Learning and generalization of within-category representations in a rule-based category structure

Shawn W. Ell
Department of Psychology
Graduate School of Biomedical Sciences and Engineering
University of Maine

David B. Smith
Department of Psychology
University of Maine

Rose Deng
Department of Psychology
University of Maine

Sébastien Hélie
Department of Psychological Sciences
Purdue University

March 6, 2020

Running Head: WITHIN-CATEGORY REPRESENTATIONS

Address correspondence to:

Shawn W. Ell
Psychology Department
University of Maine
5742 Little Hall, Room 301
Orono, ME 04469-5742
email: shawn.ell@maine.edu
phone: 207.581.2037
fax: 207.581.6128

Abstract

The task requirements during the course of category learning are critical for promoting within-category representations (e.g., correlational structure of the categories). Recent data suggests that for unidimensional rule-based structures, only inference training promotes the learning of within-category representations, and generalization across tasks is limited. It is unclear if this is a general feature of rule-based structures, or a limitation of unidimensional rule-based structures. The present work reports the results of three experiments further investigating this issue using an exclusive-or rule-based structure where successful performance depends upon attending to two stimulus dimensions. Participants were trained using classification or inference and were tested using inference. For both the classification and inference training conditions, within-category representations were learned and could be generalized at test (i.e., from classification to inference) and this result was dependent upon a congruence between local and global regions of the stimulus space. These data further support the idea that the task requirements during learning (i.e., a need to attend to multiple stimulus dimensions) are critical determinants of the category representations that are learned and the utility of these representations for supporting generalization in novel situations.

Keywords: knowledge representation; training methodology; generalization; category learning; rule-guided behavior

Learning and generalization of within-category representations in a rule-based category structure

Categories are central to virtually all cognitive processes. Much effort has been devoted to understanding how categories are represented and the particular training features that might influence how they are learned (e.g., Markman & Ross, 2003). Outside of the laboratory, however, learning a category representation is not typically an end in and of itself. Instead, the utility of category representations lie in their ability to support other functions (e.g., decision making in novel situations - Hoffman & Rehder, 2010; Markman & Ross, 2003) and the generalizability of category representations depends upon the nature of the representation itself (Carvalho & Goldstone, 2014; Ell, Smith, Peralta, & Helie, 2017; Hélie, Shamloo, & Ell, 2017; Hoffman & Rehder, 2010; Levering & Kurtz, 2015). Thus, it is important to understand the limits of different types of category representations and to identify training features that promote those representations that are most successful for generalization.

Category representations that focus on within-category similarities (e.g., prototypicality, covariation/range of stimulus dimensions within a category) have been argued to be more versatile in supporting generalization than representations that focus on between-category differences (e.g., learn what dimensions are relevant for classification, along with decision criteria or category boundaries) (Chin-Parker & Ross, 2002, 2004; Ell et al., 2017; Helie, Shamloo, & Ell, 2018; Hélie et al., 2017; Kattner, Cox, & Green, 2016; Yamauchi & Markman, 1998). For instance, within-category representations can support both generalization to novel stimuli and generalization to a novel task (Chin-Parker & Ross, 2002; Ell et al., 2017). Furthermore, within-category representations can be applied to novel categorization problems (Hélie et al., 2017; Kattner et al., 2016) and be reconfigured to form new category representations (Helie et al., 2018).

Although a number of methodological factors have been identified as being important for promoting within-category representations (e.g., blocked training - Carvalho & Goldstone, 2014;

concept learning - Hélie et al., 2017; observational training - Levering & Kurtz, 2015; family-resemblance category structures - Markman & Ross, 2003), the emphasis of the present work is on the goal of the task (Goldstone, 1996; Hoffman & Rehder, 2010; Love, 2005; Markman & Ross, 2003; Minda & Ross, 2004; Yamauchi & Markman, 1998). The task goal of classifying a stimulus into one of a number of contrasting categories has been argued to lead to a between-category representation (Erickson & Kruschke, 1998; Hélie et al., 2017; Maddox & Ashby, 1993; Nosofsky, Palmeri, & McKinley, 1994; Smith & Minda, 2002) whereas the task goal of inferring a missing stimulus feature from a partial stimulus and a category label has been argued to lead to a within-category representation (Chin-Parker & Ross, 2002; Ell et al., 2017; Markman & Ross, 2003).

The importance of classification versus inference in promoting within-category representations has been argued to depend upon the category structure (Ell et al., 2017; Hélie et al., 2017). Information-integration category structures (in which information from multiple dimensions needs to be integrated prior to making a categorization response) generally promote within-category representations (Ashby & Waldron, 1999; Ell et al., 2017; Hélie et al., 2017; Thomas, 1998). In contrast, although rule-based category structures (in which logical rules are applied to the stimulus dimensions diagnostic of category membership) can promote within-category representations when learned by inference, rule-based structures may be incapable of promoting within-category representations when learned by classification (Ell et al., 2017)[1].

An inability to learn within-category representations, however, may not be a general feature of rule-based category structures. Ell et al. (2017) used a unidimensional, rule-based structure in which the stimuli varied along two continuous-valued dimensions, but only a single stimulus dimension was diagnostic of category membership. Thus, successful classification depended upon a single dimension, but the within-category representation (i.e., knowledge of the

[1] Although logical rules can be based on either within- or between-category representations (e.g., large or larger than), the subset of logical rules learned when classifying rule-based structures tends to depend upon between-category representations (Casale, Roeder, & Ashby, 2012; Ell & Ashby, 2012; Ell, Ing, & Maddox, 2009; Hélie et al., 2017).

correlational structure of the categories) depended upon both stimulus dimensions. The inability to learn and generalize within-category representations when classifying a rule-based structure was interpreted as reflecting a limitation of the between-category representation (i.e., the logical rule used for classification). While this may be true when the classification rule depends upon a single stimulus dimension, it is also possible that within-category representations could be learned if the classification rule depended upon the same number of dimensions as the within-category representation.

The following experiments investigate this issue using a two-dimensional, rule-based category structure (i.e., exclusive-or, Figure 1, bottom). In this category structure, successful classification (i.e., classifying stimuli as a member of category A or B) requires attention to both stimulus dimensions. Similarly, successful inference (i.e., inferring a missing stimulus feature when given one feature and the category label) also requires attention to both stimulus dimensions. Although the between-category representation (i.e., the logical rule: members of category A either have larger circles and steeper lines, or smaller circles and shallower lines, than members of category B.) would convey some rudimentary information about the within-category correlations, it is not at all clear if this information would be sufficient to support generalization from classification to inference.

Briefly, across three experiments, participants were trained on classification or inference and subsequently tested on inference. If it is not possible to learn within-category representations when classifying a rule-based structure, only participants trained by inference should evidence knowledge of the within-category correlations at test. In contrast, if attending to multiple stimulus dimensions during training is a critical factor promoting the learning of within-category representations, participants in both conditions should evidence knowledge of the within-category correlations at test. To foreshadow, the results support the latter hypothesis suggesting that within-category representations can be learned in a rule-based task and generalized to a novel task (i.e., from classification to inference).

Experiment 1

Method

Participants and design

One-hundred nineteen participants were recruited from the University of Maine student community and received partial course credit for participation. Sample size (approximately 30 participants/condition) was estimated based upon a similar experiment in our lab (Ell et al., 2017). Data collection was continued beyond this target (until the end of the semester) in order to provide sufficient research opportunities for participants in an introductory psychology research pool. Participants were randomly assigned to one of two experimental conditions: classification or inference training. A total of nine participants were excluded from analyses: seven due to a software error and two participants did not complete the task within the hour-long experimental session, resulting

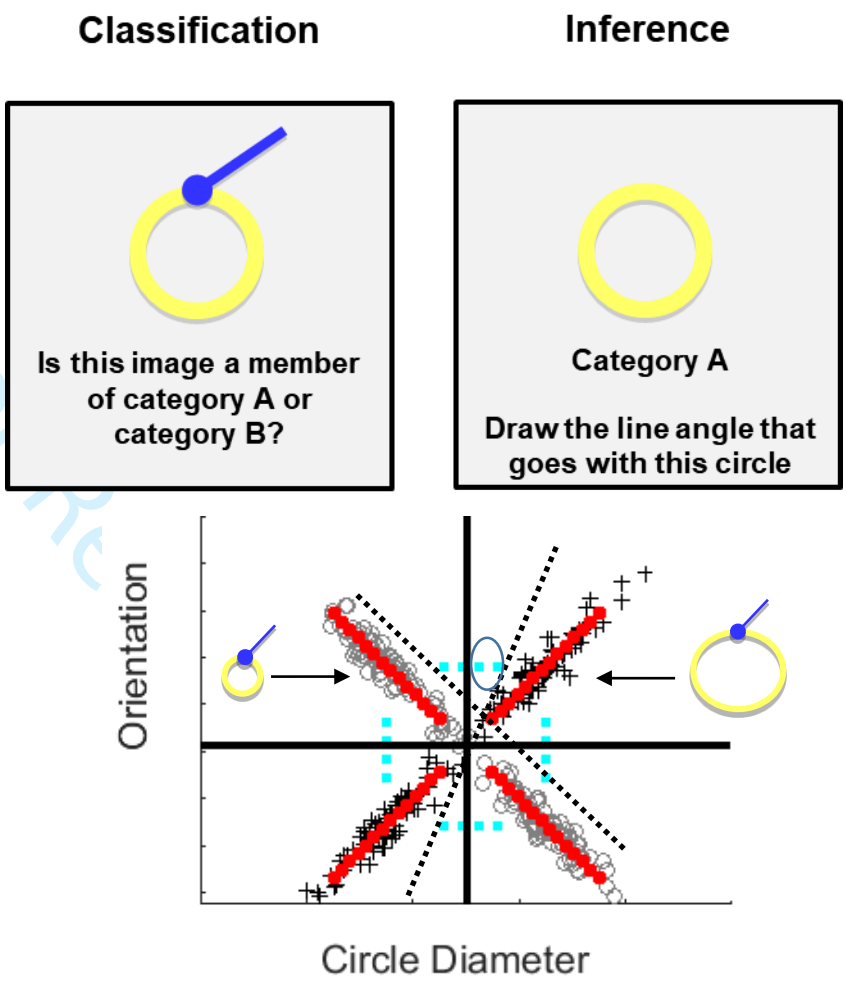

Figure 1. (Top) Example displays for the two training methodologies. (Bottom) Rule-based category structure used in Experiments 1 and 3. Category A (crosses) and B (circles) stimuli used during the training phase. The insets are example stimuli. The solid black boundaries represent the optimal conjunctive decision strategy. The dashed black boundaries represent an alternative decision strategy (see text for details). Stimuli used during the test phase are plotted as filled red circles. Probe stimuli used during the final block of training are plotted as blue squares. See text for details. (Color figure online).

in sample sizes of 55 in each condition. All participants reported normal (20/20) or corrected to normal vision.

Stimuli and apparatus

The stimuli comprised circles (varying continuously in diameter) and an attached line (varying continuously in orientation from horizontal) (Figure 1, top). The category structures were created using a variation of the randomization technique (Ashby & Gott, 1988) in which the stimuli were generated by sampling from bivariate normal distributions defined in a diameter × angle (from horizontal) space in arbitrary units. The category means for the stimuli in each of the four quadrants of Figure 1 (two per category) were μ_{A1} = [650, 250], μ_{A2} = [350, -50], μ_{B1} = [350, 250], and μ_{B2} = [650, -50]. The covariance matrices were $\Sigma_A = \begin{bmatrix} 3875 & 3625 \\ 3625 & 3875 \end{bmatrix}$ and $\Sigma_B = \begin{bmatrix} 3875 & -3625 \\ -3625 & 3875 \end{bmatrix}$ (i.e., a correlation of 1 between diameter and angle for each quadrant assigned to category A and -1 for each quadrant assigned to category B).

On each trial a random sample (x, y) was drawn from category A or B and used to create a stimulus with a circle of $\frac{x}{2}$ pixels in diameter and a line $\frac{180y}{800}$ degrees (counterclockwise from horizontal) with a length of 200 pixels. The line was always connected to the highest point of the circle. For the training phase, 80 stimuli (40 from each category, 20 from each quadrant) were generated for each of the 4 blocks of trials (Black symbols in Figure 1). For the test phase, 56 stimuli (28 from each category, 14 from each quadrant) were used for the single test block (red circles in Figure 1). The experiment was run using the Psychophysics toolbox (Brainard, 1997; Kleiner et al., 2007; Pelli, 1997) in the Matlab computing environment. Each stimulus was displayed on a 1600 x 1200 pixel resolution 20-inch LCD with a viewing distance of 20 in.

Participants were expected to use a conjunctive strategy in the classification task (e.g., the solid black decision boundaries plotted in Figure 1 – If the circle is large and high on orientation or if the circle is small and low on orientation, respond A; otherwise respond B.), but given the large separation between stimuli in the four quadrants, other strategies could also result in high

levels of accuracy. For instance, a strategy assuming participants integrate the stimulus values prior to any decision process would also predict high levels of performance during training (e.g., the linear classifier plotted as dashed black boundaries in Figure 1; see the Appendix for more details). To address this issue, probe stimuli (16 total) were included in the final block of classification training, resulting in a total of 96 trials during the final block (Light blue squares in Figure 1, Table 1). The example conjunctive and linear classifiers plotted in Figure 1 would predict different categorization responses for a subset of the probe stimuli. For instance, for the two circled probe stimuli, the conjunctive classifier (solid) would predict a category A response whereas the linear classifier (dashed) would predict a category B response. In order to equate the similarity between conditions, probe stimuli were also included in the inference condition. The probe stimuli were not members of either category, thus there is no correct or incorrect response to these stimuli. Thus, no feedback was provided on probe trials and the probe stimuli were excluded from accuracy analyses. Probe trials were only used to estimate individual participant decision strategies in the classification condition.

Table 1. Probe stimuli coordinates (arbitrary units)

Diameter	Angle
650	40
650	74
650	108
650	142
560	250
526	250
491	250
457	250
350	40
350	74
350	108
350	142
560	-50
526	-50
491	-50
457	-50

Procedure

Each participant was run individually. At the beginning of the training phase, participants were informed that stimuli would comprise a circle with a line connected at the top, and that the stimuli would be presented individually, but would vary across trials in circle diameter and line angle. In the classification condition, participants were instructed that their goal was to learn to distinguish between members of category A and B by trial and error. On each trial, participants were shown a stimulus and prompted with “Is this image a member of category “A” or category “B?” and instructed to select a category for each stimulus by pressing a button labeled “A” or a button labeled “B” on the keyboard to indicate which category was selected.

In the inference condition, participants were instructed that their goal was to learn to draw the missing stimulus component by trial and error (Figure 1). On each trial, either a circle or a line was presented with the category label. Participants were instructed to draw the missing stimulus component. On half of the trials they were asked "Draw the circle that goes with this line angle" and on the other half they were asked to "Draw the line that goes with this circle". To draw the circle, participants used the mouse to indicate the location of the bottom of the circle (indicating the diameter of the circle relative to the dot at the beginning of the line). To draw the line, participants used the mouse to indicate the location of the end of the line (indicating the orientation of the line relative to horizontal). The circle or line was then drawn to match the participant's selection with a line beginning at the dot at the top of the circle (at a constant length of 200 pixels). Subsequently, participants were able to fine-tune the circle diameter or the line angle using the arrow keys on the keyboard. Any selected stimulus values outside the allowable range (diameter 10 to 600 pixels, angle: 50 to 110 degrees) were reset to the nearest allowable value.

Stimulus presentation was response terminated with an upper limit of 60 s. After responding, feedback was provided. In the classification condition, the screen was blanked and the word "CORRECT" (in green, accompanied by a 500 Hz tone) or "WRONG" (in red, accompanied by a 200 Hz tone) was displayed. In the inference condition, the correct circle or line was overlaid upon the participant's response (in black). In all conditions, feedback duration was 2 s and the screen was then blanked for 1 s prior to the appearance of the next stimulus.

In addition to the trial-by-trial feedback, summary feedback was given at the end of each training block. For the classification condition, proportion correct for the block was shown (participants were informed that higher numbers are better) and for the inference condition the root-mean-square-error between the drawn and correct stimulus was shown (participants were informed that lower numbers are better). The presentation order of the stimuli was randomized within each block, separately for each participant. Participants completed several practice trials

prior to beginning the training phase to familiarize themselves with the task using stimuli randomly sampled (with equal probability) from the training categories.

During the test phase, all participants performed the inference task (1 block of 56 trials). Instruction was provided to all conditions and participants completed several practice trials using stimuli randomly sampled from the test phase stimuli (with equal probability). No feedback was provided during the test phase.

Results

Training Phase: Performance on Classification and Inference

In the inference condition, the diameter-angle correlations within each quadrant were in the appropriate direction (i.e., positive for the upper right and lower left, negative for the lower right and upper left), thus the following analyses average across the quadrants[2]. Data were also averaged across quadrants in the classification condition, and for all subsequent analyses, unless otherwise noted.

The dependent measure was different for the classification (proportion correct) and the inference (correlation between the given and produced stimulus values) conditions, therefore the data from each condition were analyzed separately. Performance generally improved across blocks for both conditions (Figure 2, Table 2). Consistent with this observation, separate paired-samples t-tests indicated significant increases from block 1 to block 4 in proportion correct for the classification condition: [$t(54) = -4.491$, $p < .001$, $d = .72$] and the diameter-angle correlation for the inference condition: [$t(54) = -4.454$, $p < .001$, $d = .52$].

[2] For inference training, and the test phase, the observed diameter-angle correlations for the category with the negative within-category correlation (i.e., category B in Experiments 1 and 3, category A in Experiment 2) was multiplied by -1. This was done in order to allow for the aggregation with the data from the category with the positive within-category correlation. For presentation purposes, the training data are plotted prior to multiplying by -1.

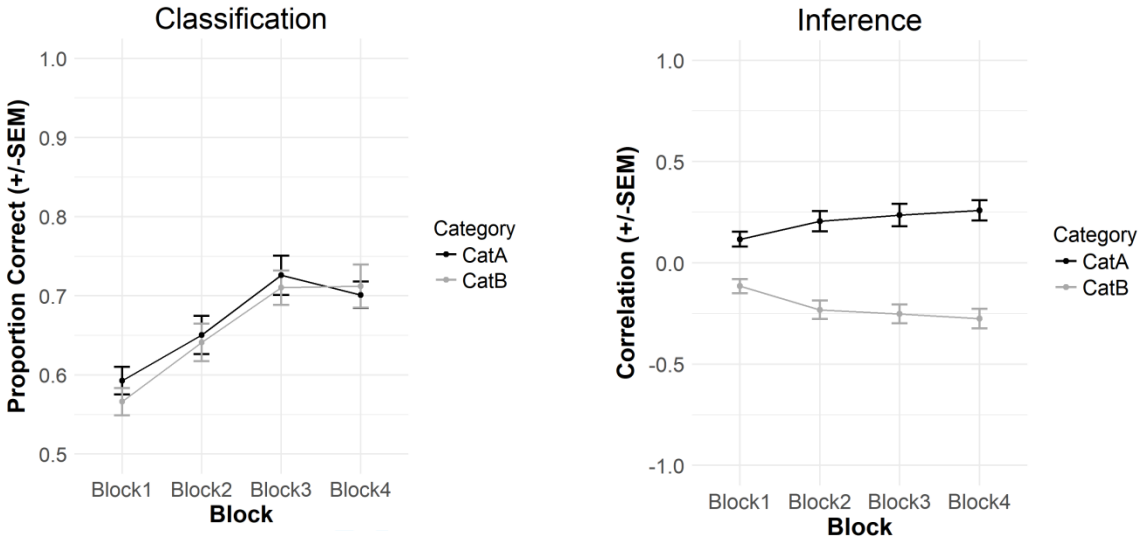

Figure 2. Training performance in the classification (proportion correct) and inference (correlation between the given and produced stimulus values) conditions of Experiment 1.

Table 2. Training and test phase performance in Experiment 1.

	Training Block 1		Training Block 4		Test	
	M	SD	M	SD	M	SD
Classification	.58	.10	.71	.14	.19	.23
Inference	.12	.29	.27	.45	.32	.40

Note: Performance during classification is indexed by proportion correct whereas performance during inference training, and test, is indexed by the diameter-angle correlation.

Training Phase: Classification Decision Strategy

Participants were expected to learn conjunctive strategies in the classification condition. In order to confirm this, a number of decision bound models (Ashby, 1992a; Maddox & Ashby, 1993) were fit to the individual participant data from the classification condition. Four different types of models were evaluated in order to assess an individual's strategy during the final training block. Unidimensional models assume that the participant sets a single decision criterion on one stimulus dimension (e.g., if the circle is large, respond A; otherwise respond B). Conjunctive models assume separate decision criteria on both dimensions (e.g., If the circle is large and high on orientation or if the circle is small and low on orientation, respond A; otherwise respond B. Figure 1). Information-integration models assume that the participant integrates the stimulus

information from both dimensions prior to making a categorization decision (Figure 1). Finally, random responder models assume that the participant guessed. Each model was fit separately to the final block of training (including the probe stimuli), for each participant, using a standard maximum likelihood procedure for parameter estimation (Ashby, 1992b; Wickens, 1982) and the Bayes information criterion for goodness-of-fit (Schwarz, 1978) (see the Appendix for a more detailed description of the models and fitting procedure). Based upon our previous work (Ell et al., 2017), participants using information-integration strategies during classification training, but not unidimensional strategies or guessing, would be expected to promote within-category representations that could be used to support performance on the test phase inference task. If simply attending to both stimulus dimensions during classification training is sufficient to promote within-category representations, then participants using conjunctive strategies would also be expected to perform well during the test phase inference task. Consistent with expectations, the majority of participants learned a task-appropriate, conjunctive strategy (64%) with the remaining participants being best fit by either the unidimensional (9%) or random responder (27%) models.

Test phase

Initial inspection of the correlations during test phase suggests the learning of the correlational structure of the categories in both the inference and classification training conditions (Figure 3). To analyze these data, one-sample t-tests (within each condition), an independent samples t-test comparing the conditions, and the scaled JZS Bayes Factor, B_{01}, (Jeffreys, 1961; Kass & Raftery, 1995; Rouder, Speckman, Sun, Morey, & Iverson, 2009) were computed. Consistent with the inspection of the Figure 3 data, the correlation during test was significantly greater than zero in both the inference [$t(50) = 6.22$, $p < .001$, $d = .87$; $B_{01} = 142785.4$ to 1 in favor of the alternative hypothesis] and classification [$t(53) = 5.88$ $p < .001$, $d = .80$; $B_{01} = 53171.69$ to

1 in favor of the alternative hypothesis] conditions[3]. For the classification condition, this result was driven primarily by participants using a task-appropriate, conjunctive strategy (conjunctive: *M* = *.28, SD* = .21; unidimensional: *M* = *-.10, SD* = .11; random responder: *M* = *.05, SD* = .17). An independent samples t-test comparing the two conditions, however, indicated superior test-phase performance in the inference condition [$t(103) = 2.26$, p = .03, $d = .44$; $B_{01} = 1.96$ to 1, weakly favoring the alternative hypothesis].

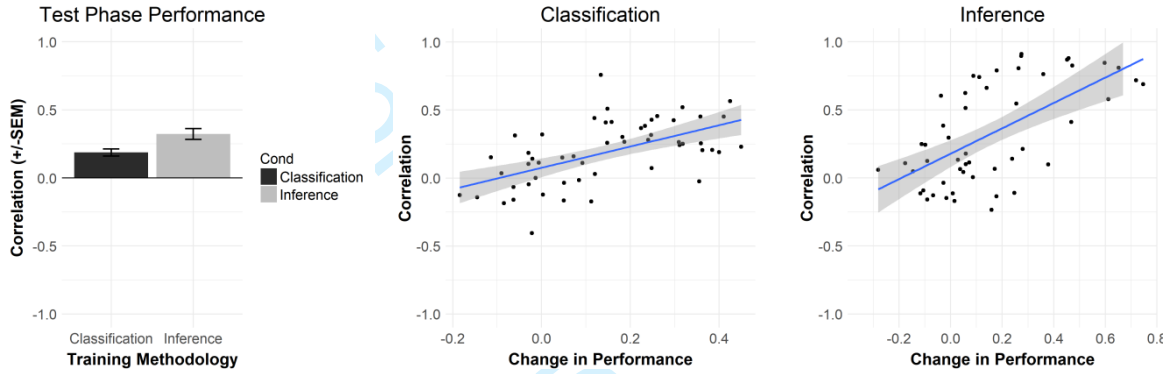

Figure 3. Performance on the inference task during the test phase (Left). Note that the diameter-angle correlations from category B are multiplied by -1 prior to averaging with the diameter-angle correlations from category A, thus positive values suggest learning of the within-category correlations. Relationship between learning during training and test phase performance in the classification (middle, $r = .57$) and inference (right, $r = .62$) conditions. The grey area represent a 95% confidence interval.

If test-phase performance is driven by learning during the training phase, the amount of learning during the training phase should be predictive of test-phase performance. To assess this, in the classification condition, the Pearson correlation was computed between the change in

[3] Given the considerable individual variability in the change in performance during training, we conducted a follow-up analysis focusing on participants who performed above chance during the final block of training (i.e., accuracy > 60% correct for classification; correlation > .11 or < -.11 for inference). This subgroup of participants evidenced stronger knowledge of the within-category correlations in both the classification. [$n = 38$, $M = .25$, $SD = .22$, $t(37) = 7.08$, $p < .001$, d = 1.15; $B_{01} = 608390.3$ to 1 in favor of the alternative hypothesis] and inference [$n = 24$, $M = .53$, $SD = .37$, $t(23) = 6.99$, $p < .001$, $d = 1.43$; $B_{01} = 41629.9$ to 1 in favor of the alternative hypothesis] conditions. Performance in the two conditions was significantly different [$t(60) = 3.63$, $p < .001$, $d = .95$; $B_{01} = 48.27$ to 1 in favor of the alternative hypothesis].

accuracy across blocks (block 4 minus block 1) and the observed diameter-angle correlation during the test phase. In the inference condition, the Pearson correlation was computed between the change in the observed diameter-angle correlation (block 4 minus block 1) and the observed diameter-angle correlation during the test phase. There was a significant positive relationship between learning during training and test-phase performance in both the classification: [$r(52) = .57$, $p < .001$] and inference: [$r(49) = .62$, $p < .001$] conditions. The strength of this relationship, however, did not differ between the classification and inference conditions [Fisher's $z = 0.36$, $p = .72$]. In sum, these data suggest learning of the within-category representations for both the classification and inference conditions, with a possible advantage for participants in the inference condition.

Summary

The goal of Experiment 1 was to determine if classification of a two-dimensional, rule-based category structure was sufficient to support the learning of within-category representations or if an inability to learn within-category representations is a more general feature of rule-based structures (Ell et al., 2017). Consistent with the former, participants demonstrated knowledge of the within-category correlations at test in both the inference and classification conditions, although test phase performance in the inference condition was superior. That being said, training phase performance was positively associated with test phase performance to a similar extent in both conditions. In sum, these data suggest that learning to classify a rule-based structure that requires attention to multiple stimulus dimensions is sufficient to support the learning of within-category representations that can be generalized to a novel task (i.e., from classification to inference).

Experiment 2

The results of Experiment 1 suggest that inference training may be superior to classification in promoting the learning of within-category representations, but there was evidence that within-category representations were learned in the classification condition as well. This latter result may be a consequence of the need to attend to both stimulus dimensions for successful performance during training, but there is another possible explanation. The Experiment 1 analysis

computed the observed diameter-angle correlation within each quadrant of the stimulus space, and then averaged these results across the two quadrants assigned to each category, in order to estimate the within-category representations. There was, however, a congruence between the local diameter-angle correlation within each quadrant and the global correlation within each category (e.g., positive within the two quadrants assigned to category A and positive within category A, across the stimulus space). Thus, another possibility is that this local-global congruence facilitated learning of the within-category representations. Experiment 2 investigates this question using a category structure in which the local diameter-angle correlation within each quadrant is incongruent with the global within-category correlation (e.g., negative within the two quadrants assigned to category A and generally positive within category A, across the stimulus space – see Figure 4). If learning within-category representations in the classification condition is dependent solely upon a need to attend to both stimulus dimensions, participants should still evidence knowledge of the within-category correlations at test. If, instead, the local-global congruence is critical, it should be difficult for participants to learn the within-category correlations. It is expected that participants in the inference condition will still be able to learn the within-category correlations with the Figure 4 structure, but it is possible that inference too would be sensitive to a local-global incongruence.

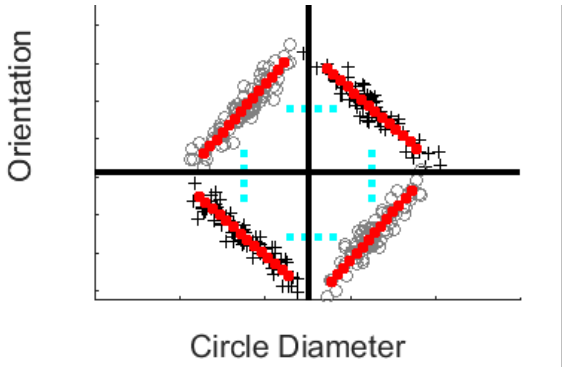

Figure 4. Conjunctive category structure used in Experiment 2. Category A (crosses) and B (circles) stimuli used during the training phase. Stimuli used during the test phase are plotted as filled red circles. Probe stimuli used during the final block of training are plotted as blue squares. (Color figure online).

Method

Participants and design

Seventy-four participants were recruited from the University of Maine student community and received partial course credit for participation. Participants were randomly assigned to one of two experimental conditions: classification or inference training. Two participants were excluded

from analyses due to software error, resulting in sample sizes of 34 (classification) and 38 (inference). All participants reported normal (20/20) or corrected to normal vision.

Stimuli, apparatus, and procedure

The stimuli and procedure were identical to Experiment 1 with one exception. The stimuli within each quadrant of the stimulus space were rotated 45 degrees (about the quadrant mean), in order to reduce the congruence between the diameter-angle correlation within each quadrant and diameter-angle correlation within each category (Figure 4).

Results

Training Phase

Only participants in the classification training condition showed learning of the category structures as accuracy was higher in block 4 than in block 1: [$t(33) = -4.36$, $p < .001$, $d = .76$] (Figure 5, Table 3). There was no significant increase in the correlation learned from block 1 to block 4 in the inference condition: [$t(37) = -0.28$, $p > .78$, $d = .06$].

The decision-bound models described in Experiment 1 were fit to the final training block in the classification training condition. A majority of the participants in the classification condition learned a task-appropriate, conjunctive strategy (53%) with the remaining participants being best fit by either the unidimensional (10%), information-integration (3%), or random responder (33%) models.

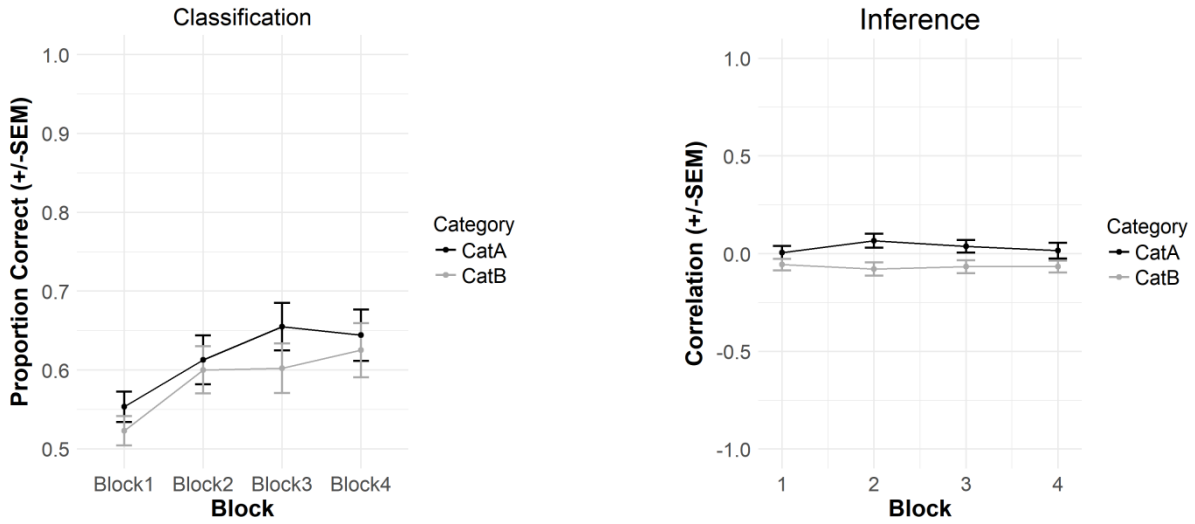

Figure 5. Training performance in the classification and inference conditions of Experiment 2.

Table 3. Training and test phase performance in Experiment 2.

	Training Block 1		Training Block 4		Test	
	M	SD	M	SD	M	SD
Classification	.54	.11	.63	.19	.10	.24
Inference	.03	.20	.04	.22	.04	.30

Note: Performance during classification is indexed by proportion correct whereas performance during inference training, and test, is indexed by the diameter-angle correlation.

Test Phase

Inspection of the test phase data revealed that participants often mis-estimated the direction of the diameter-angle correlation across quadrants of the stimulus space in both conditions (Table 4). Due to this issue, the correlations were not averaged across quadrants. Instead, diameter-angle correlations were evaluated within each quadrant against a critical M = +/-.11 [estimated using $\alpha = .05$ two-tailed, $t(33) = 2.04$ and an average $SD = .31$]. The diameter-angle correlations in the lower-left quadrant (classification) and the upper left quadrant (inference) were significantly different from 0 and in the opposite direction of the actual correlation. The correlations in the remaining quadrants were not significantly different from 0. In sum, there was no evidence that participants were able to learn the within-category correlations with the Figure 4 category structures.

Table 4. Within-category correlations by quadrant

	Classification			Inference		
Quadrant	M	SD	B_{01}	M	SD	B_{01}
Upper Right	0.08	0.33	2.32, null	0.03	0.35	4.92, null
Lower Right	-0.05	0.27	2.91, null	0.01	0.34	5.61, null
Lower Left	0.18	0.35	8.21, alternative	-0.01	0.43	5.79, null
Upper Left	-0.08	0.28	1.52,null	-0.15	0.33	8.98, alternative

Note. B_{01} is the JZS Bayes Factor for the one-sample t-test comparing the mean in each quadrant to 0. "null" and "alternative" refer to the hypothesis favored by B_{01}.

Summary

The goal of Experiment 2 was to investigate if the learning of within-category representations while classifying was dependent upon a congruence between the local, diameter-angle correlations within each quadrant and the global diameter-angle correlations within each category. The results suggest that this was the case. Although participants learned to classify the Figure 4 structure, there was no evidence that within-category representations were learned at test. Unexpectedly, this was also true in the inference condition. The results of Experiment 1, along with previous work from our lab (Ell et al., 2017), suggested that inference training facilitated the learning of within-category representations regardless of the category structure. The results of Experiment 2 suggest that even for inference training, there is a limit to the learning of within-category representations. In sum, the learning of within-category representations, with the rule-based structures investigated here, is dependent upon a local-global congruence regardless of the task goal.

Experiment 3

The results of Experiments 1 and 2 suggest that inference training more strongly promotes the learning of within-category representations, at least when there is a congruence between local and global regions of the stimulus space. This advantage may be driven by a practice effect given that participants in the inference condition performed the same task during the training and test

phases whereas participants in the classification condition performed different tasks during the training and test phases. Experiment 3 addresses this issue using a two-alternative, forced-choice version of inference training that more closely matches classification training and enables the investigation of generalization to a novel task in the inference condition (i.e., from forced-choice to a production task). In addition, the forced-choice procedure in Experiment 3 is more similar to inference training procedures used in previous work (e.g., Yamauchi & Markman, 1998). The vast majority of previous work with the forced-choice procedure, however, has used discrete-valued dimensions with a small number of stimuli. Experiment 3 extends this work to a category structure to continuous-valued dimensions with a large number of stimuli.

Method

Participants and design

Seventy-one participants were recruited from the University of Maine student community and received partial course credit for participation. Participants were randomly assigned to one of two experimental conditions: classification or inference training. One participant was excluded from analyses due to software error. The resulting sample sizes by condition were classification: 37; inference 33. All participants reported normal (20/20) or corrected to normal vision.

Stimuli and apparatus

The stimuli were identical to Experiment 1 with the exception of the inference condition. Two response alternatives were presented 325 pixels below the stimulus, one offset 325 pixels left of center and the offset 325 pixels right of center (Figure 6). One of the response alternatives was correct. The incorrect alternative was generated by selecting the corresponding value for the missing dimension from the contrasting category. The location of correct/incorrect alternatives were counterbalanced.

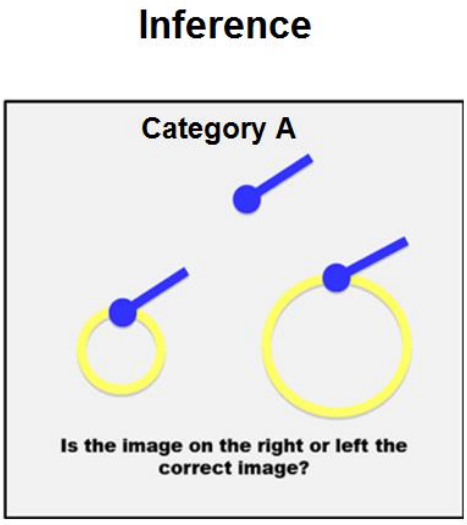

Figure 6. (Example display for the inference training methodology. (Color figure online).

Procedure

The procedure was identical to Experiment 1 with the exception that during training, participants in the inference condition were asked to choose from one of the two response alternatives rather than drawing the missing stimulus dimension. In addition, trial-by-trial feedback in the inference condition was presented in the same way as in the classification condition. The test phase was identical to Experiment 1 (i.e., all participants were instructed to draw the missing stimulus dimension and no feedback was provided).

Results

Training Phase

Learning was evident in both conditions. (Figure 7, Table 5). Although the dependent measure (proportion correct) was now the same across conditions, training performance was analyzed separately for the two conditions to maintain consistency with the analyses in the previous experiments. Separate paired-samples t-tests indicated significant increases from block 1 to block 4 in proportion correct for the classification condition: [$t(36) = -4.163$, $p < .001$, $d = .87$] and for the inference condition: [$t(32) = -2.920$, $p = .006$, $d = .56$].

Table 5. Training and test phase performance in Experiment 3.

	Training Block 1		Training Block 4		Test	
	M	SD	M	SD	M	SD
Classification	.58	.11	.69	.19	.13	.23
Inference	.52	.09	.60	.18	.10	.28

Note: Performance during classification and inference training is indexed by proportion correct whereas performance during test is indexed by the diameter-angle correlation.

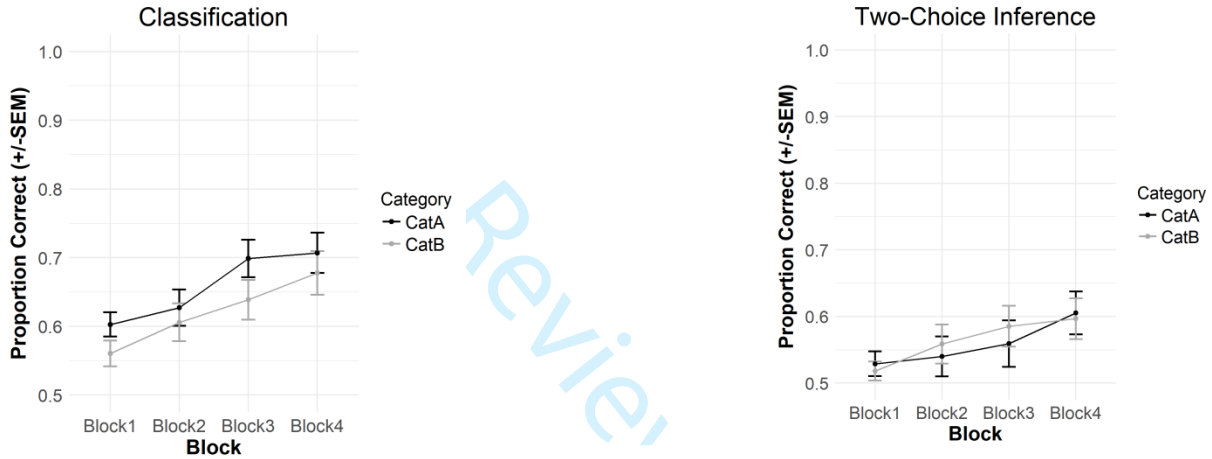

Figure 7. Training performance in the Classification and (forced-choice) Inference conditions of Experiment 3.

The decision-bound models described in Experiment 1 were fit to the final training block in the classification training condition. Consistent with expectations, the majority of participants learned a task-appropriate, conjunctive strategy (57%) with the remaining participants being best fit by the unidimensional (5%), information-integration (5%), or random responder (33%) models.

Test Phase

Participants in the classification condition evidenced knowledge of the within-category correlations [$t(36) = 3.53$, $p = .001$, d = .58; $B_{01} = 1/27.44$, favoring the alternative hypothesis] whereas participants in the inference condition performed marginally better than chance [$t(32) = 1.99$ $p = .06$, d = .35; $B_{01} = 1.07$, equivocal support for the null and alternative hypotheses]. The two conditions, however, were not significantly different from each other [$t(68) = .53$, $p = .60$, d

= .13; B_{01} = 3.6 in favor of the null hypothesis] (Figure 8)[4]. For the classification condition, this result was driven primarily by participants using a task-appropriate, conjunctive strategy (conjunctive: *M* = *.23, SD* = .24; unidimensional: *M* = *.05, SD* = .14; random responder: *M* = *.003, SD* = .16).

In both conditions, however, greater learning during the training phase was associated with higher performance during the test phase [classification: *r(36)* = .47, *p* = .003; inference: *r(32)* = .67, *p* < .001]. A re-analysis of the data from the inference condition excluding five potential multivariate outliers [robust Mahalanobis squared distances were calculated and values that exceeded $x(1)^2_{critical} = 5.02, \alpha = .025$ were considered outliers] indicated an association identical in magnitude to that of the classification condition [*r*(25) = .47, *p* = .01].

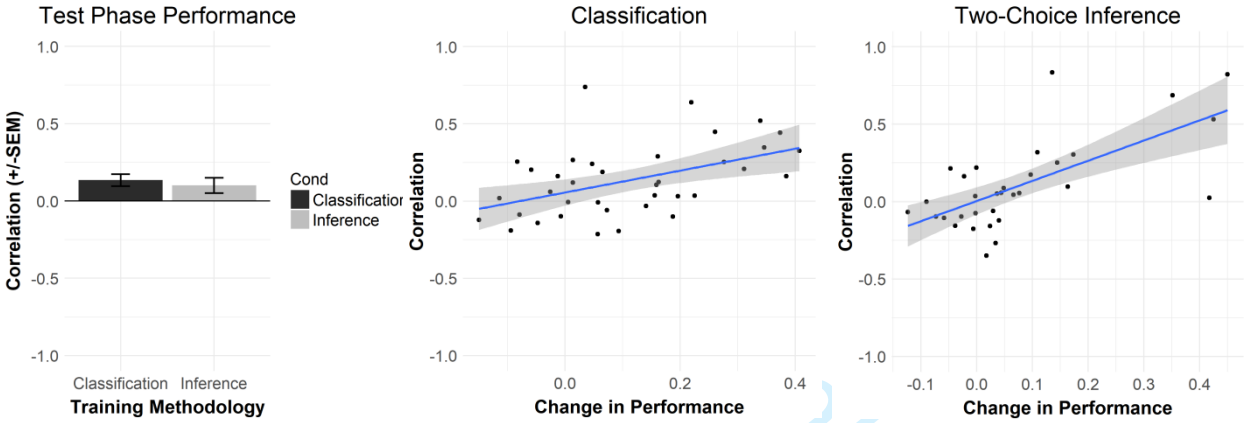

Figure 8. Performance on the inference task during the test phase (Left). Note that positive values suggest learning of the within-category correlations. Relationship between learning during training and test phase performance in the classification (*r* = .47) (Middle) and inference (*r* = .67) conditions. Note that the diameter-angle correlations from category B are multiplied by -1 prior to averaging with the diameter-angle correlations from category A. The grey bars represent a 95% confidence interval.

[4] Participants who performed above chance during the final block of training (i.e., accuracy > 60% correct) evidenced stronger knowledge of the within-category correlations in both the classification. [*n* = 24, *M* = .20, *SD* = .24, *t*(23) = 4.09, *p* < .001, d = .83; B_{01} = 70.87 to 1 in favor of the alternative hypothesis] and inference [*n* =15, *M* = .25, *SD* = .34, *t*(14) = 2.08, *p* = .014, *d*= .72; B_{01} = 1.4 weakly favoring the alternative hypothesis] conditions. Performance in the two conditions, however, were not significantly different [*t*(37) = -.52, *p* = .60, *d* = .17; B_{01} = 2.82 to 1 in favor of the null hypothesis].

Summary

The primary goal of Experiment 3 was to investigate the extent to which a two-choice version of inference training would support the learning of within-category representations. Participants in the classification condition were able to learn the test-phase correlations. Participants in the two-choice inference condition, however, performed only marginally better than chance. That being said, test phase performance was not significantly different in the classification and inference conditions. Similar to Experiment 1, there was a positive correlation between training and test phase performance which did not differ by condition. Taken together, these results suggest that by eliminating a potential practice effect by introducing a two-choice version of the inference task, participants in both conditions learned the within-category correlations equally well.

General Discussion

Previous research suggests that the between-category representations (i.e., logical rules) thought to support the learning of rule-based tasks do not also support the learning of within-category representations (Ell et al., 2017). This work, however, focused on a rule-based structure for which learning required attention to a subset of the stimulus dimensions that were critical for the within-category representation. The present work investigated the extent to which an inability to learn within-category representations is a general limitation of rule-based structures or a more specific limitation resulting from a mismatch between the information necessary for learning between- and within-category representations. The results of Experiment 1 were consistent with the latter hypothesis. More specifically, participants were able to learn to classify a two-dimensional, rule-based structure and this knowledge was able to support the learning of within-category representations that could be generalized to a novel task (i.e., inference). This result was dependent upon a congruence between local and global features of the category structure (Experiment 2). Although participants who learned by inference in Experiment 1 demonstrated stronger knowledge of the within-category representations at test, this advantage seems to have

reflected a practice effect (Experiment 3). In sum, these results suggest that a task goal thought to promote the development of between-category representations (i.e., classification) can promote the development of within-category representations, but such learning is sensitive to characteristics of the category structure.

Learning and Generalization of Within-Category Representations

Consistent with previous work (Anderson & Fincham, 1996; Ell et al., 2017; Thomas, 1998), within-category correlations could be learned during categorization. These representations could also be generalized across tasks with knowledge of the within-category correlations learned during classification training being able to support inference at test. Both learning and generalization, however, depended upon a congruency between the local, diameter-angle correlations within each quadrant and the global diameter-angle correlations within each category (Experiment 2). Disrupting this congruency seems to have impaired the ability to learn within-category representations while sparing category learning, suggesting a different type of category representation may have supported the learning of the Experiment 2 categories. Although we do not have a direct measure of the category representation learned in Experiment 2, the model-based analyses suggest that nearly half of the participants learned a between-category representation (i.e., logical rules) and previous work suggests that rule-based strategies are used with other exclusive-or category structures (Kurtz, Levering, Stanton, Romero, & Morris, 2013; Nosofsky et al., 1994). Nevertheless, we cannot rule out the possibility that participants learned a different type of within-category representation (e.g., exemplars, prototypes, within-category range).

A related, and important, question is how exactly are within-category representations learned from classification (Experiments 1 and 3)? If most participants are learning between-category representations during classification of the Figure 1 category structure, as suggested by the model-based analyses, are these between-category representations facilitating the development of within-category representations? The optimal conjunctive rule (i.e., members of

category A either have larger circles and steeper lines, or smaller circles and shallower lines, than members of category B.) conveys some basic information about the within-category correlations. It may be the case that this information was sufficient to support generalization from classification to inference. The results of Experiment 2 suggest, however, suggest that this is unlikely. In Experiment 2, the optimal rule was the same, but participants only learned within-category correlations consistent with this rule in one quadrant of the stimulus space. That being said, our method does not allow for distinguishing between participants that are good at using this rule to perform inference versus participants that have a richer knowledge of the within-category correlations, thus more work is needed to address this possibility.

Alternatively, perhaps there is a learning system operating that is acquiring within-category representations that could be used to support both classification and inference. For instance, the DIVA (Kurtz, 2007) and SUSTAIN (Love, Medin, & Gureckis, 2004) models of category learning, other kinds of models that learn multiple category prototypes (e.g., Ashby & Waldron, 1999), or hybrid models that combine exemplar and prototype processes (Minda & Smith, 2001; Smith & Minda, 1998) would, in principle, be able to estimate within-category correlations. Indeed, SUSTAIN has been successful in accounting for different patterns of performance across linearly separable and nonlinearly separable category structures in inference versus classification (Love et al., 2004). Given that within-category representations can be used to mimic rule-like behavior (e.g., Hélie, Ell, Filoteo, & Maddox, 2015), this would provide a possible means by which within-category representations could support a wide range of observable behavior.

Boundary Conditions on the Learning of Within-Category Representations

The aim of Experiment 2 was to determine if disrupting the congruence between the local, diameter-angle correlations within each quadrant and the global diameter-angle correlations within each category would impair the learning of within-category representations with classification training. Participants were able to learn during classification training, albeit at lower levels of accuracy than in Experiment 1, where there was local-global congruence. Unlike

Experiment 1, however, this learning did not promote the knowledge of within-category representations that could be used to support inference during the test phase. Surprisingly, the local-global incongruence also impaired the learning of within-category representations with inference training, suggesting the learning of within-category representations may be generally sensitive to characteristics of the category structure. That being said, we cannot rule out the possibility that our approach to introducing incongruence altered some other factor that may be critical for learning within-category representations. For instance, although there is minimal overlap between the categories, the overlap occurs in different parts of the stimulus space in Experiments 1 (center of the stimulus space) and 2 (edge of the stimulus space), but it is not clear why this would make it impossible to learn the within-category correlations with inference training while preserving learning during classification training.

The results of Experiment 1 suggest that inference may be superior to classification for promoting the learning of within-category representations. In the inference condition, the training and test phases were identical with the exception of the removal of feedback during the test phase. Thus, the test phase advantage in the inference condition may reflect a practice effect. The goal of Experiment 3 was to address this issue using a two-alternative, forced-choice version of inference training that more closely matches classification training and is more similar to inference training procedures used in previous work (e.g., Yamauchi & Markman, 1998). Test phase performance in the inference and classification conditions did not significantly differ in Experiment 3 suggesting that the Experiment 1 inference advantage may reflect a practice effect. In comparison to the Experiment 1 inference task, the Experiment 3 inference task discretized the response and feedback. Although these methodological changes increased the similarity between the classification and inference conditions, it is possible that they contributed to the relatively weak test phase performance in the inference condition of Experiment 3. This is somewhat surprising given the success of two-alternative, forced choice inference tasks (Markman & Ross, 2003). The vast majority of previous work, however, has used discrete-valued dimensions with a small

number of stimuli.. It is possible that forced-choice inference works well in promoting within-category representations having discrete-valued dimensions with a small number of stimuli, but is not well suited to a category structure having continuous-valued dimensions with a large number of stimuli.

A common theme in the research on the kinds of category representations learned during training is that participants learn what is necessary to perform the task at hand (Ell et al., 2017; Hélie et al., 2017; Love, 2005; Markman & Ross, 2003; Pothos & Chater, 2002; Yamauchi & Markman, 1998). For example, with the rule-based structure used by Ell et al. (2017), successful performance during classification training did not depend upon learning the relationship between diameter and angle. Instead, participants needed only to attend selectively to a single, diagnostic stimulus dimension in order to achieve perfect classification performance. The present results suggest that selective attention to a single stimulus dimension may hinder the ability to learn the two-dimensional, within-category correlations. Attention to multiple stimulus dimensions would seem to be a necessary, but not sufficient to promote the learning of this kind of within-category representation.

That being said, it is possible to learn and generalize other types of within-category representations when learning to categorize based upon a single stimulus dimension. A seemingly minor tweak of the typical classification instructions (i.e., concept training - participants learn categories by classifying stimuli as a member/nonmember of a target category, Maddox, Bohil, & Ing, 2004; Posner & Keele, 1968; Reber, 1998; Smith & Minda, 2002; Zeithamova, Maddox, & Schnyer, 2008) shifts the emphasis from between-category differences to within-category similarities (Casale & Ashby, 2008; Hélie et al., 2017). In Hélie et al., participants learned two rule-based category structures (simultaneously) along a single diagnostic stimulus dimension (category A vs. category B and category C vs. category D). Participants were subsequently tested on a novel categorization problem using the same categories (i.e., category B vs. category C). Participants were successfully able to generalize the knowledge when receiving concept training,

but not when receiving traditional classification training, suggesting that concept training promoted a representation based on the categories themselves rather than between-category differences (see also Hoffman & Rehder, 2010; Kattner et al., 2016). Thus, it may be the case that concept training promotes a minimal within-category representation that is sufficient to support classification on a novel rule-based categorization problem (e.g., the range of values on the stimulus dimensions), but not so rich so as to include knowledge that was not required during training (e.g., the correlational structure of the categories).

Conclusions

In sum, taken together with previous work, the current results suggest that the demands of learning may be the most critical factor in promoting within-category representations. If the task requires participants to learn about the relationship between dimensions, they can learn within-category representations. Such demands can be imposed by the nature of the category structure (e.g., the exclusive-or structure used here, the information-integration structure used by Ell et al., 2017) or by the goal of the task (e.g., inference with unidimensional rule-based structures). These data also suggest important boundary conditions on the learning of within-category representations. For instance, even when learning about the relationship between stimulus dimensions, incongruency between local and global regions of the stimulus space can disrupt the learning of within-category representations. Knowledge of this limitation may be an important factor to consider when developing training regimens to promote the knowledge of within-category representations. These results complement the growing body of work highlighting the impact of category structure and task goal on category representations (Carvalho & Goldstone, 2015; Hammer, Diesendruck, Weinshall, & Hochstein, 2009; Levering & Kurtz, 2015). These results also build upon previous work by investigating the relationship between these factors and the generalization of categorical knowledge (Carvalho & Goldstone, 2014; Chin-Parker & Ross, 2002;

Hoffman & Rehder, 2010) thereby providing a window into the cognitive utility of category representations in novel situations.

Acknowledgements

This work was supported by the National Science Foundation under Grants #1349677-BCS and #1349737-BCS to SH and SWE (respectively). The authors would like to thank Dr. Chin-Parker, Dr. Church, Dr. Minda, Dr. Ragó, and two anonymous reviewers for their valuable comments and suggestions. We would also like to thank Anna Driscoll, Renee Savoie, and Elizabeth Schreiber for their help with data collection.

Open Practices Statement

The data and materials for all experiments are available upon request. The experiments were not preregistered.

References

Anderson, J. R., & Fincham, J. M. (1996). Categorization and sensitivity to correlation. *Journal of Experimental Psychology: Learning, Memory, and Cognition, 22*, 259-277.

Ashby, F. G. (1992a). Multidimensional models of categorization. In F. G. Ashby (Ed.), *Multidimensional models of perception and cognition*. Hillsdale, NJ: Erlbaum.

Ashby, F. G. (1992b). Multivariate probability distributions. In F. G. Ashby (Ed.), *Multidimensional models of perception and cognition* (pp. 1-34). Hillsdale: Lawrence Erlbaum Associates, Inc.

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., & Gott, R. E. (1988). Decision rules in the perception and categorization of multidimensional stimuli. *Journal of Experimental Psychology: Learning, Memory, and Cognition, 14*, 33-53.

Ashby, F. G., & Lee, W. W. (1993). Perceptual variability as a fundamental axiom of perceptual science. In S. C. Masin (Ed.), *Foundations of percpetual theory* (pp. 369-399). Amsterdam: Elsevier.

Ashby, F. G., & Townsend, J. T. (1986). Varieties of perceptual independence. *Psychological Review, 93*, 154-179.

Ashby, F. G., & Valentin, V. V. (2017). Multiple systems of perceptual category learning: Theory and cognitive tests. In H. Cohen & C. Lefebvre (Eds.), *Handbook of Categorization in Cognitive Science* (2 ed., pp. 157-188): Elsevier.

Ashby, F. G., & Waldron, E. M. (1999). The nature of implicit categorization. *Psychonomic Bulletin & Review, 6*, 363-378.

Ashby, F. G., Waldron, E. M., Lee, W. W., & Berkman, A. (2001). Suboptimality in human categorization and identification. *Journal of Experimental Psychology: General, 130*, 77-96.

Brainard, D. H. (1997). Psychophysics software for use with MATLAB. *Spatial Vision, 10*, 433-436.

Carvalho, P. F., & Goldstone, R. L. (2014). Putting category learning in order: Category structure and temporal arrangement affect the benefit of interleaved over blocked study. *Memory & Cognition, 42*, 481-495. doi:10.3758/s13421-013-0371-0

Carvalho, P. F., & Goldstone, R. L. (2015). The benefits of interleaved and blocked study: different tasks benefit from different schedules of study. *Psychonomic Bulletin & Review, 22*, 281-288. doi:10.3758/s13423-014-0676-4

Casale, M. B., & Ashby, F. G. (2008). A role for the perceptual representation memory system in category learning. *Perception & Psychophysics, 70*, 983-999.

Casale, M. B., Roeder, J. L., & Ashby, F. G. (2012). Analogical transfer in perceptual categorization. *Memory & Cognition, 40*, 434-449.

Chin-Parker, S., & Ross, B. H. (2002). The effect of category learning on sensitivity to within-category correlations. *Memory & Cognition, 30*, 353-362.

Chin-Parker, S., & Ross, B. H. (2004). Diagnosticity and prototypicality in category learning: a comparison of inference learning and classification learning. *Journal of Experimental Psychology: Learning, Memory, & Cognition, 30*, 216-226. doi:10.1037/0278-7393.30.1.216

Ell, S. W., & Ashby, F. G. (2012). The impact of category separation on unsupervised categorization. *Attention, Perception, & Psychophysics, 74*, 466-475.

Ell, S. W., Ing, A. D., & Maddox, W. T. (2009). Criterial noise effects on rule-based category learning: The impact of delayed feedback. *Attention, Perception, & Psychophysics, 71*, 1263-1275.

Ell, S. W., Smith, D. B., Peralta, G., & Helie, S. (2017). The impact of category structure and training methodology on learning and generalizing within-category representations. *Attention Perception Psychophysics, 79*, 1777-1794. doi:10.3758/s13414-017-1345-2

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

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

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

Hammer, R., Diesendruck, G., Weinshall, D., & Hochstein, S. (2009). The development of category learning strategies: what makes the difference? *Cognition, 112*, 105-119. doi:10.1016/j.cognition.2009.03.012

Hélie, S., Ell, S. W., Filoteo, J. V., & Maddox, W. T. (2015). Criterion learning in rule-based categorization: simulation of neural mechanism and new data. *Brain and Cognition, 95*, 19-34. doi:10.1016/j.bandc.2015.01.009

Helie, S., Shamloo, F., & Ell, S. W. (2018). The impact of training methodology and category structure on the formation of new categories from existing knowledge. *Psychological Research*. doi:10.1007/s00426-018-1115-3

Hélie, S., Shamloo, F., & Ell, S. W. (2017). The effect of training methodology on kowledge representation in perceptual categorization. *PLoS One*. doi:10.1371/journal.pone.0183904

Hoffman, A. B., & Rehder, B. (2010). The costs of supervised classification: The effect of learning task on conceptual flexibility. *Journal of Experimental Psychology: General, 139*, 319-340. doi:10.1037/a0019042

Jeffreys, H. (1961). *Theory of probability* (3 ed.). Oxford: Oxford University Press.

Kass, R. E., & Raftery, A. E. (1995). Bayes factors. *Journal of the American Statistical Association, 90*, 773-795.

Kattner, F., Cox, C. R., & Green, C. S. (2016). Transfer in rule-based category learning depends on the training task. *PLoS One, 11*, e0165260. doi:10.1371/journal.pone.0165260

Kleiner, M., Brainard, D., Pelli, D., Ingling, A., Murray, R., & Broussard, C. (2007). What's new in Psychtoolbox-3? *Perception, 36*, ECVP Abstract Supplement.

Kurtz, K. J. (2007). The divergent autoencoder (DIVA) model of category learning. *Psychonomic Bulletin & Review, 14*, 560-576. doi:10.3758/BF03196806

Kurtz, K. J., Levering, K. R., Stanton, R. D., Romero, J., & Morris, S. N. (2013). Human learning of elemental category structures: Revising the classic result of Shepard, Hovland, and Jenkins (1961). *Journal of Experimental Psychology: Learning, Memory, & Cognition, 39*, 552-572. doi:10.1037/a0029178

Levering, K. R., & Kurtz, K. J. (2015). Observation versus classification in supervised category learning. *Memory & Cognition, 43*, 266-282. doi:10.3758/s13421-014-0458-2

Love, B. C. (2005). Environment and goals jointly direct category acquisition. *Current Directions in Psychological Science, 14*, 195-199.

Love, B. C., Medin, D. L., & Gureckis, T. M. (2004). SUSTAIN: A network model of category learning. *Psychological Review, 111*, 309-332.

Maddox, W. T., & Ashby, F. G. (1993). Comparing decision bound and exemplar models of categorization. *Perception & Psychophysics, 53*, 49-70.

Maddox, W. T., Bohil, C. J., & Ing, A. D. (2004). Evidence for a procedural learning-based system in category learning. *Psychonomic Bulletin & Review, 11*, 945-952.

Maddox, W. T., Filoteo, J. V., Hejl, K. D., & Ing, A. D. (2004). Category Number Impacts Rule-Based but not Information-Integration Category Learning: Further Evidence for Dissociable Category Learning Systems. *Journal of Experimental Psychology: Learning, Memory, and Cognition, 30*, 227-235.

Markman, A. B., & Ross, B. (2003). Category use and category learning. *Psychological Bulletin, 129*, 529-613.

Minda, J. P., & Ross, B. H. (2004). Learning categories by making predictions: an investigation of indirect category learning. *Memory & Cognition, 32*, 1355-1368.

Minda, J. P., & Smith, J. D. (2001). Prototypes in category learning: The effects of category size, category structure, and stimulus complexity. *Journal of Experimental Psychology: Learning, Memory, & Cognition, 37*, 775-799.

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

Pelli, D. G. (1997). The VideoToolbox software for visual psychophysics: Transforming numbers into movies. *Spatial Vision, 10*, 437-442.

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

Pothos, E. M., & Chater, N. (2002). A simplicity principle in unsupervised human categorization. *Cognitive Science, 26*, 303-343.

Reber, P. J., Stark, C. E. L., and Squire, L. R. (1998). Cortical areas supporting category learning identified using functional MRI. *Proceedings of the National Academy of Sciences of the United States of America, 95*, 747-750.

Rouder, J. N., Speckman, P. L., Sun, D., Morey, R. D., & Iverson, G. (2009). Bayesian t tests for accepting and rejecting the null hypothesis. *Psychonomic Bulletin & Review, 16*(2), 225-237. doi:10.3758/PBR.16.2.225

Schwarz, G. (1978). Estimating the dimension of a model. *The Annals of Statistics, 6*, 461-464.

Smith, J. D., & Minda, J. P. (1998). Prototypes in the Mist: The Early Epochs of Category Learning. *Journal of Experimental Psychology: Learning, Memory, & Cognition, 24*, 1411-1436.

Smith, J. D., & Minda, J. P. (2002). Distinguishing Prototype-Based and Exemplar-Based Processes in Dot-Pattern Category Learning. *Journal of Experimental Psychology: Learning, Memory, & Cognition, 28*, 800-811.

Thomas, R. D. (1998). Learning correlations in categorization tasks using large, ill-defined categories. *Journal of Experimental Psychology: Learning, Memory, & Cognition, 24*, 119-143.

Wickens, T. D. (1982). *Models for behavior: Stochastic processes in psychology*. San Francisco: W. H. Freeman.

Yamauchi, T., & Markman, A. B. (1998). Category learning by inference and classification. *Journal of Memory and Language, 39*, 124-148.

Zeithamova, D., Maddox, W. T., & Schnyer, D. M. (2008). Dissociable prototype learning systems: evidence from brain imaging and behavior. *Journal of Neuroscience, 28*, 13194-13201. doi:10.1523/JNEUROSCI.2915-08.2008

Appendix

Model-Based Analyses

To get a more detailed description of how participants categorized the stimuli, a number of different decision bound models (Ashby, 1992a; Maddox & Ashby, 1993) were fit separately to the final block training data for each participant in the classification conditions. These data included 80 stimuli from categories A and B as well as 16 probe stimuli that were used to help differentiate between the models described below. Decision bound models are derived from general recognition theory (Ashby & Townsend, 1986), a multivariate generalization of signal detection theory (Green & Swets, 1966). It is assumed that, on each trial, the percept can be represented as a point in a multidimensional psychological space and that each participant constructs a decision bound to partition the perceptual space into response regions. The participant determines which region the percept is in, and then makes the corresponding response. While this decision strategy is deterministic, decision bound models predict probabilistic responding because of trial-by-trial perceptual and criterial noise (Ashby & Lee, 1993).

This Appendix briefly describes the decision bound models. For more details, see Ashby (1992a) or Maddox and Ashby (1993). The classification of these models as either *rule-based* or *information-integration* models is designed to reflect current theories of how these strategies are learned (e.g., Ashby, Alfonso-Reese, Turken, & Waldron, 1998) and has received considerable empirical support (see Ashby & Valentin, 2017 for a review).

Rule-Based Models

Unidimensional Classifier (UC). This model assumes that the stimulus space is partitioned into two regions by setting a criterion on one of the stimulus dimensions. Two versions of the UC were fit to the data. One version assumes that participants attended selectively to diameter and the other version assumes participants attended selectively to angle. The UC has two free

parameters, one corresponds to the decision criterion on the attended dimension and the other corresponds to the variance of internal (perceptual and criterial) noise (σ^2). A special case of the UC, the *Optimal Unidimensional Classifier,* assumes that participants use the unidimensional decision bound that maximizes accuracy. This special case has one free parameter (σ^2)

Conjunctive Classifier (CC). An alternative rule-based strategy is a conjunction rule involving separate decisions about the stimulus value on the two dimensions with the response assignment based on the outcome of these two decisions (Ashby & Gott, 1988). The CC assumes that the participant partitions the stimulus space into four regions. Based on an initial inspection of the data, two versions of the CC were fit to these data. One version assumes that individuals assigned a stimulus to category A if it was either low on diameter and low on angle or high on diameter and high on angle; otherwise the stimulus would be assigned to category B. The other version assumes that individuals assigned a stimulus to category B if it was high on diameter and low on angle or low on diameter and high on angle; otherwise the stimulus would be assigned to category B. An example of a conjunctive classifier is plotted in Figure 1 (solid black lines). The CC has three free parameters: the decision criteria on the two dimensions and a common value of σ^2 for the two dimensions.

Information-Integration Models

The Linear Classifier (LC). This model assumes that two linear decision boundaries partition the stimulus space into four regions (see Figure 1 for an example). The LC differs from the CC in that the LC does not assume decisional selective-attention (Ashby & Townsend, 1986). This produces an information-integration decision strategy because it requires linear integration of the perceived values on the stimulus dimensions prior to invoking any decision processes. An example of a linear classifier is plotted in Figure 1 (dashed black lines)The LC assumed two linear decision bounds of opposite slope (five parameters, slope and intercept of each linear bound and a common value of σ^2).

The Minimum Distance Classifier (MDC). This model assumes that there are a number of units representing a low-resolution map of the stimulus space (Ashby & Waldron, 1999; Ashby, Waldron, Lee, & Berkman, 2001; Maddox, Filoteo, Hejl, & Ing, 2004). On each trial, the participant determines which unit is closest to the perceived stimulus and produces the associated response. The version of the MDC tested here assumes four units because the category structures were generated from two multivariate normal distributions. Because the location of one of the units can be fixed, and because a uniform expansion or contraction of the space will not affect the location of the minimum-distance decision bounds, the MDC has six free parameters (five determining the location of the units and σ^2).

Random Responder Models

Equal Response Frequency (ERF). This model assumes that participants randomly assign stimuli to the two response frequencies in a manner that preserves the category base rates (i.e., 50% of the stimuli in each category). This model has no free parameters.

Biased Response Frequency (BRF). This model assumes that participants randomly assign stimuli to the two response frequencies in a manner that matches the participant's categorization response frequencies. This model has one free parameter, the proportion of stimuli in category A. Although the ERF and BRF are assumed to be consistent with guessing, these models would also likely provide the best account of participants that frequently shift to very different strategies.

Model Fitting

The model parameters were estimated using maximum likelihood (Ashby, 1992b; Wickens, 1982) and the goodness-of-fit statistic was

$$\text{BIC} = r \ln N - 2 \ln L,$$

where *N* is the sample size, *r* is the number of free parameters, and *L* is the likelihood of the model given the data (Schwarz, 1978). The BIC statistic penalizes a model for poor fit and for extra free parameters. To find the best model among a set of competitors, one simply computes a BIC value for each model, and then chooses the model with the smallest BIC.

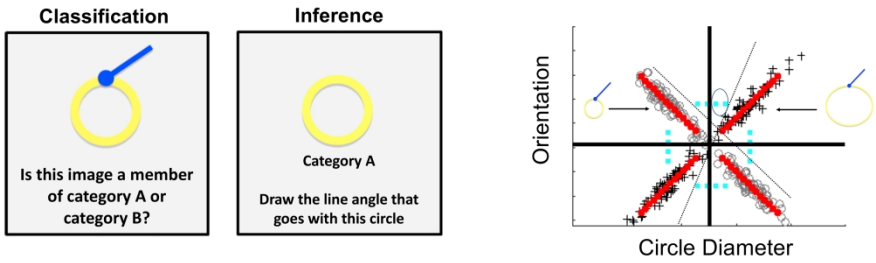

Figure 1. (Top) Example displays for the two training methodologies. (Bottom) Rule-based category structure used in Experiments 1 and 3. Category A (crosses) and B (circles) stimuli used during the training phase. The insets are example stimuli. The solid black boundaries represent the optimal conjunctive decision strategy. The dashed black boundaries represent an alternative decision strategy (see text for details). Stimuli used during the test phase are plotted as filled red circles. Probe stimuli used during the final block of training are plotted as blue squares. See text for details. (Color figure online).

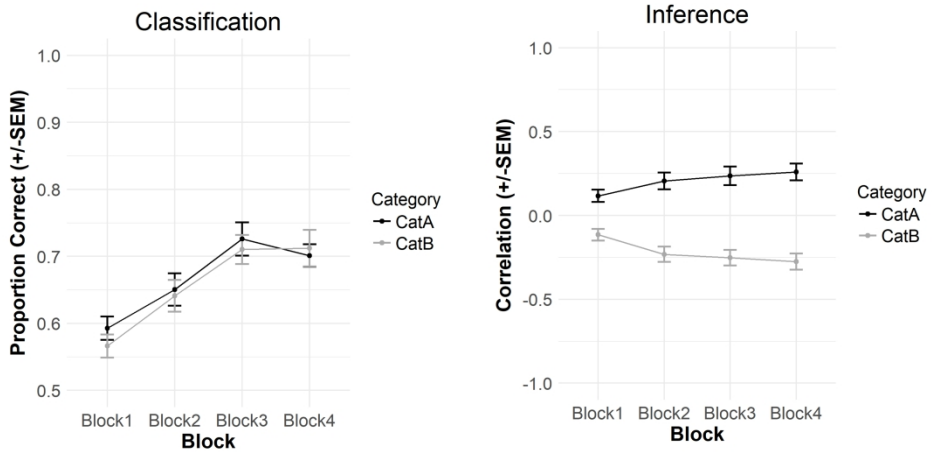

Figure 2. Training performance in the classification (proportion correct) and inference (correlation between the given and produced stimulus values) conditions of Experiment 1.

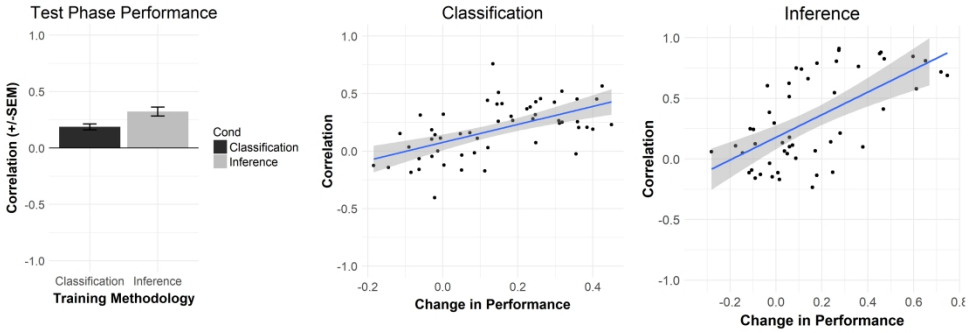

Figure 3. Performance on the inference task during the test phase (Left). Note that the diameter-angle correlations from category B are multiplied by -1 prior to averaging with the diameter-angle correlations from category A, thus positive values suggest learning of the within-category correlations. Relationship between learning during training and test phase performance in the classification (middle, r = .57) and inference (right, r = .62) conditions. The grey area represent a 95% confidence interval.

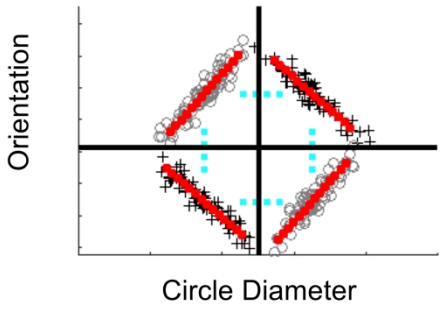

Figure 4. Conjunctive category structure used in Experiment 2. Category A (crosses) and B (circles) stimuli used during the training phase. Stimuli used during the test phase are plotted as filled red circles. Probe stimuli used during the final block of training are plotted as blue squares. (Color figure online).

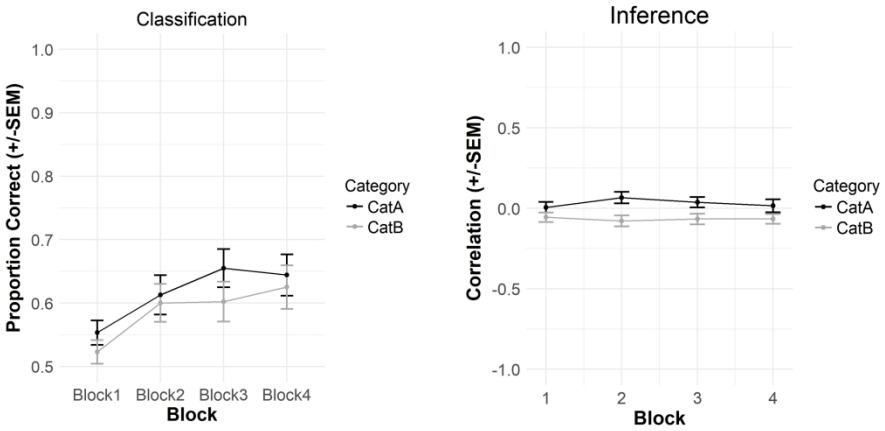

Figure 5. Training performance in the classification and inference conditions of Experiment 2.

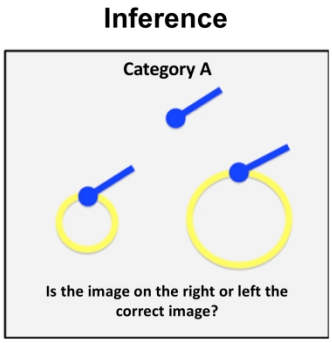

Figure 6. (Example display for the inference training methodology. (Color figure online).

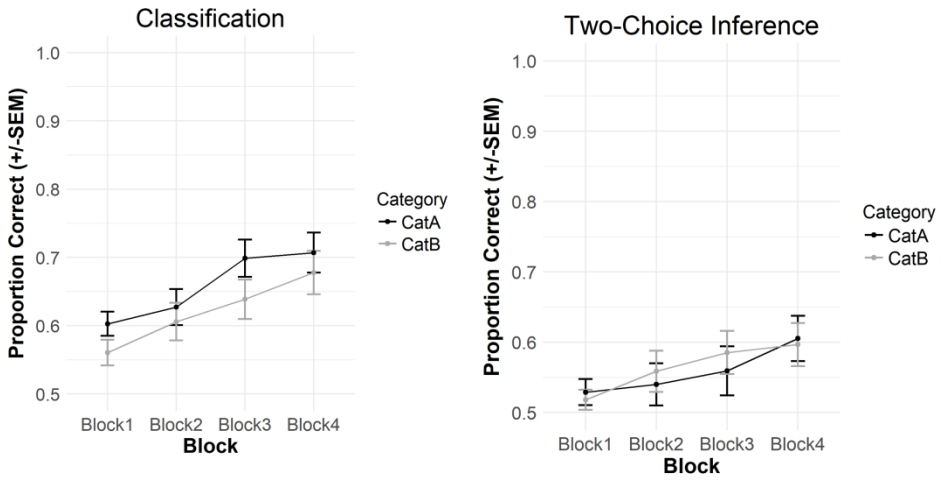

Figure 7. Training performance in the Classification and (forced-choice) Inference conditions of Experiment 3.

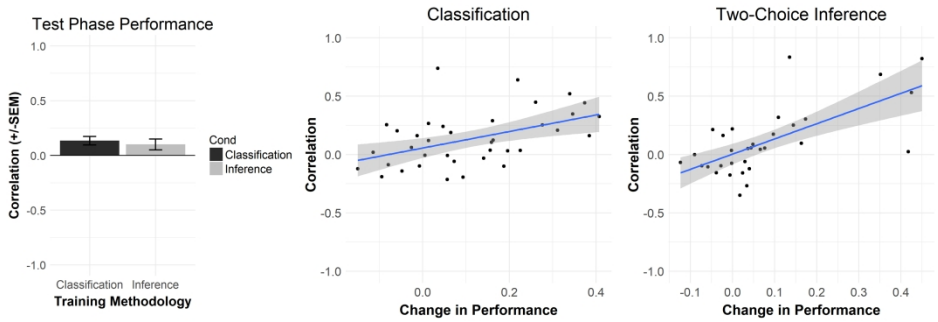

Figure 8. Performance on the inference task during the test phase (Left). Note that positive values suggest learning of the within-category correlations. Relationship between learning during training and test phase performance in the classification (r = .47) (Middle) and inference (r = .67) conditions. Note that the diameter-angle correlations from category B are multiplied by -1 prior to averaging with the diameter-angle correlations from category A. The grey bars represent a 95% confidence interval.