Attentional and Representational Flexibility of Feature Inference Learning

Aaron B. Hoffman (aaron.hoffman@mail.utexas.edu)
The University of Texas at Austin
Department of Psychology, 1 University Station
Austin, TX 78712 USA

Bob Rehder (bob.rehder@nyu.edu)
New York University
Department of Psychology, 6 Washington Place
New York, NY 10001 USA

Abstract

Previous research has shown that inference learning may motivate learners to acquire more within-category and prototypical information compared to standard supervised classification. We hypothesized that as a result inference learners would build flexible representations that could facilitate making novel category contrasts. An experiment tested the flexibility of category representations across inference and classification tasks by (1) having people make novel contrasts with categories learned earlier in the experiment and (2) recording eye movements as participants acquired and made various category contrasts across four categories. Significant differences in the attention patterns were observed in the eye movement data. Differences in attention were coupled with an advantage for inference learners in classifying among novel category contrasts.

Keywords: Categories, Concepts, Learning, Inference, Eye movements, Eye tracking, Task demands

Introduction

Two assumptions have guided the study of concept learning ever since Hull (1920). The first is that category learning amounts to learning a common label for sets of objects. This assumption is explicit in the ubiquitous *supervised classification task*, in which people label a series of visually presented stimuli. Learning is supervised, as experimenters wait for a response and then provide feedback. This kind of paradigm has been used to determine, for example, whether prototype models are superior to exemplar models (Smith & Minda, 1998, or vice versa, Zaki, Nosofsky, Stanton, & Cohen, 2003). Over the years, researchers have taught people to group objects into (usually two) sets and have explored the resulting representations.

A second assumption has been that information about a category learned in one context, transfers to another. Consider the goal of distinguishing roses from raspberry bushes. If the most diagnostic feature is the presence of berries, then people will learn that the berry feature should receive the most attention weight (since both plants have thorns) (Nosofsky, 1984; Rehder & Hoffman, 2005a; Shepard et al. 1961). However, when one later has to distinguish raspberry from cranberry bushes, thorns suddenly become diagnostic, because while both have red berries, only the raspberry bush has thorns.

The problem is that optimizing attention for one category contrast (raspberry vs. rose) is not always optimal for another (raspberry vs. cranberry). The consequence of ignoring irrelevant dimensions for one set of category contrasts means that the learner has to re-attend (and learn about) those dimensions when familiar categories are contrasted in novel ways. That is, the learner has to relearn about raspberries. In this manner, the heralded powers of selective attention assumed by present theories may actually harm performance when previously irrelevant dimensions become relevant.

The questions we ask in the present study are, how do people know what information is diagnostic across contexts, and how much should they attend to different sources of information? The mechanisms of attention allocation in many computational models of category learning (Kruschke, 1992; Erickson & Kruschke, 1998; Kruschke & Johansen, 1999) suggest that people learn to attend to only that information needed to distinguish the two categories being acquired. The problem we raise is that after learning one classification in which, say, cue A is most diagnostic, people should have trouble learning a second classification in which B is the good cue, because prior classifications have taught people to ignore it (Kruschke, Kappenman, & Hetrick, 2005). Thus, while people don't seem to have trouble classifying in different real-life contexts, our literature would suggest that they should. How do we resolve the mismatch between the flexibility of real-life categorization and the rigidity of the types of categories acquired via experimental classification tasks? The hypothesis we explore here is that perhaps other modes of category learning provide people with the flexibility needed to use their conceptual knowledge in different situations.

Inference and Classification

Other tasks, where the goal is not to classify, but to learn about the properties of categories, may yield a flexible representation that can handle novel contrasts. Research that has expanded the array of concept acquisition tasks (Markman & Ross, 2003) led us to consider a task that may produce flexible conceptual representations. Whereas classification involves predicting the category label from features, *feature inference learning* involves predicting a

missing feature from other features and the category label. So rather than determining that a plant is a raspberry bush, the inference task asks learners to determine whether a raspberry bush has thorns, or some other property.

The feature inference task has received a substantial amount of recent interest, because there is evidence that classification and inference learners process conceptual information differently. Chin-Parker and Ross (2002) found that whereas classification learners acquired diagnostic features, inference learners were also sensitive to within-category correlations of features (e.g., berries go with thorns, a leaf type and stem). Similarly, Chin-Parker and Ross (2004) found that inference learners were more sensitive than classification learners to nondiagnostic, prototypical features. (also see Anderson, Ross, & Chin-Parker, 2002; Sakamoto & Love, 2006; Yamauchi & Markman, 2000a; b). Finally, Yamauchi and Markman (1998) found that inference learners were more likely than classification learners to infer prototypical category features, but less likely to infer features associated with actual training items. Inference learners were able to learn linearly separable categories in fewer trials than classification learners. Thus, in spite of inference and classification tasks being formally identical (Anderson, 1991), it is possible that the resulting category representations can differ.

The above-cited evidence suggests that whereas classification learning may foster attention to the diagnostic dimensions that serve to distinguish between categories, inference learning may focus categorizers on within-category information. Our hypothesis is that because the within-category information acquired by inference learners is not tied to any particular set of contrast categories, such knowledge yields a more general and flexible representation. As a consequence, with respect to novel contrasts, inference learners may be at an advantage over classification learners.

Experiment

Across two training phases participants learned about categories A, B, C, and D in Table 1 via inference or classification. Eye tracking was used throughout to monitor participants' attention to the three feature dimensions and the category label. A test phase examined classification performance and attention profiles as people made novel category contrasts. From prior research we expected classification subjects to learn to ignore the irrelevant dimensions during training; this attention optimization should lead to a difficulty in making novel classifications. In contrast, prior research has demonstrated a tendency for inference learners to acquire within-category information, suggesting a general motivation to learn about all the dimensions in the inference task. Such motivation can potentially produce flexible category representations—that is—ones that support novel contrasts. Measuring eye movements during training will help explain differences in concept flexibility between groups.

In contrast to previous studies comparing inference and classification, a change was introduced to our inference training procedure: One of the dimensions, the contrast dimension 3, was never queried (and inference subjects were informed of this fact before the start of training). This change was made to better equate the two tasks; it allowed inference participants a chance to optimize their attention away from dimension 3, just as the classification learners could. We can therefore test whether inference learners are in fact generally motivated to learn about category features, or whether the demands of the task, i.e., querying the features, draws learners' attention. To improve certainty about which location subjects had to respond to, and to eliminate any unnecessary fixations, we also added a dashed line from the center of the symbol to the queried dimension for inference, or to the category label for classification.

Method

Participants Twenty-four New York University students participated for course credit. They were assigned to standard classification or to an inference task. Subjects were also assigned to one of six counterbalancing conditions corresponding to different ways of distributing the three dimensions to screen locations.

Materials Subjects learned categories of "ceremonial symbols." The features of the symbols were 2 degrees of visual angle in diameter. Examples are shown in Figure 1. The top left of each symbol contained the category label. The other locations contained features.

The eye tracker was an SMI Eyelink I, 250 Hz. We programmed a gaze-contingent window of 4 x 4 degrees of visual angle to center on subjects' gaze, when gaze was directed near a feature it was visible, but if their gaze was away from a feature, it became jumbled. Gaze-contingence ensured that subjects could only extract feature information when fixating it.

Classification Task 1: A versus B Table 1 presents a three-dimensional structure with categories A, B, C, and D. Subjects were trained on these categories using different contrasts. First, categorizers learned to contrast As versus Bs. To classify As and Bs, they needed to use dimension 1, in which feature-value 1 predicted category A and 0 predicted category B. Dimension 2 was irrelevant, with 1s

Table 1: Category structure.

	Dimension		
Category	1	2	3
A	1	1	1
A	1	0	1
B	0	1	1
B	0	0	1
C	1	1	0
C	0	1	0
D	1	0	0
D	0	0	0

and 0s occuring in each category equally. Dimension 3 contained a 1 for all category A and B members (i.e., it had perfect category validity), so it could also not be used to discriminate A from B.

Figure 1 shows the experimental procedure. Before each trial, we presented a drift correction, in which the subject fixated the point in the center to display the stimulus. As per Figure 1 (classification), the top left of the stimulus contained the possible category labels, presented as "A B." This indicated to the subject that he or she must decide whether the item is an A or a B. The label order indicates whether the left or right button corresponds to category A or B. Here, the left button corresponds to A and the right button corresponds to B. On some trials the opposite label ordering is presented as "B A;" on those trials, the button assignment is reversed.

There was no response deadline for classifying. However, immediately after a response, the category location was replaced with the correct category label, producing a chime for a correct response, or a buzzer if incorrect. The stimulus with the correct category label would remain for 4 s.

Each subject made classifications with both label orderings, which disassociated buttons and labels. The two unique category A items and two unique category B items were presented four times each, in random order (16 trials), with half of the items having the A B label order and the other half with the B A label order. Training continued for five blocks.

Classification Task 2: C versus D Subjects next learned a second contrast, between Cs and Ds. As Table 1 shows, this contrast required use of dimension 2, with 1s predicting Cs, and 0s Ds. Dimensions 1 and 3 are irrelevant. Thus, the task was identical to the A versus B task, but with the relevant dimension switched. (Note that the additional block in Task 1 was to allow learners to acclimate to the procedure.)

Inference Task 1: Category A and B The inference condition was similar to the classification condition, but instead of classifying, inference learners predicted missing features. Figure 1 also provides an example inference trial, where the bottom left of the stimulus, contains a feature option; the subject must decide which feature belongs there. The relative positions of features indicated which button to press for each option. The left button selected the feature on the left, and the right button selected the feature on the right.

Inference learning on categories A and B lasted for five blocks. Every exemplar was presented with two dimensions queried twice (once for each feature order). Exemplars were presented in random order, for a total of 16 trials per block.

Inference Task 2: Category C and D Inference learning continued with the second set of categories in Table 1, for four blocks. As for the first inference task, perfect performance is attainable on only 2 of the three dimensions.

Switch Task After the first two tasks, both classification and inference subjects were presented with classification trials involving contrasts between categories that were unpaired during training. For example, they would be presented with a member of category A or C and asked to classify it into the correct category. Other novel contrasts involved category B versus C, B versus D, and A versus D. Importantly, correct responding on these novel contrasts required the contrast dimension 3 which had been previously irrelevant during training. Dimensions 1 and 2 yield a maximum accuracy of only 75% accuracy and thus alone cannot be used to attain perfect performance on these classification trials.

Additional instructions were provided to the inference group since the switch classification task was different than their inference task from previous trials.

Feedback was provided. Subjects completed two blocks of 16, switch-contrast trials. Each block was constructed by randomly sampling with replacement from the 16 unique switch trials.

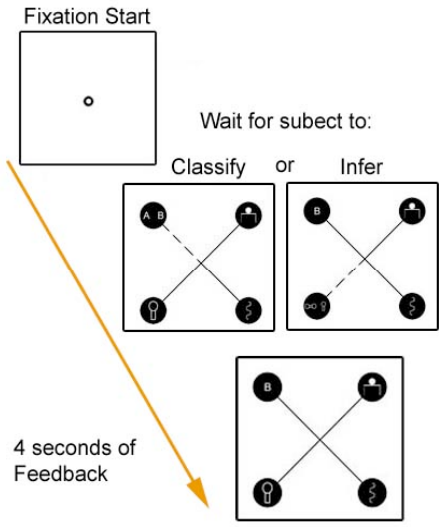

Figure 1. Materials and procedure.

Results and Discussion

Learning AB (blocks 1-5) and CD training performance (6-9) are shown in Figure 2. The figure shows average classification performance for the classification group and relevan cue inference performance. Both classification and inference groups improved over training blocks, but classification training was easier than inference training,

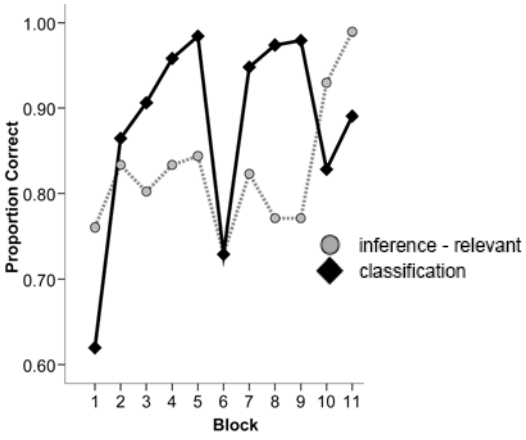

Figure 2. Proportion correct by block.

with a higher proportion correct over blocks. The inference learners performed above chance levels in predicting the valid cue, $t(11) = 4.46$, $p < .01$, but were marginally lower than the classification group, $t(22) = 1.81$, $p < .10$ on the last AB training block. The CD training blocks were similar.

Fixations The crucial question was whether inference learners fixated the non-queried dimension during learning. If inference is a more natural learning task than classification, it should motivate a general interest in learning about the category dimensions; fixations should be distributed to all dimensions, regardless of whether those dimensions are queried. However, if it is the attentional demands of the inference task that drive learning about dimensions (and not a general interest in the category dimensions), then fixations should shift away from the non-queried dimension, since it is no longer immediately relevant for the task. The latter result would suggest that differences in what is learned via inference and classification are from different attentional requirements, and not motivational factors. Eye fixations will be used to distinguish between these two possibilities.

Figure 3 shows proportion of fixations to category label and dimensions over AB and CD blocks, as a function of task. Replicating our earlier work, at the beginning of learning, the average classification learner (top) fixated dimensions about equally. We also observed the expected shift in fixations from irrelevant to relevant dimensions, until irrelevant dimensions were fixated rarely or not at all. At the onset of CD training in block six, there is uneven attention distribution resulting from the learned fixation patterns from AB training, so that in the first trial of CD training, classification learners were not fixating the contrast dimension or the CD relevant dimension. A second attention optimization obtained for classification subjects.

We next examined whether the inference condition yielded any kind of attention optimization. Recall that in the current experiment, the contrast dimension was never queried. If inference motivates a general interest in the category features, we should observe continued fixations to the contrast dimension, in spite of it now being task-irrelevant. However, Figure 3 (bottom) shows that throughout learning, inference learners largely ignored the contrast dimension. Although attention to dimensions 1 and 2 remained high throughout learning, even in the first learning block inference learners largely ignored the contrast dimension. In fact, in the first block of learning, the amount of time fixating the contrast dimension was already significantly less than that of fixating the other two dimensions and the category label (all *p*s < .01). Apparently, inference learners do in fact optimize their attention away from task-irrelevant cues.

Attention optimization in the inference task contradicts the idea that inference motivates a general interest in the category features beyond what is strictly necessary. Rather, the results of Figure 3 support the idea that what distinguishes classification from inference is the attentional demand it places on the learner. Learners fixate dimensions because the task requires it and not because of motivational factors. Any motivation there may have been to learn about all of the category features extinguished quickly.

Switch-trial performance Eye fixation data have ruled out the idea that inference motivates general interest in category features. By not querying the contrast dimension in the inference condition, we allowed inference learners the opportunity to optimize their attention, just as the classification learners could. In fact, inference learners optimized their attention to just those queried dimensions, ignoring the never-queried contrast dimension. As a result of this manipulation, the inference learners may now struggle to include the contrast dimension, since they largely ignored it during training. On the other hand, although the inference learners never directed their attention *to* the contrast dimension, because it was not part of the task, they never had to learn to direct their attention *away* from that dimension either. Rather, the task directed their attention for them. It is this fact that may still allow inference learning to nevertheless produce flexible attention allocation. By not *learning* to ignore the contrast dimension, inference learners may be free to use it during the switch trials.

Blocks 10 and 11 of Figure 2 show proportion correct for switch-classification. In spite of not deploying significant fixations to the contrast dimension during training, the inference condition nevertheless showed an advantage during the switch trials. In the first block of switch trials, the inference group ($M = 0.93$, $SD = 0.09$) outperformed the classification group ($M = 0.83$, $SD = 0.16$), $t(22) = 1.95$, $p = .064$. Likewise, during the second block of switch trials, the inference group ($M = 0.99$, $SD = 0.02$) outperformed the classification group ($M = 0.89$, $SD = 0.15$), $t(22) = 2.26$, $p < .05$. Spending a large amount of time fixating a dimension

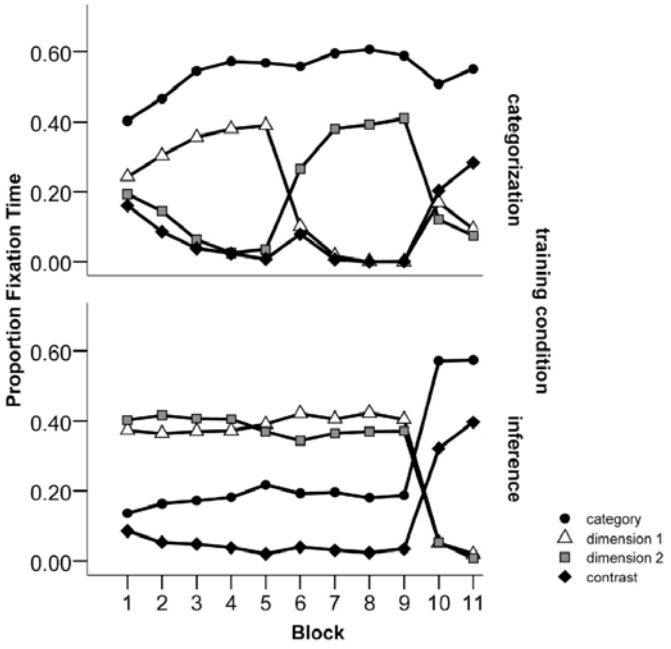

Figure 3. Fixations by task, block, and location.

during learning does not seem necessary for using that dimension later in a flexible way. Whatever inference subjects learned (or didn't learn) during training allowed them to perform well during switch trials.

The eye movement results from training showed that classification and inference learners both largely ignored the contrast dimension. It makes sense then that classification learners should fail to use the contrast dimension during the switch trials, but what allowed inference learners to have more flexible category representations than the classification group? Figure 4 shows learners' attention allocation to the contrast dimension as a function of trial for the first block of AB (trials 1-16) and CD training (trials 81-96). The figure shows that at the trial level, the largest attentional difference between the two conditions was that the classification learners allocated more attention to the contrast dimension *early* in learning. In the first 16 trials of AB training, the classification condition allocated about twice as much fixation time to the contrast dimension (M = 0.16, SD = 0.07) than the inference condition (M = 0.09, SD = 0.06), $F(1, 18) = 4.87$, $MSE = 0.073$, $p < .05$, $\eta^2_p = 0.21$, and was slower to ignore the contrast dimension, as indicated by a trial by task interaction, $F(15, 270) = 2.06$, $MSE = 0.008$, $\eta^2_p = 0.10$, $p < .05$.

A similar pattern obtained in CD training. The classification condition allocated more fixations (M = 0.08, SD = 0.05) to the contrast dimension than the inference condition (M = 0.04, SD = 0.04), $F(1, 18) = 2.92$, MSE = .034, $\eta^2_p = 0.14$, $p = .104$, and was slower to ignore the contrast dimension, $F(15, 270) = 2.67$, $MSE = 0.008$, $p <$ .01. (More accurately, this interaction reflects an inverted u-shaped pattern, in which the classification condition first increased fixations and then decreased fixations to the contrast dimension.) The different patterns of attention reflect different reasons the two groups probably ignored the contrast dimension. Inference learners ignored it because the task explicitly indicated to the subject which dimensions were important. Classification learners were not told which dimension should be ignored, but rather they *learned* to ignore the contrast dimension, as they gradually discovered that the contrast dimension didn't help them classify As from Bs or Cs from Ds. We suspect that this is why there is an initial increase in fixations to the contrast dimension in the first CD block, because classification learners attended to it, and then learned that it was useless in classifying Cs and Ds. Classification learners' fixation results reflected a learned inattention to the contrast dimension, which probably caused their difficulty in attending to the contrast dimension during the switch trials.

General Discussion

We began with the observation that real-life categorizers can make novel category contrasts and that information learned about one set of categories transfers to another without difficulty. This observation seemed to be at odds with the robust finding that people in classification experiments tend to optimize their attention to the fewest necessary dimensions. Such optimization would necessarily force learners to reallocate attention when previously irrelevant dimensions at once become relevant.

To resolve the contradiction that people can make novel category contrasts on one hand and but also tend to optimize attention on the other, we looked to other types of learning tasks they may produce classification performance that is less optimal (in a specific context) but more flexible overall. Inference training seemed like the best candidate. There were two reasons for this. The first was based on evidence that inference yields a special type of processing in humans; although the exact source of this special processing was until now not entirely clear, classification learning has been found to cause humans to attend to diagnostic information and inference learning can cause learners to focus on within-category correlations and prototypical features. We imagined that such differences may reflect that inference is a more typical learning task than classifying, and it isn't hard to imagine how familiarity in the learning task can lead to greater ease and flexibility in using the acquired information.

Our second hypothesis for how inference learning could yield flexible category representations was based on differences in attentional demands of inference and classification. Whereas most classification tasks allow learners to ignore some of the irrelevant dimensions, in the typical inference learning experiment, all of the dimensions are queried several times throughout training. Focusing people's attention on all of the dimensions in this way may cause people to look at all dimensions on every trial, in order to prepare for future queries. In fact, the eye tracking results from this study (and also in Rehder, Colner, & Hoffman, in press) show that never querying one of the dimensions allows the inference learner to optimize their attention to only those task-relevant dimensions, i.e., those dimensions that are sometimes queried.

As it turned out, our initial hypotheses about inference learning were not exactly right. Our data showed that inference subjects very quickly ignored the never-queried dimension. Significant differences in fixations to the contrast dimension were found within the first learning block. Apparently, attending to the contrast dimension during training was not necessary for creating flexible category representations. Rather, what gave subjects the advantage in switch-classification trials is that they never had to learn to *ignore* the contrast dimension, as the classification subjects did, as evidence by the much larger drop in attention to the contrast dimension from the beginning to the end of training in the classification condition. In other words, classification subjects were harmed in their task by their learned attention profiles, but the inference subjects were not.

Such a finding is in fact consistent with theories of attention and category learning. Several models, for example, RASHNL Kruschke and Johansen (1999), and EXIT Kruschke (2001), which are based on Macintosh's (1975) theory of learned attention, propose that attention

weights are learned for a given set of inputs. In these models, if feature inputs are irrelevant, or if for other reasons the features increase the number of classification errors committed, the attention system will direct attention away from those features in favor of others. These attention mechanisms help the models explain a large array of blocking and highlighting phenomena in addition to benchmark category learning data. They also explain why it is that our classification subjects failed to redirect attention to the contrast dimension during the switch trials.

Beyond supporting certain theories of categorization and attention, our results underscore an important difference between the attention profiles acquired through trial and error learning and those that arise out of the direction of the task or experimenter. It seems that ignoring features as a result of discovering that they are irrelevant over numerous trials is qualitatively different than being told explicitly not to look somewhere. Thus, *how* the learner acquires an attention profile is as important as the attention profile itself.

Conclusion We conclude that much of the observed difference in learning between inference and classification is likely because of differences in how attention is directed towards certain features by the demands of the task. Before we do, however, we would first emphasize that we do not think this is a trivial discovery.

With regards to the larger question of how it is that people build up flexible representations to learn novel category contrasts, it is clear that inference training does a much better job at this than classification training does. Thus based on our findings and the previous studies comparing learning tasks, we think it would be a mistake to generalize too broadly about category representations or about how people allocate attention when classifying based on the classification task alone. If a significant proportion of people's experience with categories involves inference and or experience, e.g., communication and problem solving, as we think it probably does, then it is critical that we understand better how tasks interact with what is learned. Finally, in the service of this goal, we believe that methods such as eye tracking, that allow researchers to access information processing online will continue to prove invaluable.

Acknowledgments

This material is based upon work supported by the National Science Foundation under Grant No. 0545298.

References

Anderson, J.R. (1991). The adaptive nature of human categorization. *Psychological Review, 98*, 409-429.

Anderson, A. L., Ross, B. H., & Chin-Parker, S. (2002). A further investigation of category learning by inference. *Memory & Cognition, 1*, 119-28.

M. A. Erickson and J. K. Kruschke, Rules and exemplars in category learning. (1998) *Journal of Experimental Psychology: General, 127*, 107–140.

Hull, C. L. (1920). Quantitative Aspects of the evolution of concepts. *Psychological Monographs, XXVIII.*

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, 812-863.*

Kruschke, J. K., & Johansen, M. K. (1999). A model of probabilistic category learning. *Journal of Experimental Psychology: Learning, Memory, & Cognition*, 25, 1083-1119.

Kruschke, J. K., Kappenman, E. S., & Hetrick, W. P. (2005). Eye gaze and individual differences consistent with learned attention in associative blocking and highlighting. *Journal of Experimental Psychology: Learning, Memory & Cognition, 31*, 830-845

N. Macintosh. (1975) A theory of attention: variations in the associability of stimuli with reinforcement. *Psychological Review, 82*, 276–298.

Markman, A. B., & Ross, B. H. (2003). Category use and category learning. *Psychological Bulletin, 4*, 592-613.

Nosofsky, R. M. (1984). Choice, similarity, and the context theory of classification. *Journal of Experimental Psychology: Learning, Memory, & Cognition*, 10, 104-114.

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

Sakamoto, Y., & Love, B. C. (2006). Sizable sharks swim swiftly: Learning correlations through inference in a classroom setting. *Proceedings of the Cognitive Science Society*. Mahwah, NJ: Lawrence Erlbaum Associates.

Shepard, R. N., Hovland, C. L., & Jenkins, H. M. (1961). Learning and memorization of classifications. *Psychological Monographs, 75 3* (13, Whole No. 517).

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.

Yamauchi, T., & Markman, A. B. (1998). Category learning by inference and classification. *Journal of Memory and Language, 39*, 124-48.

Yamauchi, T., & Markman, A. B. (2000a). Inference using categories. *Journal of Experimental Psychology. Learning, Memory, and Cognition, 3*, 776-95.

Yamauchi, T., & Markman, A. B. (2000b). Learning categories composed of varying instances: The effect of classification, inference, and structural alignment. *Memory & Cognition, 28*, 64-78.

Zaki, S.R., Nosofsky, R.M., Stanton, R.D., & Cohen, A.L. (2003). Prototype and exemplar accounts of category learning and attentional allocation: A reassessment. *Journal of Experimental Psychology: Learning, Memory, and Cognition, 2*, 1160–1173.