit to each participant's data, and the estimates of a were averaged to obtain each group's learning curve. Figure 2 shows the curves for each condition by time. A one-way ANOVA showed no reliable differences among conditions, $F(2, 72) = 0.58$, $p = .561$, n.s. When analyzed by classifications, rather than time, there was no significant main effect of condition on learning rates, $F(2, 72) = 1.70$, $p = .189$, but there were marginal advantages of both classification conditions over the All Comparisons condition ($t(48)=1.789$, $p = .080$ for the Dual Comparison condition, and $t(48) = 1.699$, $p = .096$ for Single Classification).

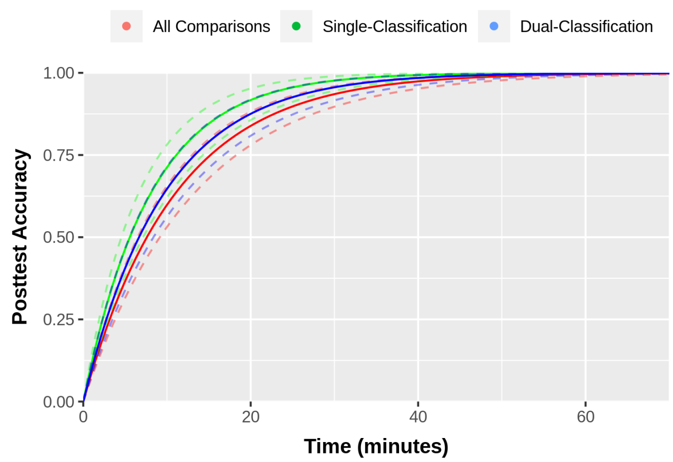

Figure 3. Learning curves by condition based on time. Dashed lines show curve fits for each condition based on a learning rate parameter that is +/- 1 standard error of the group mean learning rate parameter.

Discussion

Prior research has demonstrated that perceptual learning can be enhanced through the inclusion of comparison trials. However, little research has investigated the efficacy of learning exclusively through paired-comparison trials. The results of this experiment demonstrate that learning with all comparisons may be an equally viable approach to perceptual category learning as the more common classification-based trials.

Participants in all learning conditions showed significant learning and transfer to 22 distinct face categories. While prior research has traditionally been conducted with significantly fewer categories (e.g. 2 or 3), this finding provides evidence for the effectiveness of perceptual and adaptive learning methods in handling large quantities of unknown categories at once.

Use of all comparison trials led to faster learning as measured through classifications made and time invested, and time-based efficiency measures suggested an advantage of learning with all comparison trials relative to the other conditions. However, subsequent learning curve analyses did not show underlying differences, especially when learning was considered in terms of time invested. These results suggest that the learning trial formats tested here may be best characterized in terms of the same or similar, rather than divergent, learning rates.

The higher measured efficiency of the All Comparisons condition likely derived from the higher chance accuracy on comparison trials and the use of the same mastery criteria for all conditions. While this chance accuracy disparity was a necessary consequence of comparing two fundamentally different trial formats, participants in the All Comparisons condition were likely assessed on posttests before they had learned the material to the same strength as those in the other conditions. As the modeled learning rates suggest, had participants in the All Comparisons condition invested as much time or as many classifications in learning as those in the other conditions, they likely would have performed more similarly on the posttest as well. Future work will address rates of learning under equivalent durations of learning and conditions of category retirement.

The performance of the All Comparisons condition may be particularly impressive given the structure of the posttests. As described, the posttest trials were identical in format and task to the single classification learning trials (and thus also very similar to the dual classification trials); consequently when learners in these conditions reached the posttest, they were already familiar with both the layout of the trial and name options and had received repeated practice with the particular task. For those in the All Comparisons condition to have been able to perform as strongly as they did at testing, it suggests that learners were able to pick up the relevant information for face identification in a way that was long-lasting and transferable to a new task format.

Surprisingly, despite the reported benefits of simultaneous presentations, we observed no benefit of the dual classification trials relative to the single classification, sequential trials. There are a few possible explanations for this result. First, as previously mentioned, while the design of the dual classification trials does allow for comparisons between the presented categories, it is also possible to approach each classification independently. If participants focused on each face separately without considering the relation between the faces, potential benefits of the simultaneous presentation were likely not utilized. Recent research has also provided evidence to suggest that in classification-based learning, the goal of making successful classifications on a trial may actually detract learners from comparing presented items, thus undercutting the benefit of simultaneous presentation (Patterson & Kurtz, 2020). As previous research has shown an advantage to comparisons that direct attention to relevant features (Hammer, et al., 2008; Kurtz, Boukrina & Gentner, 2013; Kurtz & Gentner, 2013), it is unsurprising that simultaneous presentations that do not incentivize comparison do not produce discrimination benefits in learning.

An alternative explanation for why a difference was not detected may have to do with the interleaved presentation of categories used in this experiment. While we have focused on comparisons as occurring between simultaneously presented stimuli, it has been suggested that in paradigms with single, sequential trials, participants may make sequential comparisons between the currently viewed item and the most recently viewed item, or with prior information relevant to the structural relations between categories (Carvalho & Goldstone, 2015, 2017; Kang & Pashler, 2012; Kurtz & Gentner, 2013). Interleaved presentation of categories in the present experiment may have allowed participants in the Single Classification condition to attain some of the benefits of comparisons.

Despite significant research on the benefit of comparison in perceptual learning, category learning paradigms and applications in learning technology have relied heavily on single classification trials. The present research offers another effective approach to learning: paired comparison trials. The critical information learned through this trial type is shown to be long-lasting and generalizable. Intuitively, this format is comfortable and probably lower in cognitive load than considering multiple categories in single classification trials.

This study has limitations. While the modeled learning rates across conditions served to mitigate some concern regarding differences in terminal learning strength, they relied on the assumption that learning increases in accordance with the exponential model across all conditions and at all points in learning. If learning gains follow an alternative pattern, it may not be adequately accounted for in the present analyses.

Although the use and structure of comparison trials in perceptual learning deserves further research, the results of this study already suggest interesting implications for future designs of learning interventions for multi-category perceptual classification. More generally, as the role of perceptual learning in the development of expertise has become better recognized in a variety of learning domains, advances in understanding this form of learning and the conditions that optimize it have important implications for instruction and learning technology. The finding that effective learning can be based exclusively on paired-comparisons is both theoretically interesting and likely to be practically useful in applications of learning technology to real-world learning settings.

Acknowledgments

We gratefully acknowledge support for this research from National Science Foundation grant DRL 1644916 and National Institutes of Health grant R01CA236791 to PK and CM. Any opinions, findings, and conclusions are those of the authors and do not necessarily represent the views of the NSF or NIH.

References

Andrews, J. K., Livingston, K. R., & Kurtz, K. J. (2011). Category learning in the context of co-presented items. *Cognitive Processing,* 12(2), 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.

Carvalho, P. F., & Goldstone, R. L. (2015). What you learn is more than what you see: what can sequencing effects tell us about inductive category learning? *Frontiers in Psychology*, *6*, 505.

Carvalho, P. F., & Goldstone, R. L. (2017). The sequence of study changes what information is attended to, encoded, and remembered during category learning. *Journal of Experimental Psychology: Learning, Memory, and Cognition*, *43*(11), 1699.

Cochrane, A., & Green, C. S. (2021). Assessing the functions underlying learning using by-trial and by-participant models: Evidence from two visual perceptual learning paradigms. *Journal of Vision*, *21*(13), 5. https://doi.org/10.1167/jov.21.13.5

Dosher, B. A., & Lu, Z. L. (2007). The functional form of performance improvements in perceptual learning: learning rates and transfer. *Psychological Science*, 18(6), 531-539.

Gentner, D., & Namy, L. L. (1999). Comparison in the development of categories. *Cognitive Development*, *14*(4), 487–513.

Gibson, E.J. (1969). *Principles of Perceptual Learning and Development.* New York: Appleton-Century-Crofts.

Goldstone, R. L. (1998). Perceptual learning. *Annual Review of Psychology*, *49*(1), 585-612.

Hammer, R., Bar-Hillel, A., Hertz, T., Weinshall, D., & Hochstein, S. (2008). Comparison processes in category learning: From theory to behavior. *Brain Research*, *1225*, 102–118. https://doi.org/10.1016/j.brainres. 2008.04.079

Heathcote, A., Brown, S., & Mewhort, D. J. (2000). The power law repealed: The case for an exponential law of practice. *Psychonomic Bulletin & Review*, *7*(2), 185-207

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.), *Proc. of the 33rd Annual Conference of the Cognitive Science Society (pp. 1388–1393)*. Austin, TX: Cognitive Science Society.

Homa, D., & Chambliss, D. (1975). The relative contributions of common and distinctive information on the abstraction from ill-defined categories. *Journal of Experimental Psychology: Human Learning and Memory*, *1*(4), 351.

Homa, D., Powell, D., & Ferguson, R. (2014). Array training in a categorization task. *Quarterly Journal of Experimental Psychology*, *67*(1), 45-59.

Jacoby, V. L., Massey, C. M., & Kellman, P. J. (2021). Enhancing perceptual learning through adaptive comparisons. *Journal of Vision*, *21*(9), 2845-2845.

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. (2002). Perceptual learning. (3rd ed). In H. Pashler, & C. R. Gallistel (Eds.), *Stevens' Handbook of Experimental Psychology* (Vol. 3, pp. 259–299). New York: John Wiley & Sons.

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.

Kurtz, K. J., & Gentner, D. (2013). Detecting anomalous features in complex stimuli: The role of structured comparison. *Journal of Experimental Psychology: Applied*, *19*(3), 219–232.

Loewenstein, J., Thompson, L., & Gentner, D. (1999). Analogical encoding facilitates knowledge transfer in negotiation. *Psychonomic Bulletin & Review*, *6*(4), 586–597. https://doi.org/10.3758/BF03212967

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

Mettler, E., & Kellman, P.J. (2014). Adaptive response-time-based category sequencing in perceptual learning. *Vision Research*, *99*, 111-123.

Mettler, E., Massey, C. M., & Kellman, P. J. (2011). Improving adaptive learning technology through the use of response times. In L. Carlson, C. Hölscher, & T. Shipley (Eds.), In Proc. of the 33rd Annual Conference of the Cognitive Science Society (pp. 2532-2537). Austin, TX: Cognitive Science Society.

Mettler, E., Massey, C.M., & Kellman, P.J. (2016) A comparison of adaptive and fixed schedules of practice. *Journal of Exp. Psychology: General*, *145*(7), 897-917.

Min, R., Kose, N., & Dugelay, J. L. (2014). Kinectfacedb: A kinect database for face recognition. *IEEE Transactions on Systems, Man, and Cybernetics: Systems, 44*(11), 1534-1548.

Mundy, M. E., Honey, R. C., & Dwyer, D. M. (2007). Simultaneous presentation of similar stimuli produces perceptual learning in human picture processing. *J. 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.

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.

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.