UNIVERSITY OF CALIFORNIA SAN DIEGO

Learning to Categorize Objects

A dissertation submitted in partial satisfaction
of the requirements for the Doctor of Philosophy

in

Experimental Psychology

by

Sin-Heng Lau

Committee in charge:

Professor Timothy F. Brady, Co-Chair
Professor Harold E. Pashler, Co-Chair
Professor Kamalika Chaudhuri
Professor Scott Klemmer
Professor Edward Vul

2019

The Dissertation of Sin-Heng Lau is approved, and it is acceptable in quality and form for publication on microfilm and electronically:

_____

_____

_____

_____
Co-Chair

_____
Co-Chair

University of California, San Diego

2019

DEDICATION

I would like to dedicate this work to my grandmothers, who taught me about love, perseverance, and integrity; and to my parents, who have been supportive during my pursuit of scientific research.

TABLE OF CONTENTS

LIST OF FIGURES & TABLES

ACKNOWLEDGEMENTS

I would like to acknowledge the members of the Pashler lab and the Brady lab for their support over the past 5 years. It is a privilege to have worked along with all the talented and humble people in the labs. I would like to thank my advisors, Hal Pashler, for encouraging me to do research in cognitive psychology; and Tim Brady, for his unconditional support and guidance in my research. Furthermore, I thank all the undergraduate researchers for their work that make all these scientific discoveries possible.

I would also like to acknowledge the generous support from the Department of Psychology at UC San Diego.

Chapter 1, in full, is a reprint of the material submitted to a journal in revision: Lau, J. S. H., Casale, M. B., & Pashler, H. (in revision). Mitigating Cue Competition Effects in Human Category Learning. The dissertation author was the primary investigator and author of this paper.

Chapter 2, in full, was prepared to be submitted to a journal: Lau, J. S. H. & Pashler, H.. Simultaneous and Interleaved Training in Perceptual Category Learning. The dissertation author was the primary investigator and author of this paper.

Chapter 3, in full, was prepared to be submitted to a journal: Lau, J. S. H., Pashler, H. & Brady, T. F.. Low Target-Distractor Discrimination in Visual Search Promotes Detailed Target Template Generation. The dissertation author was the primary investigator and author of this paper.

VITA

2008 Bachelor of Cognitive Science, University of Hong Kong

2010 Masters of Philosophy in Psychology, Chinese University of Hong Kong

2019 Doctor of Philosophy in Experiment Psychology,
University of California San Diego

PUBLICATIONS

Rickard, T. C., Lau, J. S. H., & Pashler, H. (2008). Spacing and the transition from calculation to retrieval. Psychonomic Bulletin & Review, 15(3), 656-661.

Lau, J. S. H., & Huang, L. (2010). The prevalence effect is determined by past experience, not future prospects. Vision Research, 50(15), 1469-1474.

Lau, J. S. H., & Brady, T. F. (2018). Ensemble statistics accessed through proxies: Range heuristic and dependence on low-level properties in variability discrimination. Journal of Vision, 18(9), 3-3.

Lau, J. S. H., & Brady, T. F. (submitted). Intuitive Physics under Cognitive Load: Multiple Object Tracking Benefits from Realistic Physics. Manuscript submitted for publication. https://doi.org/10.31234/osf.io/6t859

Lau, J. S. H., Casale M. B., & Pashler, H. (under revision). Mitigating Cue Competition Effects in Human Category Learning

MANUSCRIPT IN PREPARATION

Lau, J. S. H., Pashler, H., & Brady, T. F. (in preparation). Low Target-Distractor Discrimination in Visual Search Promotes Detailed Target Templates Generation.

Lau, J. S. H., & Pashler, H. (in preparation). Simultaneous and Interleaved Training in Perceptual Category Learning.

ABSTRACT OF THE DISSERTATION

Learning to Categorize Objects

by

Sin-Heng Lau

Doctor of Philosophy in Experimental Psychology

University of California San Diego, 2019

Professor Timothy F. Brady, Co-Chair

Professor Harold Pashler, Co-Chair

Many people perform societally important categorization tasks as their full-time jobs, such as airport security personnel checking baggage, and dermatologists deciding whether skin moles are cancerous. These tasks usually require lengthy training to master. In my research program, I looked at real-world-inspired category learning tasks, and examined how the training process could be improved.

In Chapters 1 and 2, I had participants learn to classify objects into two categories. In Chapter 3, I had participants perform a visual search task and examined the learning effect in the task.

In Chapter 1, I studied people's category learning efficacy in the case of overshadowing. Overshadowing effect occurs when an object feature, while strongly associates with a category during training, does not appear in the transfer test. Learners tend to rely on the feature in the training phase, and do poorly due to its absence in the transfer test. I examined whether

overshadowing effect could be eliminated through top-down instructions.

Chapter 2 examined a new type of training schedule. The literature on category learning has primarily focused on using massed or interleaved learning as the training schedules, and discovered that each has its strengths and weaknesses. Here I proposed a new way to train learners, by presenting them with exemplars from each of the two categories simultaneously within each trial. Learning efficacy of the new training schedule is comparable to that of the interleaved schedule, while greatly reduces learners' frustrations during training. I will present data using artificial and real-world stimuli.

In the last chapter, I examined the effect of implicit learning during visual search. Participants were asked to search for a small set of targets among the same set of distractors repeated across trials. Three sets of distractors were manipulated across participants. After participants had been trained on a set of distractors, a new set of distractors was introduced. Search performance was worse when the target-distractor discriminability was higher in the second phase, indicating that distractor properties were learned during repeated visual search.

INTRODUCTION

The study of category learning has always been a central part of psychology (for a review, see Goldstone, Kersten, & Carvalho, 2017), and it spans across different disciplines within psychology. Scientists studying animal cognition try to pin-point the mechanisms of learning down to the molecular level, and understand the learning mechanisms that are common across species (e.g., Kosower, 1972). Developmental psychologists investigate how the ability to categorize objects develop over life span (e.g., Quinn, Slater, Brown, & Hayes, 2001). Cognitive psychologists develop strategies that optimize learning (e.g., Carvalho, & Goldstone, 2017). And social psychologists examine categorization known to affect social interaction, such as how people read other's emotion from their facial expression (e.g., Tajfel & Forgas, 2000).

Categorization from an Evolutionary Point of View

From an evolutionary perspective, the ability of categorization can hardly be undermined. Survival of an animal critically depends on some abilities to categorize objects, as earlier as a few hours after the animal is born. For one thing, human neonates born of a few hours old can distinguish stimuli that look like human faces, compared to non-face stimuli (Farroni, Johnson, Menon, Zulian, Faraguna, & Csibra, 2005). Their ability to recognize their primary caretakers, especially their mothers, develop within the first few days in their life. In the naturalistic environment, an animal's survival depends greatly on their ability to distinguish objects that are safe from those which are dangerous. An animal in the wild who cannot tell a nutritious food from poisonous ones is not expected to live for very long. There is also time pressure associated with making categorization judgments quickly, especially in the face of a predator. An inability to recognize the threat, or to trigger appropriate responses swiftly, can cost serious harm to an animal, if not its life.

Societally Important Categorization Task

In a modern human society, people develop skills to categorize objects as professions. Taxonomists, for instance, scrutinize the relationships between different types of living organisms, and establish systems that group organisms into categories. They define categories of organisms based not only on visual their visual features, but also draw references from phylogenetics, cladistics, and systematics to study the evolutionary history of the organisms. Multiple disciplines in medicine require years of training in categorizing various types of diseases. To name a few, dermatologists screen patients with skin conditions using mainly visual features. This skin conditions indicate benign deviations from health skin or malignant diseases that require special attention. Radiologists typically screen dozens of images a day. These images, captured with x-rays, ultra-sound, or other imaging methods, reveal the physiological conditions that are not visible from the body's surface. Accuracy in detecting each diseases is often correlated with experience in a highly specialized field. In the society, categorization tasks are often carried out to ensure public safety. Patrolling police officers are vigilant of suspicious pedestrians, making sure people with tendency to commit crimes are further scrutinized and removed from the streets. Security personnel at airport baggage checkpoints attempt to spot out restricted items from cluttered bags, while ensuring passengers do not get held for too long at the checkpoints that would lead to missing their flights.

Artificial Learning Systems and Human Categorization

Recent advance in computer science has refreshed scientists' interests in human category learning. The development of deep artificial neural network shows great promise in categorization tasks that has previously been deemed unachievable (LeCun, Boser, Denker, Henderson, Howard, Hubbard, & Jackel, 1989; Schmidhuber, 2015). In particular, with an enormous amount of training data, artificial systems can learn to categorize objects with accuracy approaching that of a seasoned expert trained in a specific domain. In the

development of artificial classifiers, human performance provide a few insights. First, it serves as a gold standard for an artificial system to compare its performance with. Second, the human perceptual system often inspires the development of the artificial ones. For one thing, the deep artificial neural network was once thought to mimic the hierarchical structure of an organic system. The convoluted neural network, on the other hand, resemble how receptive fields in human vision are organized (Ciresan, Meier, Masci, Maria Gambardella, & Schmidhuber, 2011). The development of artificial systems also inform psychologists and biologists about the human information processing.

Current Approach to Human Category Learning

In my research program, I study varies aspects of human category learning inspired by real-world scenarios. This approach allows me to study questions that have deep roots in situations that regular people encounter in the society every day. This approach has been increasing successful in generating theoretically interesting questions, which also have strong practical implications.

This dissertation has three parts. In the first two chapters, I had participants learn to classify objects into two categories. In Chapter 3, I had participants perform a visual search task and examined how target templates, a representation that participants hold in working memory, develops over time and under different training conditions.

Chapter 1 examines people's category learning efficacy in the case of overshadowing. The study of overshadowing effect has a long history in psychology, dating back to the first set of experiments reported by Palov (1927). The premise of category learning, either in human or non-human animals, is that the learner associate object features to a particular category in the learning process. Overshadowing effect occurs when an object feature, while strongly associates with a category during training, does not appear in the transfer test (Mackintosh, 1976; Rescorla & Wagner, 1972). Overshadowing happens frequently in daily-life learning

situations, when a learner is given many features that can potentially be used to learn about the categories. Often times, many of these features are not available once the learning phase is over (Rohrer, Dedrick, & Burgess, 2014). For example, in medical textbook, a disease is defined by many symptoms, including circumstantial ones. In the clinic, most patients do not show all of these symptoms. I am interested to examine how the absence of diagnostic symptoms affect learners' performance in a transfer test. A worse categorization performance indicate an overshadowing effect. It is important that this overshadowing effect is minimized in the training phase. I examined the effectiveness of possible top-down instructions that serve exactly this purpose.

Chapter 2 examines a new type of training schedule for category learning. The literature on category learning in the past two decades has primarily focused on using massed or interleaved learning as the training schedules, and discovered that each has its strengths and weaknesses (Carvalho & Goldstone, 2017). Massed training, which shows learners exemplars of the same category within a block of trials. Interleaved training, on the other hand, shows exemplars of different categories within a training block. Learners in the latter condition are often required to decide the category membership of a given stimulus. In short, massed training was found to facilitate creation of a category prototype, or the mean representation of a category. Interleaved training, on the other hand, allows generation of more flexible category boundaries. Here I propose a new way to train learners, by presenting them with exemplars from each of the two categories simultaneously within each trial. In this new training scheme, learners have the opportunity to compare features across categories. Memory and attention load is likely reduced compared to massed or interleaved training schedules. Learning efficacy of the new training schedule is comparable to that of the interleaved schedule, while greatly reduces learners' frustrations during training. I present data using artificial and real-world stimuli.

In the last chapter, I look into the dynamics of target template generation in visual search. Target templates are representations in the working memory that facilitate target

identification (Wolfe, Horowitz, Kenner, Hyle, & Vasan, 2004). The visual search literature has indicated that target templates can be created dynamically according to current task requirements (Barvo & Farid, 2012). In a set of experiments, I manipulated the discriminability between the targets and distractors, and examined how that affect search performance. In particular, three sets of distractors were manipulated across participants, while keeping the target set consistent. After participants had been trained on a set of distractors, a new set of distractors was introduced. Search performance was maintained when the target-distractor discriminability was higher in the training phase, regardless of the change in the test phase. On the contrary, if target-distractor discriminability was low (task was easy) in the training phase, performance dropped substantially once the task requirements were elevated.

The set of studies reported in the dissertation do not only answer theoretical questions that have not been addressed before, they also have immediate implications on real-world scenario. Understanding factors that promote and hinder human learning best allow educators and policy makers improve current education programs and curriculums, and as a result, an improved learning efficiency would benefit students of various kinds.

References

Bravo, M. J., & Farid, H. (2012). Task demands determine the specificity of the search template. Attention, Perception, & Psychophysics, 74(1), 124-131.

Carvalho, P. F., & Goldstone, R. L. (2017). The most efficient sequence of study depends on the type of test. Proceedings of the 39th Annual Conference of the Cognitive Science Society. (pp. 198-203). London, England: Cognitive Science Society.

Ciresan, D. C., Meier, U., Masci, J., Maria Gambardella, L., & Schmidhuber, J. (2011, July). Flexible, high performance convolutional neural networks for image classification. In IJCAI Proceedings-International Joint Conference on Artificial Intelligence (Vol. 22, No. 1, p. 1237).

Farroni, T., Johnson, M. H., Menon, E., Zulian, L., Faraguna, D., & Csibra, G. (2005). Newborns' preference for face-relevant stimuli: Effects of contrast polarity. Proceedings of the National Academy of Sciences, 102(47), 17245-17250.

Goldstone, R. L., Kersten, A., & Carvalho, P. F. (2017). Categorization and Concepts. In J. Wixted (Ed.) Stevens' Handbook of Experimental Psychology and Cognitive neuroscience, Fourth Edition, Volume Three: Language & Thought. New Jersey: Wiley. (pp. 275-317).

Kosower, E. M. (1972). A molecular basis for learning and memory. Proceedings of the National Academy of Sciences, 69(11), 3292-3296.

LeCun, Y., Boser, B., Denker, J. S., Henderson, D., Howard, R. E., Hubbard, W., & Jackel, L. D. (1989). Backpropagation applied to handwritten zip code recognition. Neural computation, 1(4), 541-551.

Mackintosh, N. J. (1976). Overshadowing and stimulus intensity. Animal learning & behavior, 4(2), 186-192.

Pavlov, I. P. (1927). Conditioned reflexes. An Investigation of the physiological activity of the cerebral cortex.

Quinn, P. C., Slater, A. M., Brown, E., & Hayes, R. A. (2001). Developmental change in form categorization in early infancy. British Journal of Developmental Psychology, 19(2), 207-218.

Rescorla, R. A., & Wagner, A. R. (1972). A theory of Pavlovian Conditioning. Variations in the effectiveness of Reinforcement and Nonreinforcement. In A. H. Black & W. F. Prokasy (Eds.), Classical conditioning II: Current research and theory. New York: AppletonCentury-Crofts.

Rohrer, D., Dedrick, R. F., & Burgess, K. (2014). The benefit of interleaved mathematics practice is not limited to superficially similar kinds of problems. Psychonomic Bulletin & Review, 21, 1323-1330.

Schmidhuber, J. (2015). Deep learning in neural networks: An overview. Neural networks, 61, 85-117.

Tajfel, H., & Forgas, J. P. (2000). Social categorization: Cognitions, values and groups. In C. Stangor (Ed.), Key readings in social psychology. Stereotypes and prejudice: Essential readings (pp. 49-63). New York, NY, US: Psychology Press.

CHAPTER 1: Mitigating Cue Competition Effects in Human Category Learning

Abstract

When people learn perceptual categories, if one feature makes it easy to determine the category membership, learning about other features can be reduced. In three experiments, we asked if this cue competition effect could be fully eradicated with simple instructions. For this purpose, in a pilot experiment, we adapted a classical overshadowing paradigm to demonstrate a robust cue competition effect with human learners. In Experiments 1 and 2, we created a new warning condition that aimed at eradicating the cue competition effect. The study shows that the cue competition effect can be mitigated through a top-down process. It also suggests that cue competition effect can be a stubborn roadblock in human category learning. Theoretical and practical implications are discussed.

Introduction

With adequate training, people can often master fairly sophisticated perceptual classification tasks, such as telling the species of a bird or diagnosing cancer from x-ray images. With perceptual categorization tasks arising in the real world, multiple features of an object are often each partly predictive of category membership. As an example, a grizzly bear in the wild can be distinguished from black bears based on its overall size, color, size of the shoulders, profile of the face and length of the claws. However, these attributes are not always available to the person who seeks to categorize objects in a naturalistic environment. For example, at dusk, color of the animal may not be well perceived, and a perceiver has to utilize other attributes she knows about the animal.

This raises a rather basic question: does having a highly salient predictor of category membership present during training hamper a learner's ability to learn other, less-predictive attributes? Putting this question into our grizzly bear example: if color of a bear is the most salient feature in distinguishing grizzly bears from black bears, do learners acquire less

information about less predictive features such as overall size, size of the shoulders, profile of the face, or length of the claws, if they learn in a situation in which color differences are available? This phenomenon is generally related to the cue competition effects, first reported by Pavlov (1927). Since then, the cue competition effects have been extensively studied. This broad family of effects include blocking effect (e.g., Kamin, 1968, 1969), and overshadowing effects due to frequency or saliency (Wagner, Logan, Haberlandt, and Price, 1968).

Cue Competition Effects

Within the framework of Pavlovian learning (Pavlov, 1927), multiple conditioned stimuli (CS) can be predictive of an unconditioned stimulus (US). As a concrete example, a whistle (CS1) and a tone from tuning fork (CS2) occurred in advance of the presentation of some food (US) to induce a dog's salivation (unconditioned response). The whistle and the tuning fork tone had similar psychological intensity. Upon pairing, the whistle and the tuning fork tone, presented in isolation, led to the same amount of salivation. In an Overshadowing condition, the whistle was made more salient than the tuning fork tone. After a considerable amount of training with the paired stimuli, when the dog was presented with the whistle alone, salivation was elicited. Presenting the weak tone alone, on the other hand, induced little salivation. This was taken as evidence that the conditioned stimuli, or cues, compete with each other for association with the unconditioned stimulus.

Other types of competition effects work similarly. In the well-known blocking paradigm (Kamin, 1968, 1969), instead of having stimuli with different levels of saliency, all stimuli have similar psychological intensity. One stimulus may be paired with the unconditioned stimulus prior to the rest of the stimulus set. Once the association has been established, the stimulus prevents new ones from pairing with the unconditioned stimulus.

Decades of research has shown that cue competition effects are quite robust in Pavlovian conditioning, human casual learning paradigms (e.g., Gluck & Bower, 1988; Shanks,

1991; Price & Yates, 1993, Vogel, Glynn, & Wagner, 2014), and human categorization tasks (e.g., Soto & Wasserman, 2010). Attributes that are less salient, occurs less frequently, or paired later with the unconditioned stimulus are less likely to be learned. Having variations in attribute qualities is undesirable if the goal is to learn as many features as they would be helpful in distinguishing the categories. On the other hand, the ability to select relevant features that are highly predictive of the outcome is crucial for survival in many naturalistic settings. It is perhaps an important skill for a learner to ignore the less predictive features for effective learning. The dynamics between the feature selection process and the top-down control over the process are the focus of the current study.

We set out to study a far less well explored question in the learning literature, which is whether the cue competition effects can be eradicated using top-down control. To answer the question, we employed a category learning paradigm. This question does not seem to be addressed in either animal or human studies (Blair, Watson, & Meier, 2009).

Top-down Control over Cue Competition Effects

In a classification task which some attributes have stronger psychological strength due to attribute saliency, frequency, or learning history, cue competition effects are likely to occur. It would be curious to know whether or not the effects can be controlled through top-down process.

In the animal learning literature, there has been a great effort attempting to understand the mechanisms of cue competition effects through computational modeling since Rescorla and Wagner (1972, also see Mackintosh, 1976). In the human learning literature, various computational models have been successful in explaining changes in learning outcomes due to base-rates (Gluck & Bower, 1988), cue validity or saliency (Kruschke, 1992; Nosofsky, Palmeri, & McKinley, 1994), or serial position of stimulus presentation. Through feedback, the weight

associated with each attribute is assumed to be adjusted. These weights are often termed attention and memory strength.

Take ALCOVE as an example (Kruschke, 1992): it has parameters to explain why some features are attended to, or memorized better from a bottom-up fashion. The attention parameter, in particular, explains how features with similar psychological strength can sometimes be attended to differently. For example, if the training is preceded by other form of learning, attention to some new features may be blocked. In the current instantiations, however, these models do not implement any parameters for top-down control. This would suggest that in the case of top-down involvement, the effects would be incorporated into the weights for attention and memory strength. An alternative explanation is to assume that top-down control has no effects on the learning outcomes. Our experiments would directly test these propositions.

Practical Motivation

Perceptual category learning has always been a central part of cognitive psychology due to its practical implications in education. The practical implications of cue competition effects extend even beyond explicitly perceptual tasks. As an example, it has been suggested that some variability in mathematical performance can be attributed to differences in ability to distinguish different problem types (Rohrer, Dedrick, & Burgess, 2014; Rohrer, Dedrick, & Stershic, 2015). In typical math textbooks and exercises, however, it is common to have similar problem types grouped together. As a consequence, the type of problem can be easily inferred from the chapter they are presented in, or from neighboring questions. This format for presentation deprived students of the opportunity to sort different problems into their corresponding types. In this classification problem, the chapter number serves as an overshadowing feature with complete validity, and the actual cues in the problems are not

utilized during practice. According to our hypothesis, students do not learn much about the cues to differentiate various problem types, and hence they perform poorly in their final exam.

It is fairly easy to envision learning mechanisms that would imply a positive or negative answer to this question. For example, it is often suggested that learning may be *error-driven* (e.g., Gluck & Bower, 1988; Kruschke, 1992; Nosofsky, Palmeri, & McKinley, 1994), in the sense that learners' internal representation of a category is adjusted only when they received feedback indicating that they misclassified an object. Having access to an easy predictor will presumably reduce the incidence of errors, and that in turn may prevent learning from taking place. On the other hand, it is easy to see how there might be functional value to learning about *all* potential predictors, given that some predictors may be available at one time and not at another.

A further question that we will examine here is whether people have voluntary control over how much they learn about weak predictors in the presence of a strong predictor. This question has not been examined before in any psychological research design, as far as we can tell. The basic question is quite concrete: If the presence of a strongly predictive feature inhibits learning about weaker predictors, can that inhibition be voluntarily "turned off" if the leaner is consciously motivated to learn as much as possible about *all* predictors that are present? This question has immediate practical importance because in many training situations, easy predictors are included (inadvertently or otherwise). For example, medical students learn to distinguish melanoma from benign skin lesions by examining photographs of different lesions that are depicted in different chapters of the book. The figure captions that label the images, and the title of the chapter the images are presented in, are all strong predictors making it immediately evident to the learner which category each training stimulus belongs to. Obviously, however, those cues will not be available in real clinical situations. If learners can completely suppress any interference from the labels, then there may be little cost to having them present

in training. If they cannot be suppressed, this might imply that standard training regimens could be usefully modified.

The potential for voluntary control is also of theoretical relevance. Acquisition-focused models such as the Rescorla-Wagner Model would seem to predict no voluntary control at all. According to those models, associations between the overshadowed conditioned stimulus (CS) and the unconditioned stimulus (US) are simply not learned during training. Performance-focused models, on the other hand, seem potentially compatible with top-down control over the overshadowing CS (Miller & Escobar, 2001). One formulation (the Comparator Hypothesis of Miller & Matzel, 1988) suggests that the association between the overshadowed CS and the US is acquired, but not expressed, in the presence of another stronger overshadowing CS (Denniston, Savastano, Blaisdell, & Miller, 2003). Top-down control is presumed to operate at the test, however.

Many readers will have noted that the questions posed here have important connections to a number of different experimental designs and literatures explored over the history of experimental psychology. These include studies of overshadowing in animal learning research going back as far as Pavlov's work, as well as in various studies of human concept and category learning. In the General Discussion, we will briefly describe some of these connections.

In the remainder of this introduction, we provide readers with a concrete overview of the designs and procedures used in the experiments reported below.

The Current Research

All the experiments reported in the current paper require human subjects to learn to classify visually presented stimuli into different categories. The stimuli varied on a number of continuous dimensions. Each study involved two phases, a learning phase and a transfer test phase.

As we were using a new human category learning paradigm, we would like to make sure that a clear cue competition effect was established with our paradigm. This worry was partly a response to the fact that cue competition effects in human category learning paradigms seem weaker and more inconsistent as compared to those found in traditional Pavlovian condition or human casual learning paradigms (Bott, Hoffman, and Murphy, 2007; Maes, Boddez, Alfei, Krypotos, D'Hooge, De Houwer, & Beckers, 2016; Murphy & Dunsmoor, 2017; but see Soto, 2018).

In our pilot experiment, human learners went through a supervised learning phase to learn to distinguish two artificial categories of objects. Subjects were assigned to one of several between-subjects conditions. In the Control condition, multiple features (values on specific continua) were independently and probabilistically predictive of category membership. In the Overshadowing condition, variation on an extra feature was included, which predicted category membership with 100% accuracy. Thus, if the learner learned the predictive significance of this feature they could potentially achieve 100% performance in the learning phase of the study without giving any weight whatsoever to the other features. A final test was included in which the extra feature was no longer present. In this pilot experiment, classification accuracy was markedly reduced in the final test for the Overshadowing condition, compared to the Control condition. This shows that people's learning of the less predictive features is not effective in the presence of a highly predictive one, indicating an overshadowing effect.

Having established a paradigm that shows a strong overshadowing effect, Experiment 1 was performed with two goals in mind. The first was to replicate the findings in the pilot experiment. The second was to test whether the overshadowing effect could be mitigated by top-down control. To this end, an additional Warning condition was included. The Warning condition was identical to the Overshadowing condition in every aspect, except that subjects were informed in advance of the learning phase that the perfectly predictive feature, while present in training, would not be available during the transfer test. In the transfer test phase,

subjects in all conditions were presented with the same types of test stimuli. Their task was to classify new specific examples that had never before been shown in training. Only the probabilistic features were present in this phase, and the deterministic predictor feature was not present. If the overshadowing effect could be mitigated through top-down control, performance of the Warning group should resemble that of the Control group, and be better than the Overshadowing condition. If top-down control was not effective, performance of the Warning and Overshadowing groups should be similar, and both groups should perform worse than the Control group.

Experiment 2 had the same design as Experiment 1. We replaced the stimuli in Experiment 1 with a set of visually more naturalistic ones. We expected to see the same general pattern of results as in Experiment 1. This would be evidence to indicate that the effect we found is not specific to a particular set of stimuli.

If the warning manipulation in Experiments 1 and 2 is effective, it would suggest that cue competition effects can be mitigated by top-down control. Such a procedure might have the potential to enhance people's flexibility in learning. Classifications of learned categories would be less dependent on a small subset of attributes and idiosyncratic environmental factors.

Pilot experiment

The goal of the pilot experiment was to establish a paradigm that shows a robust overshadowing effect in perceptual category learning. We also made use of the effect size of obtained to estimate the number of participants needed for the next 2 experiments. In subsequent experiments, we would try to eradicate this overshadowing effect by imposing top-down control.

Subjects were given two categories of cartoon-like artificial stimuli (referred to as "demons"), and their task was to classify the stimuli into two categories. They were given

feedback during the training phase, so that they could adjust their classification rules. In the transfer test, subjects classified a new set of stimuli without being given feedback.

Subjects were randomly assigned to the Control or Overshadowing condition in the training phase.

Two categories of "demons", labelled Old World and New World, were generated. Figure 1.1 shows examples of the demons. For the Control condition, the demons varied in three features: eye color, eye width, and horn height. The distribution of each feature within a category was Gaussian and the values were partially predictive of category membership. The separation between the means for the two categories was 1 standard deviation for each of the three features. These features varied independently. In general, Old World demons tended to have shorter horns, larger eyes, and their eyes tended to be blue or purple. New World demons tended to have taller horns and smaller eyes that tended to be green or blue.

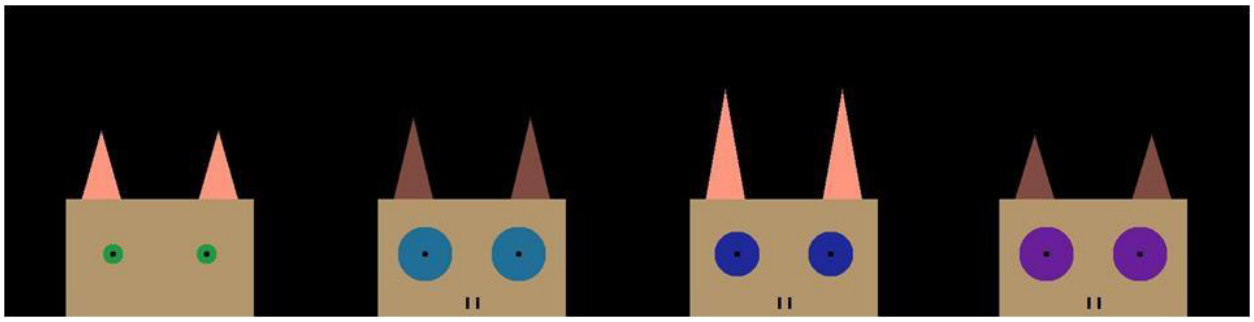


Figure 1.1. Exemplars used in the pilot experiment. The demon-like stimuli differed in eye color, eye size and horn size in the Control condition. In the training phase of the Overshadowing condition, the horn color was indicative of the category identities. The feature was not available during the transfer test.

For the Overshadowing condition, variations in the three features described above (for the Control Condition) was present. In addition, there were variations on one extra feature. Specifically, the horns for Old World demons were always in pink color, the horns of the New World demons were always in brown color.

The transfer test phase consisted of stimuli that were similar to those used in the Control condition, but the exact stimuli presented in the test had not appeared in the training phase.

Each demon had three probabilistic features that predicted its category membership. During transfer, the overshadowing feature had no variabilities: all demons had horns of the same color. Category membership was defined only with the three probabilistic predictors. No feedback was given in the transfer test.

Methods

Design.

The experiment contained two phases, a training phase and a transfer test. Two training conditions, Control and Overshadowing, were manipulated between-subjects. All subjects took a transfer test. The transfer test was identical to all subjects regardless of training conditions.

The proportion of correct classifications in both training and transfer test was recorded to assess subjects' learning.

Participants.

Seventy-six human subjects were recruited from our laboratory's online research subject pool, which includes adults of various ages living in a variety of countries. The pool has been pre-screened for excellent comprehension of English and careful attention to instructions. Subjects were randomly assigned into two training conditions: Control (n = 34) and Overshadowing (n = 42).

Materials.

The experimental environment was controlled by code written in Adobe Flash, and delivered through subjects' web browser. Subjects completed the experiment online using their own devices, so the absolute sizes of the stimuli on the screen could vary.

A total of 700 unique image files were generated, each depicting a 400 × 400-pixel demon. Three hundred images were used for training in the Control condition, another 300 images were used for training in the Overshadowing condition. The remaining 100 images were used in the transfer test for both conditions. Depending on whether the demon came from the

Old World or the New World, the values of the demon's eye color, eye size and horn height were randomly selected (with replacement) from their respective Gaussian distributions. Means of the two distributions for each feature were 1 standard deviation apart. To determine how that would translate to the predictive power of a probabilistic feature, we ran a simple simulation. If a subject relied on a single probabilistic feature for categorization and she acquires the distribution perfectly, she would achieve an accuracy of around 69% at the transfer test.

For training images used in the Control condition and the transfer test images, the value of the horn color was held constant. Training images used in the Overshadowing condition were identical to those in the Control condition, except that color of the horn was determined based on the demon's category membership.

A demon might or might not have a nose, but this feature did not predict whether it was Old World or New World.

Procedure.

The experiment started with a written introduction explaining the task to the subjects. Subjects were told to distinguish Old World from New World demons. They were told that Old World demons share some common features, as did the New World demons. They were encouraged to learn the category properties through feedback in the training phase, and advised that there would be no feedback in the transfer test. Subjects were also instructed to spend no more than 3 seconds with each demon. A short multiple-choice quiz then followed the introduction to ensure that subjects understood the task.

The training phase was divided into 6 blocks, with 50 trials in each block. Depending on the subject's condition, a training stimulus was drawn randomly from the respective training stimulus set without replacement. During each trial, a demon was shown at the center of the screen on a black background. The subject pressed one of the two keys on the keyboard to indicate the demon's category membership. The demon disappeared upon a response, or when 3 seconds had elapsed. A new trial began after a 1-second auditory feedback period and a 2-

second inter-trial interval (ITI). Subjects were encouraged to take short breaks in between the blocks.

Immediately after the training phase, subjects received instructions for the transfer test. They were told that feedback would not be given. They were also encouraged to use the same strategies they employed during training to classify the demons.

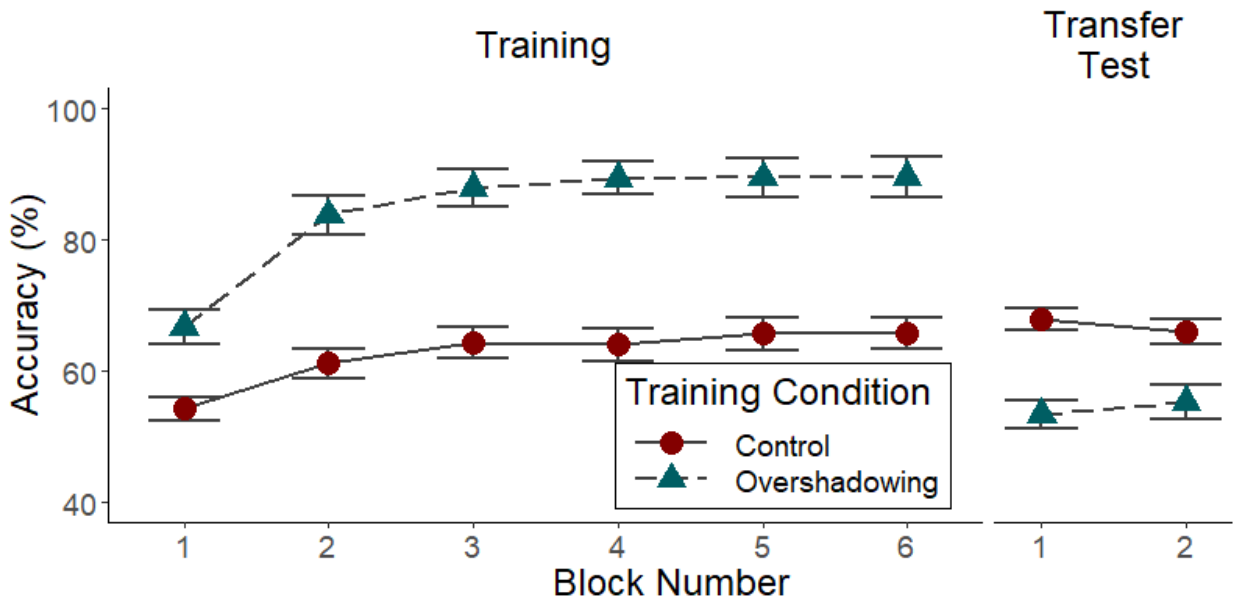


Figure 1.2. Performance in the Control and Overshadowing conditions over blocks in the pilot experiment. Subjects in the Overshadowing condition outperformed those in the Control condition during training, but the pattern reversed in the transfer test. Error bars denote between-subjects standard errors of the means.

Subjects in both conditions had the same set of test stimuli in the transfer test. The order of stimulus presentations in this phase was randomized for each subject. Each test trial began with the presentation of a test stimulus. Subjects could take as long as they needed to decide which category the demon belonged to. When the subject responded, no feedback was given and a new trial began after a 2-second ITI. Each subject completed 100 transfer test trials, broken down into 2 blocks.

Results

Training Phase.

The average accuracy for both Control and Overshadowing conditions increased steadily from Block 1 to Block 3, reaching an asymptote for its respective condition (Figure 1.2, left panel). Overall accuracy for the Overshadowing condition was higher than the Control condition throughout the training phase. This suggests that subjects in the Overshadowing condition made use of the horn color -- the deterministic predictor, in the classification process. Comparing the last two blocks of trials between conditions, subjects in the Overshadowing condition ($M = 89\%$, $SD = 19\%$) significantly outperformed those in the Control condition ($M = 66\%$, $SD = 13\%$), $t(74) = 6.16$, $p < 0.001$. As a measure of effect size for overshadowing, we computed Cohen's d which proved to be very large, $d = 1.42$, 95% C.I. = [0.90, 1.94].

Transfer Test.

The right panel of Figure 1.2 shows the overall performance for each condition during the transfer test blocks. Data from the last two training blocks and the transfer test was entered into a 2 (Training Condition) × 2 (Last 2 Training Blocks, Test Phase) ANOVA. A significant interaction between the two factors suggested that knowledge acquired in training with the two conditions was differentially transferred into the transfer test, $F(1,146) = 56.45$, $p < 0.001$. Specifically, subjects in the Overshadowing condition showed a substantial drop in performance from training ($M = 89\%$, $SD = 19\%$) to the transfer test ($M = 54\%$, $SD = 14\%$), $t(41) = 9.29$, $p < 0.001$. The drop, measured by effect size, was large, Cohen's $d = 1.43$, 95% C.I. = [0.94, 1.93]. Their performance at test, with an average 54% accuracy, was not different from chance, $t(41) = 2.02$, $p = 0.05$.

On the other hand, subjects in the Control condition showed no reliable differences in performance between training ($M = 66\%$, $SD = 13\%$) and transfer test ($M = 67\%$, $SD = 10\%$), $t(33) = 0.67$, $p = 0.52$. The change was negligible as measured by effect size, Cohen's $d = 0.11$, 95% C.I. = [-0.38, 0.60], indicating a successful transfer of knowledge learned in training to the transfer test. Difference between the two conditions was reliable, $t(74) = 4.42$, $p < 0.001$, Cohen's $d = 1.02$, 95% C.I. = [0.52, 1.51].

Discussion

In the pilot experiment, subjects in the two conditions received slightly different training. In the Control condition, subjects learned to classify exemplars of the two categories potentially relying on up to three different probabilistic predictor variables. In the Overshadowing condition, in addition to the probabilistic features that were available to the Control subjects, subjects also had a fourth feature available that had a completely deterministic connection to category membership. While this determinative feature enhanced performance in training, it impeded the learning of the other, partially predictive, features. Hence, in the transfer test, when the deterministic feature was no longer available, subjects in the Overshadowing condition performed poorly to a point that their performance was not reliably different from the 50% chance level. As we will discuss in the General Discussion, similarly large overshadowing effects have been observed in some animal studies of overshadowing (e.g., Pavlov, 1927), albeit using quite different task designs. Unlike other animals, however, human subjects can be instructed to try to learn despite the presence of the overshadowing feature. Experiment 1 examined the effect of this kind of top-down instruction.

Experiment 1

In Experiment 1, we first attempted to replicate the overshadowing effect observed in the pilot experiment with stimuli that were generated at the runtime of the experiment. This procedure made sure the stimulus set adhered to the generative process of the categories. Any idiosyncratic stimulus for a particular trial was unlikely to repeat. With this procedure, all participants got a different set of stimuli, while all of them belong to the same categories. The results obtained are more generalizable to a related context. This procedure is different from our pilot experiment that a set of images was pre-generated. It is also different from the standard practice in the field that a small set of stimuli (generally fewer than 10 exemplars in each category) was used.

Having established a robust overshadowing effect in the pilot experiment, we aimed to test if the effect could be mitigated by top-down attention control. In addition to the Control and Overshadowing conditions in the pilot experiment, a Warning condition was utilized. The Warning condition was identical to the Overshadowing condition, except that subjects were informed of the overshadowing feature before the training. They were instructed to ignore the overshadowing feature, the horn colors of the demons in this experiment. Subjects were told to try to learn other features that would help them classify the demons into the two categories.

If the top-down control could be deployed effectively, performance in the Warning condition could potentially resemble that of the Control condition in both training and the transfer test (or conceivably just in transfer test). On the other hand, if top-down control was ineffective, performance of the condition should resemble that of the Overshadowing condition.

Methods

The design of Experiment 1 was very similar to that of the pilot experiment, except that there was an additional between-subjects Warning condition. In addition, the study was run in the laboratory rather than online. Other changes made in Experiment 1 are detailed below.

Design.

As in the pilot experiment, the experiment had two phases, training and transfer test. Three between-subjects conditions were compared: Control, Overshadowing, and Warning. The Control and Overshadowing conditions were identical to those in the pilot experiment. In the Control condition, three probabilistic features were independently predictive of the stimuli's category memberships. In the Overshadowing condition, an additional binary feature was indicative of the stimuli's category memberships. The new Warning condition was identical to the Overshadowing condition, except that subjects were informed about the deterministic predictor before the training phase. All subjects took a transfer test in which new stimuli were shown.

Participants.

One hundred twenty participants (83 females, mean age = 21) from the University of California, San Diego Psychology Subject Pool participated for course credits. The number of participants was determined by a power analysis using the data from the pilot experiment. With a Cohen's d of 1.4 between the Control and Overshadowing conditions, an estimate of 9 participants per condition was needed to achieve a power of 80%. To ensure normality assumptions for the statistics tests, we decided to run at least 30 participants per condition.

All participants took a computerized color test, one participant did not pass the test and was replaced. Subjects were randomly assigned into one of the three conditions: Control (n = 41), Overshadowing (n = 36), and Warning (n = 43). Informed consent was obtained before the experiment began. All participants passed a computerized color test.

Materials.

As in the pilot experiment, two categories of demons were created. In both the training phase of Control condition and the transfer test of all conditions, the two categories of demons varied in eye color, eye size, and horn height. Demons in the training phase of the Overshadowing and Warning conditions had an additional binary feature -- each of the two horn colors mapped perfectly onto one of the categories. Once we generated the values of these features, we created a brown square on the screen which denoted the face of the demon. We then map the features onto the brown square to create an image of a demon.

Instead of having a fixed set of images for all the subjects, the stimuli were created with a generative process. The stimuli were created at the runtime of the experiment. Hence, no two stimuli were identical except by coincidence. The experimental environment was written in JAVA programming language.

Procedure.

The Control and Overshadowing conditions were identical to the corresponding conditions in the pilot experiment, except that each block was 100 trials long. The total numbers of trials in the training and transfer test phases were again 300 and 100, respectively. Subjects were seated in a normally-lit, sound attenuated room. In the training phase, they were told that the task was to classify the visual stimuli into two categories. They were encouraged to spend no more than 3 seconds on each trial, and the stimulus would be replaced by a white blank screen if no responses were given within 10 seconds. Audio feedback was given upon response.

For the Warning condition, the basic procedure was identical to the Overshadowing condition. Subjects were trained with the presence of the deterministic feature (horn color). The only difference between the two conditions was that subjects in the Warning condition were informed about the deterministic feature. They were told that one of the horn colors signified one category, the other horn color signified another category. They were instructed to ignore the horn color, and try to learn about the category properties through other the probabilistically defined features. Specifically, they were told:

> Old- and New-world Demons you see in Training will differ in their horn colors: Old-world Demons have one horn color and New-world Demons have another horn color. However, you should know that this horn color difference will not be present in the test. To maximize your performance in the final Test phase, please try to learn about the intrinsic differences between the Old- and New-world Demons. Just ignore their horn colors and attend to their other properties.

The same message was repeated before each block began. They were not told which other features defined the categories.

All subjects performed a transfer test which the demons' horns were always gray in color. Again, all stimuli were created according to the generative process. All subjects encountered different sets of stimuli. Hence, subjects had to classify the demons using the three probabilistic predictors.

Results

Training Phase.

Performance of the Control and Overshadowing conditions during training phase was very similar to that in the pilot experiment (Figure 1.3, left panel). Accuracy increased steadily across blocks. Performance of the Warning condition was superior compared to the Control condition, but inferior compared to the Overshadowing condition.

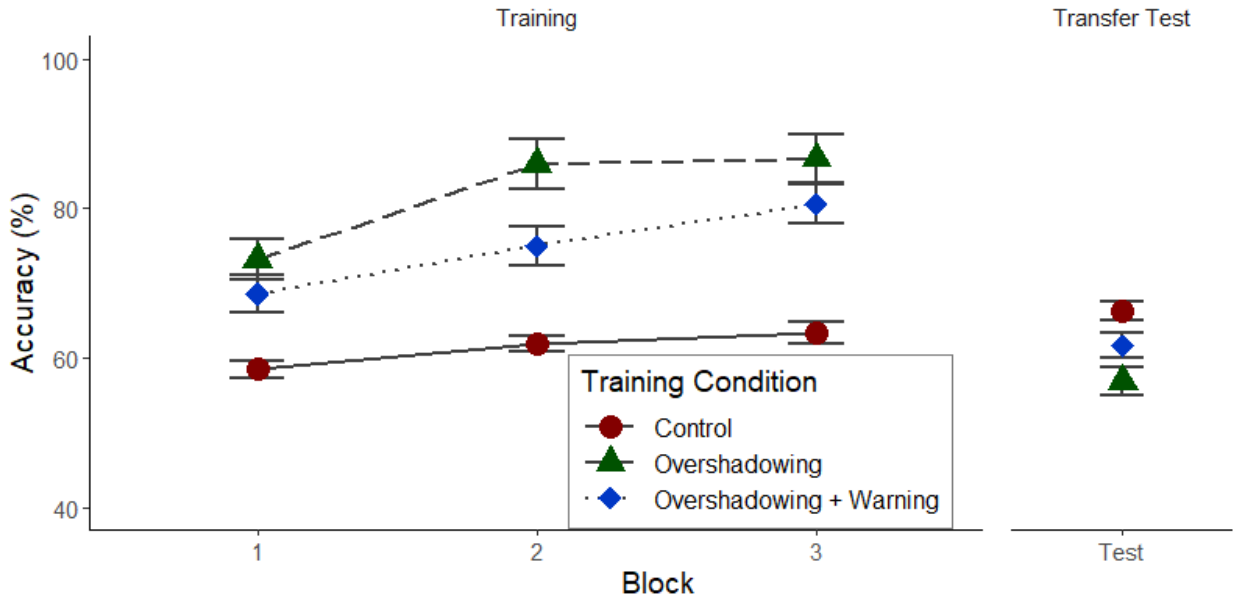


Figure 1.3. Subjects' performance in Experiment 1. In the last training block, those in Overshadowing condition performed the best, followed by those in the Warning condition. Those in the Control condition performed the worst. In the transfer test when overshadowing feature was no longer available, the pattern reversed. Error bars denote between-subjects standard errors of the means.

A 3 (Block) × 3 (Training Condition) mixed-design ANOVA was employed to analyze the training performance. A significant main effect of Block [$F(3,351) = 47.19$, $p < 0.001$] indicates improvements in performance across training blocks. A significant main effect of Condition [$F(2,117) = 15.71$, $p < 0.001$] indicates differential performance across training conditions. A significant interaction between the two factors [$F(6,351) = 21.42$, $p < 0.001$] provides evidence that the rates of improvement in training were different in the three conditions. Specifically, the

Overshadowing condition reached a performance asymptote quickly, while the other two conditions improved at slower rates.

In the last training block, the Control condition had the lowest performance and the least variability (M = 63%, SD = 9.5%), the Overshadowing condition had the highest performance but also greatest variability within condition (M = 87%, SD = 19.7%). Performance of the Warning condition was in between the two (M = 81%, SD = 16.7%). A one-way ANOVA [$F(2,117) = 23.42$, $p < 0.001$] indicates significant differences between these conditions. Tukey's honest significance test indicates reliable differences between the Control and Overshadowing conditions ($p < 0.001$, Cohen's $d = 1.54$, 95% C.I. for d = [1.02, 2.07]), and Control and Warning conditions ($p < 0.001$, Cohen's $d = 1.26$, 95% C.I. for d = [0.78, 1.75]). However, there was no reliable difference between Overshadowing and Warning conditions ($p = 0.20$, Cohen's $d = 0.34$, 95% C.I. for d = [-0.12, 0.79]).

Transfer Test.

Subjects' performance of the Control and Overshadowing conditions during transfer test (Figure 1.3, right panel) was very similar to that in the pilot experiment. Contrary to the high level of performance in the training phase, subjects in the Overshadowing condition (M = 56%, SD = 11%) did the worst, while those in the Control condition did the best (M = 66%, SD = 7.8%). Subjects in the Warning condition scored in between the two conditions (M = 62%, SD = 11.1%). Transfer test performance for all three conditions were significantly different from chance ($ps < 0.001$).

A one-way ANOVA [$F(2,117) = 8.43$, $p < 0.001$] indicates significant differences between these conditions. Tukey's honest significance test indicates a reliable pairwise comparison between Control and Overshadowing conditions ($p < 0.001$, Cohen's $d = 1.00$, 95% C.I. for d = [0.51, 1.49]), while the pairwise differences between Control and Warning conditions ($p = 0.086$, Cohen's $d = 0.49$, 95% C.I. for d = [0.04, 0.93]), and between Overshadowing and Warning conditions ($p = 0.098$, Cohen's $d = 0.43$, 95% C.I. for d = [-0.03, 0.89]) are not significant.

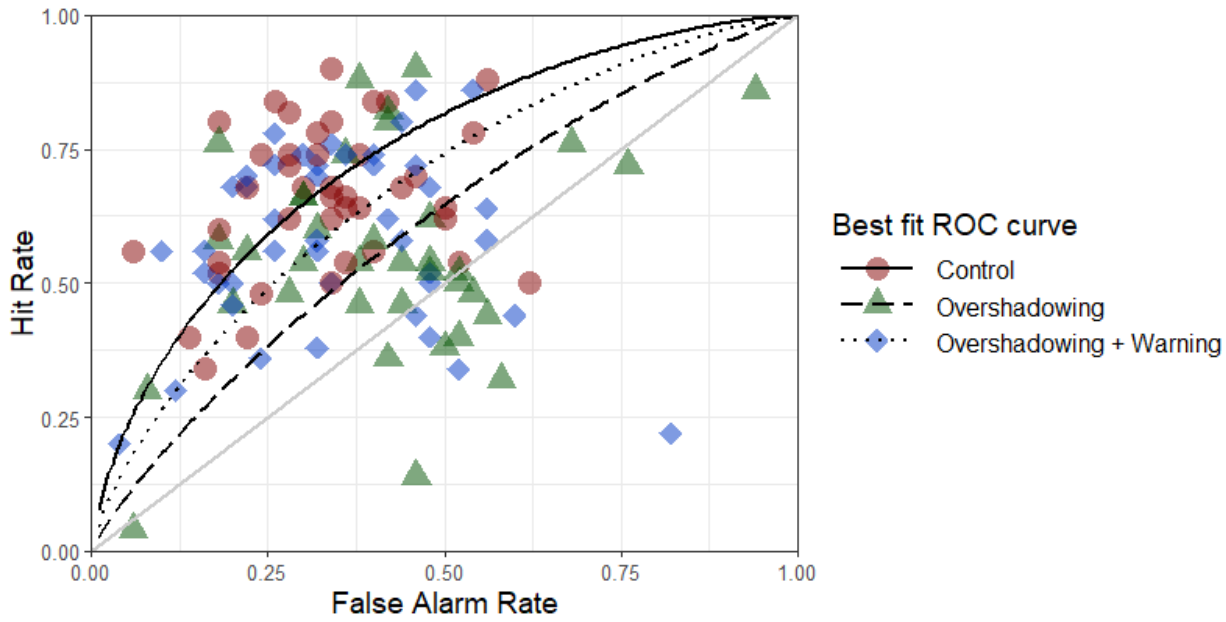


Figure 1.4. Receiver Operating Characteristic (ROC) analysis for test blocks in Experiment 1. Each point on the graph represents a subject's performance in the transfer test blocks. We denote hit rate as the proportion of correct classifications to the Old-word demons. The false alarm rate represents the proportion of incorrect classifications to New-word demons. The diagonal line represents chance level in discriminating Old World demons from New World ones.

Comparing performance of the final training block and the transfer test, a 2 (Final Training Block, Test Block) × 3 (Training Condition) mixed-design ANOVA was conducted. It suggests that the changes in performance from the final training block to the transfer test block in the three conditions were different [$F(2,117) = 32.54$, $p < 0.001$]. While performance of the Control condition improved slightly in the transfer test [$M_{different} = 3.0\%$, $t(40) = 2.77$, $p < 0.001$], a drop in performance was seen for both Overshadowing [$M_{different} = -29.8\%$, $t(35) = 7.78$, $p < 0.001$] and Warning conditions [$M_{different} = -18.9\%$, $t(42) = 6.08$, $p < 0.001$].

Receiver Operating Characteristic (ROC) analysis.

To better quantify how effective the three training conditions were independent of possible variability in criteria, we used Receiver Operating Characteristic (ROC) analysis (Green & Swet, 1966; Macmillan & Creelman, 2004). Each point on Figure 1.4 denotes the proportion of correct classification of Old World demons and incorrect classification of New World demons for a learner. Correctly classifying an Old World demon as Old World was considered a hit,

whereas incorrectly classifying an Old World demon as New World was considered a miss, and so forth. Discriminability (d') was calculated for each subject assuming an equal variance signal detection model. The value of d' measures how well a subject was able to tell apart the two categories.

In line with our conclusions from the pilot experiment, subjects in the Control condition achieved the highest mean discriminability (d' = 0.90, SD = 0.45), followed by the Warning condition (d' = 0.65, SD = 0.61), and the Overshadowing condition (d' = 0.38, SD = 0.63). A one-way ANOVA confirmed that discriminability differed across conditions [$F(2,117) = 8.24$, $p < 0.001$]. Tukey's honest significance test suggested that subjects in the Control condition could better discriminate the two categories compared to those in the Overshadowing condition ($p < 0.001$, Cohen's $d = 0.97$, 95% C.I. for d = [0.48, 1.46]). However, discriminability of the Warning condition was not reliably different from either the Control condition ($p = 0.11$, Cohen's $d = 0.47$, 95% C.I. for d = [0.02, 0.92]) or the Overshadowing condition ($p = 0.09$, Cohen's $d = 0.44$, 95% C.I. for d = [-0.02, 0.90]).

Classification strategies in transfer test.

To look in more detail at which features were relied upon by individual subjects in transfer test, we performed a logistic regression for each subject using his or her transfer test data. Features of the stimuli, namely eye color, eye width, horn height, and nose were used to predict a subject's classification decisions in the transfer test. Weightings of features were standardized. The relative strength of a subject's standardized weightings gives an indication of which features were utilized. Each row in Figure 1.5 denotes a subject's data. The color of each of the four tiles indicates the level of reliance on each of the features in classifying a demon. A deep green tile means that a feature had a heavy weight in the appropriate direction, a white tile suggests no reliance, and a deep red tile indicates a weight in the wrong direction. For the nose feature, any non-zero weighting, regardless of color, would be inappropriate since it was not

predictive of category membership. Subjects were sorted in descending order according to their classification accuracy in the transfer test, so the higher rows refer to the better-performing subjects.

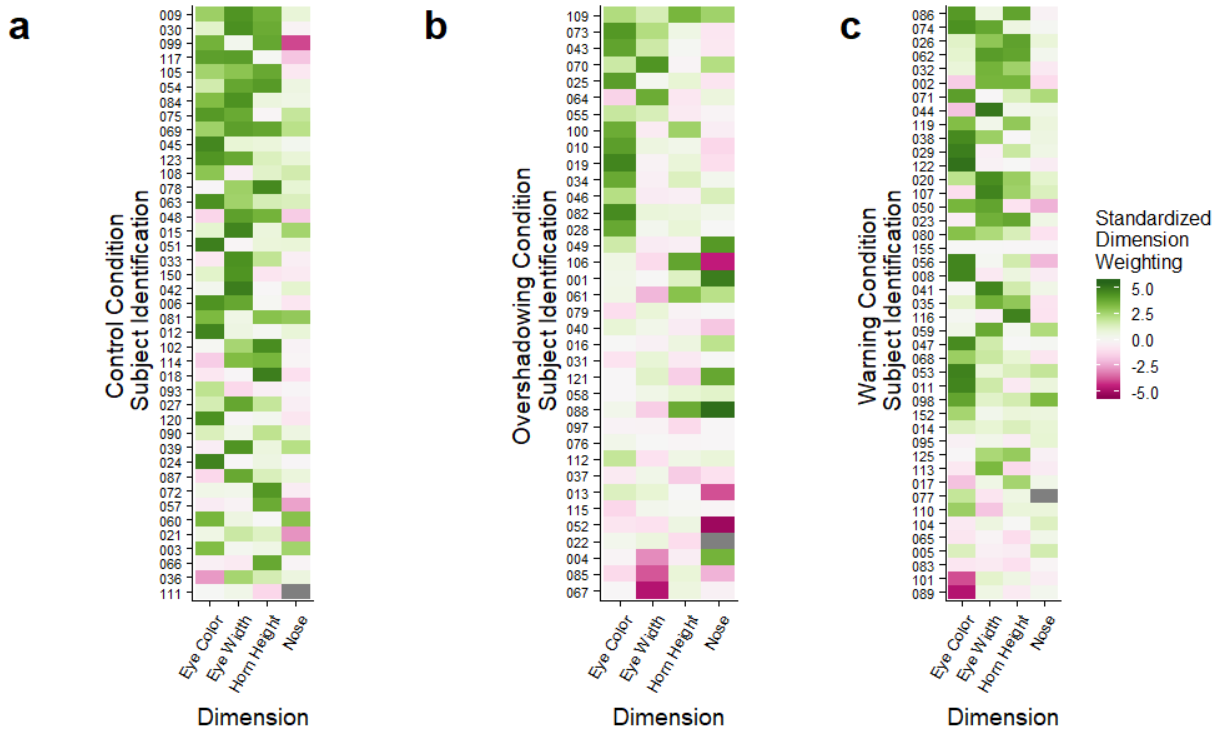


Figure 1.5. Classification strategies adopted by individual subjects in the transfer test. Each row represents the data of a subject. The color of a tile represents the weight a subject put on a feature (deep green indicates correct utilization, white indicates no utilization, and red indicates incorrect utilization). A non-white color for nose column indicates an incorrect weighting of that feature in the decision rule. Subjects were grouped by condition, (a) Control condition, (b) Overshadowing condition, and (c) Warning condition.

Regardless of conditions, it can be seen that subjects with higher classification accuracy tended to use a mix of multiple defining features in their classifications. More subjects in the Control condition (Figure 1.5a) and Warning condition (Figure 1.5c) appeared to utilize multiple defining features compared to those in the Overshadowing condition (Figure 1.5b). Around half of the subjects in the Overshadowing condition did not show any obvious dominant classification strategies.

Discussion

Experiment 1 successfully replicated the overshadowing effect found in the pilot experiment. When a highly predictive feature was available in the training phrase, learning of other probabilistic predictors of category membership was markedly reduced. The results also show that the overshadowing effect seems to be quite stubborn. Warning against the overshadowing feature appeared to shrink the effect, but it did not eradicate the overshadowing feature's influence on learning. The ROC analysis suggests that subjects' ability to discriminate the two demon categories was reduced by the overshadowing feature. Warning against the feature did not reliably improve discriminability.

In the following experiment, we attempted to replicate the findings of Experiment 1 using stimuli that are simpler but more naturalistic in appearance. As seen in Figure 1.6, the stimuli appeared a little bit like living cells. The category membership was predicted by a single probabilistic feature in the Control condition. We tested whether the overshadowing effect demonstrated in the preceding experiments would still stand in such a case, and if so, whether warning would help in mitigating the effect.

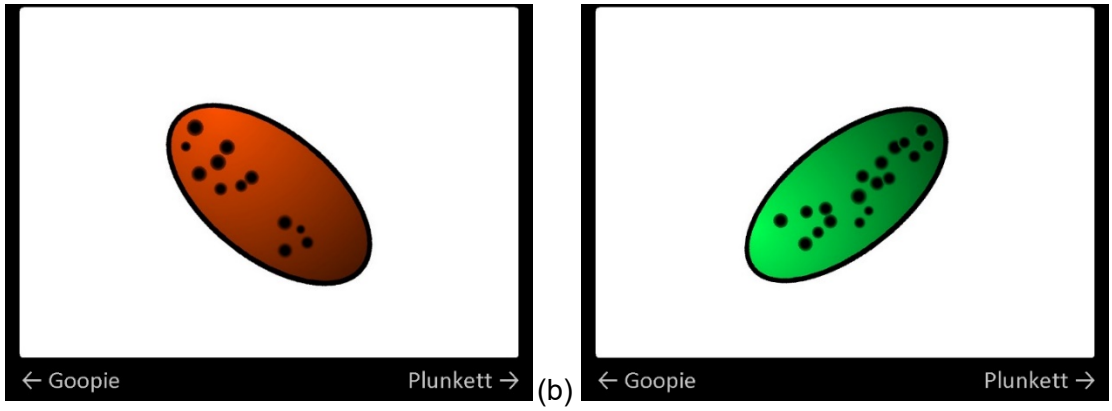


Figure 1.6. Sample stimuli used in Experiment 2. In general, Goopies had fewer dots and Plunketts had more dots. In the training phase of the Control condition, all stimuli were presented horizontally. In those of Overshadowing and Warning conditions, Goopies were tilted to the left and Plunketts to the right. All stimuli were presented horizontally in the transfer test trials.

Experiment 2

Methods

Design.

The basic design of Experiment 2 was identical to that of Experiment 1. The experiment consisted of two phases, training and transfer test. Three training conditions were compared between-subjects: Control, Overshadowing, and Warning. All subjects took the same transfer test which new stimuli were shown. As in the previous experiments, proportion of correct classifications in both training and transfer test was recorded to assess subjects' learning.

Participants.

Ninety participants (61 females, mean age = 20.2) from the University of California, San Diego participated for course credits. Subjects were randomly assigned into the three conditions: Control (n = 31), Overshadowing (n = 31), and Warning (n = 28). Informed consent was obtained before the experiment began. All subjects passed a computerized color test.

Materials.

Cell-like stimuli were used in the experiment (see Figure 1.6). Two categories of cell-like stimuli were created, which we named Goopies and Plunketts. Stimuli in the Control condition and the transfer test of all conditions varied in color, height, and number of black dots, but only the number of dots was predictive of category identity. In general, Goopies had fewer dots, and Plunketts had more dots. The number of dots was drawn at the runtime of the experiment, from a discretized Gaussian distribution with a mean that differed for the two categories. The means of the two distributions were 12 and 22, with a standard deviation of 5. We ran a simulation and determined that each probabilistic feature would have a predictive power of 84% to its corresponding category.

In the training phase in the Overshadowing and Warning conditions, the stimuli were defined as in the Control condition. The only difference was that these stimuli were tilted to one

orientation. Goopies were tilted to the left, while Plunketts were tilted to the right. The mean orientation of the two distributions were -45° and +45°, with a standard deviation of 10°. We jittered the orientation such that they did not stay at the same orientation across trials. This also made make orientation a slightly less salient feature. The stimuli were presented on a white background.

Procedure.

The procedure of Experiment 2 was identical to that of Experiment 1, except that each experimental session consisted of 600 trials. Subjects went through 400 training trials, and 200 transfer test trials. After completing every 100 trials, subjects were given a 30-second break.

In each of the training trials, subjects in the Control condition saw a cell-like stimulus in the middle of the computer screen. The stimulus was always in a horizontal position. Subjects were encouraged to make a keyboard response within 3 seconds. The stimulus disappeared in 10 seconds if no responses were made. A feedback tone indicated whether the response was correct.

The training phase of the Overshadowing and Warning conditions was identical to the Control condition except that the stimuli were tilted differently in the two categories. Goopies were always tilted to the left, and Plunketts to the right. In the Warning condition, subjects were told that the stimuli would always be lying horizontally in the transfer test, and they were encouraged to learn other orientation-invariant properties of the categories that would help them do well in the transfer test.

After 400 training trials had elapsed, all subjects took the same transfer test. In the transfer test, one stimulus was presented at a time. The stimulus was positioned horizontally. Subjects had no time constraints for their responses. The experiment ended with a debrief after 200 transfer test trials were completed.

Results

Training Phase.

Performance of the all three conditions improved from Block 1 to Block 2 and stayed at a high level until the end of the training phase (Figure 1.7, left panel). Accuracies for the three conditions were quite different even in the first block, suggesting utilization of the overshadowing feature commenced early in both the Overshadowing and the Warning conditions. Performance was different across conditions in the last training block, as indicated by a one-way ANOVA [$F(2,87) = 18.16$, $p < 0.001$]. Tukey's honest significance test indicated that subjects in the Control condition performed significantly worse than those in both the Overshadowing ($p < 0.001$, Cohen's $d = 1.73$, 95% C.I. for $d = [1.12, 2.34]$) and the Warning conditions ($p < 0.001$, Cohen's $d = 1.10$, 95% C.I. for $d = [0.53, 1.67]$). However, performance in the Overshadowing and Warning conditions was not reliably different from each other ($p = 0.54$, Cohen's $d = 0.25$, 95% C.I. for $d = [-0.28, 0.79]$).

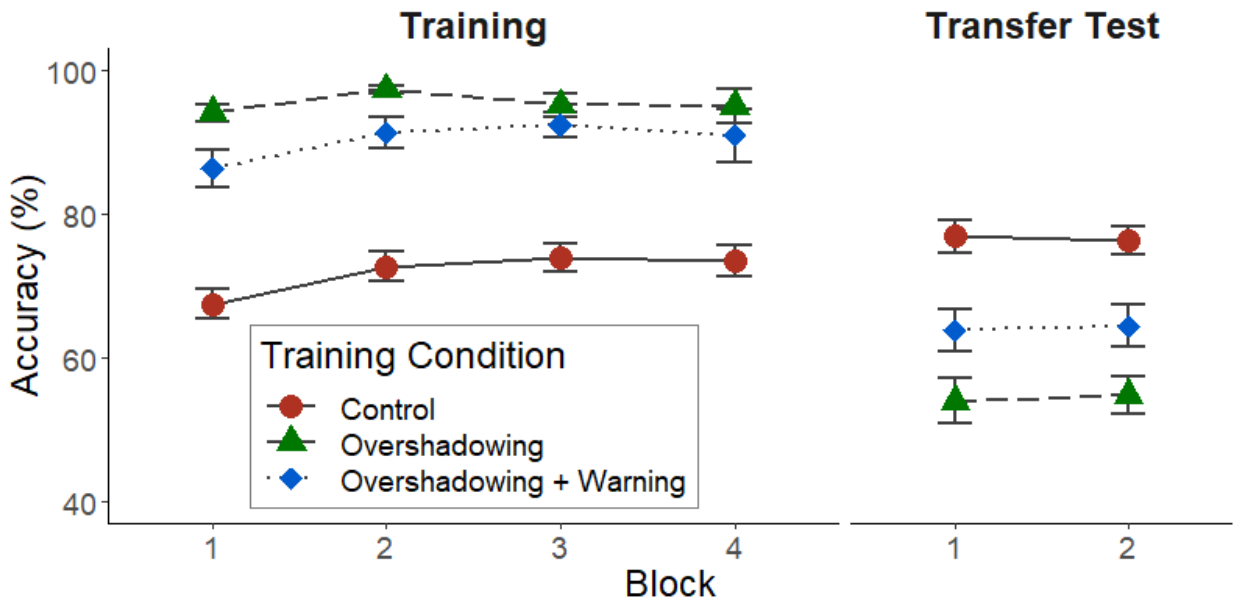


Figure 1.7. Training and test accuracies in Experiment 2. Subjects in the Overshadowing and Warning conditions reached an asymptote within the first training block, but their performance was worse than those in the Control condition in the transfer test. Error bars denote within-subjects standard errors of the means.

Transfer Test.

As in the previous experiments, subjects in the Control condition performed the best (M = 77%, SD = 11.5%). Subjects in the Overshadowing condition performed the worst (M = 54%, SD = 15.1%). Subjects in the Warning condition performed at an intermediate level (M = 64%, SD = 15.1%). Performance of the Control and Warning groups were above chance (ps < 0.001), while that of the Overshadowing group was not, $t(30) = 1.64$, $p = 0.11$.

A one-way ANOVA suggested that subjects in the three conditions performed differently in the transfer test (Figure 1.7, right panel), $F(2,87) = 19.68$, $p < 0.001$. Tukey's honest significance test indicated that all pairwise comparisons were statistically significant (ps < 0.05). Specifically, there was a 22% difference between the Control and Overshadowing conditions ($p < 0.001$, Cohen's $d = 1.65$, 95% C.I. for d = [1.05, 2.25]), a 13% difference between the Control and Warning conditions ($p = 0.003$, Cohen's $d = 0.94$, 95% C.I. for d = [0.38, 1.50]), and a 9.7% difference between the Overshadowing and Warning conditions ($p = 0.02$, Cohen's $d = 0.64$, 95% C.I. for d = [0.10, 1.19]).

A 2 (Last 2 Training Blocks vs Test Blocks) × 3 (Training Condition) mixed-design ANOVA was conducted to examine the changes from training to transfer test phases in the three conditions. Accuracy in the transfer test was generally lower than that in the last training blocks, $F(1, 87) = 147.3$, $p < 0.001$. Importantly, a significant interaction between the two factors suggested that the changes were different in the three conditions, $F(2, 87) = 54.4$, $p < 0.001$. Specifically, subjects in the Control condition showed a slight improvement in the transfer test [$M_{difference} = 2.9\%$, $t(30) = 3.84$, $p < 0.001$]. A huge drop in performance was seen in both the Overshadowing [$M_{difference} = -40.7\%$, $t(30) = 11.87$, $p < 0.001$] and the Warning conditions [$M_{difference} = -27.6\%$, $t(27) = 6.58$, $p < 0.001$] when the deterministic binary feature was no longer available. The main effect of Condition was not significant when accuracy was averaged across the final two training blocks and the transfer test blocks, $F(2,87) = 0.90$, $p = 0.41$.

Receiver Operating Characteristic (ROC) analysis.

Figure 1.8 shows subjects' performance in terms of hit and false alarm rates. It can be seen that most of the points of the Control condition clustered in the upper-left-hand corner, indicating an overall high discriminability (d' = 1.55, SD = 0.70). A sizable portion of the Warning condition also clustered in the same region, indicating high discriminability for those subjects. However, the remaining subjects clustered around the gray line, which denotes chance-level performance. This brought the mean discriminability down (d' = 0.86, SD = 0.87), compared to the Control condition. Discriminability of the Overshadowing condition was the lowest (d' = 0.25, SD = 0.87), which was in fact not reliably different from chance level, $t(30) = 1.63$, $p = 0.11$. A one-way ANOVA indicates significant differences between the conditions, $F(2,87) = 19.82$, $p < 0.001$. All pairwise comparisons are significant ($ps < 0.05$). Specifically, there was a 1.30 difference in discriminability between the Control and Overshadowing conditions ($p < 0.001$, Cohen's d = 1.65, 95% C.I. for d = [1.05, 2.25]), a 0.70 difference between the Control and Warning conditions ($p = 0.004$, Cohen's d = 0.88, 95% C.I. for d = [0.33, 1.44]), and a 0.61 difference between the Overshadowing and Warning conditions ($p = 0.01$, Cohen's d = 0.70, 95% C.I. for d = [0.15, 1.25]).

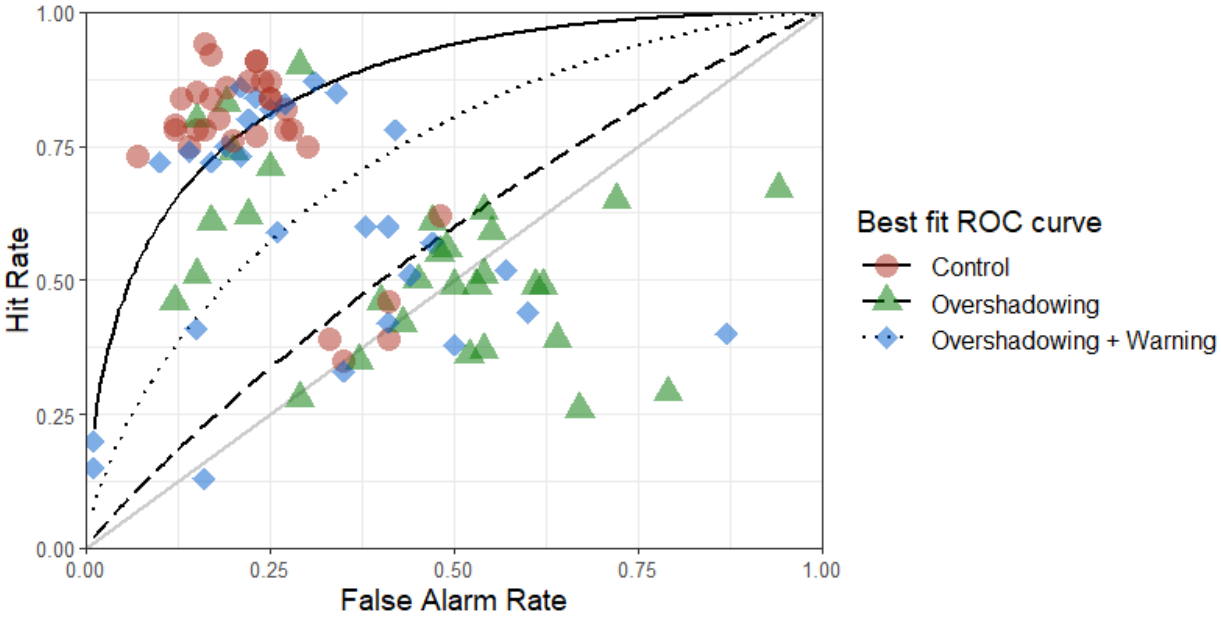


Figure 1.8. ROC analysis for Experiment 2. The ability to discriminate the two categories was high for the Control condition. Performance of the Warning condition was more diverse, while those in the Overshadowing condition did not differ from chance level.

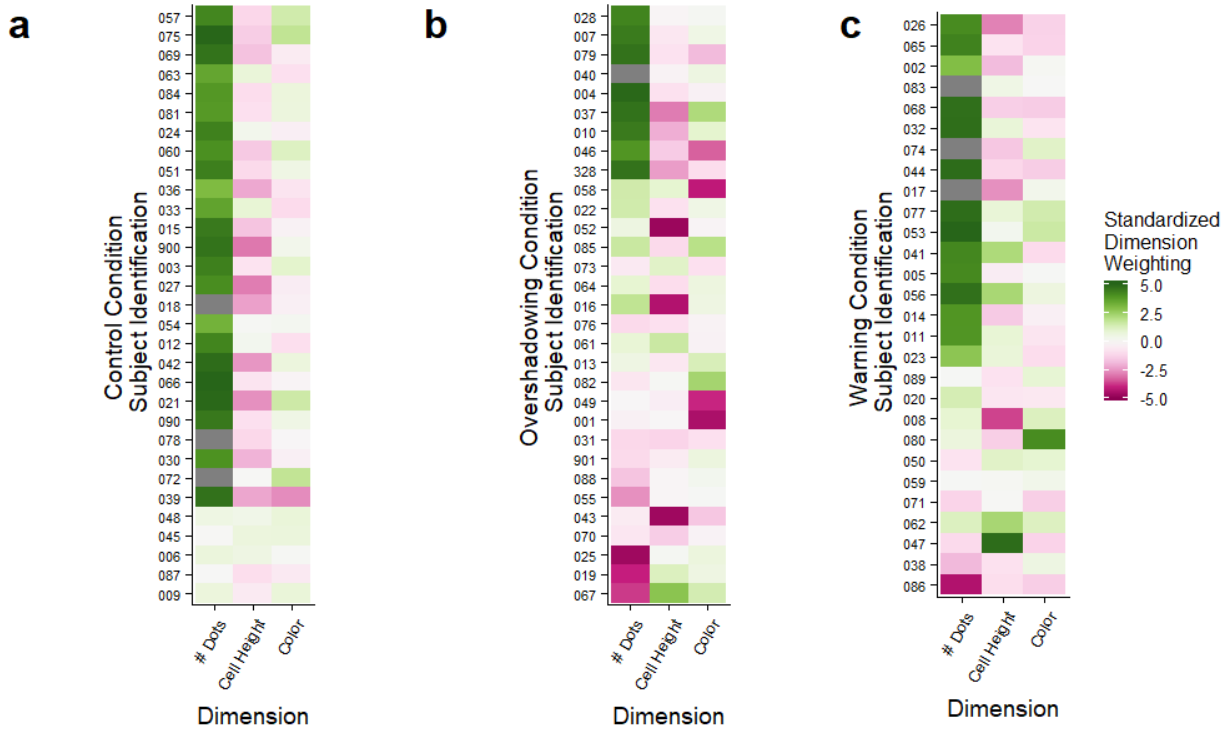


Figure 1.9. Classification strategies adopted by individual subjects in the transfer test. Subjects were grouped by condition, (a) Control condition, (b) Overshadowing condition, and (c) Warning condition. Subjects who performed well were more likely to make use of the number of dots (shown in deep green in the first columns in each figure), while not utilizing other non-defining features.

Classification strategies in transfer test.

As in Experiment 1, we performed a logistic regression to determine which features were utilized by individual subjects in the transfer test. Number of dots, and height and color of the cell were entered into the logistic regression to predict individual subject's responses.

In Experiment 2, only the number of dots predicted category membership in the transfer test. We predicted that best-performing subjects would utilize this feature strongly in their decisions. As shown in Figure 1.9a, most of the subjects in the Control condition made use of the number of dots in classification, as shown in deep green in the first column. Best-performing subjects did not incorrectly rely much upon other non-defining features in the decision rule. More than half of the subjects in the Warning condition also utilized the number of dots (Figure

1.9c), while substantially fewer subjects in the Overshadowing condition reliably did so (Figure 1.9b).

Discussion

Experiment 2 suggests the overshadowing effect reported in the previous experiments may be quite robust. Here, using a simpler mapping of features to categories but more naturalistic-appearing stimuli, we still see that when a deterministic binary predictor feature was made available in training, performance on a transfer test was impaired. From the logistic regression analysis, it appears that a large portion of subjects in the Warning condition were able to effectively ignore the overshadowing feature, or use it to their advantage during training. This ability allowed the Warning group to outperform the Overshadowing group in terms of overall accuracy.

General Discussion

In three experiments, we showed a powerful overshadowing effect in human perceptual category learning (a point that had been somewhat in doubt given the small pre-existing literature on the question). Specifically, when a stimulus feature highly predictive of its category membership was made available during training (but not in the transfer test), subjects acquired much less ability to perform the classification based on one or multiple probabilistic cues. In the three experiments we reported, this overshadowing effect showed up as a strong interaction between training versus transfer test × Control versus Overshadowing conditions. The effect is quantitatively impressive, both in terms of proportional change of performance and in terms of Cohen's d (effect size relative to inter-individual variability in performance).

At the group-level, the overshadowing effect was so powerful that performance of the overshadowing groups in Experiments 1 and 2 was close to chance level in the transfer test. This is a sharp contrast compared to the Control groups, which clearly retained most of the category information. At an individual level, our analyses through ROC curves and logistic

regression show that a considerable amount of subject variability. Some subjects in the Overshadowing conditions were able to pick up the probabilistic features, and to utilize them when the overshadowing feature was no longer available. Our logistic regression analysis shows how individual participants utilized each feature in the transfer test. To our knowledge, the use of logistic regression to understanding individual participant has not been done in the field.

Our observation mirrors that of Kelmer Nelson (Experiments 1 & 2, 1984), who showed that some subjects might take a more analytic approach in category learning, while others might take a more holistic approach.

The current study was partly motivated by Bott, Hoffman, and Murphy (2007). In their Experiments 1 and 2, the researchers had participants learn to associate features of a car with one of the category labels. Half of the participants were in the blocking group. The group went through a pre-training phase where they were told that a particular feature always predicted the group identity. In an actual training trial, the participants were shown a list of car features, and had to classify the car into one of the two categories. In the test phrase, the features were shown one at a time, and the participants had to decide which category the feature belonged to. The authors predicted that participants in the blocking group would classify the features at chance in the test. This was not the case. The blocking group learned nearly as much about the features as the control group. This was taken as an evidence that blocking effect was weak in human learning in their study.

There were a number of differences between Bott et al's and our paradigm, which might lead to the cue competition effect that we found in all three experiments. Instead of focusing on whether cue competition effect is reliable (Maes, Boddez, Alfei, Krypotos, D'Hooge, De Houwer, & Beckers, 2016; but see Soto, 2018), we were primarily interested in whether human learners had voluntary control over the effect (Mitchell, De Houwer, & Lovibond, 2009). The basic question is concrete: If a strongly predictive feature inhibits learning about weaker predictors,

can that inhibition be voluntarily reduced if the learner is consciously motivated to learn as much as possible about *all* predictors that are present?

In the rest of the Discussion, we will first explore top-down voluntary control in the Recorla-Wagner model and its variants, and then consider how the comparator hypothesis may explain some of the findings. We will then end with some practical implications of our finding.

Cue Competition Effects in Computational Models

The results seem at least qualitatively compatible with the hypothesis of error-driven learning mentioned in the Introduction. A majority of subjects learned in these tasks when they made mistakes, presumably because they adjusted their classification rule in response to the feedback they received. When subjects were explicitly instructed not to rely upon the overshadowing feature in training, this overshadowing effect appeared to be reduced. The group means may not tell the whole story in this regard, however. The ROC analysis in Experiment 2 showed that around half of the subjects in the Warning condition seemed to show a dramatically shrunken overshadowing effect, while others seemed not to show any such reduction (solid dots in Figure 1.8).

According to the Rescorla-Wagner Model, associations between unconditioned and conditioned stimuli were enhanced because doing so would reduce the error in future categorization trials. In general, such acquisition-focused models have little problems explaining the results of cue competition effects.

By taking a bottom-up approach in explaining how the stimulus structure affects learning, however, it is not entirely clear how they can incorporate voluntary control during category learning. Showing top-down voluntary control over the stimuli, here we explore the implications on currently popular models of categorization.

A number of category-learning models that rely on error-driven learning, such as ALCOVE or RULEX (Kruschke, 1992; Nosofsky, Palmeri, & McKinley, 1994) seem well able to

account for overshadowing and blocking of the kind reported here. Fundamentally, these networks learn at the feature level, not at the exemplar level, and they change their weights only when errors are made. In the present task, subjects performing in the training phase of the Overshadowing condition did so with few errors, and thus the ALCOVE model would predict that little useful adjustment would take place in irrelevant or redundant feature attention weights, resulting in little or no learning for irrelevant or redundant features.

The same appears true for error-driven hypothesis testing models such as RULEX. These models assume that individuals are biased toward learning simple rules for exemplar categorization. Importantly, like ALCOVE, these models adjust the weights on the given set of exemplar features to find the ones that are best predictive of category membership. Put in the context of the current task, RULEX would impose a heavy weight on the Overshadowing feature. Once an individual learned about the Overshadowing feature and stopped making errors, the model leaves little room for learning other less valid features.

Less clear is how the current set of results fit with clustering models of categorization such as SUSTAIN (Love, Medin, & Gureckis, 2004). Although SUSTAIN can selectively weigh certain exemplar features more than others like ALCOVE and RULEX, it is not necessarily the case that learning about irrelevant or redundant feature information would be profoundly dampened. According to the model, exemplars are grouped into various clusters based on feature similarity, which in turn drives decisions about category membership. Although a single feature like the horn color may drive classification performance during training, it is possible that SUSTAIN would still learn to categorize test exemplars that lack information about horn color. This is because cluster membership in SUSTAIN is based on similarity matching. In our Experiment 1, features varied independently. Test exemplars would continue to be associated with clusters of similar looking exemplars. This may explain why a minority of subjects in the Overshadowing condition were able to perform well in the transfer test.

Other theoretical models of categorization, such as exemplar models (e.g., Hintzman, 1986; Medin & Schaffer, 1978; Nosofsky, 1984) and prototype models (e.g., Homa, Rhoads, & Chambliss, 1979; Posner & Keele, 1968; Rosch, 1973; Smith & Minda, 1998) do not appear to fit with the data we have. During learning, these models construct a boundary that would separate the categories. The boundary is determined by two main factors, the selective-attention weightings and the memory strength of training items. For the selective-attention component, the more relevant features are likely to be stretched and given more weight. In our Overshadowing condition, it means that the completely valid feature is likely to be attended, and gain a heavy weight. Other features are far less valid, and hence the psychological space of these features would be shrunk. The resulting decision rule is likely to resemble a single-dimension one, with the overshadowing feature as the sole component. Obviously, this rule would fail miserable during the transfer test, when the overshadowing feature is no longer available for classification. These models may be able to capture the group-level at-chance accuracy patterns of the Overshadowing condition, but unlikely to explain why some individual subjects were able to utilize the less valid features in the transfer test.

Finally, one multiple-systems account of category learning, COVIS (Ashby et al., 1998) posits that category learning takes place by one of two types of learning systems: a rule-based hypothesis testing system and a procedural learning system that relies on the physical similarity of category exemplars. Learning via the procedural learning system involves integrating information across multiple exemplar features, so that categorization judgments are made based on the overall physical similarity of test objects to the learned category. Thus, it might be the case that learning via the procedural learning system in COVIS would learn all predictive features of the presented exemplars. However, since the procedural learning system of COVIS is agnostic with respect to attentional allocation of exemplar features, it is unclear whether or not this is true. In contrast, the rule-based system of COVIS involves testing rules that could determine category membership by learning about one or a conjunction of multiple, independent

features. Once a rule is discovered, it is presumably used until an error is made. Again, however, the model does not specify formal rules for the functioning of this attentional filter so it is unclear whether or not learning about other irrelevant or redundant features would take place.

All of these formal models mentioned above are bottom-up and stimulus-driven. They assume no top-down intervention in biasing the attention towards a particular feature. According to these models, if two features are independently and equally predictive of a stimulus's category membership, they will end up having equal, or very similar weightings. This is a logical assumption in early development of a formal model of learning using artificial stimuli. In naturalistic learning environment, however, learners surely often have biases towards certain particular feature dimensions, reflecting previous learning experience, instructions from educators, and so forth. The popular models mentioned above do not have any independent parameters to bias top-down attention. As we mention below, top-down attention modulation is likely to be prevalent in day-to-day life. Therefore, modifications of these computational models are needed that they are capable of explaining top-down effects such as the one reported here.

Alternative Models Explaining the Cue Competition Effects

The computational models mentioned above fit reasonably well with means of group performance. They explain how much participants in a group, as a whole, learn about a particular feature. While looking at performance of individual participants within a condition, these models do not work as well. As we can see in the ROC analyses and logistic regression analyses in Experiments 1 and 2, behaviors of some of the participants in the Overshadowing condition resembled that of those in the Control condition, while others resembled those in the Warning condition. These individual differences can be captured by a set of models that is known as performance-focused models (Miller & Escobar, 2001).

Performance-focused models posit that learners encode the events which stimuli are paired. This contrasts with the acquisition models discussed above. From the perspective of the

acquisition models, the learners are assumed to maintain some summary statistics of the groups that they are trying to categorize, such as the mean, standard deviation of each category, or the category boundary. Information encountered during the learning process beyond the summary statistics is not encoded. In our experiments, that would mean that non-deterministic features never entered the memory.

Data from our Overshadowing conditions suggest otherwise. Even when the deterministic features were not helpful in the training phase, some of the learners clearly encoded the information. This observation appears potentially compatible with the view of performance-focused models. One formulation of the performance-focused models, the Comparator Hypothesis (Miller & Matzel, 1988), suggests that the association between the overshadowed CS and the US is acquired, but not expressed, in the presence of another stronger overshadowing CS (Denniston, Savastano, Blaisdell, & Miller, 2003). When the overshadowing CS does not appear, the overshadowed CS has a chance of being expressed. Our data, at least for a proportion of participants in the Overshadowing condition, fit well with this description. They showed clear learning of some of the probabilistic features in the transfer test, when the deterministic feature no longer appeared.

As one of our reviewers pointed out, our results would also align with the propositional approach to learning (Mitchell, De Houwer, & Lovibond, 2009). This approach argues that a subject generates explicit propositions about the relationships between different elements in the environment and creates associations between them through a controlled reasoning process. It contrasts with the dual-system approach to learning, which states that the associate between the CS and US is formed automatically, often outside of consciousness. Colgan (1970) trained participants to associate a light and electric shock. After pairing, participants skin conductance went up to the light. Half of the participants were told how to predict the shock. The group showed little increased skin conductance for the no-shock trials, indicating effects of the instructions. Given the evidence that associations between stimuli could be modified by

instructions, the effect of warning on the overshadowing effect fits naturally with propositional approach to learning.

Why Was Warning Not Fully Effective?

Although our results showed an effect of warning, it did not improve performance up to the level of the control condition. Why not?

Many researchers in category learning suggested that attention has to be deployed to a dimension before it is utilized in category learning tasks (see Goldstone, 1998, for a review). In fact, many early computational models of category learning included a parameter that regulates how attention is weighed across features (e.g., Nosofsky, 1991). A feature that is predictive of category identity would be assigned a higher attentional weighting in the decision rule. This larger attentional weighting also makes differentiation within the feature more fine-grained.

If variations within a feature are not predictive of category identity, attentional weighting to that feature is tuned down, denying maximal influence to the feature in the decision rule. This modulation is believed to be not completely voluntary (Rehder & Hoffman, 2005). For example, Shiffrin and Schneider (1977) showed that when stimuli previously served as targets, they may capture attention automatically. Subjects in our Warning conditions might have faced a similar dilemma. They were instructed that the overshadowing feature should be ignored, yet it was highly predictive of category identity. The instruction might be overwritten by the automatic attention deployment, preventing the overshadowing feature from being fully ignored.

Alternatively, the overshadowing effect may be affecting the calculation and storage of predictive relationships. In both the Overshadowing and Warning conditions, once subjects figured out the mapping of the overshadowing feature to category identity, they did not need to make any estimations based on other features. On the other hand, to do well in training, subjects in the Control condition had to make a decision taking account of all the features. Long-term memory of the to-be-remembered items, the defining features of the categories in the

current experiments, is believed to improve with repeated retrieval (as in the retrieval practice effect, e.g., Carrier & Pashler, 1992; Gates, 1917; Roediger & Karpicke, 2006).

The retrieval practice effect has been shown to apply to human perceptual category learning. Jacoby, Wahlheim and Coane (2010) had subjects classify two families of birds with study-only blocks or with repeated testing blocks. Those in the repeated testing with feedback condition outperformed the study-only condition in the transfer test. In the experimental design examined here, subjects in the Overshadowing and Warning condition may be deprived of opportunities to test themselves, leading to suboptimal learning assessed by the transfer test.

Implications for Real-World Category Learning

We conclude with some very tentative suggestions on possible implications for real-world category learning and training. Our studies contained relatively low-dimensional stimuli with continuous feature values, which we believe have at least a moderate resemblance to a certain proportion of real-life human category learning tasks. The three experiments showed that when an easy discrete predictor feature was present, learning of simultaneously present probabilistic predictors was dramatically impaired. Future research could test whether our finding applies to the real-world setting. Here we provide some possible directions.

First, training conditions might be profitably structured in a way that obligate the learner to perform the actual categorization task using the predictors that will be available in the field (but no additional “easy predictors”). If a powerful predictor will not be available, it should not be present in the training process.

Second, individual differences in overshadowing appear strikingly large and important as seen in Experiments 1 and 2. Warning of overshadowing features in Experiments 1 and 2 appeared to have quite different effects for different people. Akin to the instructions given to her subjects in Kelmer Nelson (1984), warning is a type of top-down modulation (see also Wills, Inkster, & Milton, 2015). Subjects with better attentional control may be well able to ignore the

overshadowing feature. As a result, this makes their performance in the transfer test resemble that in the Control condition. Selective attention capacity, short-term and long-term memories might all play a role in the differential performance. Identifying these factors using standardized tests may help educators devise useful auxiliary training interventions.

Lastly, it might be worth exploring training regimes in which learners are trained with one feature at a time. In our classification task, features are independently manipulated. That is, the value of one feature does not predict a value of another feature. Some classification tasks in real-life may have similar properties. In those cases, independent features can be trained by withholding other features of the stimuli during the training phase. In the case of learning to distinguish two kinds of birds, for example, exemplars of beaks can first be trained, followed by the legs, and so on. Learners might then acquire an independent weighting for each feature, and omission of some features in the test stimulus set should have a less detrimental effect on classification accuracy.

Chapter 1, in full, is a reprint of the material submitted to a journal in revision: Lau, J. S. H., Casale, M. B., & Pashler, H. (in revision). Mitigating Cue Competition Effects in Human Category Learning. The dissertation author was the primary investigator and author of this paper.

References

Allan, L. G. (1993). Human contingency judgments: Rule based or associative?. Psychological Bulletin, 114(3), 435.

Arcediano, F., Matute, H., & Miller, R. R. (1997). Blocking of Pavlovian conditioning in humans. Learning and Motivation, 28(2), 188-199.

Ashby, F. G., Alfonso-Reese, L. A., Turken, A. U., & Waldron, E. M. (1998). A neuropsychological theory of multiple systems in category learning. Psychological Review, 105, 442-481.

Ashby, F.G., Paul, E.J., & Maddox, W.T. (2011). COVIS. In E. M. Pothos & A.J. Wills (Eds.), Formal approaches in categorization, 65-87. New York: Cambridge University Press.

Baetu, I., Baker, A. G., Darredeau, C., & Murphy, R. A. (2005). A comparative approach to cue competition with one and two strong predictors. Learning & Behavior, 33(2), 160-171.

Blair, M. R., Watson, M. R., & Meier, K. M. (2009). Errors, efficiency, and the interplay between attention and category learning. Cognition, 112(2), 330-336.

Bott, L., Hoffman, A., & Murphy, G. L. (2007). Blocking in category learning. Journal of Experimental Psychology: General, 136, 685-699.

Carrier, M., & Pashler, H. (1992). The influence of retrieval on retention. Memory & Cognition. 20: 632–642.

Colgan, D. M. (1970) Effect of instructions on the skin conductance response. Journal of Experimental Psychology, 86:108–12.

Denniston, J. C., Savastano, H. I., Blaisdell, A. P., & Miller, R. R. (2003). Cue competition as a retrieval deficit. Learning and Motivation, 34(1), 1-31.

Dickinson, A., Shanks, D., & Evenden, J. (1984). Judgement of act-outcome contingency: The role of selective attribution. The Quarterly Journal of Experimental Psychology, 36(1), 29-50.

Gates, A. I. (1917). Recitation as a factor in memorizing. Archives of Psychology, 6 (40).

Gluck, M. & Bower, G. (1988). From conditioning to category learning: An adaptive network model. Journal of Experimental Psychology: General, 117, 227-247.

Goldstone, R. L. (1998). Perceptual learning. Annual review of psychology, 49(1), 585-612.

Green, D. M., & Swets, J. A. (1966). Signal detection theory and psychophysics. New York: Wiley.

Hintzman, D, L. (1986). "Schema abstraction" in a multiple-trace memory model. Psychological Review, 93, 411-428

Homa, D., Rhoads, D., & Chambliss, D. (1979). Evolution of conceptual structure. Journal of Experimental Psychology: Human Learning and Memory, 5, 11-23.

Jacoby, L. L., Wahlheim, C. N., & Coane, J. H. (2010). Test-enhanced learning of natural concepts: effects on recognition memory, classification, and metacognition. Journal of Experimental Psychology: Learning, Memory, and Cognition, 36(6), 1441.

Kemler Nelson, D. G. (1984). The effect of intention on what concepts are acquired. Journal of Verbal Learning and Verbal Behavior, 23(6), 734-759.

Kruschke, J. K. (1992). ALCOVE: An exemplar-based connectionist model of category learning. Psychological Review, 99, 22-44.

Kruschke, J. K. (2001). Toward a unified model of attention in associative learning. Journal of mathematical psychology, 45(6), 812-863.

Love, B. C., Medin, D. L, and Gureckis, T. M. (2004). SUSTAIN: A Network Model of Category Learning. Psychological Review, 11, 309-332.

Mackintosh, N. J. (1976). Overshadowing and stimulus intensity. Animal learning & behavior, 4(2), 186-192.

Macmillan, N. A., & Creelman, C. D. (2004). Detection theory: A user's guide. Psychology press.

Maes, E., Boddez, Y., Alfei, J. M., Krypotos, A. M., D'Hooge, R., De Houwer, J., & Beckers, T. (2016). The elusive nature of the blocking effect: 15 failures to replicate. Journal of Experimental Psychology: General, 145(9), e49.

Matute, H., Arcediano, F., & Miller, R. R. (1996). Test question modulates cue competition between causes and between effects. Journal of Experimental Psychology: Learning, Memory, and Cognition, 22(1), 182.

Medin, D. M. & Schaffer, M. M. (1978). Context theory of classification learning. Psychological Review, 85, 207-238.

Miller, R. R., & Escobar, M. (2001). Contrasting acquisition-focused and performance-focused models of acquired behavior. Current Directions in Psychological Science, 10(4), 141-145.

Miller, R. R., & Matzel, L. D. (1988). The comparator hypothesis: A response rule for the expression of associations. In G. H. Bower (Ed.), The psychology of learning and motivation (Vol. 22, pp. 51–92). San Diego, CA: Academic Press.

Miller, R. R., & Schachtman, T. R. (1985). Conditioning context as an associative baseline: Implications for response generation and the nature of conditioned inhibition. In R. R. Miller & N. E. Spear (Eds.), Information processing in animals: Conditioned inhibition (pp. 51–88). Hillsdale, NJ: Lawrence Erlbaum Associates.

Mitchell, C. J., De Houwer, J., & Lovibond, P. F. (2009). The propositional nature of human associative learning. Behavioral and Brain Sciences, 32(2), 183-198.

Murphy, G. (2004). The big book of concepts. Cambridge: MIT press.

Murphy, G. L., & Dunsmoor, J. E. (2017). Do salient features overshadow learning of other features in

category learning?. Journal of Experimental Psychology: Animal Learning and Cognition, 43(3), 219.

Nosofsky, R. M. (1991). Tests of an exemplar model for relating perceptual classification and recognition memory. Journal of Experimental Psychology: Human Perception and Performance, 17(1), 3.

Nosofsky, R. M., Palmeri, T. J., & McKinley, S. C. (1994). Rule-plus-exception model of classification learning. Psychological Review, 101, 53-79.

Pashler, H., & Mozer, M. C. (2013). When does fading enhance perceptual category learning?. Journal of Experimental Psychology: Learning, Memory, and Cognition, 39(4), 1162.

Pavlov, I. P. (1927). In G. V. Anrep (Ed.), Conditioned reflexes: An Investigation of The Physiological Activity of The Cerebral Cortex. Oxford University Press: London. Retrieved from: http://psychclassics.yorku.ca/Pavlov/lecture8.htm

Posner, M. I., Keele, S.W. (1968). On the genesis of abstract ideas. Journal of Experimental Psychology, 77, 353-363.

Rehder, B., & Hoffman, A. B. (2005). Eyetracking and selective attention in category learning. Cognitive Psychology, 51(1), 1-41.

Rescorla, R. A., & Wagner, A. R. (1972). A theory of Pavlovian Conditioning. Variations in the effectiveness of Reinforcement and Nonreinforcement. In A. H. Black & W. F. Prokasy (Eds.), Classical conditioning II: Current research and theory. New York: AppletonCentury-Crofts.

Roediger, H. L., & Karpicke, J. D. (2006). Test-Enhanced Learning: Taking Memory Tests Improves Long-Term Retention. Psychological Science. 17 (3): 249–255.

Rohrer, D., Dedrick, R. F., & Burgess, K. (2014). The benefit of interleaved mathematics practice is not limited to superficially similar kinds of problems. Psychonomic Bulletin & Review, 21, 1323-1330.

Rohrer, D., Dedrick, R. F., & Stershic, S. (2015). Interleaved practice improves mathematics learning. Journal of Educational Psychology, 107, 900-908.

Rosch, E. H. (1973). Natural categories. Cognitive Psychology, 4, 328-350.

Shanks, D. R. (1991). Categorization by a connectionist network. Journal of Experimental Psychology: Learning, Memory, and Cognition, 17(3), 433.

Smith, J. D., & Minda, J. P. (1998). Prototypes in the mist: The early epochs of category learning. Journal of Experimental Psychology: Learning, Memory, and Cognition, 24, 1411-1436.

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.

Soto, F. A. (2018). Contemporary associative learning theory predicts failures to obtain blocking. Comment on Maes et al. (2016). Journal of Experimental Psychology: General, 147(4), 597-602.

Soto, F. A., & Wasserman, E. A. (2010). Missing the forest for the trees: Object discrimination learning

blocks categorization learning. Psychological Science, 21(10), 1510-1517.

Wagner, A. R., Logan, F. A., & Haberlandt, K. (1968). Stimulus selection in animal discrimination learning. Journal of Experimental Psychology, 76(2p1), 171.

Williams, D. A., Sagness, K. E., & McPhee, J. E. (1994). Configural and elemental strategies in predictive learning. Journal of Experimental Psychology: Learning, Memory, and Cognition, 20(3), 694.

Wills, A. J., Inkster, A. B., & Milton, F. (2015). Combination or differentiation? Two theories of processing order in classification. Cognitive psychology, 80, 1-33.

CHAPTER 2: Simultaneous and Interleaved Training in Perceptual Category Learning

Abstract

The great majority of studies of perceptual category learning have presented learners with category members one at a time. In two experiments, we explored whether simultaneously presenting two stimuli of different categories led to superior learning. Learners were trained either with sequential or simultaneous stimulus presentations, and tested either with sequential or simultaneous displays. Experiment 1 used relatively simple, artificial stimuli, and Experiment 2 used real images of skin moles (some showing cases of melanoma, others not). When the final test involved sequential presentation of novel stimuli, neither simultaneous nor sequential training was superior. However, when the transfer test involved simultaneous presentation of stimuli, differences due to the two types of training were observed. We discuss practical and theoretical implications.

Introduction

When children are taught to distinguish different categories of objects, they are often shown examples of the two categories simultaneously present. For example, two similar-appearing types of dogs might appear together in a dog park, and a caretaker might take the opportunity to point out similarities and differences between the breeds. Similarly, in many children's picture books, objects of different categories are often placed within the same page, or spread across two pages that can be viewed at the same time (e.g., Kovecses, 2015; Scarry, 1998). Children's books laid out in this manner allows caretakers to contrast objects with different features while explaining them to the children. However, simultaneous presentation as a training strategy has not frequently been discussed in the psychological literature on category learning, despite its common occurrence in real-world settings. In the classic supervised perceptual category learning paradigm, stimuli are shown to the human learners one at a time

(e.g., Gauthier & Tarr, 1997; Medin & Schaffer, 1978; Nosofsky, 1991; Posner & Keele, 1968). On each trial, a stimulus from one of the categories is shown, and the learner makes a decision. Afterwards, feedback is given to the learner. Supposedly, learners modify their hypotheses about categorization rules upon receiving feedback, until their hypotheses are confirmed by subsequent trials. Real-life category learning may not follow the same trajectory.

When learning to categorize objects into different categories, a learner can employ different strategies. One can focus on the similarities within a category. Oranges are round, they are orange in color, and often have shiny surface. Apples are red or green, they have a heart shape, and they have waxy surfaces. Using this strategy to learn has clear benefits: the learning is highly generalizable. When learners are asked to distinguish oranges from a new category, say pears, they only have to learn about the features of the new category. This strategy is akin to the prototype theory (Posner & Keele, 1968), which learners generate a prototype for each category they are to learn. When they encounter a new object, it is matched with the corresponding prototype for categorization. Categorization based on similarity was shown to be an effective strategy, even for complicated ones such as skin lesion diagnosis (Aldridge, Glodzik, Ballerini, Fisher, & Rees, 2011; Brown, Robertson, Bisset, & Rees, 2009).

On the other hand, learners can employ a learning strategy that is more specific to the few categories they are asked to learn. They can focus on the differences between easily confusable categories. When asked to distinguish grapefruits from oranges, learners can focus on several dimensions that would help separate the two categories. Grapefruits are usually bigger than oranges, their peel color is also less saturated. Numerous studies have shown that opportunities to compare exemplars from different categories benefit category learning (e.g., Kang & Pashler, 2012; Kornell & Bjork, 2008). In recent years, there has been an interest to tease apart the relative contributions of the two processes in category learning. The way that training items are presented affects whether learners focus on the similarities within categories or differences between categories.

In the rest of our Introduction section, we first briefly review studies that utilized different modes of presenting training items, and the consequences of such manipulations. We then describe our simultaneous training procedure.

Blocked and Interleaved Training Schedules

During a typical perceptual category learning experiment, learners are usually shown instances of multiple categories. These instances are usually presented to a learner one after the other. The learner's task is to indicate which category each stimulus belongs to, and feedback is given after each response.

When multiple instances of each categories are shown in succession, different presentation schedules can be used. For an experiment with 3 categories, for example, instances from each of the three categories can be shown in an uninterrupted sequence. Thus, stimuli can be represented as [A1, A2, A3, ..., An, B1, B2, B3, …, Bn, C1, C2, C3, …, Cn] where the letter A, B, or C represents the category, and n examples of each category are displayed. This is generally referred to as a blocked schedule. Blocked schedules appear to be quite common in educational setting. For instance, in Dermatology textbooks, lesions are grouped either by their causes or physiology (Weller, Hunter, & Mann, 2015). Pictures of the same disorder, emerged in different parts of the body, or in different ethnic populations, are then shown in the same section (Wolff, Johnson, Saavedra, & Roh, 2017). The same information is unlikely to be shown in any other portions of the book. Learners are expected to acquire a generalized concept upon engaging with these pictures.

In contrast with blocked learning schedule, instances of different categories can also be presented in an interleaved fashion. Here, instances of the different categories are interspersed. One example of such a sequence of presentation would be [A1, B1, C1, B2, A2, C2, …, An, Bn, Cn,]. This schedule may occur in some informal learning environments. To help toddlers learn the concepts of trucks, buses, and trains, caregivers may show them one toy object at a time,

and provide the category name verbally. By showing instances of different categories successively, this kind of instruction forms an interleaved training schedule.

In their experiment, Kornell and Bjork (2008, Experiment 1a) had participants learn painting styles of 12 artists. Some paintings were presented using a blocked schedule, in which paintings of the same artist were presented one after another with the artist's name. Other paintings were presented using an interleaved schedule, in which paintings of different artists were interspersed with one another. In a transfer test, learners were presented with some new paintings that they had not seen during the training. Their task was to match paintings with the artists who created them. The learners did better when they learned the artists' styles with interleaved training.

Carvalho and Goldstone (2015) reviewed literature comparing blocked versus interleaved training schedules. They concluded that blocked design and interleaved design benefit different types of learning tasks. Blocked training, with very few alternations between categories across time, tends to promote creation of category prototypes. Associations of instances within each category tend be strong. Features that are subtle but strongly predictive of a category are more likely to be noticed (Goldstone, 1996). Hence, blocked training may be suitable for learning categories when categories are dissimilar with each other. On the contrary, two features associated with interleaved training, namely temporal spacing and temporal juxtaposition between instance of different categories, give rise to other forms of learning patterns. With interleaved design, instances of the same categories are interspersed among those of other categories, creating temporal spacing between repetitions of the same category. Temporal spacing between repetitions has been repeatedly shown to aid memory retrieval (Dempster, 1996; Cepeda, Pashler, Vul, Wixted, & Rohrer, 2006). In addition, temporal juxtaposition facilitates comparisons between categories. Interleaved training yields better learning outcomes when stimuli between and within categories are highly similar, and when there is a substantial delay between training and test.

Simultaneous Training Schedule

Mundy, Honey, and Dwyer (Experiment 4, 2007) had learners discriminate pairs of highly similar face stimuli. Each pair of face stimuli was generated by morphing two distinct faces at a slightly different level. A total of four face pairs (A and A', B and B', and so on) were generated. The experiment had two phases, exposure and discrimination. During the exposure phase, two pairs of faces were used in the simultaneous condition, the other two were used in the successive condition. In the simultaneous condition, a pair of stimuli (e.g., A and A') was presented side-by-side. Learners could compare the differences between the pair of images. In the successive condition, two identical faces were presented side-by-side. Learners were told to count the number of occurrences of each image, without being told the purpose of the study.

After all four pairs of faces were presented for 5 times, the experiment entered the discrimination phase. Each stimulus in a pair was given an arbitrary label, either "married" or "unmarried". One face was presented at a time, and learners had to indicate whether the face denoted a married or unmarried person. Feedback was given upon a response. Learners achieved higher accuracies with faces that were simultaneous presented in the exposure phase, compared to those presented successively.

Mundy et al's experiment was an explicit recognition task. Within each arbitrary category (i.e., married and unmarried), there were no common structures among instances. To perform the task well, all eight instances of the stimuli had to be maintained in the memory. Nevertheless, it showed simultaneous presentation aided discrimination between highly similar pairs of stimuli.

Other than this explicit recognition task, there has been evidence suggesting that superior simultaneous training effects compared to blocked schedule. In an attempt to tease apart the effects of temporal juxtaposition and spacing in interleaved learning, Kang and Pashler (Experiment 2, 2012) conducted a category learning experiment with blocked, interleaved, and

simultaneous training schedules. Three groups of learners learned painting styles of three artists. In the training phase, learners were given 10 paintings of each artist, together with the names of the artists attached to them. The three groups differed in the way that the paintings were presented. In the massed condition, paintings of each artists were presented successively in a block, before paintings of another artists were shown. In the interleaved condition, paintings were also presented successively, but those of different artists follow one another. In the simultaneous condition, a painting of each of the three artists was presented side-by-side in each trial. Critically, the transfer test included 90 paintings that the learners had not seen during training. Each painting was presented alone, and learners had to identify the artists who created the paintings. While accuracy in interleaved and simultaneous training schedules did not differ, learners in the two group outperformed those in the massed condition. The authors concluded that both interleaved and simultaneous training schedule encourage category discriminations, which leads to superior training outcomes compared to massed training.

The Current Study

The focus of the current study is the simultaneous presentation of examples from different categories. This mode of teaching has potential to be an effective strategy for category learning. It possesses features of blocked and interleaved schedules that are known to benefit learning. First, as in the blocked schedule, instances of all categories appear in every learning trial. Since learners engage with similar features across trials, a prototype of each category can be more readily generated. Second, as in interleaved schedule, simultaneous training schedule allows learners to interact with instances from different categories within close temporal approximation. Instances from different categories can be directly compared. Third, simultaneous presentation is ubiquitous in real-world educational settings. When a learner acquires knowledge about a concept, there are usually multiple examples from different categories. As mentioned above, this happens mainly in informal environments such as child-

and-caretaker interactions, but it is not completely novel even if implemented in formal educational settings.

Mundy et al (2007) and Kang and Pashler (2012) suggested that simultaneous stimulus presentation aids discrimination among highly similar categories. Their conclusions regarding simultaneous over successive (interleaved) advantage, however, were quite different. Mundy et al (2007) suggested that merely exposing learners with contrasting stimuli simultaneous, compared to successively, leads to improved categorization accuracy in subsequent supervised learning. On the other hand, Kang and Pashler suggested that performance of learners trained with simultaneous presentations was no better or worse than that with interleaved presentations. This discrepancy might arise from methodological differences, training duration, stimulus employed, choice of performance test, or a combination of these factors.

In the current study, we compare interleaved (sequential) schedule with simultaneous presentation. Learners engaged in a supervised learning task: they were trained to classify two categories of artificial stimuli (Experiment 1) and skin moles into benign and Melanoma (Experiment 2). Both experiments adopted a 2 × 2 design, with training and test schedules as factors. Each factor had two levels: stimuli were either presented sequentially or simultaneously. Learners were randomly assigned into the four independent conditions. The first two groups were *trained* with sequential schedule, in which only one stimulus was presented to the learner in each trial. Learners had to indicate which of the two groups the stimulus belonged to. Feedback was given upon response. The other two groups were trained with simultaneous presentation. Two stimuli were shown in each trial, and a learner's task was to indicate which of the two belonged to a predefined category. The overall time all learners spent engaging with instances from the two categories was controlled. Among the two groups trained with the same schedule, one of the groups was *tested* with the sequential schedule, the other was tested with a simultaneous schedule. No feedback was given to the learners during the test phase.

The two test schedules tap into different types of knowledge learners acquired during the training phase. In the sequential testing environment, only one stimulus was presented in each trial. To perform the task well, learners must maintain independent representations of both concepts. The test stimulus can then be compared with the stored representations. In contrast, the simultaneous testing environment requires learners to tell the differences between the two categories, with instances of both categories present. The design greatly reduces memory requirements of the categorization task. If one of the training schedules facilitates generations of prototype representations, the group should perform better on the sequential test. If a training schedule facilitates discrimination, the group should perform better on the simultaneous test.

Regardless of which training conditions a learner was assigned to, all of them had the same total amount of time to engage with instances from different categories. This way, any differences in performance in the transfer test reflect training efficiency between the two training schedules.

An Ideal Observer at Transfer Test

In order to test whether classification with simultaneous presentation is objectively easier than that with sequential presentation, or vice versa, we ran a simulation on a simplified version of the task. Instead of a visual categorization task, we turned the task into a numerical categorization task. The computer serves as an ideal observer in this task, assuming category structure is learned prior to performing the task, and the learning is perfect.

Stimuli were numbers, and they were drawn from two distributions. In real life, these numbers can represent some properties of an object, such as height, color, or shape. Category A had a mean value of -5, and Category B had a mean value of +5. The two distributions both had a standard deviation of 10. That is, the two categories were defined by a single dimension, and the two distribution means were separated by 1 standard deviation (Figure 2.1).

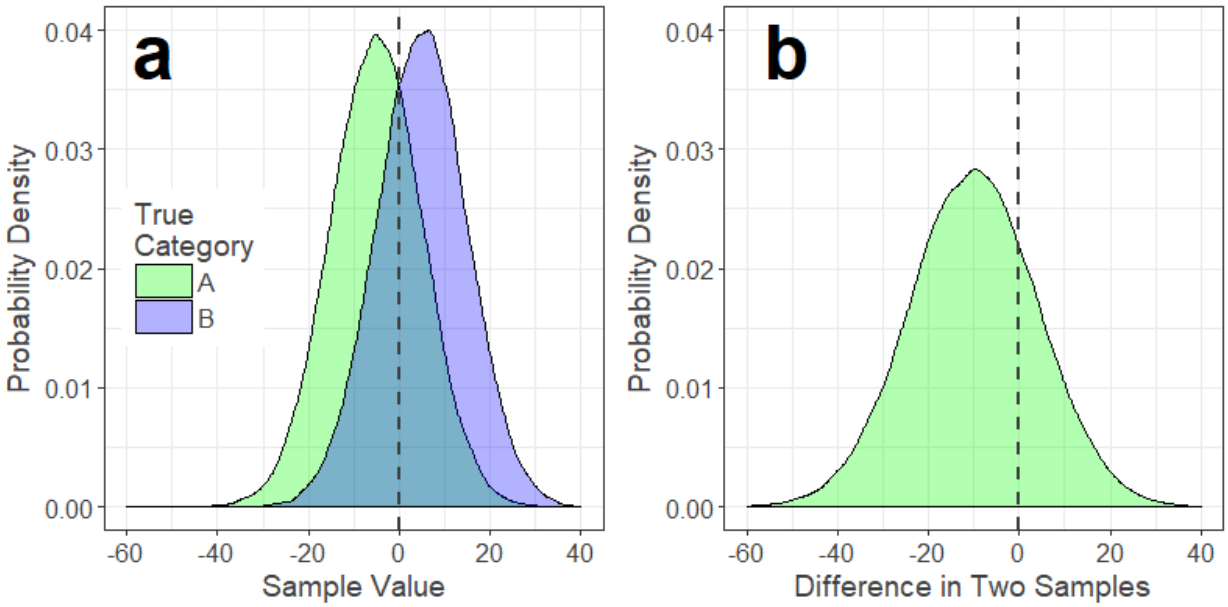


Figure 2.1. Simulations of categorization performance in the (a) Sequential and (b) Simultaneous condition. In the sequential display condition, the task of the observer was to categorize the number into Category A or B, using a criterion of 0. Average accuracy is 69%. In the simultaneous display condition, two sampled number were shown to the observer, and the tasks was to pick the number that belonged to Category A. Average accuracy was 76%.

The numbers were shown to the observer under two conditions. In the Sequential condition, one number was shown in each trial. The numbers were equally likely to be drawn from Category A and Category B. The observer had to decide which category the number was from. A logical boundary to separate the two categories would be the mid-point of the two distributions, which was zero in this case. For the observer, the task was straightforward: If the given number had a value smaller than zero, it would classified as Category A, otherwise, it would be classified as Category B.

In the Simultaneous condition, two stimuli were shown in each trial. One of the two stimuli was drawn from the Category A distribution, the other from the Category B distribution. The observer's task was to decide which of the two stimuli belong to Category A. To perform the task, the observer compared the values of the two stimuli, and assigned the one with a smaller value as Category A. This is because, by definition, Category A had a smaller overall mean value.

All simulation was done in R. The observer performs 100,000 trials in each condition. It scores an accuracy of 69% in the Sequential condition, and 76% in the Simultaneous condition. A sensitivity measure d' (Green & Swets, 1966; Macmillan & Creelman, 2004) was also calculated for each condition. The observer attained a d' of 0.98 in the Sequential condition, and 1.42 in the Simultaneous condition. This analysis clearly shows that Simultaneous presentation is superior in telling the two categories apart, provided that the category structures are optimally learned in the prior learning session.

Having established that Simultaneous presentation aids classification at transfer test, we asked whether the same presentation mode also improve training efficiency for human learners. Experiments 1 and 2 were designed to test the hypothesis.

Experiment 1

Methods

Design.

A 2 (training schedule: sequential versus simultaneous) × 2 (testing schedule: sequential versus simultaneous) factorial design was used in the experiment. Learners were randomly assigned into one of the four between-subjects conditions. Two conditions sharing an identical training schedule had different test conditions. In the *sequential training* conditions, one stimulus was presented at a time during the training phase. The two other conditions shared another training schedule: In the *simultaneous training* conditions, a pair of stimuli, one from each of the two categories, was presented during the training phase.

For the two conditions sharing the same training schedule, they differed in the way stimuli were presented during the *transfer test*. In the *sequential test* conditions, a single stimulus was presented in each trial. In the *simultaneous test* condition, a pair of stimuli, one from each of the two categories, were presented during the transfer test.

Participants.

One hundred and twenty participants (104 female) were recruited from the University of California, San Diego Psychology Subject Pool. Mean age of the participants was 20.3 years. The learners participated for course credit. They were randomly assigned into the four conditions, with 30 participants in each condition.

Stimuli.

The stimuli were cartoon-like figures which we called “demons” (Figure 2.2). Two categories of demons, namely “Old-word” and “New-world”, were used. All demons had the same box-shaped face in brown color. They varied in eye color, eye size, and horn height. For each demon, values of its eye color, eye size, and horn height were independently drawn from predefined distributions pertaining to its category. The mean difference for each dimension between the two categories was one standard deviation.

Stimuli were generated during experiment’s runtime. Hence, each learner encountered a unique set of demons, and no two demons in the stimulus set was identical except by coincidence. The experimental environment was written in JAVA programming language.

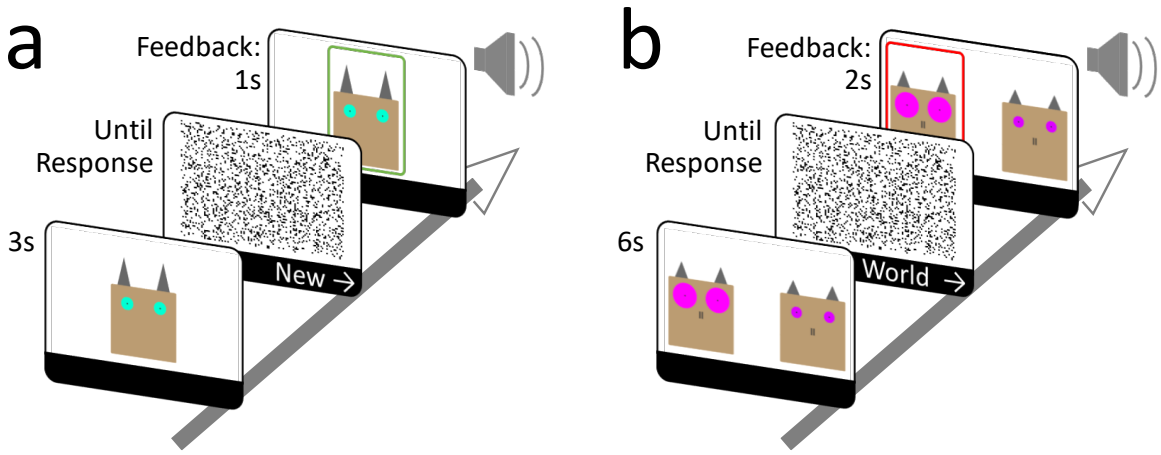


Figure 2.2. Stimuli used in Experiment 1. The two categories of demons differ in eye color, eye size and horn height. (a) A single demon was shown in each trial in the sequential schedule. Learners had to decide whether the demon came from the Old- or New-world. (b) A pair of demons, one from the Old-world, the other from the New-world, were shown in each trial in the simultaneous schedule. Learners’ task was to decide which of the two demons came from the Old-world.

Procedure.

All learners were first provided with an introduction to the experiment. They were told they would be learning to distinguish demons from the two categories based on their features. Learners were encouraged not to memorize specific demons, as the transfer test would include demons that had not been seen during the training phase. All learners were given a short quiz prior to the beginning of the training phase, to ensure they understood the experimental instructions. Depending on the conditions, stimuli were displayed differently in the training and transfer test.

In the *sequential training* conditions, learners were shown one demon at a time during the training phase. The demon belonged to either the Old-world or the New-world. The demon appeared at the center of the screen against a white background for 3 seconds. After the time elapsed, a static visual mask replaced the demon. The learner was then asked which category, Old- or New-world, the demon belonged to. Once the learner made a keyboard response, the same demon then reappeared for 1 second with a feedback tone and a colored frame surrounding the demon. The feedback tone and colored frame indicated whether a correct response was made. Half of the demons were from the Old-world, the other half from the New-world. The order in which the Old- and New-world demons showed up was randomized. The training phase contained 320 trials, divided into 4 blocks. A 30-second break was enforced between blocks.

In the *simultaneous training* conditions, learners were shown a pair of demons at a time during each training trial. Among the pair of demons, one demon was from the Old-world, the other was from the New-world. The pair were positioned side-by-side, with the Old-word demon on the left in half of the trials. The order in which the Old-world demon was on the left or on the right was randomized. The pair of demons appeared for 6 seconds, before a static visual mask showed up. The learner had to indicate which of the two demons, the one on the left or on the right, was an Old-world demon. The pair of demons then reappeared for 2 second with a

feedback tone and a color frame surrounding the Old-world demon. The feedback tone and frame denoted whether the learner made a correct response. Learners completed 160 trials in divided into 2 blocks. A 30-seconds break was enforced between the two blocks. The total time a learner in the sequential training engaged in the learning was 1280 seconds, and the same amount of training was given to the simultaneous training group.

Upon completing the training phase, the subject was given the transfer test. Here, subjects in the *sequential test* condition saw a single demon in each test trial. They had to decide whether the demon was Old-world or New-world. Responses were unspeeded and feedback was not given. Learners in the *simultaneous test* condition saw a pair of demons in each trial. One of those came from the Old-world, the other came from the New-world. Their task was to decide which demon belonged to the Old-world by pressing the corresponding key on a standard keyboard. A new trial began once the learner made a response using the keyboard. None of the demons in the transfer test was shown during the training phase. Learners in all conditions performed 100 transfer test trials without feedback.

Results

For data analysis purpose, we combined every two sequential training blocks to form an *epoch*. All learners, regardless of condition, had the same amount of exposure to instances in both categories in each epoch. The left panel of Figure 2.3 shows the results of those trained by a sequential schedule. The right panel shows the results of those trained by a simultaneous schedule. The last data point in each line denotes performance of transfer test.

Training accuracy.

We ran a 2 × 2 × 2 analysis of variance (ANOVA), with training schedule, test schedule, and epoch as the factors. Since the two training blocks within each training condition were identical, we did not expect any group difference within each training condition. A lack of significant main effect of test schedule ($F = 0.02$) supports the prediction.

To quantify the effectiveness of our training, we looked at the main effect of epoch. An improvement in classification was seen across training epochs [$F(1,116)$, $p < 0.001$]. Across training conditions, the simultaneous training was easier than the sequential training. Learners in the simultaneous training achieved a higher training accuracy, $F(1,116) = 56.64$, $p < 0.001$, $d = 1.19$, 95% C.I. on d = [0.80, 1.59]. This is not very surprising as each display in the simultaneous training schedule contained more information compared to one in the sequential training schedule, as mentioned in the simulation above.

Transfer test accuracy.

The overall time learners spent on the two categories of demons was controlled for across all training conditions, so performance in the transfer test is a good indication of learning effectiveness. Transfer test accuracies of the four conditions were submitted to a 2 × 2 ANOVA, with training and test schedules as factors. The main effect of training schedule is not significant [$F(1,116) = 1.72$, $p > 0.1$], suggesting that, in general, neither training schedule was more effective than the other.

There is a significant main effect of test schedule, $F(1,116) = 13.06$, $p < 0.001$, $d = 0.65$, 95% C.I. = [0.27, 1.02], and a marginal interaction between training and test schedules [$F(1,116) = 3.76$, $p = 0.05$]. Aggregating learners' accuracies for each condition, those trained and tested with simultaneous displays attained much better performance in the transfer test (75.6%) than those trained with sequential displays (69.9%, $p = 0.10$, $d = 0.61$, 95% C.I. on d = [0.07, 1.14]), according to Tukey Honest Significant Difference (Tukey HSD) Test. It also outperformed the two other groups tested with sequential displays (mean accuracy = 66.4%, ps < 0.01). The rest of the pairwise comparisons are not significant (ps > 0.3).

To conclude, learners attained the highest classification accuracy when they were trained and tested in the simultaneous schedules, while training effectiveness of sequential schedule was simiar regardless of the type of test used.

Comparing Training and Transfer Test.

The simulation we did in the introduction reminds us that sequential test and simultaneous test have different sensitivities, even if learning is optimal. To test whether sequential or simultaneous training is better, we should compare the transfer test performance of each condition with the first training epoch in a training condition that had identical design. As shown in Table 1, the transfer test of Group 1 should be compared to the first training epoch of the same group. The transfer test of Group 2 should be compared to the first training epoch of group 3, and so on.

First, we examine the conditions when the training involved sequential presentations of stimuli. Comparison 1 yields a significant difference between the transfer test and first training epoch, $t(29) = 5.49$, $p < 0.001$, $d = 1.00$, 95% C.I. = [0.44, 1.56]. In other words, sequential training is effective when the transfer test also presents stimuli sequentially. Comparison 2 does not yield any significant differences, $t(58) = 0.01$, $p = 0.99$. It suggests that sequential training was task specific, such that the knowledge learned during the training was not detected by a transfer test with simultaneous stimulus presentations.

We then compare the conditions which presented stimuli simultaneously during training. Regardless of the structure of the transfer test, learning was detected. For Comparison 3, a sequential transfer test indicates learning using simultaneous training, $t(58) = 3.13$, $p < 0.01$, $d = 0.8$, 95% C.I. = [0.26, 1.36]. For Comparison 4, a simultaneous transfer test was able to detect learning when the training also involved simultaneous stimulus presentations, $t(29) = 3.22$, $p < 0.01$, $d = 0.59$, 95% C.I. = [0.05, 1.13].

Table 2.1. Comparison scheme. Transfer test of a condition was compared to the first epoch of a condition that had the same training structure.

Comparison	Group	Training	Transfer Test	Compared with Epoch 1 of
1	1	Sequential	Sequential	Group 1
2	2	Sequential	Simultaneous	Group 3
3	3	Simultaneous	Sequential	Group 2
4	4	Simultaneous	Simultaneous	Group 4

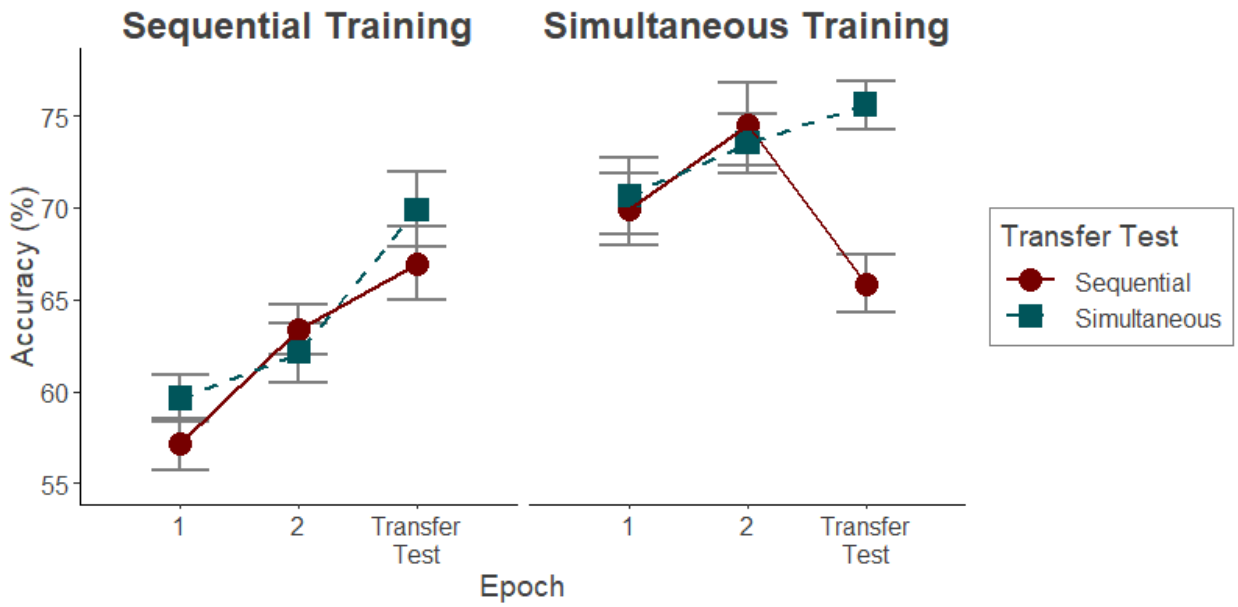


Figure 2.3. Results of Experiment 1. Left panel: sequential training conditions separated by transfer test schedules. Right panel: simultaneous training conditions separated by transfer test schedules. The group being trained and tested with simultaneous schedule performed the best. The two groups being tested with sequential schedule performed the worst in the transfer test, and their performance did not differ from each other. Error bars show the standard errors of the means.

Discussion

Training using simultaneous displays produced better performance during learning in terms of accuracy, which was expected given our ideal observer analysis. When two stimuli are simultaneously presented in the same trial, there is more information a learner can use for the categorization, compared to a sequential display. As shown in our simulation, when the task is to tell the difference between the two, it is objectively easier than trying to label a single stimulus in the display.

Experiment 1 replicated the results from Kang and Pashler (2012), who showed that learners trained with either sequential or simultaneous presentations did not differ in a transfer test using sequential displays. In addition, the current experiment expands our understanding on what had been learned during the training phase. While sequential training was only helpful when the transfer test also involved sequential displays, but not when the transfer test utilized

simultaneous displays. It shows that sequential training failed to help learners utilize the extra information in the simultaneous transfer test. This is perhaps related to context effect reported by Godden and Baddeley (1980), in which knowledge learned in a particular context cannot be applied in a new environment. On the contrary, when learners were trained with simultaneous displays, improvements were seen regardless of the format of the transfer test. It means that the learners did not only acquire knowledge about the differences of the two categories, they also had knowledge about the defining features of each categories.

The artificial stimuli used in the experiment have several desirable features in the investigation of perceptual. For one thing, the complexity of the category structures can be quantitatively manipulated. We could specify the number of features that defined each category, the distributions of feature space, the interactions between features, and the degree of similarity between categories. In many real-world situations, categories are usually less well-defined. To test the generalizability of the simultaneous training advantage to a more naturalistic and perceptually challenging task, images of malignant or nonmalignant skin moles were used in Experiment 2.

Experiment 2

Methods

Participants.

Participants were recruited in the same way as in Experiment 1. One hundred and sixty participants (109 female) were recruited in this experiment. Mean age of the participants was 20.7 years. They were randomly assigned into the four conditions, with 40 in each condition.

Materials.

Stimuli were obtained from an online database for a Machine Learning challenge (ISIC, 2016). The data set contains 1000 annotated images, 727 benign skin mole and 173

Melanomas. Duplicated images were removed from the image set. Images that contain obvious labels were cropped, such that the resulting image contains only the area of interest. For each participant, 80 images from each category were randomly selected for the training phase, 90 images from each category were randomly selected for transfer test phase. Due to a relatively small set of Melanoma images, the training images were cycled twice in different orders. None of the training images were used in the transfer test.

The images were shown on a Dell E173FPc 19-in LCD monitor with 4:3 aspect ratio. The resolution of the screen was set to 1024 × 768. The images were rotated such that its longer axis aligned with the height of the screen. Each image had a maximum width of 500 pixels and maximum height of 666 pixels.

Design.

The basic design of the experiment was similar to Experiment 1. The experiment was a 2 × 2 between-subjects design, with training schedule being one factor, and testing schedule being another factor. As in Experiment 1, each factor was divided into two levels: sequential and simultaneous displays. Participants were randomly placed into one of the four conditions.

Procedure.

All learners were informed about the purpose of the study in the introduction. They also received some brief information about Melanoma and how it is usually detected. The learners completed a quiz before the training phase.

The structure of the training phase depended on the condition a learner was in. The first two groups were trained with sequential displays (Figure 2.4a). The training contained 320 trials in 4 blocks. During training, learners saw a single image denoting either a Melanoma or a benign skin mole for 3 seconds. A static mask then showed up, replacing the images. The learner gave an unspeeded response once the static mask shows up. Audio and visual feedback were given on a trial-by-trial basis. The same image showed up again for 1 second

during the feedback screen before a new trial began. The order in which a Melanoma or benign skin mole showed up was randomized.

The two other groups were trained with simultaneous displays (Figure 2.4b). Two images, one benign mole and another Melanoma, showed up side-by-side for 6 seconds, before a static visual mask replaced the images. The positions of the benign mole and Melanoma images were randomly determined in each trial. Learners had to indicate which of the two images was Melanoma by pressing a key on the keyboard. Upon response, an audio feedback was given to the learner. The same two images reappeared for 2 seconds, with a colored frame indicating the Melanoma image. A new trial began after the feedback was shown.

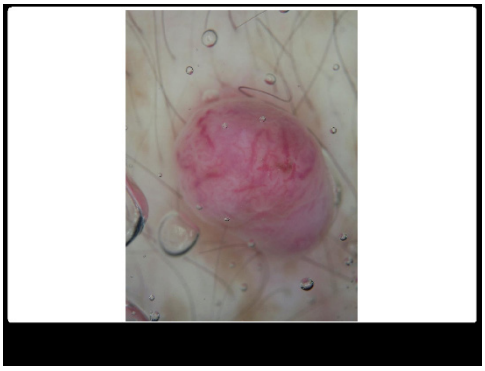
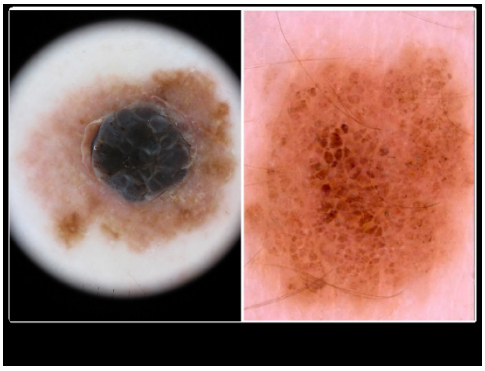
(a)  (b) 

Figure 2.4. Examples of displays used in the Experiment 2. (a) Display for sequential schedule. Learners had to decide whether the image shown was benign or Melanoma. (b) Display for simultaneous schedule. Learners had to decide which of the two images was a Melanoma.

In the transfer test, images were displayed individually or in pairs depending on the condition a learner was in. Those assigned to the sequential test schedule saw one image at a time. They had unlimited time to look at an image before arriving at a decision. Learners indicated whether the image shown was benign or Melanoma by pressing a key on a standard keyboard. Learners assigned to the simultaneous test schedule saw a paired of images on each trial, one benign and one Melanoma. Their task was to decide which of the two images was a Melanoma. The images remained on the screen until learner responded. No feedback was given during the transfer test. The experiment ended after learners went through 90 trials.

Results

Figure 2.5 shows the results of Experiment 2. As in Figure 2.3, the left panel shows performance of the groups that were trained with sequential displays, the right panel show performance of the groups that were trained with simultaneous displays. The first four data point on each line denote training accuracy. The last data point of each line denotes the transfer test performance.

Training accuracy.

Accuracies in the training phase were submitted to a 2 × 2 × 4 ANOVA, with training schedule, transfer test schedule, and block as the factors.

In the training blocks, there is a main effect of training schedule, $F(1,156) = 178.8$, $p < 0.001$, $d = 2.13$, 95% C.I. on d = [1.73, 2.52]. It shows that learners performed better in simultaneous training than in sequential training in terms of accuracy. Again, this was expected given our simulation, as the task in simultaneous training was objectively easier. However, performance does not differ within each training condition. In the last training block, the two sequential groups attained a mean accuracy of 62.2%, the two groups did not differ from each other, $t(78) = 0.29$, $p = 0.77$). The two simultaneous groups attained a mean accuracy of 71.9%, the two groups did not differ from each other, $t(78) = 0.27$, $p = 0.78$.

In general, learners' performance improved across blocks [$F(3,468) = 7.748.0$, $p < 0.001$]. This improvement could be due to repetition of stimulus used in Blocks 3 and 4. Even when the improvement is significant, the actual size of the improvement is small, as shown in Figure 2.5.

Transfer test accuracy.

The accuracy data was submitted to a 2 × 2 ANOVA, as in Experiment 1. The two factors are training schedule and test schedule. The ANOVA indicates a main effect of transfer

test schedule, $F(1,156) = 53.8$, $p < 0.001$, $d = 1.16$, 95% C.I. = [0.82, 1.50], but not a main effect of test schedule, $F < 1$, nor an interaction between the two factors, $F < 1$.

Comparing Training and Transfer Test.

We conducted the same analysis we did in Experiment 1, comparing performance of the transfer test in each condition with the corresponding training blocks. The plan for comparisons is detailed in Table 1.

Comparisons 1 and 2 look at the learning with sequential displays. When both the training and the transfer test involved sequential displays, learners showed clear improvements, $t(39) = 3.91$, $p < 0.001$, $d = 0.62$, 95% C.I. on d = [0.16, 1.08]. When the transfer test involved simultaneous displays, the test was not able to detect learning, $t(78) = 0.61$, $p = 0.54$.

Comparisons 3 and 4 look at the learning with simultaneous displays. When the transfer test showed learners one stimulus at a time, they showed clear learning, $t(78) = 2.42$, $p = 0.02$, $d = 0.54$, 95% C.I. on d = [0.08, 1.00]. When the transfer test involved simultaneous presentation of stimuli, the transfer test failed to detect changes in learning, $t(39) = 0.36$, $p = 0.72$.

To summarize, learners showed improvements only when they encountered a sequential transfer test, but not a simultaneous one. Improvements were seen regardless of the type of training the learner received.

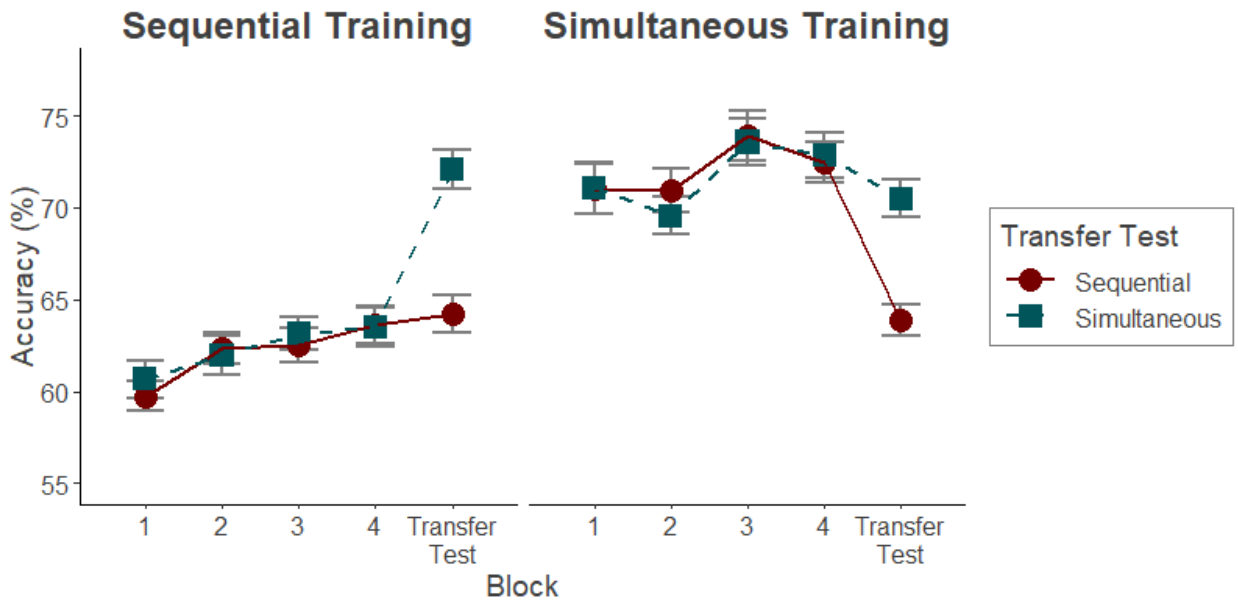


Figure 2.5. Results of Experiment 2. Left panel: sequential training conditions separated by transfer test schedules. Right panel: simultaneous training conditions separated by transfer test schedules. The group being trained with sequential schedule, and tested with simultaneous schedule performed the best. The two groups being tested with sequential schedule did not differ in the transfer test. Error bars show the standard errors of the means.

Discussion

The learning captured by an hour of training with the skin mole stimuli are moderate, and the rate of learning in Experiment 2 was slower than in Experiment 1. This is clearly expected: the stimuli used in Experiment 2 are less well-defined and of higher dimensions. Interactions between different feature dimensions, e.g., color and shape, are also expected, which potentially hindered learning.

Nevertheless, benefits of training were seen over the relatively short period of time. Consistent with Experiment 1, performance of the two groups tested with sequential displays did not differ (Figure 2.6, right panel). Both groups performed reliably better in the transfer test, compared to the first training blocks of their corresponding control conditions. When the transfer test explicitly required learners to distinguish the differences between two stimuli, no clear benefits were manifested. It shows that neither training method has a clear advantage over the other.

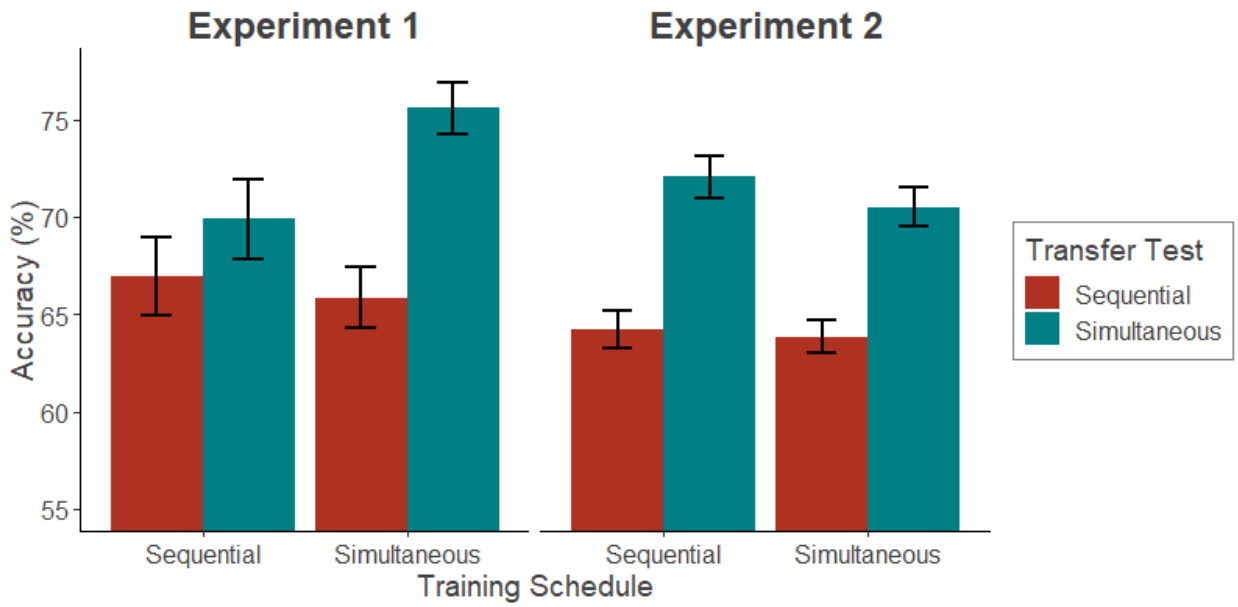


Figure 2.6. Performance of transfer tests in Experiments 1 and 2. In both experiments, learners' performance in the transfer test did not differ when it was in sequential schedule. In Experiment 1, when the stimulus structure was relatively simple, and learners were trained with stimuli from two opposite categories, simultaneous training helped learners tell the two categories apart. The same benefit did not manifest in Experiment 2.

General Discussion

Relationship to Previous Findings

In two experiments, we examined the effectiveness of sequential and simultaneous training schedules. The first experiment illustrates a relatively ideal case in studying the question, as the stimuli were novel to learners, and the distributions of features were well defined. Stimuli in the training phase never repeated, nor did those in the transfer test. The epoch-wise accuracy curves illustrate how expertise develop under different conditions. Experiment 2 asked whether the main finding in Experiment 1 holds for learning more naturalistic stimuli.

We asked whether the two training schedules produce the same transfer test outcomes. In general, our experiment results were consistent with Kang and Pashler (2012) that sequential and simultaneous training schedules are not different in effectiveness when stimuli were presented using a sequential schedule in the transfer test. In addition, our experiments showed

that an added benefit to train learners using simultaneous training schedules. The training helps learners to discriminate the two categories better when the stimuli have a relatively simple structure.

Our experiments differ from Mundy et al (2007) and Kang and Pashler (2012) in the way learners acquired knowledge. Mundy et al report experiments that stimuli were first exposed to learners without informing them about the task. In their experiments, a small set of stimulus were utilized. There were also no structural similarities among members of the same category. To perform recognition task well, learners had to remember details of each stimulus presented to them. In both of our experiments, there is a shared underlying structure among members of the same category, albeit less well-defined ones in Experiment 2. Stimulus used in the transfer test were also not utilized in the training, to encourage learning of the category structure rather than idiosyncratic features of individual stimulus.

Our experiment also differ from Kang and Pashler (2012) in an important way. In their experiment, stimuli were presented together with the category label in the training phase. In contrast, our experiments required learners to associate features of the stimuli with the categories through feedback in the training phase. Hence, the tasks in both the training and test phases were structurally more similar compared to the Kang and Pashler's experiments.

Sequential versus Simultaneous Training Schedules

With the rich literature on category learning, there is little doubt that different training schedules lead to differential training outcomes. In general, blocked training schedule is beneficial when similarity, both within and between categories, is low. In such cases, it is likely that a prototype for each category is generated when instances encountered (e.g., Minda and Smith, 2000). Blocked schedule facilitates the process because instances of the same category are encountered without interruptions in between. When new instances are encountered, they can be compared to the prototypes of each category generated during the training phase. The

goal for interleaved training is largely the same: it encourages learners to generate a prototype, or some representations for each category. The temporal juxtaposition of categories allows slightly easier comparisons between categories. As a result, interleaved training schedule is beneficial when differences between categories are more subtle, such that finer discriminations between them can be made (Kang & Pashler, 2012; Kornell & Bjork, 2008).

Simultaneous training schedule, compared to the sequential (interleaved) schedule, further promotes discriminations between categories. Instances of two categories were shown in the same trial, and learners' task was to tell the difference between the two. To perform the task well, prototypes or representations of each category are not necessary, as long as the learners can associate the relative strength of the features with the categories. Therefore, simultaneous training schedule has the potential to improve discrimination between categories that have finer differences, beyond what learners using blocked or interleaved schedule are able to achieve. Our results support the argument: simultaneous stimulus presentation is at least as effective as sequential stimulus presentation as a training method, and is superior when the category structures are simple and well-defined.

To an educator or trainer, regardless of training schedules, the goal is help learners distinguish contrasting concepts. The training schedules are likely to invoke different categorization heuristics. Blocked and interleaved training schedule promote generation of category prototypes, while interleaved training further encourages discrimination between categories. For learners trained with simultaneous schedule, they can achieve the training goals by either generating category prototypes or attend to the differences between categories. The two methods are not mutually exclusive, and they can be used interchangeably. Therefore, simultaneous training schedule should promote more routes to meet the goals.

Difference between Experiments 1 and 2

Experiments 1 and 2 show that sequential and simultaneous training do not differ when the transfer test shows stimulus sequentially. It means that both methods are equally effective to promote generations of category prototypes.

In the simultaneous test displays, learners were provided with some added information. The structure of the transfer test allow us to assess whether learners can utilize the added information. In Experiment 1, subjects were able to utilize the information, but not in Experiment 2. It is important to explore why this pattern emerges.

Experiment 1 differs from Experiment 2 in a number of ways. We suspect that some of the differences, or a combination of them, may be responsible for the patterns. First, the stimuli employed in the two experiments were fundamentally different as intended. The cartoon stimuli used in Experiment 1 were generated in a highly controlled manner. The distribution of each feature, as well as the difference between categories, are well-defined. For each defining feature, the means between the two categories were exactly one standard deviation. Features were also equally predictive of the categories. Variations between features are independent of each other. On the other hand, skin moles used in Experiment 2 were less well-defined. It was less clear how well the categories, benign or Melanoma, was defined by the visual features of the stimuli. It is known that some visual features are predictive of the categories, such as symmetry, border, color, and size (Rigel, Russak, & Friedman, 2010), but their relative strength of prediction cannot be reliably measured. It is also unclear how these features interact with each other. Hence, Experiment 2 is likely to be a high-dimensional categorization task, compared to Experiment 1.

Second, the length of the training sections in both experiments appeared to be suboptimal, but to a different degree. It is clear that learners were not trained to their full capacity Experiment 1. In the conditions which the training and the transfer tests had the same

structure, the transfer test performance was higher than those in the last training epoch. In comparison, in Experiment 2, performance seemed to have reached an asymptote by the last training block. This accuracy is much lower than an accuracy of 79.3% reported in the literature using the same categorization method (Anessi, Bono, Sampogna, Faragginana, & Abeni, 2007). It shows that there is room for improvement if our learners were extensively trained. To conclude, learners in both experiments were not trained to their full capacity, but those in Experiment 1 were better trained for their task than those in Experiment 2 for theirs. The less "competent" learners in Experiment 2 might have chosen a suboptimal strategy during the transfer test.

Third, training stimuli in Experiment 1 never repeated, but those in Experiment 2 were cycled through twice during the training phase. It is possible that some idiosyncratic features in the training stimulus set were incorrectly attributed to the categories. These features may not generalize when a new set of stimuli was used in the transfer test. Lastly, Experiment 1 was a pure bottom-up supervised learning task, while Experiment 2 was not. Prior to Experiment 1, learners had no knowledge about the artificial stimuli. This is not the case in Experiment 2. When asked whether they have any knowledge about Melanoma, a majority of learners indicated that they have at least heard of the disease. Some indicated that they have family members suffering from the disease. Due to the complexity of the categorization task in Experiment 2, we provided learners with some general diagnostic criteria about Melanoma before the experiment began. Some learners might have tried to apply these rules during the training phase. These factors were reflected in the above-chance performance in the first training block. Hence, learners in Experiment 2 possessed both top-down knowledge through previous encounters of the disease and the instructions, and bottom-up experience through supervised training. It is unclear how this affect learners' choice of strategies in the transfer test, but the prior knowledge has likely lowered the effect size of our training program.

Moving Forward

In the two experiments we report here, we showed that sequential and simultaneous training schedules lead to different training outcomes, expanding Kang and Pashler's (2012) finding. The experiments also illustrate that determining the best strategy for a categorization task is not straightforward. It depends on a number of factors. We will discuss a few of these issues for future investigation.

First, decision of training schedule should be based on the goals of the training program, being assessed with a transfer test. When the goal is identification, i.e., to tell the categorize given a stimulus, either sequential or simultaneous training schedule does not seem to matter. If the goal is to discriminate two contrasting stimulus, the benefits of sequential or simultaneous training are less clear. When the categorization task is relatively simple and has low-dimensionality, simultaneous training is likely to work better. When the task has high dimensions, other factors in the learning program would probably have to be considered. Regardless, it is worth noting that simultaneous training produces learning outcomes that are at least comparable to sequential training.

Second, the time course for training deserves more scrutiny. In the current experiments, overall training time was around 30 minutes. While we show improvements over time during the training phase and such short training is not uncommon in real-world education, there are merits to examine effects of these training schedules for a more extensive training. The short training period taps into how category representations are generated during initial learning. More extensive training would probably inform us about memory retention of these representations over time.

In the current experiments, we controlled for the overall time learners interacted with the stimuli. Learners in the four conditions within an experiment had the same amount of time to study the stimuli. Related to the issue above, a train-to-criterion strategy can be used instead. In

such case, the training phase terminates once learners achieve a predefined block-wise accuracy. Performance of the transfer test would then indicate the degree of transfer and memory retention.

Third, our results in the two experiments strongly suggest that category structure plays a role in training effectiveness. In Experiment 1, stimuli were relatively simple, with no interactions between features. Simultaneous training schedule led to a superior transfer test result when the test presented stimuli in pairs. Maddox and Filoteo (2011) suggested that different category structure may better support categorization using prototypes or exemplars, the same argument may apply in the case of different training schedules. It is also possible that these category structures activate different memory systems (Ashby & O'Brien, 2005), which are better trained with either sequential or simultaneous schedules.

Fourth, we used only two categories in our two experiments. In real-world settings, learners probably engage in more challenging situations with multiple categories. Given that our experiments replicated that of Kang and Pashler (2012), it is likely that our results would extend to a multiple-category situation. Future experiments can confirm this prediction.

Conclusion

The ability to categorize objects forms the basis of many higher order cognitive functions. For instance, one has to be able to differentiate various mathematical problems before attempting to solve them with the correct methods (e.g., Rohrer, Dedrick, & Burgess, 2014; Rohrer, Dedrick, & Stershic, 2015). Categorization also serves many societally important functions. For one thing, we have to be able to classify a newly met person as a male or female, before interacting with him or her in a socially appropriate manner. The importance of high quality categorization can hardly be dismissed: distinguishing life-threatening items from safe ones in airport baggage search, or benign body tissues from malignant ones, to name a few. Our experiments show that simultaneous training schedule, i.e., putting instances of contrasting

categories, could be a more effective strategy depending on the task demand. Throughout training procedure, learners also attained a higher accuracy. As learners are making fewer errors, they may be less frustrated in a prolonged training program. It has the potential to motivate learning. Future studies should look into those situations and assess when it benefits human learners the most.

Chapter 2, in full, was prepared to be submitted to a journal: Lau, J. S. H. & Pashler, H. Simultaneous and Interleaved Training in Perceptual Category Learning. The dissertation author was the primary investigator and author of this paper.

References

Aldridge, R. B., Glodzik, D., Ballerini, L., Fisher, R. B., & Rees, J. L. (2011). Utility of non-rule-based visual matching as a strategy to allow novices to achieve skin lesion diagnosis. Acta dermato-venereologica, 91(3), 279-283.

Ashby, F. G., & O'Brien, J. B. (2005). Category learning and multiple memory systems. Trends in cognitive sciences, 9(2), 83-89.

Annessi, G., Bono, R., Sampogna, F., Faraggiana, T., & Abeni, D. (2007). Sensitivity, specificity, and diagnostic accuracy of three dermoscopic algorithmic methods in the diagnosis of doubtful melanocytic lesions: the importance of light brown structureless areas in differentiating atypical melanocytic nevi from thin melanomas. Journal of the American Academy of Dermatology, 56(5), 759-767.

Brown, N. H., Robertson, K. M., Bisset, Y. C., & Rees, J. L. (2009). Using a structured image database, how well can novices assign skin lesion images to the correct diagnostic grouping?. The Journal of investigative dermatology, 129(10), 2509.

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.

Cepeda, N. J., Pashler, H., Vul, E., Wixted, J. T., & Rohrer, D. (2006). Distributed practice in verbal recall tasks: A review and quantitative synthesis. Psychological Bulletin, 132, 354–380.

Chandrasekaran, B., Yi, H. G., & Maddox, W. T. (2014). Dual-learning systems during speech category learning. Psychonomic bulletin & review, 21(2), 488-495.

Contraire , B. (2016). Undercover: One of These Things is Almost Like The Others. UK: Phaidon Press.

Dempster, F. N. (1996). Distributing and managing the conditions of encoding and practice. In E. L. Bjork & R. A. Bjork (Eds.), Handbook of perception and cognition: Memory (pp. 317–344). San Diego, CA: Academic Press.

Gauthier, I., & Tarr, M. J. (1997). Becoming a "Greeble" expert: Exploring mechanisms for face recognition. Vision Research, 37(12), 1673-1682.

Godden, D. R., & Baddeley, A. D. (1975). Context-dependent memory in two natural environments: On land and underwater. British Journal of psychology, 66(3), 325-331.

Goldstone, R. L. (1996). Isolated and interrelated concepts. Memory & cognition, 24(5), 608-628.

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

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

Kost, A. S., Carvalho, P. F., Goldstone, R. L. (2015). Can You Repeat That? The Effect of Item Repetition on Interleaved and Blocked Study. Proceedings of the 37th Annual Conference of the Cognitive

Science Society.

Kovecses, A. (2015). One Thousand Things: learn your first words with Little Mouse (Learn with Little Mouse). Wide Eyed Editions.

Maddox, W. T., & Filoteo, J. V. (2011). Stimulus range and discontinuity effects on information-integration category learning and generalization. Attention, Perception, & Psychophysics, 73(4), 1279-1295.

Medin, D. L., & Schaffer, M. M. (1978). Context theory of classification learning. Psychological review, 85(3), 207.

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.

Nosofsky, R. M. (1991). Relation between the rational model and the context model of categorization. Psychological Science, 2(6), 416-421.

Posner, M. I., & Keele, S. W. (1968). On the genesis of abstract ideas. Journal of experimental psychology, 77(3p1), 353.

Rigel, D. S., Russak, J., & Friedman, R. (2010). The evolution of melanoma diagnosis: 25 years beyond the ABCDs. CA: a cancer journal for clinicians, 60(5), 301-316.

Rohrer, D., Dedrick, R. F., & Burgess, K. (2014). The benefit of interleaved mathematics practice is not limited to superficially similar kinds of problems. Psychonomic Bulletin & Review, 21, 1323-1330.

Rohrer, D., Dedrick, R. F., & Stershic, S. (2015). Interleaved practice improves mathematics learning. Journal of Educational Psychology, 107, 900-908.

Scarry, R. (1998). Richard Scarry's Cars and Trucks and Things That Go. Golden Books.

Weller, R. B., Hunter, H. J. A., & Mann, M. W. (2015). Clinical Dermatology, Fifth Edition. Oxford, UK: John Wiley & Sons Ltd.

Wolff, K., & Johnson, R. A., Saavedra, A. P., & Roh, E. K. (2017). Fitzpatrick's color atlas and synopsis of clinical dermatology. McGraw-Hill Education.

CHAPTER 3: Low Target-Distractor Discriminability in Visual Search Promotes Detailed Target Template Generation

Abstract

When you search repeatedly for a set of items among very similar distractors, how does this impact the way you perform search? To address this, we ask in the context of a real-world object visual search task how people's visual search performance improves over time when distractors are repeated across trials, and then what happens to search performance when distractors change suddenly. Our results suggest that in the presence of repeated distractors, participants do not generate target templates specific to the exact distractors. However, the difficulty of the search task does impact the precision of the target template participants use. In particular, a coarse target template is created when the target and distractor are easy to discriminate, and hence the search task is easy. These coarse target templates do not transfer well to new distractors. Target templates with greater detail are generated when the search task requires more difficult target-distractor discrimination, and these detailed target templates better survive a distractor change. This suggests that discriminability between target and distractors has a large impact on long-term search performance by impacting the level of detail contained in the target template people use to search.

Introduction

Imagine an airport security employee has been screening bags for several hours. She is searching for prohibited items in passengers' carry-on bags as they pass through an x-ray machine. She is regularly finding targets, since the scanning system occasionally inserts prohibited items in the digital images of the bags in order to keep the employee vigilant. The task is known to be mentally demanding, and prone to high error rates (Wolfe, Horowitz, & Kenner, 2005). During her shift, the staff member performs visual search on thousands of bags -- but they all have something in common. In particular, the 'distractors' tend to be the same in all

cases, as all of the bags mainly contain clothing. If a passenger puts their prohibited item in a bag filled not with clothes, but among unusual objects (e.g., stuffed animals), would this make the airport security employee more likely to miss the target? Has she learned a distractor-specific search template for prohibited items, or a general target template that works equally well for all distractors?

To address this, we need to know both how people's visual search performance improves over time when distractors are repeated across trials (e.g., searching among clothing repeatedly), and if there is a significant cost for switching the distractors (e.g., distractors switched from clothing to stuffed animals). This question is difficult to study using a traditional visual search paradigm where targets and distractors are simple shapes (Treisman & Gelade, 1980) or letters (Wolfe, Klempen, & Dahlen, 2000) because the feature sets of the targets and distractors are small and there are only a small number of relevant features of the target. This limits the strength of their associations and the possible influence of the distractors on the learned target template. In addition, the task is often much easier than baggage screening, leading to fast response times (<1000ms) that could mask performance changes in a new context. In such situations, maximum learning effects are achieved within a few trials, and people can thus quickly disengage from the current search context. Thus, we sought to examine this question in a more naturalistic environment where the visual stimuli are complex, which allows us to address the question of how specific the target template is for natural images, and how it is tied to a specific set of distractor images.

The role of distractors in search

The issues at stake in the current work have roots in separate lines of research of visual search. In particular, it is well known that both the difference between the targets and distractors (e.g., Avraham, Yeshurun, Lindenbaum, 2008) and the heterogeneity of the distractors play an important role in visual search (Duncan & Humphreys, 1989). In fact, distractors play a

complicated role in search, both because they affect the target templates that are available to participants and because rejecting distractors may be easiest if they are homogenous. For example, replacing a set of distractors with new distractors that are more easily distinguished from the target can degrade search performance if it is done in a way that makes the distractors more heterogeneous (Rosenholtz, 2001). In general, search performance is best when people know in advance the exact features that define the distractors and the targets (Schoonveld, Shimozaki, & Eckstein, 2007; Eckstein, 2011). In fact, in some situations it has even been shown that participants often base their target template not on the optimal way of detecting the target but on the optimal way of distinguishing the target from the distractors (e.g., Navalpakkam and Itti, 2007).

Learning distractors

In addition to the influence of distractors during a given visual search, our question also depends critically on the extent to which people form memories of distractors. This question has been well studied. For example, Wolfe, Klempen, and Dahlen (Experiment 6, 2000) studied the effect of repeated distractors in visual search. In their experiment, they had participants searched for a target letter among 3 or 5 letter distractors. In the repeated condition, the locations and identities of the stimuli were the same on every trial. Participants searched for a different target on each trial, which was indicated to them at the beginning of the trial. Wolfe et al. (2000) found that participants became significantly faster in this condition than in a condition where the identities of target and all distractors changed in every trial. However, even in the repeated condition, participants never became as good as if they simply memorized the display and searched their memory. This suggests that people do form and make use of distractor information at least under some circumstances, but that even when the distractors are sometimes targets, people do not form perfect distractor memories or optimally take into account the ways in which search stays the same on every trial.

This question, in the broadest sense, has also been studied under the domain of contextual cueing (Chun & Jiang, 1998, 1999). In a typical contextual cueing paradigm, participants search for a simple pre-defined target (e.g., a T) among similar distractors (L's). Without explicitly communicated to the participants, the locations of the distractors in some trials are predictive of the location of the target. With repeated presentations of the pairing during the training, participants become faster at locating the targets when they are at a learned position relatively to the distractors, compared to when the target is in a new location. Thus, even with very simple stimuli, information about the distractors is not completely lost. In fact, which shapes tend to be the distractors for a given target is also learned in a contextual cueing-like setting (Chun & Jiang, 1999). Thus, people can associate not only spatial configurations of distractors with the location of target, but even particular sets of distractors with particular targets.

With real-world objects, it is also known that participants tend to perform reasonably well at incidentally remembering items that have been presented in search displays, either as targets or distractors -- particularly distractors that closely resemble the targets (e.g., Williams & Henderson, 2005). The effects of repeated distractors on search performance has also been studied using eye-tracking (Yang, Chen & Zelinsky, 2009). The researchers had participants search for targets in the same category, while repeating some of the distractors across trials. In the experiment, first fixations were slightly less likely to fall on an old distractor, compared to distractors that had not been repeated as frequently or new distractors. In the study, there was a high variability among the distractors, and the discriminability between the target and distractors was not manipulated. Hence, the effects of context change would be hard to assessed.

Categorical target templates

When people search for a target among distractors, a representation of the target is held in working memory to facilitate the task (Bravo & Farid, 2009; Vickery et al., 2005; Wolfe et al., 2004). This target representation can be quite flexible and is able to survive rotation (Reeder &

Peelen, 2013). It is variously referred to as the “target template,” “attentional template,” or “search template.”

A number of studies have investigated how the properties of the target templates emerge, and how they are affected by the task. For example, target templates can vary in their specificity or detail. It is perhaps not surprising that more detailed target templates can guide visual search more efficiently (Malcolm & Henderson, 2010). Wolfe, Horowitz, Kenner, Hyle, & Vasan (2004) suggested that the target template generation process could be dynamic. When participants are given a text cue prior to a search task, the template generated would be coarse. When a picture cue is used instead, the template becomes more detailed, which can improve search performance.

Highly specific target templates are not always ecologically effective, however. If a target template is general, it can be applicable to many related tasks once it is generated. On the other hand, if the target template is very specific, and it depends entirely on the current context, the target template may become less useful with slight changes to the viewpoint of the target or the search task. Barvo and Farid (2009) had participants search for a fish target among coral reef distractors. Each search trial was preceded by a cue. The cue was exactly the same as the search target in some trials. In other trials, the cue could be the same image that was rotated, or a fish image that belonged to the same species. The cue could also be an uninformative word label (i.e., "fish"). Response times to locate the target decreased as the amount of information contained in the cue increased. The results agree with that of Maxfield, Stalder, and Zelinsky (2014). In that study, the researchers presented a text cue to the participants, and had them perform a visual search for categorical objects. Participants were slower when the targets were low typicality given the text cue. In those situations, participants’ first fixations were also less likely to fall on the target. Hout and Goldinger (2015) found the same pattern in response times with targets of decreased precision. Using eye tracking, they found that participants spent more

time scanning the stimulus array and making a decision once fixation fell on the target when the target template was of low precision.

In a naturalistic environment, targets are never visually stable. When we are performing a visual search, such as looking for our car key on a cluttered desk, a slight change in perspective or time could make the target look very different (Zelinsky, 2008). Therefore, from daily-life experience, it is likely that the visual system tends to create target templates that are more general rather than highly specific. Barvo and Farid (2012) examined how a target template was generated when participants were asked to search for a fish target among distractors under two conditions. In the first condition, an exact target image was used across trials. In the other condition, fish images of the same species were used as target. When an exact image was repeated over trials in the training phase, participants generated a very specific target template. During the test when participants were asked to search for different fish targets across trials, their response times were slower than the other group who saw different images of the same species within a block. The latter condition is likely to be more representative of how people search for objects in a real-world scenario. Nevertheless, it echoes the idea that target template generation is dynamic and task specific. In an experiment that employed simple stimuli (colored rings), Goldstein and Beck (2018) showed that varying search templates across trials lengthened dwell time on distractors, and the process of target verification. It also lengthened the time that participants took to establish the target prior to initiating the search.

As noted previously, some work has also provided evidence that target templates might vary not only in their specificity, but that for cases of fixed distractors, people might build a target template that is not only detailed but also distractor-specific. That is, target templates can be based not on the optimal way of detecting the target but on the optimal way of distinguishing the target from the distractors (e.g., Navalpakkam and Itti, 2007).

Thus, overall there is significant evidence that using cues before trials can change people's target template to be more specific or more general, and that target template specificity plays an important role in the speed and generalizability of visual search. There is also evidence from simple stimuli that people might form distractor-specific target templates. However, the role of learning about distractors in shaping the specificity of the target template in real-world categorical search has not been investigated. In addition, it is unknown whether under such scenarios people form not only detailed but also distractor-specific templates (e.g., Navalpakkam and Itti, 2007).

Current Investigation

In the current study, we had participants perform a demanding visual search task to examine the nature of the target template people form and the extent to which this template becomes distractor-specific. Participants were asked to look for two categories of targets (lanterns and binoculars) among distractors. The same set of distractors were used for a given participant throughout the first half of the experiment. This provided a chance for the participants to learn about the visual features of the distractors, and to adjust their target templates over the course of training. Some groups were trained on distractors that required detailed search templates (e.g, man-made artifacts; see Figure 3.1), and others were trained on distractors that did not require detailed search templates to distinguish from the targets (e.g., mammals, or plants). Once the participants had this extended experience with the distractors, we replaced them with a new set of distractors. The new set of distractors either required a detailed search template (as they were quite similar to the targets; Experiment 1) or did not require a detailed template (Experiment 2).

The first half of the experiment allows us to examine whether people learn about the distractors during a visual search task. While encouraging high accuracy in the search task, we could examine the trajectory of search efficiency over the course of training. The second half of

the experiment allows us to examine whether the knowledge learned in training phase can be transferred to a relatively new environment. If the knowledge learned was very specific to the context, we should expect a drop in performance once a new set of distractors were introduced. If the knowledge learned was relatively general, we should expect stable performance before and after the switch.

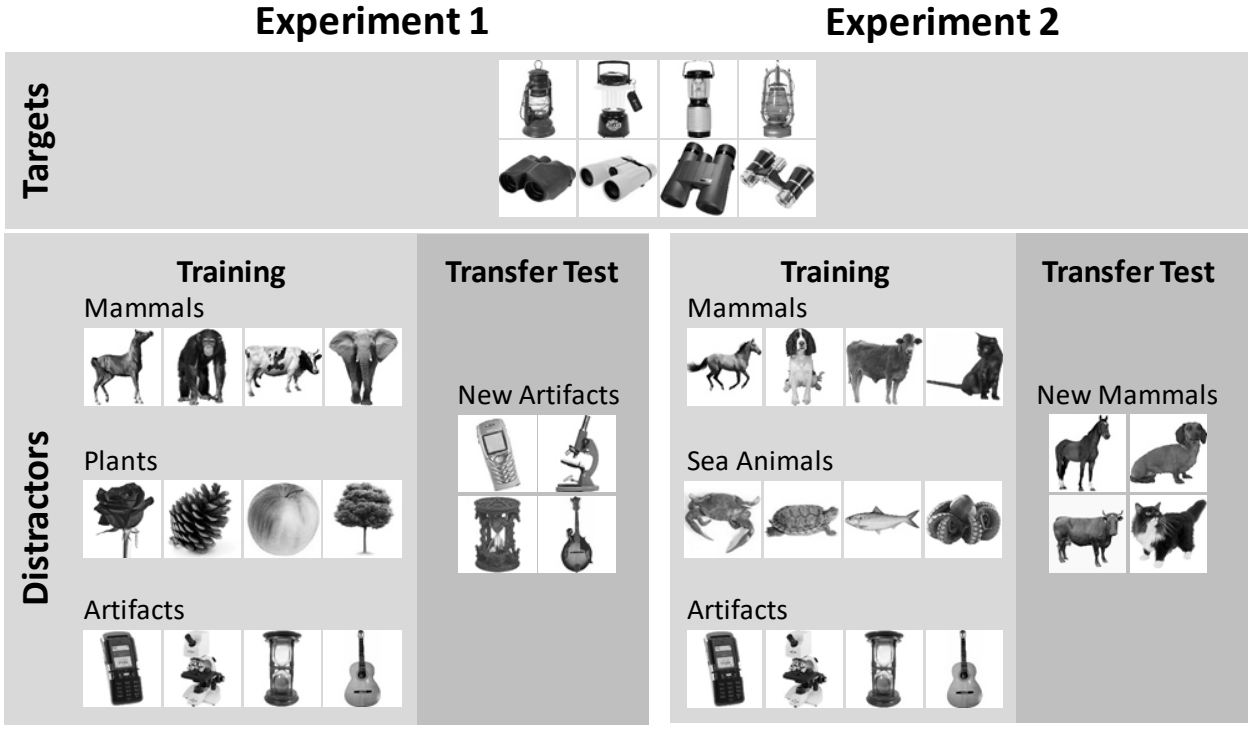


Figure 3.1. Structure of Experiment 1 and Experiment 2. Participants were asked to search for any lantern or any pair of binoculars in each trial and indicate whether they had found a lantern or a pair of binoculars. In Experiment 1, across-participants we manipulated whether participants searched for the targets among either mammal, plant, or artifact distractors. Once they had sufficient experience with the distractors, a new set of artifact distractors were introduced, replacing the ones they were trained on, to ask how well their learned target template transferred to this new search task. The structure of Experiment 2 was identical to that of Experiment 1, except that participants searched for the targets among either mammal, sea animal, or artifact distracts during training, and among new mammal distractors during the transfer test. Such an easier transfer test was designed to ask how specific the target template was to a particular set of distractors.

Experiment 1

Methods

Design

The experiment had two phases, a training phase and a transfer test phase. The tasks for the participants was to search for any lantern or any pair of binoculars among some distractors. In the training phase, participants were randomly assigned into one of the three conditions, in which they encountered a specific set of distractors. The group searched for the targets among pictures of mammals , and the other groups had plants and artifacts as distractors, respectively. The same set of distractors was repeated in each trial for a particular participant.

Behavioral changes in reaction times across blocks allow us to assess the learning trajectory during visual search. Changes across conditions indicate the degree of learning for different distractor sets.

In the transfer test phase, participants in all three conditions searched for the same set of targets among a new set of artifact distractors. Hence, all participants experienced the same set of visual stimuli as distractors. Behavioral differences between conditions can be attributed to the training participants received. In the ‘artifact’ training condition, the new distractors were distinct exemplars of artifacts but members of this same superordinate category. In the two other conditions, the distractors changed to an entirely new superordinate category (e.g., from plants to artifacts).

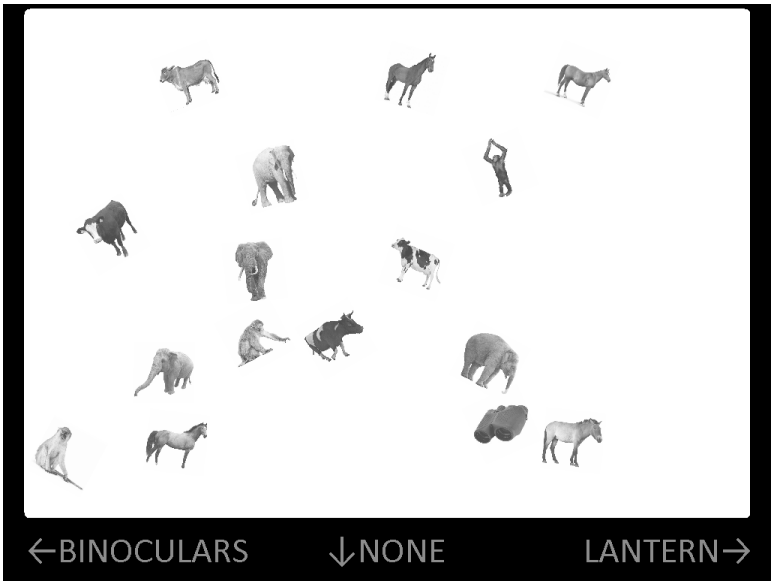


Figure 3.2. A sample visual search screen with a pair of binoculars target. Participants searched for any pair of binoculars or any lantern throughout the whole experiment. In the first part of the experiment, people searched among distractors that were either mammals (pictured), plants, or artifacts. This set of distractors was then replaced by a new set of artifacts in the second transfer phase of the study to assess the target template people had learned.

Participants

Sixty-nine participants (42 female) were recruited from University of California, San Diego's Psychology Subject Pool. They had a mean age of 20.8 at the time of participation. All participants gave informed consent prior before taking part in the study. All participants took part in the study for partial course credit.

We were interested in the reaction times between groups, and expected a large effect size for comparing the groups (Cohen's d ≈ 0.8). To achieve a power of 80% for a pairwise t-test, at least 26 participants were required per condition. The sample size obtained was determined according to our power analysis prior to participant recruitment.

Apparatus and Materials

Participants were placed in a sound attenuated room with normal indoor lighting.

Stimuli were shown on a Dell E173FPc 17-in LCD monitor with 4:3 aspect ratio. At a viewing distance of approximately 60cm, the monitor's visible area was 31.2 degrees wide and

25.4 degrees tall in visual angle. The resolution of the screen was set to 1024 x 768. The visible region of the screen was cropped, such that only the central rectangular area of the screen was utilized to show the visual stimuli. The rectangular window measured 970 pixels wide and 680 pixels tall, which translate to 29.63 degrees wide and 21.00 degrees tall in visual angle.

Images obtained online and publicly available image database (Brady, Konkle, Alvarez, & Oliva, 2008) were turned into grayscale, and resized to 80 x 80 pixels (2.50 visual degrees). Sixteen slots on an 8 x 5 imaginary grid was selected to host the images on each trial. The images were jittered with a maximum of 15 pixels left or right, 23 pixels up or down, and 30 degree clockwise or anticlockwise, to avoid appearing on a static grid.

Three labels were shown below the rectangular image window, “binoculars”, “none”, and “lantern”. Positions of the three labels corresponded to the three arrow buttons, “left”, “down”, and “right” that participants used for responses. Figure 3.2 shows an example of the search display.

Procedure

In each experimental session, the participant was first shown a page of instructions on the screen. Afterwards, a multiple-choice quiz was administered. Participants had to answer all the questions correctly to move on, otherwise he or she would be shown the instructions again. Participants were encouraged to prioritize accuracy over speed during the visual search task.

The experiment started with 9 trials in the practice block. There were 3 trials of each trial type, binoculars, lantern, or target-absent. The order of the trials was randomized.

All trials started with a fixation cross for 200ms. In a target-absent trial, four items from each of the four distractor sub-categories were shown on the screen (e.g., monkey, horse, elephant, and cow within the mammal category). Distractor categories that a participant encountered depended on the condition that he or she was in. One-third of all participants saw mammal images as distractors, another one-third had plants as distractors, and the remaining group had artifacts as distractors. These conditions are referred to as mammal training, plant

training, and artifact training conditions. The same set of distractors was shown in each trial. Once a response was detected, an audio feedback was also delivered. The next trial appeared after an 800ms blank screen.

A target-present trial was generated in a similar manner. A pair of binoculars or lantern image replaced one of the distractors. The images stayed on the screen until the participant responded. If an incorrect response was made, a hollow circle was shown at the location of the target for three seconds. This feedback screen served as a delay in the task and indirectly encouraged high accuracy throughout the experiment.

After the 9-trial practice block was over, participant completed 5 additional training blocks. Each block contained 36 search trials. There was a 30-second break in between the blocks.

Once the participant completed the 5 training blocks, a new set of distractors was drawn from the image set. All the distractors were replaced with 16 images of artifacts. Hence, participants in the artifact training condition saw new exemplars within the same superordinate distractor category (artifact), while those in mammal training, and plant training conditions were introduced new distractor categories. The target image set remained the same. There were a total of 238 pictures in the imager bank, but each participant only saw 64 of them in his or her session. This was to avoid idiosyncratic features in any particular images harming the generalizability of the experiment.

A total of 5 test blocks, each with 36 trials, were administered. Afterwards, the participant was debriefed and dismissed. Each experimental session took about 40 minutes.

Results

We were interested in how experience with certain distractors in visual search would affect search performance when there was a change in context (distractors). We first assessed whether participant learned about the visual features of the distractor categories in the training

phase. Any differences in the training phase must be attributed to differences caused by the distractors, since all three groups searched for the same targets but among different distractors. We then examined the effects of learning about the distinctions between the target and the distractors in the transfer test phase. In the transfer test phase, all participants were searching for binoculars and lantern targets among the same artifact distractors. We were particularly interested in the differences in search performance in Block 6, when a new set of distractors was first introduced. Performance in the last transfer test block could also be informative, indicating whether there were long-term effects on the training.

Training Phase

The left panel of Figure 3.3 shows participants' response times in the training phase. Only target present trials were included. Participants spent around twice as long detecting the target in the artifact training condition, compared to either mammal training or plant training conditions [$F(2,65) = 104.8$, $p < 0.001$]. A post-hoc Tukey Honest Significance Difference test indicates that the plant training group had the shortest response times. The differences between the plant training and mammal training groups was marginally significant ($p = 0.08$), and that between the plant training and artifact training, and mammal and artifact training were highly significant ($p < 0.001$).

In addition, visual search became more efficient over time [$F(4,260) = 41.6$, $p < 0.001$]. The rate of improvement was similar across conditions [$F(8, 260) = 1.1$, $p = 0.35$].

Even though participants were encouraged to attain a high accuracy in all conditions, small differences between conditions due to the distractors did arise. However, accuracy for detecting targets mirrored response times (Figure 3.4), indicating there was no speed/accuracy trade-off. Accuracy was the highest for the plant training condition (97.9%), compared to that of mammal training (96.4%) and artifact training (93.8%) conditions [$F(2,65) = 10.2$, $p < 0.001$].

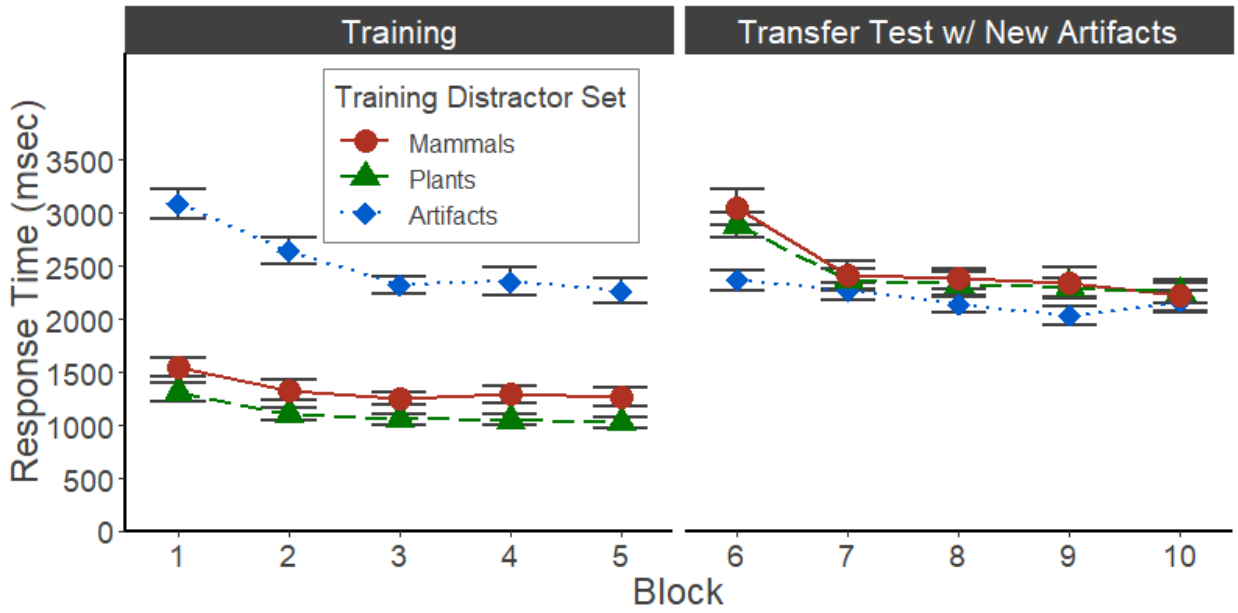


Figure 3.3. Response times of Experiment 1. Left: All groups searched for a set of 4 pairs of binoculars and 4 lanterns. However, during training, the groups differed in the distractors they needed to distinguish these items from. The artifact group had a difficult set of distractors, as, like the binoculars and lanterns, their distractors were man-made artifacts. The plant and mammal training groups had considerably easier distractors to differentiate from the targets, reflected in their much faster reaction time. Right: During the transfer test phase, beginning at block 6, all participants saw the exact same search displays, with binocular targets and lantern targets now embedded amongst a new set of artifacts that was also hard to differentiate from the targets. The distractors changed even for the artifact training group to new artifacts. There was a substantial switch cost for the groups that had been performing a search with easy to differentiate distractors when a new set of distractors was introduced that required a difficult search, but almost no cost for the artifact training group which had already been performing this difficult search task.

Transfer Test Phase

In the transfer test phase, participants in all three conditions searched for binoculars and lantern targets among new artifact distractors (that is, the distractors changed for all groups, including the artifact training group). In this phase, participants in all three conditions encountered the exact same stimuli. Therefore, differences in performance should be attributed to the kind of target template and search strategy they used during training and how this carries over to the transfer test trials. The right panel of Figure 3.3 captures the transfer test performance.

There was a clear effect of training on the transfer test performance. The artifact training group encountered a switch in the exemplars, but not the superordinate category. The group showed a negligible delay in response times (Block 5: 2265ms; Block 6: 2365ms). The other two groups showed significantly worse performance in the transfer phase (Mammal-training: 2888ms; Plant-training: 3055ms) compared to the artifact training group [$F(2,65) = 9.08$, $p < 0.001$]. Performance of the mammal and plant training groups at transfer was similar to the first training block of the artifact training group ($ps \geq 0.36$), as though they had almost no previous experience with these targets. These two groups did, however, quickly recover, improving performance dramatically in one transfer test block of search trials, such that their performance was indistinguishable from the artifact training group in the second transfer test block [$F(2,67) = 0.65$, $p = 0.52$].

Changes in accuracy again mirrored those of reaction time. It can be seen in Figure 3.4, the mammal training and plant training groups had lower accuracy in the first transfer test block [$F(2,65) = 5.2$, $p = 0.008$], whereas accuracy of all the groups were indistinguishable from each other for the rest of the transfer test phase [$F(2,67) = 1.232$, $p = 0.30$].

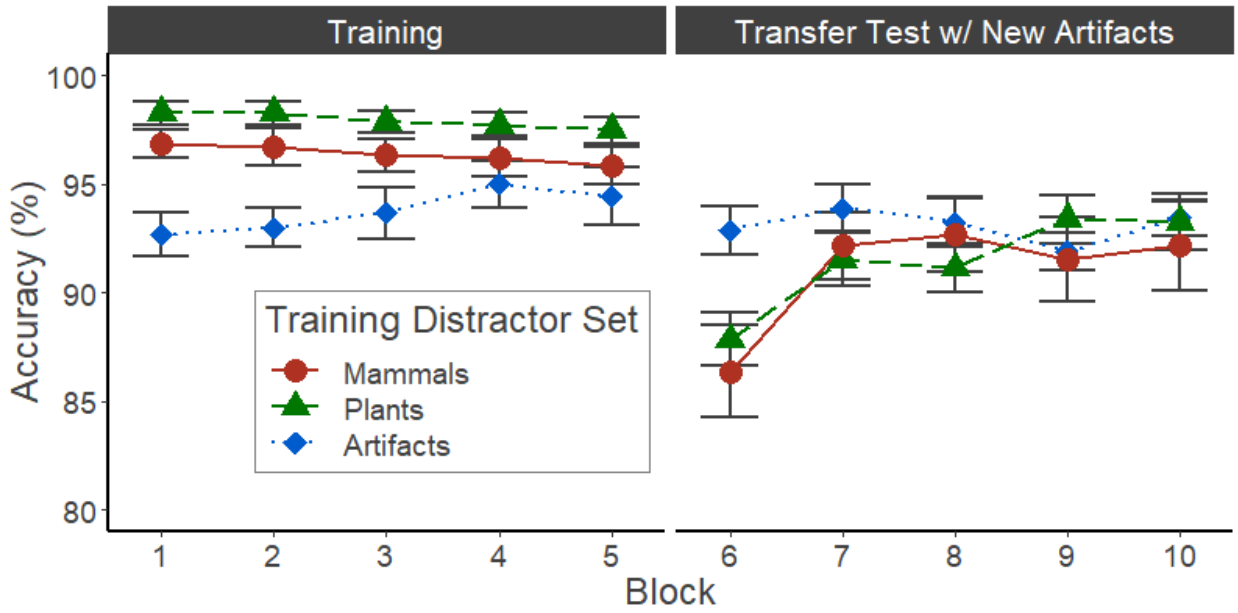


Figure 3.4. Accuracy effects in Experiment 1 mirrored those found with reaction time. While all groups performed well, the artifact training group had the most difficult distractor set during training and the least switch cost when the transfer phase began with new artifacts as distractors.

Discussion

As expected, we find that the visual features of the distractors drastically affect visual search performance in the training phase. Participants were much faster searching for the same targets among plant and mammal distractors than artifact distractors. Binoculars and lantern distractors should be easier to search for among mammal and plant distractors compared to artifacts because the targets and the artifact distractors are all man-made, and thus share more visual similarities than with the other two distractor categories (e.g., Long, Stormer & Alvarez, 2017). In addition to this search difficulty manipulation, we also found that participants improved over the course of the training phase, in part because they learned about the visual features of targets and distractors (as demonstrated by the transfer phase). Learning about visual features is hard to assess in a traditional visual search paradigm where simple stimuli are utilized. While people search for a T among L-shape distractors, or a green dot among red ones, the response times are usually within a few hundred milliseconds. Accuracy is usually at ceiling. The kind of testing environment makes it hard to study the trajectory of learning.

The transfer test phase showed the benefits of a more difficult training task. Participants in the artifact training conditions performed the worst in the training task based on both speed and accuracy measures, but they performed the best when a new set of artifact distractors was shown in the transfer test phase. This is consistent with the idea that the artifact training allows people to develop detailed target templates for the task, which allows them to distinguish the targets even from novel artifacts. By contrast, the groups performing the mammal and plant search conditions had no need for detailed target templates and so may not have developed a detailed enough representation. Thus, they had significant difficulty with the new, and more difficult-to-discriminate distractors.

In addition, we showed that this detailed target template benefit does not last very long. Within a block of transfer test trials, participants from the mammal and plant conditions caught up quickly, such that their response times and accuracy were no longer distinguishable from the artifact training group by the second transfer test block. Nevertheless, this switch cost is of practical importance. In a real-world setting, such as airport baggage checkpoints, observers do not usually have a block of trials to adjust to a new set of distractors. A change in the distractor sets would mean a delay in response time, or a drop in search accuracy.

To what extent is the learning during the training phase semantic or verbal rather than visual in nature? In theory, participants may learn labels for the distractors in the first block, and reject them in subsequent searches based on the meaning of the images. When the set of distractors was replaced, the learned semantic labels can no longer guide their search, perhaps resulting in lower performance in the transfer test phase.

This position cannot be fully dismissed with our current design, but attentional guidance through semantics is likely quite a bit weaker than guidance via visual features of the target, and cannot fully account for the effect (Bravo & Farid, 2009). For one thing, there was a difference in performance between the different conditions in the training phase, even though if the semantic labels for all items in all conditions are equally accessible, rejecting distractors based purely on

semantics should be comparable. For another thing, it is unlikely that the semantic differences between the targets and plant distractors was larger than those between the targets and mammal distractors. Future studies may equate the visual features and vary semantic differences, to allow finer-grained partitioning of the learning effects into visual vs. semantic features of the target template.

Experiment 2

In Experiment 2, participants searched for binoculars and lantern targets, as in Experiment 1. In one condition, participants searched for the targets among artifact distractors, in the other two conditions, they searched for the targets among mammal and sea animal distractors, respectively. We changed the transfer test so that a detailed target template was not required for the task. In particular, during the transfer test phase, all participants searched for the binocular and lantern targets among mammal distractors. Searching for the targets among mammal distractors required only a coarse discrimination. Compared to the transfer test between Experiments 1 and 2, the one in Experiment 2 was much easier. This allowed us to tease apart whether the transfer training benefits were general benefits that arise whenever participants performed a difficult task (e.g., Bjork, 1994) or were specific to the need for a detailed target template (which was learned only by the artifact training group). In addition, this manipulation allows us to tell whether the detailed target template learned by participants in the artifact training group was specific to the artifacts they had learned to discriminate the targets from (e.g., Navalpakkam and Itti, 2007) or whether it was simply a detailed target representation. If the template was specific to the distractors, even switching to an ‘easier’ search with different distractors would be expected to come at a cost to search performance; if it was simply a detailed target template, then no such cost would be expected.

Methods

Design

The overall design of Experiment 2 was identical to that of Experiment 1. Participants went through 1 practice block, five training blocks, and five transfer test blocks. Participants searched for binoculars and lantern targets among distractors. We made two changes to the stimuli.

We replaced plant distractors with sea animals during training. In addition, in the transfer test, we utilized only mammal distractors. The transfer test was designed to require less detailed target templates than the transfer test in Experiment 1.

Participants

Seventy participants (53 female) were recruited from the same population described above. The participants had a mean age of 20.7 at the time of their participation. All participants gave informed consent prior to their participation and they participated for partial course credits.

We obtained a sample size that was comparable to Experiment 1 so that the results of the two experiments can be compared.

Apparatus and Stimuli

Equipment used in Experiment 2 was identical to that of Experiment 1. Images used in Experiment 2 were identical to that of Experiment 1, except that some of the mammals images were replaced, and the plant images were replaced by sea animals. All images were obtained from the same source as stated in Experiment 1.

Procedure

Experiment 2 was administered in exactly the same way as in Experiment 1, except for the following changes.

The three training groups saw mammals, sea animals, and artifacts as distractors during the training phase. In the transfer test, all three groups searched for lantern and a pair of binoculars as targets, among a new set of mammal distractors. Hence, the transfer test of Experiment 2 should be notably easier than that of Experiment 1.

Results

As in Experiment 1, participants in all conditions searched for the same targets in the training phase; the conditions differed only in the distractor set. All participants encountered the same mammal distractor set during the transfer test phase and had the same targets. Thus, differences in performance during the test phase can only be attributed to differences in target template or strategy developed during the training phase. In particular, performance in Block 6 indicates the most direct influence of the training. Performance in Block 10 indicates a longer-term influence. The slope of the improvement from Block 6 to Block 10 indicates how the training affects learning of a new set of distractors.

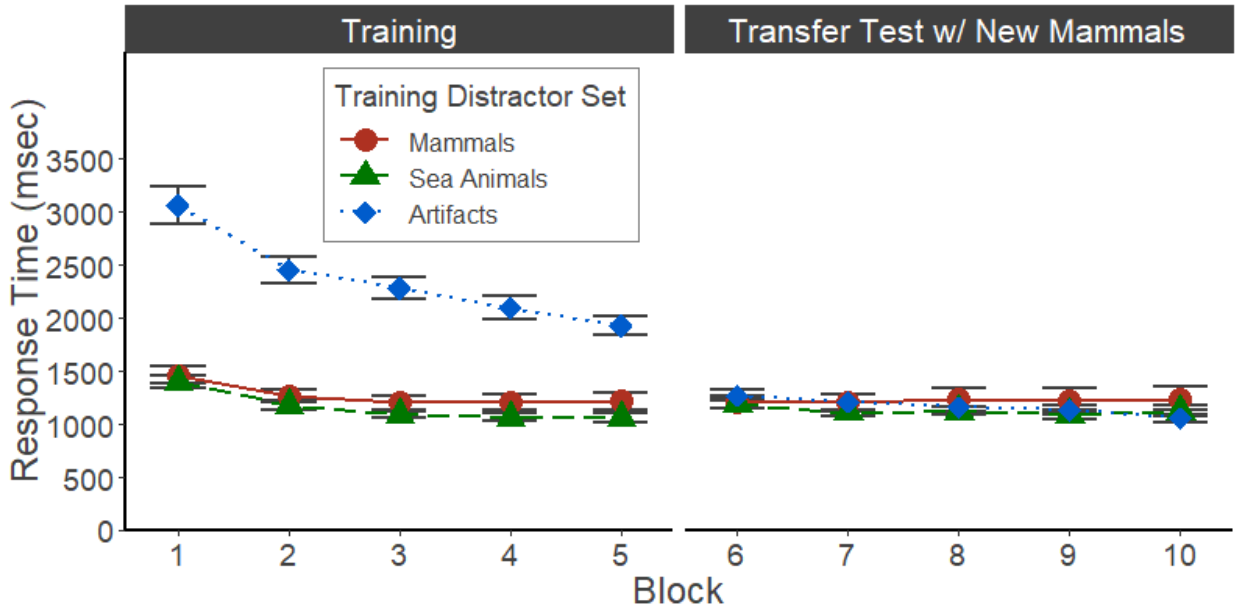


Figure 3.5. Response times by condition and block in Experiment 2. Left: Participants were faster at searching with the same set of distractors being used. Right: The cost for switching was negligible when a new set of distractors was introduced. Error bars in the graph denote between-subject standard error of the means.

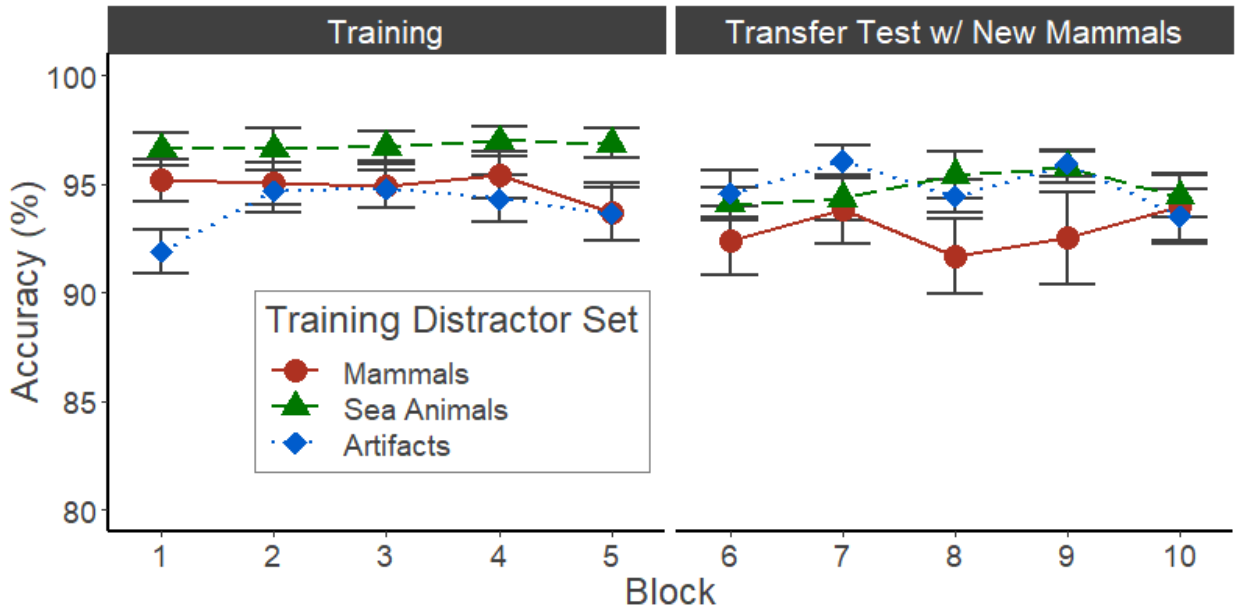


Figure 3.6. Accuracy in Experiment 2. Participants were encouraged to maintain a high accuracy. Performance did not differ by condition.

Training Phase

The left panel of Figure 3.5 shows that features of distractor sets again affected performance during training. In general, participants spent around twice as long detecting the target in the artifact training condition, compared to either mammal training or sea animal training conditions [$F(2,67) = 89.1$, $p < 0.001$]. This visual search became faster over blocks [$F(4,268) = 39.2$, $p < 0.001$]. The rate of improvement was steeper for the artifact training condition than the two other conditions, because of the relatively slower initial reaction times [$F(8, 268) = 8.5$, $p < 0.001$].

Accuracy was uniformly high, and showed no signs of a speed-accuracy trade-off, although accuracy for detecting targets was higher in the sea animal training condition (96.8%), compared to that of mammal training (94.9%) and artifact-training (93.9%) conditions [$F(2,67) = 4.4$, $p = 0.02$] (left panel of Figure 3.6).

Transfer Test Phase

In the transfer test phase, participants in all three conditions searched for binoculars and lantern targets among new mammal distractors. Since the visual stimuli were identical across conditions, any differences in the conditions should be attributed to participants' training experience. As shown in the right panel of Figure 3.5, response times across the three conditions were highly similar throughout the transfer test phase [$F(2,67) = 1.3$, $p = 0.27$]. Response times did not further improve across blocks [$F(4,268) = 1.1$, $p = 0.35$], and the two factors do not interact with each other [$F(8,268) = 1.3$, $p = 0.22$]. Importantly, the performance of all groups were better than that the first training block of the mammal-training condition ($ps \leq 0.05$, uncorrected for multiple comparisons), suggesting at least partial knowledge learned during training was successfully transferred to the test phase.

Accuracy of the three conditions in the transfer test also did not statistically differ from each other [$F(2,67) = 1.3$, $p = 0.27$)] (right panel of Figure 3.6).

Discussion

Binoculars and lantern distractors were easier to search for among mammals and sea animal distractors during training, compared to artifact distractors. This is likely because artifact distractors are all man-made, and they share more visual similarities with the targets compared to the other two distractor categories (e.g., Long, Stormer & Alvarez, 2017). Ultimately, this difference gives rise to the differences in response times between the artifact training condition and the two other conditions during training.

In the transfer test phase, we found that while artifact distractors made the visual search task harder in the training phase, they did not seem to have any long-term effects on search performance when the distractors became easier to distinguish during the transfer test phase. Thus, there was not a general desirable difficulty effect where the harder participants worked on a task, the better they learned about the target (Bjork, 1994), and therefore the better they performed during the transfer test. In addition, participants in the artifact training group did not

appear to have learned a template that was specific to the artifacts they had learned to discriminate the targets from (e.g., Navalpakkam and Itti, 2007). If the template was specific to the distractors, even switching to an 'easier' search with different distractors would be expected to come at a cost to search performance, which did not occur. Instead, given the results from Experiment 1, the current results are consistent with our hypothesis that the training groups learned different amounts of detail in their target templates -- i.e., that the artifact training group learned a more specific template. In the transfer test phase of this experiment, even a less detailed template would support performance, and so there was no additional benefits of a more detailed template; whereas in Experiment 1, a more detailed template was required, and so already having learned this template supported improved performance.

In particular, the main difference between Experiments 1 and 2 was the distractor sets used in the transfer test. Experiment 2 utilized a set of mammal images as distractors, while Experiment 1 utilized a set of artifact images as distractors. In Experiment 1, we showed that when the distractors were switched, even when the targets stayed the same, people needed time to adjust to the new context. In Experiment 2, with a less detailed template needed during the transfer test, this did not occur. The artifact training group generated a detailed enough target templates that survive change in the context.

Importantly, Experiment 2 also provides evidence against the semantic/verbal label account that is possible in Experiment 1. Experiment 1 and 2 are symmetric from the perspective of semantic/verbal labels, yet lead to a large difference in transfer performance. This is consistent with our account of a more detailed vs. less detailed target template but inconsistent with the idea of verbal labels being primarily responsible for transfer costs.

General Discussion

In two experiments, we had participants search for two categories of targets repeatedly. In both cases, the distractors were always consistent throughout an entire training phase. This

allowed participants to learn about the visual features of the distractors, and create appropriate target templates. Although participants were neither explicitly told nor encouraged to learn about the distractors, they nevertheless learned about distractors during the visual search task, and response times become shorter when they searched for targets among the same set of distractors over time.

Our critical manipulation was the introduction of a new set of distractors in the middle of the experiment. The results suggested that coarse target templates are generated during training when the targets are easily distinguishable from distractors (e.g., searching for binoculars among mammal distractors). The coarse target templates do not support performance well in a new context which requires detailed target templates (e.g., searching for binoculars among artifact distractors). On the other hand, if the training requires demanding target-distractor discriminations (e.g., searching for binoculars among artifact distractors), detailed target templates are likely to be generated, and these templates are sufficient for both hard and easy searches in transfer tasks. Importantly, changing the distractors per se did not impact search performance (e.g., the change from sea animals to mammals, or from artifacts to mammals in Experiment 2). Instead, as long as the level of detail in the target template was sufficient to support the new search task, participants performed well during the transfer test. Thus, our results suggest the importance of having detailed search templates for transfer learning, but we do not find direct evidence for distractor-specific templates in this categorical search context with a wide range of both targets and distractors.

Practical Implications

Visual search has been an instrumental tool in understanding the human mind for decades. One reason for its popularity has to do with the potential practical applications. Revisiting the baggage screening scenario we discussed earlier, airport security staff generally search thousands of bags in a shift. The distractors are generally very limited and drawn from

similar categories repeatedly. Our results support the idea that if passengers stuff their bags with unusual items, this may interfere with the security staff members' search performance.

Our research design maps on this real-world scenario closely. Participants searched for a limited set of targets among a repeated set of distractors. The baggage screening task is likely to be a demanding task, like our transfer test in Experiment 1. Our results shed lights on the change in performance over time in such a situation. First, they suggest that airport security staff are likely to improve over time. They are likely to extract features of the targets and distractors that help the search task. Second, a change in the familiar distractors are likely to impose a cost to their performance. The cost is likely to be minimal when it involves switching out exemplars within the distractor categories, but when the change involves a new set of distractor categories, the cost in performance is likely to be substantial (first transfer test block in Experiment 1).

A practical solution to emolliate the effect would involve increasing the diversity of distractors (e.g., Hout & Goldinger, 2012). That may involve digitally inserting distractors of different categories, in additional to insertion of targets that is currently in practice.

Connection to the Visual Search Literature

The current study is closely related to a number of previous results in the literature. Here we highlight some differences between the current study and this previous work.

Hout and Goldinger (Experiments 1C & 2C, 2010; Experiment 2, 2012) studied how repeated distractors leads to improvements in search performance. Participants in the studies searched for a target among the same set of distractors within each block. The study utilized grayscale images of objects like current study. Unlike the current study, Hout and Goldinger's experiments specified a new target on each trial. The length of time that participants encountered the same set of distractors was also shorter than the current study. Participants saw the same set of distractors for no more than 40 trials. In addition, the association between

the distractors was low. Differences between distractor sets were not systematically manipulated. The researchers did not find any costs when distractors switched from one block to the next.

Wolfe, Klempen, and Dahlen (2000) had participants search for a target letter among 3 or 5 letter distractors. In their repeated search condition, the same search display was visible to the participants throughout the experiment. A cue was presented to indicate the beginning of a trial and the target for the current trial. Participants became faster over the 700-trial session. Importantly, the researchers did not find any improvements in search efficiency when taking set size into account. They showed that participants spent around 40 to 60 msec per item in the search regardless of practice.

The current study asked a similar question that these researchers examined: whether having repeated distractors helps visual search. The major difference between the current study and the ones in the literature is that the target set in the search task was kept constant in the current experiment. The design allowed participants to not only learn about the distractors, but also the relationship between the target and the distractors. As Duncan and Humphrey (1989) convincing showed that one major factor for visual search efficiency is the target-distractor discriminability, our design provided an opportunity for participants to learn and improve this discrimination. As we are interested in a learning effect, we also had a longer training period for each set of distractors, compared to Hout and Goldinger's study. It appears to us that participants in certain conditions kept improving even after 180 trials with the same distractors, so a longer training is more ideal in studying learning effects of this sort.

Our results are in agreement with those discussed above, and showed that the learning effects can be enhanced with the use of real-world images, without varying the target set, and extending the training period that participants engage with the same set of distractors. They also show that memory is better utilized when the search task is sufficiently difficult (Solman & Smilek, 2012).

Possible Mechanisms

We argue that people learn not only the characteristics about the targets and the distractors during visual search, but they also fine-tune their target templates that would maximally separate the feature space between the targets and distractors. This target-distractor discriminability is likely to be task dependent and differ from condition to condition. When discrimination is easy, such as the case of finding binoculars among mammal distractors, a coarse target template may be generated. The locus of this target template possibly lies in the mid-level features, such as curves, that are more likely to appear in animals than in artifacts (e.g., Long, Stormer & Alvarez, 2017). When target-distractor discrimination is difficult, such as the case of finding binoculars among artifacts, a detailed target template is likely to generated. The locus of this target template is likely to be more detailed. There may be very specific shapes that would differentiate binoculars and lanterns from our artifact distractors. Semantic labels of the objects may also be utilized if visual features alone are highly inefficient.

While possible in some conditions, low-level priming do not seem to paint a full picture of the learning effect (Kristjánsson & Driver, 2008). Priming happens when stimuli appear prior to the current ones, and bias people to process certain stimulus. The effect is usually short-lived. The continued improvements in visual search during the training phase may be partially accounted for by priming, but not in the transfer test. For one thing, if priming can account for most of the transfer effect in Experiment 2, we should expect a learning effect in the corresponding conditions in Experiment 1. The fact that there is an asymmetric training effect in the two experiments suggested that priming is unlikely the sole driving force in learning. The same logic applies to low-level visual adaptation such as frequency and feature adaptation due to prolonged exposure to certain visual features (Kohn, 2007).

Conclusion

In two experiments, we showed that people learned about the properties of distractors in a visual search task. Their prior experience affected the nature of their target templates. While we did not find evidence that people form templates that are specific to the set of distractors, we found that when targets and distractors were more difficult to discriminate, people create target templates with more detail, which are more likely to be transferable to a new context. In a baggage screening context, this suggests that a diverse set of distractors would be beneficial to the generalizability of search. One strategy would be to be digitally inserted into the images of the bags, especially distractors that are difficult to distinguish from prohibited items. This would avoid a decrease in search performance when the search context and distractor set change.

Chapter 3, in full, was prepared to be submitted to a journal: Lau, J. S. H., Pashler, H. & Brady, T. F. (in preparation). Low Target-Distractor Discrimination in Visual Search Promotes Detailed Target Template Generation. The dissertation author was the primary investigator and author of this paper.

References

Avraham, T., Yeshurun, Y., & Lindenbaum, M. (2008). Predicting visual search performance by quantifying stimuli similarities. Journal of Vision, 8(4), 9-9.

Bjork, R. A. (1994). Institutional impediments to effective training. Learning, remembering, believing: Enhancing human performance, 295-306.

Brady, T. F., Konkle, T., Alvarez, G. A. and Oliva, A. (2008). Visual long-term memory has a massive storage capacity for object details. Proceedings of the National Academy of Sciences, USA, 105 (38), 14325-14329.

Bravo, M. J., & Farid, H. (2009). The specificity of the search template. Journal of Vision, 9(1), 34-34.

Bravo, M. J., & Farid, H. (2012). Task demands determine the specificity of the search template. Attention, Perception, & Psychophysics, 74(1), 124-131.

Chun, M. M., & Jiang, Y. (1998). Contextual cueing: Implicit learning and memory of visual context guides spatial attention. Cognitive psychology, 36(1), 28-71.

Chun, M. M., & Jiang, Y. (1999). Top-down attentional guidance based on implicit learning of visual covariation. Psychological Science, 10(4), 360-365.

Duncan, J., & Humphreys, G. W. (1989). Visual search and stimulus similarity. Psychological review, 96(3), 433.

Eckstein, M. P. (2011). Visual search: A retrospective. Journal of vision, 11(5), 14-14.

Goldstein, R. R., & Beck, M. R. (2018). Visual search with varying versus consistent attentional templates: Effects on target template establishment, comparison, and guidance. Journal of Experimental Psychology: Human Perception and Performance, 44(7), 1086.

Hout, M. C., & Goldinger, S. D. (2010). Learning in repeated visual search. Attention, Perception, & Psychophysics, 72(5), 1267-1282.

Hout, M. C., & Goldinger, S. D. (2012). Incidental learning speeds visual search by lowering response thresholds, not by improving efficiency: Evidence from eye movements. Journal of Experimental Psychology: Human Perception and Performance, 38(1), 90.

Hout, M. C., & Goldinger, S. D. (2015). Target templates: The precision of mental representations affects attentional guidance and decision-making in visual search. Attention, Perception, & Psychophysics, 77(1), 128-149.

Kanerva, P. (1988). Sparse distributed memory. Cambridge, MA: Bradford Books.

Kohn, A. (2007). Visual adaptation: physiology, mechanisms, and functional benefits. Journal of neurophysiology, 97(5), 3155-3164.

Kristjánsson, Á., & Driver, J. (2008). Priming in visual search: Separating the effects of target repetition, distractor repetition and role-reversal. Vision Research, 48(10), 1217-1232.

Long, B., Störmer, V.S., & Alvarez, G.A. (2017). Mid-level perceptual features contain early cues to animacy. Journal of Vision, 17(6), 1–20

Malcolm, G. L., & Henderson, J. M. (2010). Combining top-down processes to guide eye movements during real-world scene search. Journal of Vision, 10(2), 4-4.

Maxfield, J. T., Stalder, W. D., & Zelinsky, G. J. (2014). Effects of target typicality on categorical search. Journal of vision, 14(12), 1-1.

Navalpakkam, V., & Itti, L. (2007). Search goal tunes visual features optimally. Neuron, 53(4), 605-617.

Neider, M. B., & Zelinsky, G. J. (2008). Exploring set size effects in scenes: Identifying the objects of search. Visual Cognition, 16(1), 1-10.

Reeder, R. R., & Peelen, M. V. (2013). The contents of the search template for category-level search in natural scenes. Journal of Vision, 13(3), 13-13.

Rosenholtz, R. (2001). Search asymmetries? What search asymmetries?. Perception & Psychophysics, 63(3), 476-489.

Schoonveld, W., Shimozaki, S. S., & Eckstein, M. P. (2007). Optimal observer model of single-fixation oddity search predicts a shallow set-size function. Journal of Vision, 7(10), 1-1.

Solman, G. J., & Smilek, D. (2012). Memory benefits during visual search depend on difficulty. Journal of Cognitive Psychology, 24(6), 689-702.

Treisman, A. M., & Gelade, G. (1980). A feature-integration theory of attention. Cognitive psychology, 12(1), 97-136.

Võ, L.-H. M., & Wolfe, J. M. (2015). The role of memory for visual search in scenes. Annals of the New York Academy of Sciences, 1339(1), 72-81.

Williams, C. C., & Henderson, J. M. (2005). Incidental visual memory for targets and distractors in visual search. Perception & Psychophysics, 67(5), 816-827.

Wolfe, J. M., Horowitz, T. S., & Kenner, N. M. (2005). Cognitive psychology: rare items often missed in visual searches. Nature, 435(7041), 439.

Wolfe, J. M., Horowitz, T. S., Kenner, N., Hyle, M., & Vasan, N. (2004). How fast can you change your mind? The speed of top-down guidance in visual search. Vision research, 44(12), 1411-1426.

Wolfe, J. M., Klempen, N., & Dahlen, K. (2000). Postattentive vision. Journal of Experimental Psychology: Human Perception and Performance, 26(2), 693.

Yang, H., Chen, X., & Zelinsky, G. J. (2009). A new look at novelty effects: Guiding search away from old distractors. Attention, Perception, & Psychophysics, 71(3), 554-564.

Zelinsky, G. J. (2008). A theory of eye movements during target acquisition. Psychological review, 115(4), 787.

Conclusion

In my research program, I study various factors that would potentially affect learners' performance in visual categorization. Chapter 1 examined the overshadowing effect, when a highly predictive feature was available during the training but not the transfer test. People's ability to categorize objects decreased when an overshadowing feature is present, compared to a group which did not encounter any overshadowing features in the training. Importantly, the effect can be mitigated with top-down instructions among some learners. These top-down instructions have not been studied in this particular context, partly because the study of overshadowing effect has deep roots in the animal literature. Top-down effects like the one reported here cannot be easily manipulated on animal subjects.

The implications of the study are clear: In real-world settings, people often learn to distinguish objects of different categories with the help of some visual aids, such as a label below an example picture. These visual aids may serve as overshadowing cues that would affect performance in the transfer test, when these aids are taken away. Learners should be made aware that the aids do not usually appear in the actual categorization task, and they should learn about other features that would help them perform the categorization task.

In Chapter 2, I examined a new form of categorization training. In the training phase, learners were given two stimuli to learn about their corresponding categories. Simultaneously presenting two stimuli, one from each category, allowed learners to compare the diagnostic features. During the training phase, when learners were asked to perform the task with feedback, they were quite accurate. Importantly, overall transfer test performance of these learners was on par with those trained with interleaved schedules. It shows that simultaneous training is as effective as the most widely adopted training strategy currently in the literature. With the additional benefits of high training accuracy, this training strategy creates less

frustration to the learners, at least at the beginning of the learning. It may prove to be a preferred alternative for some types of category learning tasks.

Chapter 3 investigated the dynamics of target template generation in visual search. In any visual search task, a participant has to create a temporary representation of the target in the working memory, such that this template can be matched with items in the search array. Previous research suggested that manipulating the target across trials had an effect on target template generation. The effects of the distractors, on the other hand, have been understudied. Chapter 3 provided evidence that target template formation depends greatly on the discriminability between targets and distractors. When discriminability is high, the search task is easy, and the target template is likely to be coarse. When discriminability is low, the search task is relatively more difficult, and the target template is more detailed. A more detailed target template is more robust to change in the context, and hence is desirable in situations that search context can abruptly change. Airport baggage search is likely to be one of these situations.

To conclude, this research program involves a variety of aspects that can affect human category learning. Understanding these factors is theoretically important for scientists to paint a complete picture of human information processing. These issues also have direct implications on how educators can structure their training programs.