4-16-2012

Unsupervised Category Learning with Integral-Dimension Stimuli

Shawn W. Ell
University of Maine, shawn.ell@maine.edu

Gregory F. Ashby
University of California, Santa Barbara, ashby@psych.ucsb.edu

Steven B. Hutchinson
University of Maine, steven.hutchinson@umit.maine.edu

Quarterly Journal of Experimental Psychology, 65(8), 1537-1562. doi: 10.1080/17470218.2012.658821

Unsupervised Category Learning with Integral-Dimension Stimuli

Shawn W. Ell[1], Gregory F. Ashby[2], Steven Hutchinson[1]

[1]Psychology Department, University of Maine, Orono, Maine
[2]University of California, Santa Barbara, California

Abstract Despite the recent surge in research on unsupervised category learning, the majority of studies have focused on unconstrained tasks in which no instructions are provided about the underlying category structure. Relatively little research has focused on constrained tasks in which the goal is to learn pre-defined stimulus clusters in the absence of feedback. The few studies that have addressed this issue have focused almost exclusively on stimuli for which it is relatively easy to attend selectively to the component dimensions (i.e., separable dimensions). In the present study, we investigated the ability of participants to learn categories constructed from stimuli for which it is difficult, if not impossible, to attend selectively to the component dimensions (i.e., integral dimensions). The experiments demonstrate that individuals are capable of learning categories constructed from the integral dimensions of brightness and saturation, but this ability is generally limited to category structures requiring selective attention to brightness. As might be expected with integral dimensions, participants were often able to integrate brightness and saturation information in the absence of feedback – an ability not observed in previous studies with separable dimensions. Even so, there was a bias to weight brightness more heavily than saturation in the categorization process, suggesting a weak form of selective attention to brightness. These data present an important challenge for the development of models of unsupervised category learning.

Introduction

There has been a recent surge in research on unsupervised category learning – i.e., the ability to learn categories in the absence of corrective feedback. Studies focusing on unsupervised learning provide an important complement to the studies of supervised learning that have dominated the field as, arguably, much everyday learning occurs in the absence of trial-by-trial feedback. Given the ubiquity of unsupervised category learning, it is not that surprising that individuals can spontaneously construct categories in the absence of feedback (e.g., Medin, Wattenmaker, & Hampson, 1987).

Much of the research on unsupervised category learning, however, has focused on unconstrained tasks where participants have no knowledge that there is an optimal categorization strategy, if one exists at all (Ahn & Medin, 1992; Billman & Knutson, 1996; Clapper & Bower, 1994; Colreavy & Lewandowsky, 2008; Diaz & Ross, 2006; Handel & Imai, 1972; Love, 2002; Medin, et al., 1987; Milton, Longmore, & Wills, 2008; Milton & Wills, 2004; Pothos & Chater, 2005; Pothos & Close, 2008; Regehr & Brooks, 1995). In unconstrained tasks, the primary focus is on how participants prefer to construct categories. For instance, in the typical free sorting task, participants are presented with a number of stimuli (either simultaneously or sequentially) and asked to place the stimuli into a number of categories in any way they like. The participants are not informed that there is an underlying category structure (if one exists). Therefore, using a common performance measure such as accuracy is problematic because there is no objectively correct response.

Although unconstrained tasks have been important for understanding how characteristics of the stimuli and task influence categorization strategy, it is also important to investigate unsupervised category learning in more constrained tasks in which participants are attempting to learn the optimal categorization strategy (Ashby, Queller, & Berretty, 1999; Zeithamova & Maddox, 2009)[1]. In constrained tasks, the primary focus is on what types of category structures individuals are capable of learning. With the exception of feedback, the methodology in constrained tasks closely parallels most supervised category-learning paradigms as participants know that their goal is to learn an underlying category structure. Therefore, accuracy is an appropriate performance measure because there is an objectively correct response.

To our knowledge, all constrained tasks and the majority of unconstrained tasks have used stimuli for which it is relatively easy to attend selectively to the component dimensions (i.e., separable dimensions). Two dimensions are said to be separable if it is possible to attend to one dimension and ignore the other (e.g., hue and shape - Garner, 1974; Imai & Garner, 1965). Conversely, two dimensions are said to be integral if it is impossible to attend to one and ignore irrelevant variations in the other (e.g., brightness and saturation - Garner & Felfoldy, 1970; Torgerson, 1958)[2].

Under supervised conditions, participants can readily learn categories constructed from integral dimensions (Grau & Kemler-Nelson, 1988; McKinley & Nosofsky, 1996; Mounts & Melara, 1995; Nosofsky & Palmeri, 1996; Shepard & Chang, 1963). There is, however, an extensive literature documenting differences in the processing of separable and integral dimensions (Foard & Kemler-Nelson, 1984; Lockhead, 1972). In the context of supervised category learning, selective attention mechanisms operate less efficiently when learning categories constructed from integral, rather than separable, dimensions (Maddox, 2001; Maddox & Dodd, 2003; Nosofsky, 1986, 1987).

A number of studies using unconstrained tasks have shown that the preferred decision strategy varies as a function of whether the stimuli are constructed from separable or integral dimensions (Handel & Imai, 1972; Handel, Imai, & Spottswood, 1980; Imai & Garner, 1965)[3]. For example, Handel and colleagues (1972; Handel, et al., 1980) compared the separable dimensions of shape and color with the integral dimensions of brightness and saturation. Separable dimension stimuli were sorted using a one-dimensional strategy whereas integral dimension stimuli were sorted using a similarity-based strategy. A similar bias to use one-dimensional strategies with separable

[1] See a recent paper by Pothos and colleagues (Pothos, Edwards, & Perlman, in press) for a related distinction between constrained and unconstrained unsupervised category learning tasks.

[2] More specifically, the observation of 1) a Euclidean metric in multidimensional scaling, 2) interference when the stimuli vary orthogonally, and 3) a redundancy gain when the stimuli are correlated in the speeded classification paradigm are often considered as evidence for dimensional integrality (Garner, 1974), but see Ashby and Maddox (1994).

[3] Color naming tasks are a special case of the free sorting paradigm and have frequently been used to assess people's ability to identify color categories in a variety of color spaces (Boynton & Olson, 1987; Sturges & Whitfield, 1995). Data from this paradigm indicate that participants are quite capable of sorting color stimuli into various categories without feedback. While these studies have explored how variations along the color dimensions affect the preferred classifications of participants they have been primarily interested in variations along hue and do not provide strong predictions for stimuli varying along the integral dimensions of brightness and saturation.

dimensions has been reported with other unconstrained tasks (e.g., Colreavy & Lewandowsky, 2008; Medin, et al., 1987). Importantly, however, research from unconstrained tasks suggests that the bias to use one-dimensional strategies is critically dependent upon the particular category structures (Ahn & Medin, 1992; Pothos & Chater, 2005; Pothos & Close, 2008), spatial configuration of the stimuli (Milton & Wills, 2004), and experimental procedure (Milton, et al., 2008; Regehr & Brooks, 1995). For example, simply informing participants of the number of categories has been argued to instill a one-dimensional bias (e.g., Murphy, 2002).

Although the separable-integral distinction is often described as being discrete, such a characterization is likely to be an oversimplification (Ashby & Townsend, 1986). Studies using constrained tasks with stimuli that are strongly separable have consistently demonstrated a bias to use one-dimensional strategies (Ashby, et al., 1999; Zeithamova & Maddox, 2009). There are at least two studies using constrained tasks with stimuli that fall somewhere in the middle of the separable-integral continuum. (10 x 10 grids of randomly distributed light and dark squares - Fried & Holyoak, 1984 ; lines connecting nine randomly located dots - Homa & Cultice, 1984). Although learning was evident in both studies, there are several limitations with respect to the question of what individuals are capable of learning under unsupervised conditions on constrained tasks. For instance, the stimuli in the Fried and Holyoak (1984) study varied on 100 physical dimensions while the Homa and Cultice (1984) stimuli varied along 18 physical dimensions. The dimensionality of the psychological representation of these stimuli is not known and it is likely that there is no straightforward mapping between the psychological and physical dimensions (Shin & Nosofsky, 1992). Without knowing the psychological representation of the stimuli it is impossible to obtain an accurate estimate of the decision strategy participants were using to perform the task.

To summarize, with separable stimulus dimensions, unsupervised category learning is possible and, in some cases, there is a bias to use one-dimensional strategies. With integral dimensions, the picture is more complicated. On unconstrained tasks, individuals do not demonstrate a strong preference for one-dimensional strategies. On constrained tasks using stimuli that likely have some degree of integrality, unsupervised category learning is possible. We know, however, very little about what types of strategies individuals are capable of learning under unsupervised conditions when the categories are constructed from integral dimensions and whether the bias to use one-dimensional strategies that has been demonstrated on constrained tasks with separable dimensions extends to integral dimensions.

We investigate these questions in the present experiment using a constrained task with stimuli constructed from the integral dimensions of brightness and saturation defined in the Munsell color system (Figure 1). The structure of the Munsell color system is such that variations along the value dimension correspond to changes in brightness whereas differences in chroma reflect changes in saturation (Munsell, 1915). For simplicity, the physical dimensions of value and chroma will be referred to by the perceptual labels of brightness and saturation, respectively. Two one-dimensional (Vertical and Horizontal conditions in the top panels) and two diagonal conditions (Positive and Negative conditions in the bottom panels) were constructed by randomly sampling from a bivariate uniform distribution defined on the brightness and saturation dimensions. Each of the category structures differ only in the orientation of the optimal decision strategy. As the name implies, to learn the one-dimensional structures participants should attend to the relevant stimulus dimension (while ignoring the other, irrelevant dimension). To learn the diagonal structures participants should integrate information from the brightness and saturation dimensions.

The unsupervised category learning literature makes conflicting predictions for the Figure 1 category structures. Data from constrained tasks with separable dimensions would predict a bias to use one-dimensional decision strategies regardless of the task. As a result, participants should be able to learn the one-dimensional categories, but have difficulty with the diagonal categories. Alternatively, data from unconstrained tasks with integral dimensions would not predict a bias to use one-dimensional decision strategies. For example, similarity-based strategies may be preferred (e.g., Handel & Imai, 1972). This would predict similar performance across the one-dimensional and diagonal category structures because similarity is generally invariant to rotation (Shepard, 1964).

Experiment 1

Method

Participants and Design. Forty participants were recruited from the University of California, Santa Barbara student community and received partial course credit for participation. Ten participants were randomly assigned to each of four experimental conditions: Vertical, Horizontal, Positive, and Negative. No one participated in more than one experimental condition. All participants had normal (20/20) or corrected to normal vision and normal color vision. Each participant completed two sessions of approximately 45 minutes that were separated, on average, by 24 hrs. Participants in any condition who were more than three SD away from the average accuracy in that condition during the second day of training were omitted from all subsequent analyses. This criterion for the detection of outliers resulted in the omission of one participant from each of the Vertical and Negative conditions.

Stimuli and Apparatus

The stimuli in all experiments were Munsell color patches (Munsell, 1915; Newhall, 1940; Newhall, Nickerson, & Judd, 1943) of constant purple-blue hue (10 PB) that varied continuously along the dimensions of value (i.e., brightness) and chroma (i.e., saturation). The complete set of stimuli used in the four different experimental conditions is shown in Figure 1. Each symbol in Figure 1 denotes the value and chroma of a single color patch. Category A stimuli are denoted by the black “+” signs and category B stimuli are denoted by the gray circles. The optimal decision criteria are the vertical and diagonal lines shown in Figure 1.

The experiment used a variation of the randomization technique introduced by Ashby and Gott (1988) in which each category was defined as a bivariate uniform distribution. Each category distribution was specified by the minimum and maximum on each dimension. The exact parameter values for the Vertical condition are displayed in Table 1.On each trial, a random sample (x, y) was drawn from the category A or B distribution and these values were used to construct a Munsell color patch of value $x' = .0275x + 4.1$ and chroma $y' = .055y + 4.3$. While one of the goals of the Munsell system was to equate the perceived difference between equal steps along the value and chroma dimensions, in practice the perceived difference between two steps on the chroma dimension is approximately perceptually equal to a single step on the value dimension (Newhall, 1940; Newhall, et al., 1943; Nickerson, 1940). The choice of scale values in the above transformations was designed to preserve the 2:1 relationship between the dimensions. For each participant, a new sample of 720 stimuli

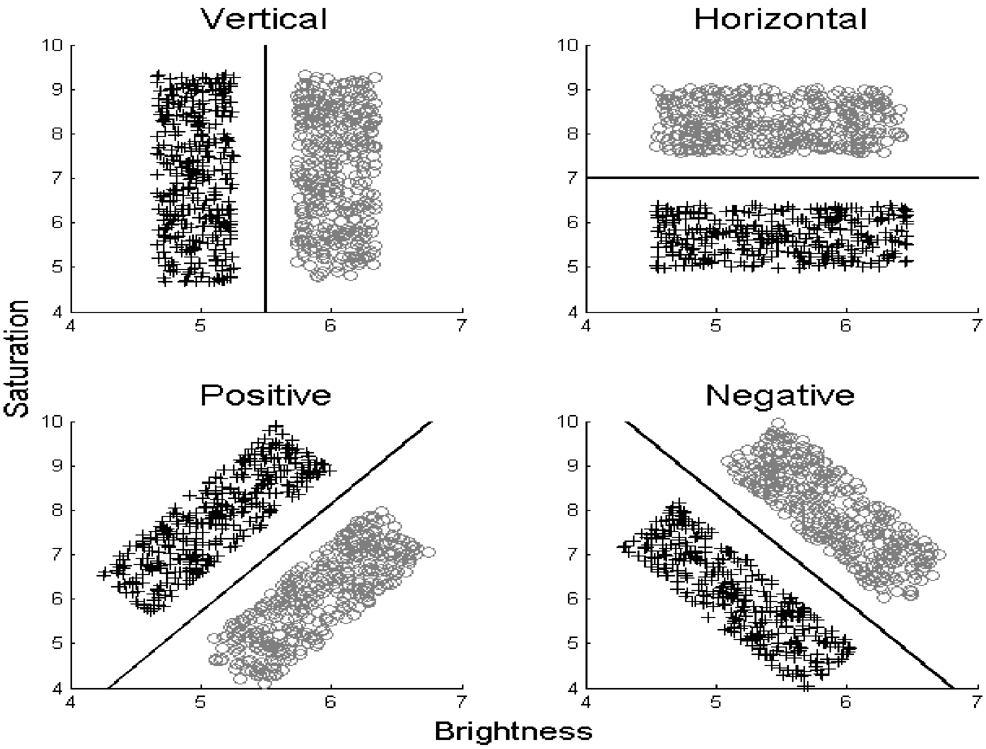

Figure 1. Scatterplots of the stimuli used in the present Experiments 1 and 2. Each point represents a rectangular, iso-hue color patch that varied continuously on Munsell value and chroma. Category A and B exemplars are depicted as black plus signs ('+') and gray circles ('o'), respectively. The solid lines are examples of decision strategies that maximize accuracy (i.e., the optimal decision strategies).

Table 1. Initial Parameters Used to Generate the Vertical Category Structure Before Transforming to the Munsell Color Space.

	Value		Chroma	
	Min	Max	Min	Max
Category A	20	40	7.5	92.5
Category B	60	80	7.5	92.5

(360 from each category) were generated. All stimuli were generated offline and a linear transformation was applied to ensure that the sample statistics matched the population parameters.

For the Horizontal, Positive, and Negative conditions, the stimuli for each participant were created by first generating a random sample of 720 stimuli (360 from each category) from the Table 1 distributions and then rotating the resulting stimuli by 90°, -45°, or 45° from vertical (respectively) about the center of the stimulus space (i.e., the point 50, 50). For the Positive and Negative category structures, the most accurate one-dimensional rule (i.e., respond A if the stimulus value is less than some criterion, otherwise respond B) yields an accuracy of approximately 85 percent correct. The presentation order of the stimuli was randomized separately for each participant in every condition and divided into nine blocks of 80 trials each.

Color monitor calibration was achieved with a PhotoResearch PR-650 spectral radiometer and the Psychophysics Toolbox software (Brainard, 1997; Pelli, 1997). The transformation from the Munsell color space to RGB values was performed in three stages. First, the value and chroma coordinates were transformed to CIE xyY chromaticities using a color lookup table obtained from the Center for Imaging Science at the Rochester Institute of Technology (http://www.cis.rit.edu/mcsl/online/munsell.php). Those value and chroma coordinates not given in the table were converted to CIE xyY chromaticities using equations given in (Wyszecki & Stiles, 1982). Second, xyY coordinates were converted to CIE XYZ tristimulus coordinates. Finally, XYZ coordinates were converted to RGB coordinates using the Psychophysics Toolbox software. The experiment was run using the Psychophysics toolbox in the Matlab computing environment. Each color patch was presented on a gray background, subtended a visual angle of approximately 6 degrees, and was displayed on a 15-inch CRT with 832 x 624 pixel resolution in a dimly lit room.

Procedure

Each participant was run individually. Participants were told that rectangular color patches varying in brightness and saturation would be presented one at a time on a monitor and their task was to learn to categorize the stimuli into two categories[4]. Following (Ashby, et al., 1999), five response blocks (blocks 1, 3, 5, 7, and 9) alternated with four observation-only blocks (blocks 2, 4, 6, and 8). During the observation-only blocks, participants were instructed to look at 80 sequentially presented stimuli and to try and learn about the categories. The stimuli in the observation-only blocks were presented for 1 s with an inter-stimulus interval of 0.5 s. The observation-only blocks were included in an effort to increase the number of stimuli that the participants were exposed to during an experimental session. The observation-only blocks do not require a response and, thus, take less time to complete than the response blocks (Ashby, et al., 1999). During the response blocks participants were instructed to select a category for each stimulus and to press a button labeled “A” or a button labeled “B” to show which category had been selected. The participants were told that the category labels were arbitrary, but were warned to be consistent with what they called a member of category A and what they called a member of category B. Given that the category labels were arbitrary, it was assumed that participants assigned the stimuli to the two categories in a manner that resulted in the highest accuracy (percent correct) for each block. Therefore, it was impossible for participants to achieve accuracy below 50% correct in any given block. The participants were told that perfect accuracy was possible, but were never given any feedback about their performance. The stimuli were response terminated (with 5 s maximum exposure duration) in the response blocks and the response-stimulus interval was 0.5 s. The break between blocks was participant paced.

Results

Accuracy-based analyses

The average learning curves for each of the four experimental conditions are shown in Figure 2. Visual inspection of Figure 2 suggests that accuracy improved across the two days of training only in the Vertical and Positive conditions and accuracy was highest in the Vertical condition. A 4 condition × 10 response-block mixed ANOVA (with block as the within-subjects factor) conducted on the accuracy data revealed significant main effects of condition [$F(3, 34) = 5.46$, $p < .01$, $MSE =$

[4] All participants indicated that they understood what brightness was, but several were unfamiliar with saturation. Thus, for all participants, saturation was described as the amount of white in a color patch with low levels of saturation indicating a large amount of white in the color patch. As an example, participants were told that pink is a desaturated red. This explanation was effective (according to verbal report) in eliminating any confusion regarding the saturation dimension.

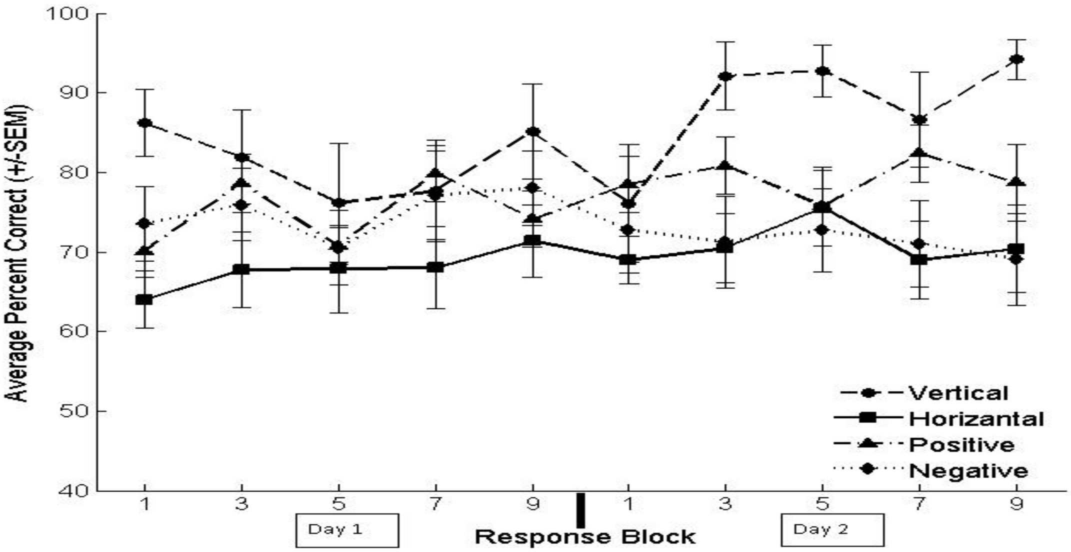

Figure 2. Average accuracy in the four conditions of Experiment 1.

.35, $\eta_p^2 = .33$] and block [F(5.88, 199.95) 2.22, p < .05, MSE = .06, $\eta_p^2 = .06$][5]. However the condition × block interaction was not significant [F(17.65, 199.95) = 1.41, p = .13, MSE =.06, $\eta_p^2 = .11$]. Further analysis of the main effects revealed that the effect of condition was driven solely by superior performance in the Vertical condition relative to the Horizontal, Positive, and Negative conditions (*p's* < .05). None of the other pairwise comparisons were significant (*p's* ≥.39). The effect of block was driven by an increasing linear trend in accuracy across conditions as evidenced by a significant linear contrast [F(1, 34) = 5.75, p < .05, MSE = .09, $\eta_p^2 = .15$]. In sum, at the group level, a high level of performance was observed only in the Vertical condition – the category structure that required participants to attend selectively to brightness while ignoring variations in saturation. The superior performance in the Vertical condition can be seen at the individual participant level as well. Table 2 lists the individual average accuracy rates by block for each participant in each condition during the second day of training. All but one participant in the Vertical condition was near optimal (> 90%) during the final block of training. In contrast, only eight participants in the remaining conditions (three in the Horizontal, four in the Positive, and one in the Negative) achieved a similar level of accuracy. Furthermore, two thirds of the participants from the Vertical condition maintained an accuracy level > 90% during the last four response blocks whereas only two participants in the other conditions (one in the Horizontal and one in the Positive) performed at this high level. Interestingly, the individual participant data suggest that while it was certainly more

[5] To meet the assumptions of ANOVA these data were first subjected to an arcsine transformation. For descriptive purposes, the non-transformed data were presented in all figures and tables. A Huynh-Feldt correction for violation of the sphericity assumption has been applied. All subsequent analyses of accuracy rates used the same transformation and correction. Post hoc comparisons were evaluated using the Student-Newman-Keuls procedure.

Table 2. Individual Participant Accuracy During the Second Day of Training in Experiment 1.

		Participant											
	Block	1	2	3	4	5	6	7	8	9	10	Avg	SEM
Vertical	1	99	58	76	55	53	51	100	95	99		76.1	7.0
	3	99	61	83	94	100	100	100	95	98		92.1	4.1
	5	96	88	76	78	99	100	100	100	99		92.8	3.1
	7	99	71	53	68	99	98	98	96	100		86.6	5.7
	9	99	95	76	91	100	100	95	91	100		94.2	2.4
Horizontal	1	55	68	76	68	62	63	81	70	63	85	69.0	2.8
	3	73	59	78	73	50	54	93	79	63	85	70.5	4.2
	5	70	51	89	81	66	64	98	84	63	90	75.5	4.5
	7	58	62	60	84	58	60	96	74	51	88	69.0	4.6
	9	51	66	90	55	60	59	95	84	54	90	70.4	5.2
Positive	1	74	98	76	80	68	70	69	80	99	73	78.5	3.5
	3	78	100	84	93	71	90	60	78	79	78	80.8	3.6
	5	98	99	88	85	71	56	63	58	74	68	75.8	8.0
	7	98	100	88	88	79	79	68	66	85	75	82.4	3.6
	9	94	100	91	63	60	94	64	74	83	66	78.7	4.8
Negative	1	76	85	51	78	79	51	93	55	88		72.7	4.8
	3	91	51	66	53	70	64	93	59	95		71.3	5.2
	5	85	71	55	89	83	51	81	53	88		72.8	4.7
	7	94	74	70	62	62	51	85	50	91		71.1	4.9
	9	80	81	51	51	67	55	85	54	98		69.1	5.2

difficult for participants to improve with training in the Horizontal, Positive, and Negative conditions in general, it was not altogether impossible.

In an ideal learning trajectory, accuracy might steadily improve across trials and peak at the completion of training. However, it is possible that participants may have peaked during some training block other than the last. In fact this was true for 29 of the 38 participants. Analyzing the accuracy in this way did not change the ordering: Vertical (*M* = 97.6, *SD* = 4.1), Positive (*M* = 91.8, *SD* = 8.5), Negative (*M* = 88.2, *SD* = 10.8), and Horizontal (*M* = 80.9, *SD* = 13.1). A one-way ANOVA using each participant's best block generally supported this conclusion. The effect of condition was significant [*F*(3, 34) = 5.88, $p < .01$, *MSE* = .06, $\eta_p^2 = .34$] with accuracy in the Vertical condition being significantly higher than the Horizontal and Negative conditions (*p's* < .05), but only marginally higher than in the Positive condition (*p* = .06). This lack of a significant difference between the best block for the Vertical and Positive conditions may reflect a ceiling effect in the Vertical condition.

Model-Based Analyses

Analysis of the accuracy data does not directly address the question of what decision strategies were used to perform the Figure 1 tasks. For instance, accuracy for many of the participants in the Positive and Negative conditions was consistent with both a one-dimensional strategy and a strategy that integrated brightness and saturation (albeit in a suboptimal manner). The following analyses represent a quantitative approach to investigating these questions.

Three different types of decision bound models were fit to the data of each individual participant, each based on a different assumption concerning the participant's strategy. First, the unidimensional classifiers assume that the participant attends selectively to one dimension (e.g., if the stimulus is bright, respond B; otherwise respond A). For the Vertical and Horizontal conditions, there were two versions of the unidimensional classifier (UC), one assuming participants used the optimal decision strategy in two top panels of Figure 1 (optimal classifier, OC) and one assuming participants used a UC with a suboptimal intercept on one of the dimensions (UC-brightness and UC-saturation). Second, the conjunctive classifier (CC) assumes that participants make independent decisions about the stimulus on both dimensions (e.g., if the stimulus is bright and saturated respond B; otherwise respond A). Third, the linear classifier assumes that participants integrate the stimulus information from both dimensions prior to making a categorization decision. For the Positive and Negative conditions, there were two versions of the linear classifier, one assuming participants used the optimal decision strategy in Figure 1 (optimal classifier, OC) and one assuming participants used a linear classifier with a suboptimal slope and/or intercept (LC).

Each of these models was fit separately to the data from every response block for all participants 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). The data from the first day of training are omitted for brevity.

The primary goal of this analysis was to investigate whether one-dimensional decision strategies dominated under unsupervised training. The distribution of best-fitting models in each of the four conditions is listed in Table 3. First consider the one-dimensional conditions. In these conditions, both the optimal and unidimensional classifiers assumed participants attended selectively to the relevant stimulus dimension. In the Vertical condition, 73% were using decision strategies consistent with selective attention to brightness. In contrast, in the Horizontal condition, only 8% were using decision strategies consistent with selective attention to saturation. Instead, participants were either attending selectively to brightness or integrating brightness and saturation. Instead, participants were either attending selectively to brightness or integrating brightness and saturation information to some extent. Given the relatively low accuracy in the Horizontal condition, it is important to verify that the linear classifier was not simply better at fitting noisy data. The high percentage of responses accounted for argue strongly against this possibility. Thus, the deficit observed in the Horizontal condition was, at least in part, driven by an inability to attend selectively to saturation in the absence of feedback. Although, this explanation does not account for the fact that average accuracy during blocks where participants were attending

Table 3. Percentage of Blocks Best Accounted for by each Model Across the Four Conditions of Experiment 1

	Models						
Condition	OC	UC-B	UC-S	CC	LC	Avg. %RA	SD
Vertical	46.7	26.7	8.9	0	17.8	95.1	7.3
Horizontal	4.0	26.0	4.0	10.0	56.0	88.2	10.9
Positive	14.0	18.0	16.0	12.0	40.0	91.3	8.2
Negative	11.1	20.0	2.2	2.2	64.4	86.9	9.5

Note. OC - optimal classifier, UC-B – one-dimensional classifier on the brightness dimension, UNI-S - one-unidimensional classifier on the saturation dimension, CC - conjunctive classifier, LC - linear classifier, %RA - percent of responses accounted for by the best-fitting model. In all conditions, the OC is a special case of one of the other models in which it is assumed that the participant used the optimal decision strategies plotted in Figure 1 (Vertical: the OC is a special case of the UC-B, Horizontal: the OC is a special case of the UC-S; Positive and Negative: the OC is a special case of the LC).

selectively to saturation was far less than optimal (*N* = 4, M = 71.3, *SD* = 17.8). Instead, the accuracy deficit during these blocks was due to the use of suboptimal decision criteria.

Next, consider the diagonal conditions. As expected from the accuracy data, participants in both the Positive and Negative conditions were responding optimally in a relatively small number of blocks. Participants were not constrained to use one-dimensional decision strategies, as evidenced by the relatively small percentage of data sets accounted for by unidimensional models (34% and 22% of the data in the Positive and Negative conditions, respectively). Instead, as was the case in the Horizontal condition, in the majority of blocks, participants were integrating brightness and saturation information, albeit not optimally.

The fact that the linear classifier accounted for a large percentage of the responses in the Horizontal, Positive, and Negative conditions may still be consistent with the use of one-dimensional decision strategies. In some cases, the best-fitting linear classifier may deviate only slightly from one-dimensional (e.g., a decision bound rotated 10° from horizontal). Such small deviations may suggest a weak form of selective attention to one of the stimulus dimensions. Plotted in Figure 3 is the distribution of slopes (in degrees) for those data sets that were best accounted for by the linear classifier in the four conditions. There were a substantial number of strategies that would be consistent with a weak form of selective attention to brightness (i.e., near 90°) in the Vertical, Horizontal, and Negative conditions, but not in the Positive condition. In addition, in the Horizontal condition, the best-fitting linear bounds were highly variable and deviated substantially from optimal. In the Positive condition, the linear bounds were also highly variable with the majority being positively sloped.

Figure 3 also suggests that when participants were not attending selectively to brightness or saturation, or using the optimal decision strategy, they used decision strategies between 0° and 90°. This preference suggests that there may be some salient decision rule that is a consequence of the integral stimulus dimensions. Indeed, inspection of the entire stimulus space (i.e., plotting all stimuli

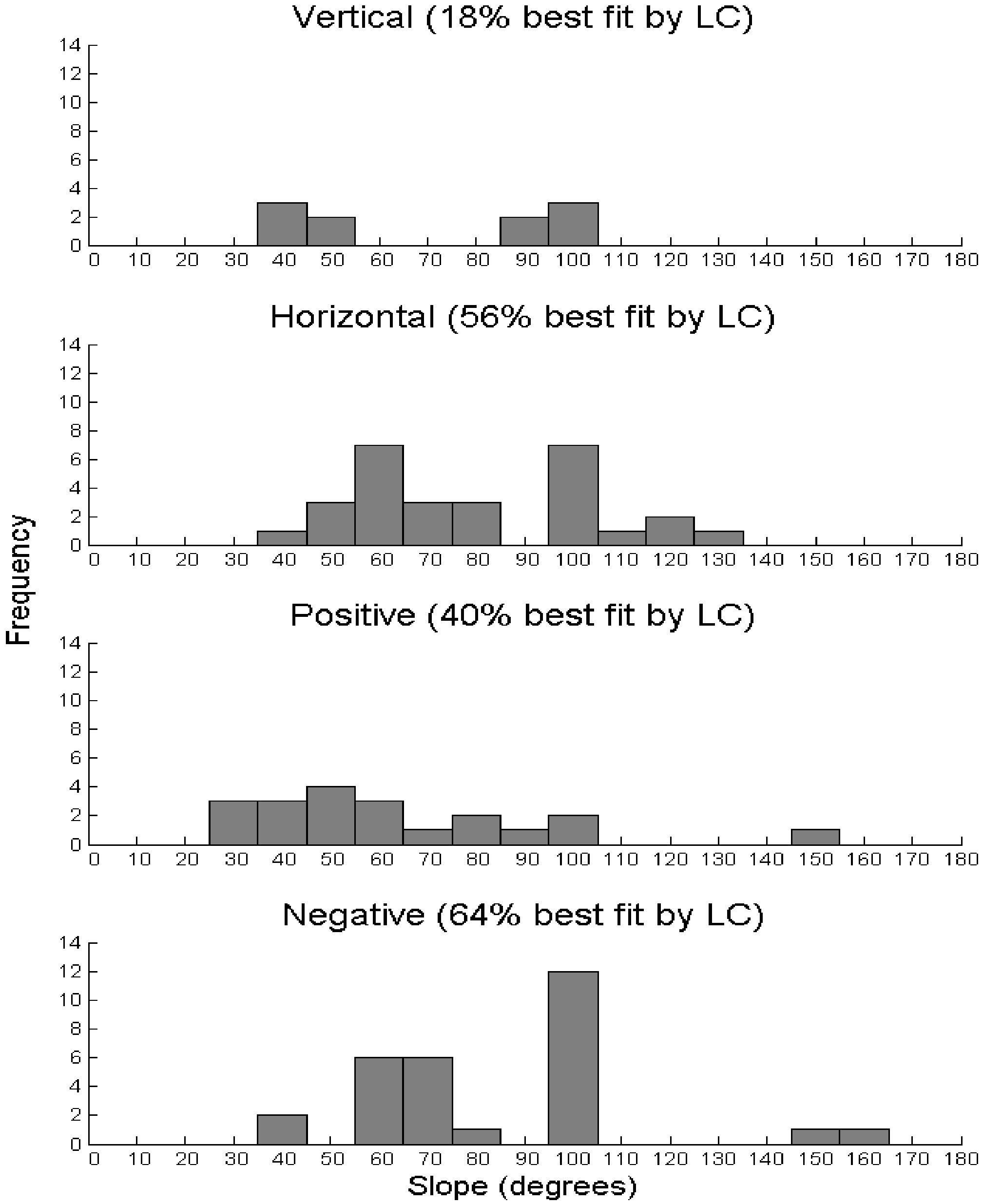

Figure 3. Distribution of the slope (in degrees) for cases where the linear classifier provided the best account of the data. The percentage of blocks for which the linear classifier provided the best fit is provided in the title for reference. For descriptive purposes, it was assumed that the slopes ranged from 0° to 180°. The optimal classifier predicts the following slopes: Vertical - 90°, Horizontal - 0°/180°, Positive - 68°, Negative - 113°. Bin-width = 10°.

simultaneously) suggests a strategy that could be best described as a “grayness” rule. More specifically, the transition from dim, saturated stimuli to bright, de-saturated stimuli could produce such a rule with a slope between 0° and 90°. If participants were truly using such a strategy, it is clear from Figure 3 that its application was highly variable across participants. Consistent with these data, inspection of written descriptions of decision strategies (collected post-experimentally) did not reveal any systematic strategy use or any mention of the word “gray”.

Recall that the decision bound models not only provide estimates of the best-fitting decision bounds, but estimates of the combined variance of criterial and perceptual noise (σ^2). A comparison of σ^2 estimates across conditions suggests that noise was lower in the Vertical condition (*Mdn* = 0.26) as compared to the Horizontal (*Mdn* = 2.9), Positive (*Mdn* = .91), and Negative (*Mdn* = 1.27) conditions. This observation was supported by an analysis of the estimates of σ^2 between all conditions using six separate t-tests (Welch’s t-test, and not ANOVA, was used due to severe violations in homogeneity of variance). All pairwise comparisons were significant (p < .0085, following Sidak correction for multiple comparisons).

Brief Summary

At least two conclusions can be drawn regarding unsupervised learning of categories constructed from the integral dimensions of brightness and saturation. First, participants demonstrate a relatively limited ability to learn such categories. Only when the category structures required selective attention to brightness were participants able to learn without feedback. Second, these data suggest that participants do not show either the preference for, or a general ability to learn, one-dimensional decision strategies under unsupervised conditions.

Experiment 2

Given the lack of learning in the Horizontal, Positive, and Negative conditions under unsupervised conditions, it is necessary to demonstrate that participants can, in fact, learn these category structures. The goal of Experiment 2 was to test the ability of participants to learn the Figure 1 categories under supervised conditions. Furthermore, the majority of research on the ability of people to learn categories constructed from integral dimension stimuli has been limited to stimuli constructed from discrete- or binary-valued rather than continuous-valued dimensions. Thus, an added contribution of Experiment 2 is that it extends previous research to continuous-valued dimensions.

Method

Participants and Design

Twenty-two participants were recruited from the University of California, Santa Barbara student community and received partial course credit for participation. The participants were randomly assigned to the four experimental conditions in the following manner: Vertical-6, Horizontal-5, Positive-6, and Negative-5. No one participated in more than one experimental condition. All participants had normal (20/20) or corrected to normal vision and normal color vision. Participants completed one session of approximately 45 minutes.

Stimuli and Apparatus

Identical to Experiment 1.

Procedure

The procedure was identical to Experiment 1 with the following exceptions. All participants were presented with nine response blocks comprising 80 trials each. Each stimulus was presented for 1s followed by a brief (0.5 s) high-pitched tone (500 Hz) if the

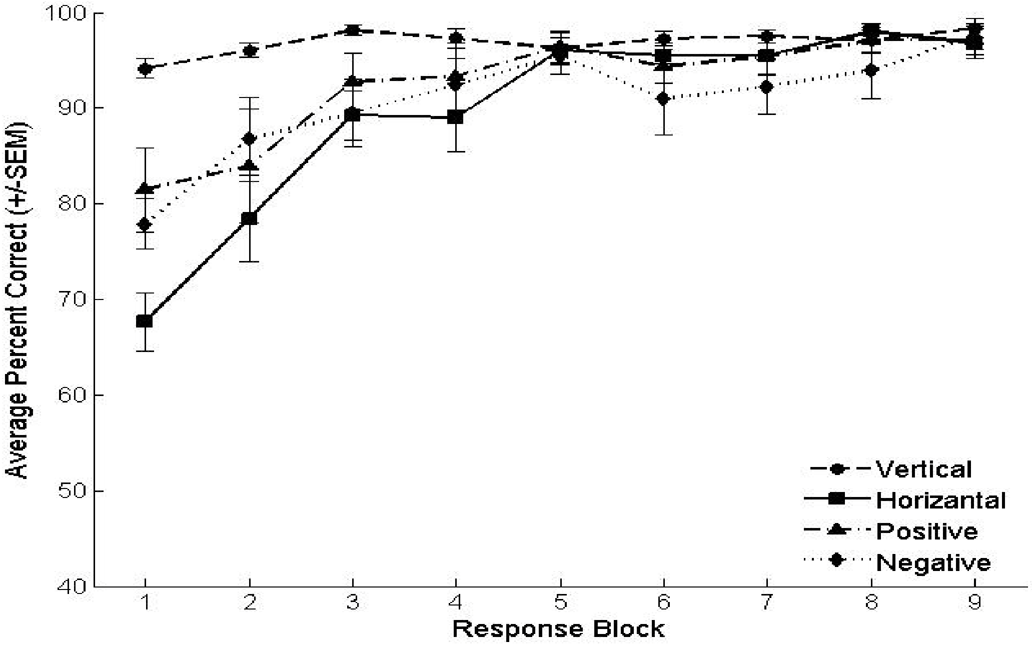

Figure 4. Average accuracy in the four conditions of Experiment 2.

Table 4. Percentage of Blocks Best Accounted for by each Model Across the Four Conditions of Experiment 2.

			Models		
Condition	OC	UC-B	UC-S	CC	LC
Vertical	77.8	13.0	0.0	1.9	7.4
Horizontal	62.2	2.2	6.7	6.6	22.2
Positive	59.3	14.8	0.0	0.0	25.9
Negative	60.0	13.3	0.0	6.7	20.0

Note. OC - optimal classifier, UC-B – one-dimensional classifier on the brightness dimension, UNI-S – one-dimensional classifier on the saturation dimension, CC - conjunctive classifier, LC - linear classifier. In all conditions, the OC is a special case of one of the other models in which it is assumed that the participant used the optimal decision strategies plotted in Figure 1 (Vertical: the OC is a special case of the UC-B, Horizontal: the OC is a special case of the UC-S; Positive and Negative: the OC is a special case of the LC).

response was correct and a low-pitched tone (200 Hz) if the response was incorrect. In addition, feedback was given at the end of each block regarding the participant's accuracy during that block.

Results

Accuracy-based analyses

The average learning curves for each of the four experimental conditions are shown in Figure 4. Visual inspection of Figure 4 suggests an ordering of the four conditions by task difficulty early in training similar to that observed at the end of training in Experiment 1. Specifically, the participants in the Vertical condition were the most accurate followed by the Positive, Negative, and Horizontal conditions. These accuracy differences, however, were nonexistent by the end of training. A 4 condition × 9 response block mixed ANOVA conducted on the accuracy data (with block as the within subjects factor) largely supported the visual inspection of Figure 4. There was a significant main effect of block [$F(8, 152) = 33.68$, $p < .001$, $MSE = .01$, $\eta_p^2 = .64$] that was qualified by a significant condition × block interaction

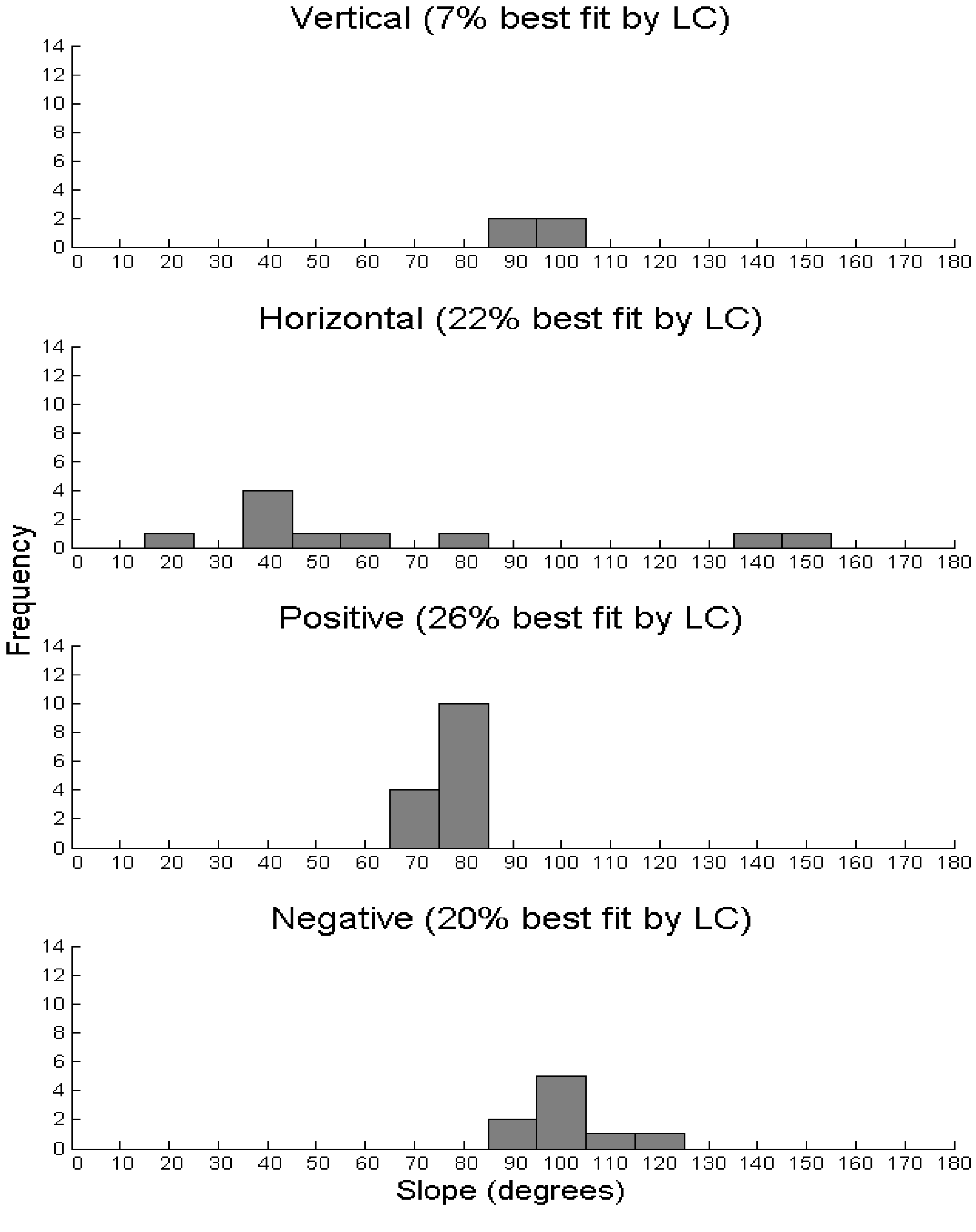

Figure 5. Distribution of the slope (in degrees) for cases where the linear classifier provided the best account of the data in Experiment 2. The percentage of blocks for which the linear classifier provided the best fit is provided in the title for reference. For descriptive purposes, it was assumed that the slopes ranged from 0° to 180°. The optimal classifier predicts the following slopes: Vertical - 90°, Horizontal - 0°/180°, Positive - 68°, Negative - 113°. Bin-width = 10°.

[F(24, 152) = 2.19, p <.01, MSE = .01, $\eta_p^2 = .26$]. The main effect of condition was not significant [F(3, 19) = 2.50, p =.09, MSE = .14, $\eta_p^2 = .28$]. A simple main effects analysis revealed a pattern of results consistent with the visual analysis of Figure 4. Specifically, accuracy in the Vertical condition during the first response block was higher than accuracy in both the Horizontal (p < .01) and Negative (p < .05) conditions, but not the Positive condition (p = .17). Accuracy in the Vertical condition continued to exceed that of the Horizontal condition (p < .05), but not the Negative condition (p = .48) during block 2. The difference between the Vertical and Horizontal conditions was no longer present during block 3 (p = .11). None of the remaining pairwise comparisons were significant.

Model-based analyses

The same models investigated in Experiment 1 were fit to each participant's responses separately for every block in Experiment 2. The distribution of best-fitting models in each of the four conditions is listed in Table 4. The first thing to note is that, in comparison to Experiment 1, the percentage of blocks in which the optimal classifier was the best-fitting model greatly increased in all conditions with the addition of feedback. In the one-dimensional conditions, decision strategies assuming participants attended selectively to brightness and saturation provided the best account of the data on 91% (Vertical) and 69% (Horizontal) of the blocks, respectively. In the diagonal conditions, the use of one-dimensional decision strategies was far less frequent than in Experiment 1.

In addition, the linear classifier provided the best fit to almost 25% of the blocks in the Horizontal, Positive, and Negative conditions. While this percentage is far less than that observed in Experiment 1, it is still worthwhile to determine whether or not these bounds were consistent with a weak form of selective attention or the integration of brightness and saturation. The distribution of slopes estimated from the linear classifier for those blocks in which the linear classifier provided the best fit is plotted in Figure 5. The results from the Horizontal and Negative conditions are quite similar to those of Experiment 1. The slopes from the Horizontal condition were highly variable and not consistent with one-dimensional decision strategies. In the Negative condition, the majority of the slopes were consistent with a weak form of selective attention to brightness (i.e., between 90° and 110°). In contrast to Experiment 1, those blocks that were best fit by the linear classifier in the Positive condition were consistent with a weak form of selective attention on brightness.

Brief Summary

The results of Experiment 2 indicate that the Figure 1 category structures can be learned with feedback and, more generally, that category structures constructed from continuous-valued, integral (i.e., brightness and saturation) dimensions are easily learned. In all conditions, participants were more accurate and there was an increase in the percentage of participants using optimal decision strategies with the addition of feedback. Perhaps not surprisingly, the ordering by task difficulty early in training mimicked that observed when feedback was omitted.

Experiment 3

The category structures used in Experiments 1 and 2 were designed to equate discriminability across brightness and saturation. Even so, in both experiments, participants were better when the categorization judgment required selective attention to brightness (the Vertical condition) than when it required selective attention to saturation (the Horizontal condition).This leaves open the possibility that the categories in the Horizontal condition were less discriminable than the categories in the Vertical condition. To address this question, we ran an unsupervised version of the Horizontal condition in which we varied category discriminability along the saturation dimension by increasing the inter-category distance (Figure 6A). If the ability to learn one-dimensional strategies on saturation is dependent upon category discriminability then accuracy should be higher, and one-dimensional strategies should be used more frequently, in the High Discriminability condition relative to the Low Discriminability condition.

One consequence of increasing the inter-category distance is that there is also an increase in the number of qualitatively different decision strategies that predict high accuracy. For example, in the High Discriminability condition, a one-dimensional strategy on saturation would be indistinguishable from many strategies that integrate saturation and brightness. To address this issue, we replaced the final response block with a test block using a uniform grid of stimuli that spanned the range of the training stimuli (Figure 6B). Fitting the models to the categorization responses from the test block will provide a stronger test of the use of one-dimensional strategies.

Method

Participants and Design

Twenty-six participants were recruited from the University of Maine student community and received partial course credit for participation. Thirteen participants were randomly assigned to each of the two experimental conditions. No one participated in more than one experimental condition. All participants had normal (20/20) or corrected to normal vision and normal color vision. Participants completed one session of approximately 45 minutes.

Stimuli and Apparatus

The stimulus generation procedures were identical to Experiments 1 and 2 with two exceptions. First, the parameters in Table 5 were used to generate the training stimuli (see Figure 6A). Second, to facilitate the identification of decision strategies in the model-based analyses, the stimuli from the final response block were replaced with a test block that included a uniform grid of stimuli selected to match the range of the training stimuli. In the Low Discriminability condition, the test stimuli were generated by all possible combinations of 8 equally spaced points on brightness (from 7.5 to 92.5) and 10 equally spaced points on saturation (Low Discriminability: from 20 to 80; High Discriminability: from 5 to 95). The test block is not included in the accuracy-based analysis because there is no objectively correct response for many of the test stimuli with respect to the trained categories.

Procedure

The procedure was identical to Experiment 1. The participants were not informed that the stimuli during the final test block differed from the stimuli during the earlier training blocks.

Results

Accuracy-based analyses

Average accuracy during training is plotted in Figure 7. A clear accuracy advantage emerged with training for participants in the High Discriminability condition. The results of a 2 condition × 4 response-block mixed ANOVA (with block as the within-subjects factor) were consistent with this claim. Specifically, the main effect of condition $F(1,24) = 46.94$, $p < .001$, $MSE = 135.23$, $\eta_p^2 = .66$] and the condition x block interaction [$F(1.99, 47.86) = 4.99$, $p < .05$,

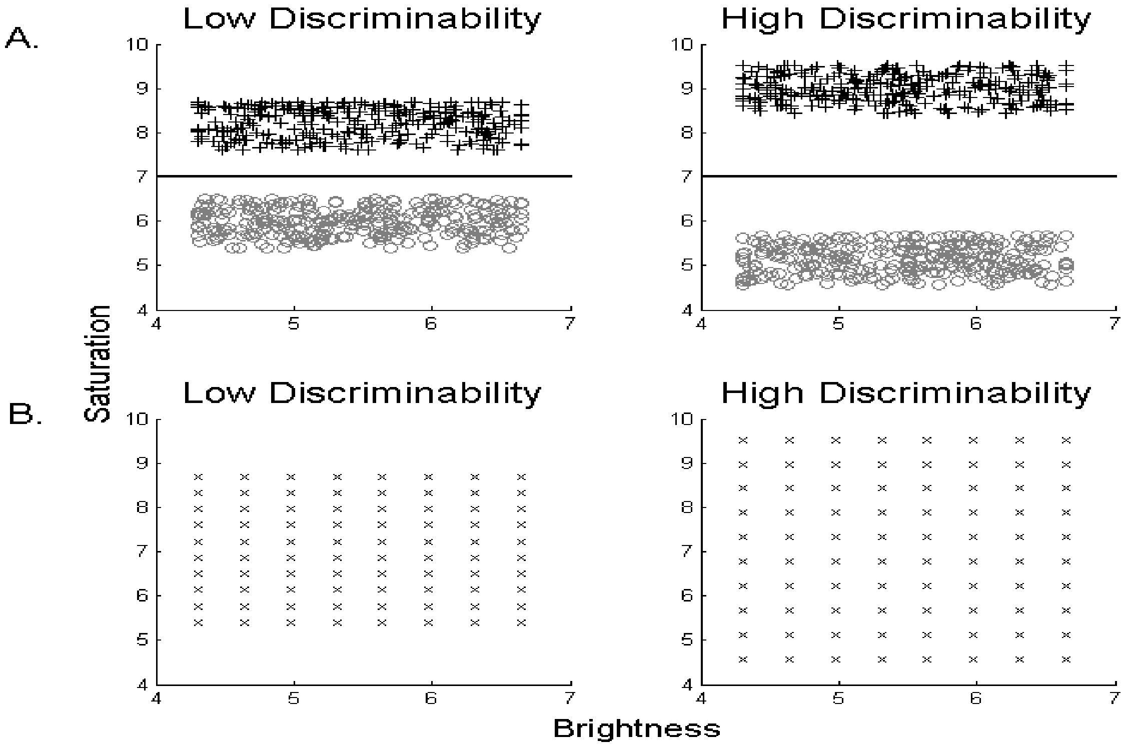

Figure 6. A. Scatterplots of the training stimuli used in Experiment 3. Each point represents a rectangular, iso-hue color patch that varied continuously on Munsell value and chroma. Category A and B exemplars are depicted as black plus signs ('+') and gray circles ('o'), respectively. The solid lines are examples of decision strategies that maximize accuracy (i.e., the optimal decision strategies). B. Scatterplots of the test stimuli used during the final response block. The stimulus coordinates are equally spaced and were selected to span the range of the training stimuli.

Table 5. Parameters Used to Generate the Low and High Discriminability Category Structures Before Transforming to the Munsell Color Space.

	Value		Chroma	
	Min	Max	Min	Max
Low Discriminability				
Category A	7.5	92.5	60	80
Category B	7.5	92.5	20	40
High Discriminability				
Category A	7.5	92.5	75	95
Category B	7.5	92.5	5	25

MSE = 95.54, $\eta_p^2 = .17$] were significant. The interaction was driven by superior performance in the High Discriminability condition during the last three response blocks (block 1: *p* = .11; blocks 3, 5, and 7: *p*'s < *.001).* The main effect of block was not significant [*F*(1.99, 47.86) = .21, *p* = .81, *MSE* = 95.54, $\eta_p^2 = .009$].

Model-based analyses

Recall that increasing inter-category distance comes with the cost of decreased identifiability of the decision strategy. Thus, the higher accuracy in the High Discriminability condition during training may be a consequence of the success of strategies that integrate brightness and

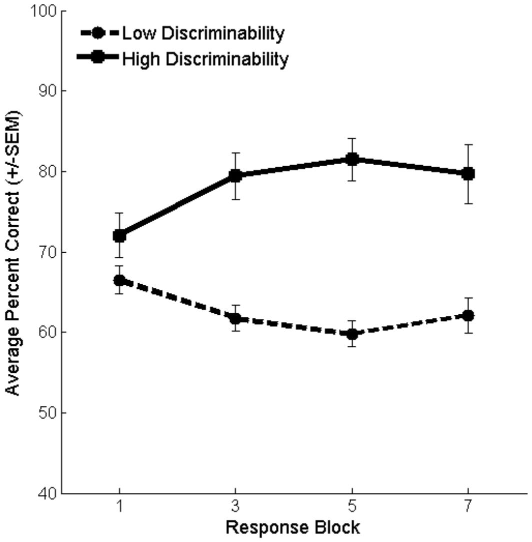

Figure 7. Average accuracy in the two conditions of Experiment 3.

saturation rather than an increase in the use of strategies assuming selective attention to saturation. To address this issue, we focused the model-based analyses on the data from the test block in which the stimuli were sampled uniformly from across the range of the training stimuli. The same models investigated in Experiments 1 and 2 were fit to each participant's responses separately for the test block and the distribution of best-fitting models is listed in Table 6. As was the case in Experiment 1, decision strategies consistent with selective attention to saturation were extremely rare (n_{low} = 2/13, n_{high} = 1/13). This pattern held across the training blocks with 12.3% and 13.8% of the blocks being best accounted for by decision strategies consistent with selective attention to saturation in the Low and High Discriminability conditions, respectively. Importantly, these data suggest that increasing category discriminability for categories defined on saturation was not accompanied by an increase in the use of one-dimensional strategies on saturation.

Table 6. Number of Blocks Best Accounted for by each Model During the Final Test Block of Experiment 3.

	Models				
Discriminiability	OC	UC-B	UC-S	CC	LC
Low	1	4	1	0	7
High	0	1	1	0	11

Note. OC - optimal classifier, UC-B – one-dimensional classifier on the brightness dimension, UNI-S – one-dimensional classifier on the saturation dimension, CC - conjunctive classifier, LC - linear classifier. The OC is a special case of the UC-S in which it is assumed that the participant used the optimal decision strategy plotted in Figure 6.

The preceding analysis implies that the increased accuracy in the High Discriminability condition occurred because the more widely separated categories in that condition did not penalize strategies in which selective attention to saturation failed as much as the Low Discriminability condition. The relative degree to which participants attended to saturation versus brightness can be assessed by examining parameter estimates from the best-fitting version of the linear classifier. Recall, that in this model the slope of the decision bound is free to vary. Slopes near 0 or 180 degrees (counterclockwise from horizontal) are consistent with a weak form of selective attention to saturation, whereas slopes near 90 degrees are consistent with a weak form of selective attention to brightness. As shown in Figure 8, the slopes for participants best fit by the linear classifier were more consistent with a weak form of selective attention to brightness (i.e., 90 degrees counterclockwise from horizontal) (M_{low} = 102.4 degrees, SE_{low} = 8.6; M_{high} = 95.2 degrees, SE_{high} = 8.7). Thus, despite the higher accuracy of participants in the High Discriminability condition, participants in this condition appeared to allocate more attention to the irrelevant brightness dimension than to the relevant saturation dimension..

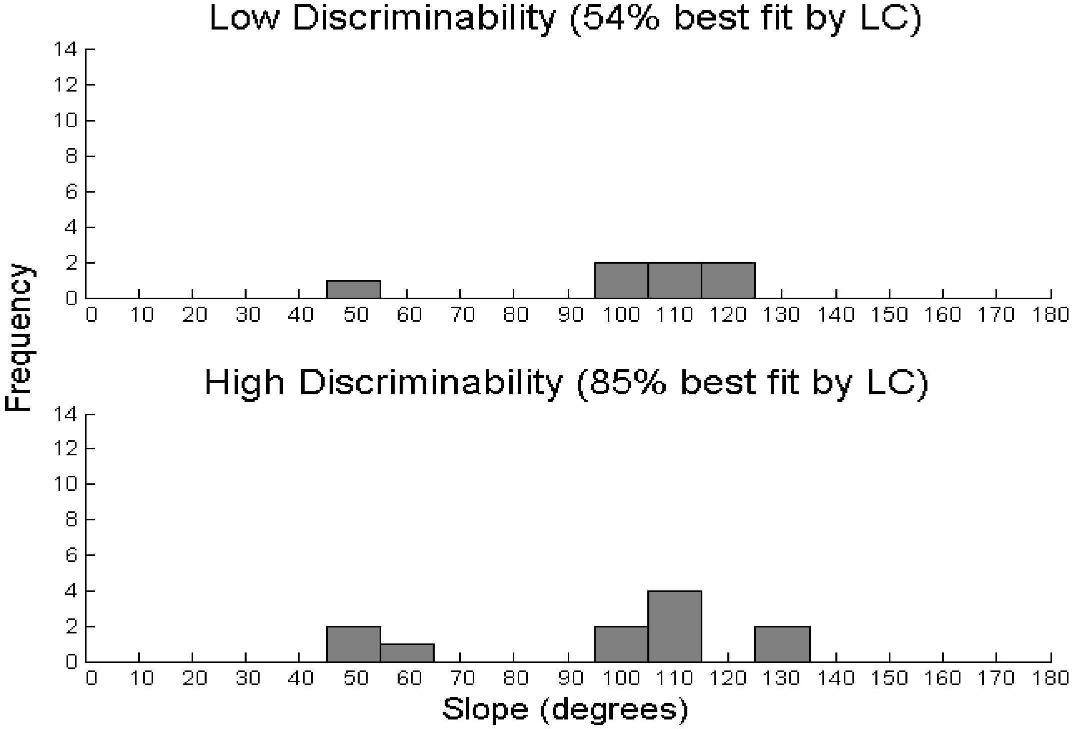

Figure 8. Distribution of the slope (in degrees) for cases where the linear classifier provided the best account of the data in Experiment 3. The percentage of blocks for which the linear classifier provided the best fit is provided in the title for reference. For descriptive purposes, it was assumed that the slopes ranged from 0° to 180°. The optimal classifier predicts a slope of 0° or 180°. Bin-width = 10°.

Brief Summary

The goal of Experiment 3 was to

determine if increasing category discriminability improved the ability of participants to learn categories defined by saturation in the absence of feedback. Although participants were more accurate when discriminability was increased, the increased accuracy was not driven by an increased ability to attend selectively to saturation. Instead, consistent with Experiment 1, participants tended to integrate brightness and saturation in a manner that suggested greater weighting of brightness.

General Discussion

Previous research on the unsupervised categorization of stimuli constructed from integral dimensions has focused primarily on categorization preferences in unconstrained tasks where no instructions are provided about the underlying category structure. Although such studies are clearly important, they do not address the ability of individuals to learn in constrained tasks where the goal is to learn a pre-defined category structure. The present study makes an important contribution to the literature by addressing this question. Experiment 1 showed that the ability to learn categories constructed from brightness and saturation under unsupervised conditions varies as a function of the category structure. Specifically, in the absence of feedback, participants were able to learn only when the category structures required attending selectively to brightness. Experiment 2 demonstrated that these category structures can be learned with feedback and that categorization based on brightness is easier than categorization based on saturation. Experiment 3 demonstrated that unsupervised categorization accuracy for categories defined by saturation can be increased by increasing category discriminability. In contrast to the Vertical condition of Experiment 1, however, higher accuracy was not driven by the use of strategies assuming selective attention to the relevant dimension, but rather by the increased accuracy associated with strategies assuming the integration of brightness and saturation.

These data are partially consistent with the predictions motivated by both constrained and unconstrained unsupervised category learning tasks. Data from constrained tasks using separable dimensions predict a bias to use one-dimensional strategies (e.g., Ashby et al., 1999). Strong evidence in support of this bias, however, was only observed when the categories were defined by brightness. In contrast, data from unconstrained tasks using integral dimensions typically predict that there is a bias to use similarity-based strategies. Indeed, the majority of data from the Experiment 1 and 3 conditions were best accounted for by strategies assuming the integration of brightness and saturation. In each experiment, however, a subset of the participants that integrated brightness and saturation used strategies that weighted brightness more heavily than saturation, suggesting a weak form of selective attention to brightness. Furthermore, the different patterns of strategy use across experimental conditions and the general advantage for the Vertical conditions of Experiments 1 and 2 contradict the assumption that there should be a general bias to use similarity-based strategies with integral dimensions. If this was true then performance should have been equivalent in all conditions since the various conditions were all rotations of each other and similarity with integral dimensions is generally thought to be rotation invariant (Shepard, 1964).

Is Brightness Privileged in the Munsell System?

Several aspects of these data suggest that brightness is more efficiently processed than saturation. First, under unsupervised conditions, participants were most accurate when the categories were defined by brightness. Second, under supervised conditions, participants learned categories defined by brightness at a faster rate. Third, participants relied upon one-dimensional rules on brightness much more frequently than saturation. Fourth, even when category discriminability along saturation was increased, participants rarely used one-dimensional strategies on saturation. Finally, as might be expected with integral dimensions, many participants integrated brightness and saturation in the absence of feedback. In general, however, these participants did not give brightness and saturation equal weighting. Instead, there was a bias to give brightness more weight in the categorization process.

Indeed, attentional mechanisms operate more efficiently with brightness than saturation under supervised conditions (Maddox & Dodd, 2003; Nosofsky, 1987). The Maddox and Dodd experiments suggest that the advantage for brightness over saturation is driven by a perceptual bias (i.e., the perceptual representation of brightness is less noisy than the perceptual representation of saturation). In spite of this difference, participants can clearly learn categories defined by saturation under supervised conditions. One possibility is that the perceptual advantage for brightness over saturation is exaggerated under unsupervised conditions.

Furthermore, in everyday life, people are constantly making discriminations based on brightness, but how common are saturation discriminations? In fact, all participants indicated that they understood what brightness was, but several were unfamiliar with saturation and had to be given specific examples. Clearly participants demonstrated an ability to attend selectively to saturation when feedback was provided, but the Horizontal condition still required more training to learn than the Vertical condition in Experiment 2. Moreover, only three of the 10 participants were ever able to achieve near optimal accuracy without the aid of feedback. It may be that with extended training more participants would have been successful learning categories defined by saturation in the absence of feedback. This, however, seems unlikely given the lack of improvement observed in the Horizontal condition of Experiment 1 and the Low Discriminability condition of Experiment 3. Thus, despite the fact that the Munsell system was designed to equate variation on brightness and saturation, there may be an advantage for making decisions based on brightness (at least when it is paired with saturation).

Experiment 3 was designed to explore whether it is possible to overcome the disadvantage for categories defined by saturation. More specifically, would increasing category discriminability (i.e., increasing inter-category distance) improve accuracy by increasing the use of one-dimensional strategies on saturation? Perhaps not surprisingly, accuracy improved with category discriminability. Critically, however, the improvement in accuracy was not driven by an increase in the use of one-dimensional strategies on saturation. Instead, the improvement in accuracy was driven by the increased success of strategies assuming the integration of brightness and saturation that resulted from increasing inter-category distance.

We chose to define our stimuli in the Munsell color system because it provides a more direct connection to previous work on the categorization of stimuli constructed from integral dimensions. Of course, there are many other color systems. Interestingly, brightness, but not saturation, is generally represented across color systems. For example, brightness in the Munsell system is monotonically related to luminance in the Natural Color System (e.g., Brainard, 2003). Moreover, it has been suggested that luminance (and not saturation) may be one of the features that guides preattentive visual processing (Wolfe, 2005). Although speculative, these arguments suggest that brightness may be a more fundamental feature than saturation in the representation of color.

Constrained Categorization of Separable versus Integral Dimensions

Data from constrained tasks demonstrates that individuals are capable of learning categories constructed from separable dimensions in the absence of feedback (Ashby, et al., 1999; Zeithamova & Maddox, 2009). This capability, however, appears to be limited as individuals were not able to learn when a one-dimensional strategy predicted poor performance (i.e., similar to the Diagonal conditions - Ashby, et al., 1999). Moreover, even when there is a highly accurate one-dimensional strategy, unsupervised learning with separable dimensions appears to be limited to categories that are highly discriminable (Ell & Ashby, in press).

Given the dominance of one-dimensional rules that had been observed in some previous studies with separable stimulus dimensions (e.g., Ashby, et al., 1999), it is surprising that participants in the Horizontal condition did not outperform those in the diagonal conditions. One possible explanation relates to our definition of a one-dimensional rule. Here we have operationally defined a one-dimensional rule as a decision bound orthogonal to the physical dimensions of value or chroma. It is possible that the psychological representation of the decision strategy does not correspond exactly to the dimensional structure we intended (e.g., Melara, Marks, & Potts, 1993). It is also possible that the mapping between the physical space and the perceptual (brightness × saturation) space is nonlinear. The Munsell system is based on scaling judgments performed on one color attribute while the remaining attributes were held constant. Thus, the relations proposed in the Munsell system are not guaranteed to hold when the dimensions are varied orthogonally (Brainard, 2003). Therefore, there is reason to expect that the psychological representation of a rule may not exactly match our operational definition of a rule in the value × chroma space.

This explanation would be more compelling if participants demonstrated some degree of consistency in their decision strategy in the saturation-relevant conditions. In contrast to the Vertical condition, there was little agreement in the best-fitting decision bounds across participants. If there was some psychological rule that did not correspond to the physical axes, then one would expect some degree of agreement between the participants. Of course, these data do not rule out this interpretation as the different participants may have each been attending selectively to different psychological rules. To the extent that categorization based on saturation is less intuitive than categorization based on brightness, the variability in strategy use is consistent with recent work by Pothos and colleagues (Pothos et al., 2011).

It would be inaccurate to say that unsupervised learning is impossible in the diagonal conditions. Although the optimal classifier was the best-fitting model for only a small number of blocks in both the Positive and Negative conditions, the data from several other blocks that were best fit by the linear classifier did not deviate substantially from the optimal decision strategy. Therefore, it appears as though it is possible to successfully integrate information from integral stimulus dimensions in the absence of feedback, but also that there are large individual differences in this ability.

The distribution of best fitting models observed in the diagonal conditions is quite different than that observed with separable dimensions. For example, Ashby et al. (Ashby, et al., 1999) found that participants almost exclusively relied upon one-dimensional decision strategies in category structures similar to the present diagonal conditions. This is not the case when the stimuli are constructed from integral dimensions. In the Positive condition, participants were integrating brightness and saturation information rather than attending selectively to one of those dimensions. In contrast, in the Negative condition, there was more frequent use of decision strategies consistent with one-dimensional rules. However, many participants were also integrating brightness and saturation information.

In addition, in the Positive condition of Experiment 1, there were a number of blocks in which decision strategies were more consistent with the optimal classifier (between 40° and 70°) than a one-dimensional strategy. A close inspection of the entire stimulus space, suggested that there may have been an alternative "grayness" rule to which participants were sensitive. Inspection of Figure 3 and written descriptions of decision strategies, however, reveals no evidence that participants used a "grayness" strategy. Furthermore, no such bias was observed when feedback was provided in Experiment 2 (Figure 5). Instead, participants whose data were best fit by the linear classifier were more likely to use a strategy consistent with a less accurate one-dimensional rule than the so-called "grayness" rule.

The few available studies on free sorting of integral-dimension stimuli suggest that people prefer to use similarity-based decision strategies (e.g., Handel & Imai, 1972).

Although testing between similarity-based and one-dimensional strategies is not the focus of this article, the similarity-based account would seem to predict that performance should be identical across the four category conditions. This is because all of our conditions are simply rotations of each other and according to most popular definitions, similarity among integral-dimension stimuli is invariant under rotation of the categories (Nosofsky, 1986).

Implications for Models of Category Learning

Many computational models have been developed that are capable of category learning in the absence of feedback (Ahn & Medin, 1992; Anderson, 1991; Ashby, Alfonso-Reese, Turken, & Waldron, 1998; Billman & Heit, 1988; Carpenter & Grossberg, 1991; Fried & Holyoak, 1984; Love, Medin, & Gureckis, 2004; Pothos & Chater, 2002). The majority of these models, however, predict that at least some aspects of performance on the Figure 1 category structures should be invariant with rotation of the categories and even fewer make explicit predictions for integral versus separable dimensions. The pattern of results observed in the accuracy- and model-based analyses therefore provides a challenge to future model development.

For example, consider a recent network model of category learning proposed by Love and colleagues (Supervised and Unsupervised Stratified Adaptive Incremental Network, or SUSTAIN - Love, et al., 2004), which has been successfully applied to data from unsupervised tasks. In short, SUSTAIN is a multi-layer neural network that maps stimulus representations to the appropriate responses via an intermediate layer of abstract category representations (or clusters). In unsupervised tasks, SUSTAIN assumes that the initial category representation comprises a single cluster and that additional clusters are created as exemplars that are highly dissimilar to existing clusters are encountered. SUSTAIN was applied to a broader class of unsupervised tasks than considered here, but it does correctly predict that participants prefer one-dimensional strategies in unconstrained tasks, at least with separable dimensions. Similar to many other models of category learning, SUSTAIN is equipped with a selective attention learning mechanism that, together with a bias to attend to brightness over saturation, would allow it to capture the higher accuracy in the Vertical condition in Experiment 1. This attentional bias, however, should be invariant with rotation of the categories, thereby predicting the frequent use of decision strategies on brightness in all conditions – a prediction that is inconsistent with the data.

Pothos and Chater's (2002) simplicity model is also particularly relevant to the issue of unsupervised categorization. Briefly, this model assumes that the preferred categorization strategy will be the simplest one (i.e., in an information-theoretic sense). A strategy's simplicity, or code length, is a function of the similarity structure of the stimuli and the costs and benefits of the constraints imposed by classifying the stimuli. The simplicity model accurately predicts a bias to use one-dimensional strategies over two-dimensional strategies in the absence of feedback and, as a result, the pattern of performance in the Vertical condition (Pothos & Close, 2008). The simplicity model, however, would predict a similar bias for the Horizontal condition and an approximately equal distribution of one- and two-dimensional strategies in the Positive and Negative conditions (Pothos & Close, 2008). Neither of these patterns was observed in the present data.

An alternative account is offered by the COVIS model of category learning (Ashby, et al., 1998). COVIS hypothesizes that category learning is a competition between separate hypothesis-testing and procedural-based systems. The hypothesis-testing system is specialized to learn abstract rules (e.g., one-dimensional rules) whereas the procedural-based system is specialized to learn stimulus-response mappings. Because learning in the procedural-based system is highly dependent upon feedback and there is an initial bias towards the hypothesis-testing system, COVIS predicts that the hypothesis-testing system should dominate in unsupervised tasks. COVIS also predicts that the hypothesis-testing system will experiment with one-dimensional rules only if two conditions are met – that selective attention can be directed to this dimension and that a salient verbal label describes the dimension (e.g., brightness). These conditions are met with most separable dimensions, so COVIS correctly predicts that one-dimensional decision strategies should dominate in unsupervised tasks with separable dimensions (Ashby, et al., 1999). Thus, COVIS correctly predicts the high prevalence of one-dimensional strategies on brightness in the Vertical condition. COVIS, however, would not predict that participants would be able to integrate brightness and saturation as evidenced by the high percentage of data sets best fit by the linear classifier in the Horizontal, Positive, and Negative conditions.

Summary

In summary, the present experiments demonstrate that individuals are capable of learning categories constructed from the integral dimensions of brightness and saturation in the absence of feedback. This ability, however, has several limitations. Consistent with the claim that integral dimensions are initially processed holistically (e.g., Kemler Nelson, 1993), participants had some success in conditions that required the integration of brightness and saturation. Consistent with the claim that integral dimensions can subsequently be processed in terms of the individual dimensions given the appropriate task demands (e.g., Garner, 1974), participants were able to learn when a one-dimensional strategy on brightness was highly accurate. In addition, participants demonstrated a general tendency to weight brightness more heavily than saturation across all three experiments, suggesting that brightness may have privileged status relative to saturation. Whether such a pattern holds when brightness is paired with other dimensions is a matter for future research.

Although current models can account for some aspects of these data, to our knowledge no current models can account for the pattern of strategy use observed In Experiment 1. In fairness, although SUSTAIN, the simplicity model, and COVIS all make predictions for unsupervised tasks, none of these models were designed to account for differences between integral and separable stimulus dimensions. Thus, it may be possible to augment these models to account for the present data (e.g., Pothos, et al., in press). Even so, because of the theoretical difficulties posed by our results, these data will provide an important benchmark for the development of theories of unsupervised category learning and may have implications for the application of cognitive science to education and training where constrained tasks are the norm (e.g., the training of medical professionals to categorize pre-defined medical conditions).

References

Ahn, W., & Medin, D. L. (1992). A two-stage model of category construction. *Cognitive Science, 16*, 81-121.

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

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., & Maddox, W. T. (1994). A response time theory of separability and integrality in speeded classification. *Journal of Mathematical Psychology, 38*, 423-466.

Ashby, F. G., Queller, S., & Berretty, P. M. (1999). On the dominance of unidimensional rules in unsupervised categorization. *Perception & Psychophysics, 61*, 1178-1199.

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

Billman, D., & Heit, E. (1988). Observational learning from internal feedback: A simulation of an adaptive learning method. *Cognitive Science, 12*, 587-625.

Billman, D., & Knutson, J. (1996). Unsupervised concept learning and value systematicity: A complex whole aids learning the parts. *Journal of Experimental Psychology: Learning, Memory, & Cognition, 22*, 458-475.

Boynton, R. M., & Olson, C. X. (1987). Locating basic colors in the OSA space. *Color Research and Application, 12*(2), 94-105.

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

Brainard, D. H. (2003). Color appearance and color difference specification. In S. K. Shevell (Ed.), *The Science of Color* (2nd ed., pp. 191-216). Washington D. C.: Optical Society of America.

Carpenter, G. A., & Grossberg, S. (1991). *Pattern recognition by self-organizing neural networks*. Cambridge, MA: MIT Press.

Clapper, J. P., & Bower, G. H. (1994). Category invention in unsupervised learning. *Journal of Experimental Psychology: Learning, Memory, & Cognition, 20*, 443-460.

Colreavy, E., & Lewandowsky, S. (2008). Strategy development and learning differences in supervised and unsupervised categorization. *Mem Cognit, 36*(4), 762-775.

Diaz, M., & Ross, B. H. (2006). Sorting out categories: incremental learning of category structure. *Psychon Bull Rev, 13*(2), 251-256.

Ell, S. W., & Ashby, F. G. (in press). The impact of category separation on unsupervised categorization. *Attention, Perception, & Psychophysics*.

Foard, C. F., & Kemler-Nelson, D. G. (1984). Holistic and analytic modes of processing: The multiple determinants of perceptual analysis. *Journal of Experimental Psychology: General, 113*, 137-159.

Fried, L. S., & Holyoak, K. J. (1984). Induction of category distributions: A framework for classification learning. *Journal of Experimental Psychology: Learning, Memory, & Cognition, 10*, 234-257.

Garner, W. R. (1974). *The processing of information and structure*. New York:

Wiley.

Garner, W. R., & Felfoldy, G. L. (1970). Integrality of stimulus dimensions in various types of information processing. *Cognitive Psychology, 1*, 225-241.

Grau, J. W., & Kemler-Nelson, D. G. (1988). The distinction between integral and separable dimensions: Evidence for the integrality of pitch and loudness. *Journal of Experimental Psychology: General, 117*, 347-370.

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

Handel, S., & Imai, S. (1972). The free classification of analyzable and unanalyzable stimuli. *Perception & Psychophysics, 12*, 108-116.

Handel, S., Imai, S., & Spottswood, P. (1980). Dimensional, similarity, and configural classification of integral and separable stimuli. *Perception & Psychophysics, 28*, 205-212.

Homa, D., & Cultice, J. (1984). Role of feedback, category size, and stimulus distortion on the acquisition and utilization of ill-defined categories. *Journal of Experimental Psychology: Learning, Memory, and Cognition, 10*, 83-94.

Imai, S., & Garner, W. R. (1965). Discriminability and preference for attributes in free and constrained classification. *Journal of Experimental Psychology, 69*, 596-608.

Kemler Nelson, D. G. (1993). Processing integral dimensions: The whole view. *Journal of Experimental Psychology: Human Perception & Performance, 19*, 1105-1113.

Lockhead, G. R. (1972). Processing dimensional stimuli: A note. *Psychological Review, 79*, 410-419.

Love, B. C. (2002). Comparing supervised and unsupervised category learning. *Psychon Bull Rev, 9*(4), 829-835.

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. (2001). Separating perceptual processes from decisional processes in identification and categorization. *Perception & Psychophysics, 63*, 1183-1200.

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

Maddox, W. T., & Dodd, J. L. (2003). Separating perceptual and decisional attention processes in the identification and categorization of integral-dimension stimuli. *Journal of Experimental Psychology: Learning, Memory, & Cognition, 29*, 467-480.

McKinley, S. C., & Nosofsky, R. M. (1996). Selective attention and the formation of linear decision boundaries. *Journal of Experimental Psychology: Human Perception & Performance, 22*, 294-317.

Medin, D. L., Wattenmaker, W. D., & Hampson, S. E. (1987). Family resemblance, conceptual cohesiveness, and category construction. *Cognitive Psychology, 19*, 242-279.

Melara, R. D., Marks, L. E., & Potts, B. C. (1993). Early-holistic processing or dimensional similarity? *Journal of Experimental Psychology: Human Perception & Performance, 19*, 1114-1120.

Milton, F., Longmore, C. A., & Wills, A. J. (2008). Processes of overall similarity sorting in free classification. *Journal of Experimental Psychology: Human Perception and Performance, 34*(3), 676-692.

Milton, F., & Wills, A. J. (2004). The

influence of stimulus properties on category construction. *J Exp Psychol Learn Mem Cogn, 30*(2), 407-415.

Mounts, J. R. W., & Melara, R. D. (1995). Classification of color dimensions in multiple contexts. *Journal of Experimental Psychology: Human Perception & Performance, 21*(257-274).

Munsell, A. H. (1915). *Atlas of the Munsell Color System*. Malden, MA: Wadsworth-Howland & Company.

Murphy, G. L. (2002). *The big book of concepts*. Cambridge: MIT Press.

Newhall, S. M. (1940). Preliminary report of the O.S.A. subcommittee on the spacing of the Munsell colors. *Journal of the Optical Society of America, 30*, 617-645.

Newhall, S. M., Nickerson, D., & Judd, D. B. (1943). Final report of the O.S.A. subcommittee on the spacing of the Munsell colors. *Journal of the Optical Society of America, 33*, 385-412.

Nickerson, D. (1940). History of the Munsell color system and its scientific application. *Journal of the Optical Society of America, 30*, 575-586.

Nosofsky, R. M. (1986). Attention, similarity, and the identification categorization relationship. *Journal of Experimental Psychology: General, 115*, 39-57.

Nosofsky, R. M. (1987). Attention and learning processes in the identification and categorization of integral stimuli. *Journal of Experimental Psychology: Learning, Memory, & Cognition, 13*, 87-108.

Nosofsky, R. M., & Palmeri, T. J. (1996). Learning to Classify integral-dimension stimuli. *Psychonomic Bulletin & Review, 3*, 222-226.

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

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

Pothos, E. M., & Chater, N. (2005). Unsupervised categorization and category learning. *Q J Exp Psychol A, 58*(4), 733-752.

Pothos, E. M., & Close, J. (2008). One or two dimensions in spontaneous classification: a simplicity approach. *Cognition, 107*(2), 581-602.

Pothos, E. M., Edwards, D. J., & Perlman, A. (in press). Supervised vs. unsupervised categorization: Two sides of the same coin? *Quarterly Journal of Experimental Psychology*.

Pothos, E. M., Perlman, A., Bailey, T. M., Kurtz, K., Edwards, D. J., Hines, P., et al. (2011). Measuring category intuitiveness in unconstrained categorization tasks. *Cognition, 121*(1), 83-100.

Regehr, G., & Brooks, L. R. (1995). Category organization in free classification: The organizing effect of an array of stimuli. *Journal of Experimental Psychology: Learning, Memory, and Cognition, 21*(347-363).

Schwarz, G. (1978). Estimating the dimension of a model. *The Annals of Statistics, 6*(2), 461-464.

Shepard, R. N. (1964). Attention and the metric structure of the stimulus space. *Journal of Mathematical Psychology, 1*, 54-87.

Shepard, R. N., & Chang, J. (1963). Stimulus generalization in the learning of classifications. *Journal of Experimental Psychology, 65*, 94-102.

Shin, H. J., & Nosofsky, R. M. (1992). Similarity-Scaling Studies of Dot-Pattern Classification and Recognition. *Journal of Experimental Psychology: General, 121*(3), 278-304.

Sturges, J., & Whitfield, T. W. (1995).

Locating basic colours in the Munsell space. *Color Research and Application, 20*, 364-376.

Torgerson, W. S. (1958). The nature of measurement. In W. S. Torgerson (Ed.), *Theory and Methods of Scaling*. New York: Wiley.

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

Wolfe, J. M. (2005). Guidance of visual search by preattentive information. In L. Itti, G. Rees & Tsotsos (Eds.), *Neurobiology of attention*. San Diego, CA: Academic Press.

Wyszecki, G., & Stiles, W. S. (1982). *Color Science: Concepts and methods, quantitative data and formulae* (2d ed.). New York: John Wiley & Sons.

Zeithamova, D., & Maddox, W. T. (2009). Learning mode and exemplar sequencing in unsupervised category learning. *Journal of Experimental Psychololgy: Learning, Memory, & Cognition, 35*(3), 731-741.

Author Notes

This research was supported in part by National Science Foundation Grant BCS99-75037 and Public Health Service Grant MH3760. We would like to thank David Brainard for the use of his spectral radiometer. We would also like to thank Michael Casale, Chris Mitchell, Emmanuel Pothos, Vivian Valentin, and an anonymous reviewer for helpful comments on an earlier draft of this manuscript. Correspondence concerning this article should be addressed to Shawn W. Ell, Psychology Department, University of Maine, 5742 Little Hall, Room 301, Orono, ME 04469-5742 (email: shawn.ell@umit.maine.edu).

Appendix

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 data for each participant from every block. 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).

The appendix briefly describes the decision bound models. For more details, see Ashby (1992a) or Maddox and Ashby (1993).

Unidimensional Classifier

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 unidimensional classifier were fit to these data: one assumed that participants attended selectively to brightness (UC-B) and the other assumed participants attended selectively to saturation (UC-S). The unidimensional classifier has two free parameters: a decision criterion on the relevant perceptual dimension and the variance of internal (perceptual and criterial) noise (i.e., σ^2). In the Vertical and Horizontal conditions, a special case of the unidimensional classifier, the optimal unidimensional classifier (OC), assumes that participants use the unidimensional decision bound that maximizes accuracy (Figure 1). This special case has one free parameter (σ^2).

Conjunctive Classifier

Another possibility is that the participant uses 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 conjunctive classifier (CC) assumes that the participant partitions the stimulus space into four regions. Based upon inspection of the data from the individual participants, two versions of the CC were fit to these data. The first assumed that individuals assigned a stimulus to category A if it was high on brightness and low on saturation; otherwise the stimulus was assigned to category B. The second assumed that a stimulus was assigned to category A it was low on brightness and low on saturation; otherwise the stimulus was assigned to category B. The CC has three free parameters: the decision criteria on the two dimensions and a common value of σ^2 for the two dimensions.

Linear Classifier

This model assumes that a linear decision bound partitions the stimulus space into two regions. The linear classifier (LC) differs from the CC in that the LC does not assume decisional selective-attention (Ashby & Townsend, 1986). Instead, the LC requires integration of the perceived values on the stimulus dimensions. The LC has three parameters, slope and intercept of the linear bound, and σ^2. In the Positive and Negative conditions, a special case of the LC, the optimal linear classifier (OC), assumes that participants use the linear decision bound that maximizes accuracy (Figure 1). This special case has one free parameter (σ^2).

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. To assess the absolute fit of the models, the percent of responses accounted for by the best-fitting model was computed for each data set. This statistic ranges from 0% to 100% with the latter implying that the model perfectly accounted for all of the participant's responses.