Testing Analogical Transfer in Pigeons (*Columba livia*) and Humans (*Homo sapiens*)

A dissertation submitted by

Muhammad A. J. Qadri

in partial fulfillment of the requirements for the degree of

Doctor of Philosophy

in

Psychology: Cognitive Science

Tufts University

February 2017

Advisor: Robert G. Cook, Ph.D.

Secondary Advisor: Matthias Scheutz, Ph.D.

Abstract

Categorization allows organisms to behave appropriately to novelty. In a recent method, categorization has been investigated with stimuli that vary parametrically in their features. Previous research has found that primates can approach this task using rules, whereas pigeons only use associative processing. Furthermore, using these stimuli, analogical transfer has been demonstrated in primates by testing stimuli in a new region of the feature space after restricted training. However, the fundamental dimensionality of the task seems to generate such "analogical transfer," regardless of rule-based processing mechanisms. Here, seven pigeons were successfully trained and tested with that original task to determine whether they show effects of analogical transfer. The pigeons were then further tested to determine the generality of their solution to the discrimination task. To explain their results, an association-based model of the task is developed using neural networks with a stimulus representation that is configural or dimensional. A method was developed to determine the likelihood of this association-based responding and evaluated against the pigeons' data. Finally, comparable human data was collected and the performance was analyzed to determine whether human procedural learning conforms to this unbound, dimensional stimulus representation. The impact of this shift in fundamental representation on our understanding of human cognition and categorization behavior more broadly is discussed.

Acknowledgments

In the name of God, the Most Gracious, the Most Merciful

This manuscript is the culmination of many decades of work by many hands. It takes a village, global in scope and diverse in nature, to generate a Ph.D., and it is my goal here to remember those who took part. Those who are living know who they are, and with one exception I won't perseverate on them and their contributions here.[1] I must, however, thank my wife Saosan, whom I met just before starting graduate school and whose unerring support was critical for these past seven years. Those who have passed, those have no voice except ours to remember them by, I honor here. First, I honor my grandfather Afzal Hussain Qadri, who started an academic legacy and inspired achievement in his society, his country, and his family. Second, I honor my grandmother Marium Khatoon and my dear friend Tanya Mitra, who contributed importantly to my personal growth, and whom I deeply regret not having cherished properly while they were still here. Third, Pax, who should have lived much longer and retired into adoption as the family pet. And of course, Boober, Jo, Lex, and every other two-legged two-winged animal that frustrated me and endeared themselves to me while I tried to understand how they saw the world. They all were close friends.

[1] For their edification, I will list some among the living here, in order of presumed shoe size: Robert Cook, Yawar Qadri, Shirwac Mohamed, Sultan Khan, Dan Brooks, Usama Qadri, Benjamin Hescott, Umar Qadri, Danish Qadri, Richard Chechile, Usman Qadri, Jamal Qadri, Maria Qadri, Fouzia Sultana, Claire Pompei, Pallabi Guha, Fahmeeda Qadri, Nayema Khan, Rukshana Niloofar, Reene Ghosh

Table of Contents

ABSTRACT ..... **I**

ACKNOWLEDGMENTS ..... **II**

TABLE OF CONTENTS ..... **III**

LIST OF TABLES ..... **V**

LIST OF FIGURES ..... **VI**

INTRODUCTION **1**

- **Discrimination & Generalization** **2**
- **Categorization** **5**
- **Categorization by Parametric Optimization** **8**
- **Multiple Systems Models** **14**

EXPERIMENT 1 ..... **18**

- **Methods** **19**
- **Results** **23**
- **Discussion** **25**

EXPERIMENT 2 ..... **26**

- **Methods** **27**
- **Results** **29**
- **Discussion** **35**

FORMAL METHODS **36**

- **Overview** **36**
- **Simulations 1** **46**
- **Simulations 2** **57**
- **Simulations 3** **62**

BEHAVIORAL MODEL FITTING **69**

Methods **69**
Model Results **69**
Discussion **71**

EXPERIMENT 3 ..... **73**

Methods **74**
Results **79**
Discussion **82**

GENERAL DISCUSSION **85**

List of Tables

Table 1. Distribution parameters for training and transfer distributions in Experiment 1. 21

Table 2. Distribution parameters for the transfer clusters in Experiment 2. 28

Table 3. Formal model fitting results. 68

Table 4. Model fitting results for the pigeon data from sessions containing transfer data in Experiments 1 and 2. 70

Table 5. Distribution parameters for the new transfer clusters in Experiment 3. . 76

Table 6. Model fitting results for the human data from Experiment 3. 82

List of Figures

Figure 1. Depiction of Signal Detection Theory. 9

Figure 2. Different bivariate normal distributions. 11

Figure 3. The optimal discriminant for two examples of a two category discrimination task. 13

Figure 4: Distributions of the stimuli investigated in Experiment 1. 16

Figure 5: Analogical transfer performance from Experiment 1. 24

Figure 6: Distributions of the stimuli investigated in Experiment 2. 29

Figure 7: Analogical transfer accuracy from Experiment 2. 31

Figure 8: Analogical transfer performance for the birds in the Rule-Based condition from Experiment 2. 32

Figure 9: Analogical transfer performance for the birds in the Information Integration condition from Experiment 2. 33

Figure 10: A depiction of stimulus representation for the configural activation model and the dimensional activation model, using a 20-unit based grid. 42

Figure 11. Neural network acquisition of the II categorization task with ten hidden units. 51

Figure 12. Final iterations for neural network simulations with reduced hidden units. 53

Figure 13. Average final weights for the 1000 neural net simulations. 56

Figure 14. Distribution of learning times for the two representations in the two tasks 59

Figure 15. Model performance on transfer stimuli. 60

Figure 16: Distributions of the novel stimuli investigated in Experiment 3..... 77

Figure 17: Baseline performance through the task for human participants in Experiment 3..... 80

Figure 18: Accuracy on transfer tests for human participants in Experiment 3. .. 81

Introduction

Animals in everyday behaviors encounter novel conspecifics, objects, perspectives, and conditions. Although no two situations are identical, animals would too frequently endanger themselves and waste energy in continuous large-scale adaptation of their behavior without the ability to generalize appropriate responses from old encounters. Generalizing the knowledge that lake water will freeze rendering food unavailable when the temperature is -5° C to nearby (especially lower) temperatures allows animals to allocate resources effectively without expending the energy necessary to observe it in that instance. Identifying conspecifics provides a more complex example: there are many features that distinguish individual humans from one another, but for each observer, a set of behaviors are appropriate to those who are male and those who are female. This many-to-few mapping or continuous-to-discrete processing is a special level of generalization termed categorization.

The investigations of generalization and categorization have taken on many forms over the past century. Three prominent techniques have been employed, each distinctly tied to the feature space of the stimulus investigated. The earliest and simplest is discrimination which relied primarily on continuous variables, which quick became deeply entrenched with generalization. Soon thereafter categorization followed relying on features with discrete and often binary values. Finally, in more recent advances, the investigation of

categorization has used stimuli with features that are continuously valued, tying the categorization literature back to its generalization roots.

Discrimination & Generalization

Studies of discrimination and generalization have been a cornerstone in psychology because they are the fundamental processes that underlie cognition. At a basic neural level, all organisms do is ascribe sameness and differentness and behave appropriately in terms of those two concepts with respect to previously learned contingencies.

Generalization is the process of taking a learned response of a specific stimulus and producing it in the context of a novel, non-identical stimulus. Ghirlanda and Enquist (2003) collected and organized the results of a near-century of research on the features and properties of generalization. This process has been investigated with rodents (Blackwell & Schlosberg, 1943), birds (Blough, 1961), and primates (Ganz & Riesen, 1962), to name a few prominent biological model. It has also been investigated in many modalities, from basic auditory psychophysics (Hovland, 1937) to more complex visual stimuli and "concepts" (Lumsden, 1977). Ghirlanda and Enquist classify the generalization tasks into two types: generalization across intensity (prothetic) dimensions and generalization across rearrangement (metathetic) dimensions. Generalization across intensity dimensions occurs when a stimulus of a given intensity supports generalization to stimuli of a different intensity, such as a loud sound generalizing to louder sounds or softer sounds. Generalization across rearrangement dimensions would occur when responding to a stimulus of a different quality that

does not change the intensity of stimulation, such as rotations or spatial shifts. A given discrimination is not necessarily limited to one type or the other of generalization; variations of size could be described as both a change in intensity and in quality or arrangement. The authors suggest these differences in the nature of the modalities correlate with differences in generalization.

In most cases, generalization is characterized by the shape of responding (either in reaction time or in amount of production) to some measurable, metrically organized stimuli. For example, the generalization to untrained frequencies of tones (transfer stimuli) can be examined by plotting the rate of responding against the frequency of tones tested. Depending on the scheme of training, this can yield flat or decreasing rates of responding with transfer stimuli more distant from the training stimulus. Generalization has been suggested to be controlled by a universal law founded on psychological spaces of stimuli (Shepard, 1987). The use of psychological space allows the dimensionalization/metric relatability of complex stimuli, accounts for organism-specific specialization of modality processing, and creates a standardized space to consider behavior within. The "universal law of generalization" predicts an exponential decay relationship between the likelihood of exhibiting the learned response to a known stimulus and the likelihood of exhibiting the learned response to a novel stimulus that is some psychological distance away. Others, have suggested other forms of this decay function, such as a Gaussian function; Ghirlanda and Enquist (2003) report only 25% of datasets are better described by

the exponential decay over the Gaussian. Whether this is reflective of underlying psychological "law" or an artifact of the training and testing process is unclear.

Discrimination is the process of responding differentially to items within a set of stimuli. Some discriminative behavior is "innate" in that it can be automatically elicited from the animal without reinforcement or artificial prior experience. This generally falls under the study of biologists, evaluating "sign stimuli" in ethological settings. In psychology, discrimination is more often examined in learning contexts, in which mechanisms of conditioning support discriminative behavior. In these learning contexts, discrimination learning is often evaluated by way of generalization testing, and thus it has a similarly broad spectrum of evaluated animals and modalities. Discrimination can be established by rewarding (or punishing) responses to one or many specific values in a dimension (e.g., bright stimulus/i) and not rewarding (or punishing) one or many distinct other values in a dimension (e.g., dark stimulus/i). Generalization curves using untrained stimulus values seem to be affected by discrimination training, most notably by introducing a "peak shift" in responding (Ghirlanda & Enquist, 2003). The "peak shift" effect is the empirical finding that the stimuli of maximal and minimal responding are not necessarily the trained reinforced and trained non-reinforced values, but these peaks are instead slightly shifted "away" from the opposite stimulus type. Discussing the generalization of discriminations introduces the concept of categorization: when a novel stimulus or stimulus region is assigned the attribute (i.e., reward contingencies) of a trained stimulus

that is distinct from some other trained stimulus with other attributes, the process can be considered one of categorization.

Categorization

Categorization is the classification of new objects to previously learned groups or labels. It is most often considered to be a many-to-few mapping, where the number of categories is often much smaller than the number of exemplars or individual items. The apparent evolutionary benefit of this categorization lies in this orders of magnitude reduction, allowing simple categories like "predator" and "food" to control behavior in a cognitively non-intensive fashion. The question of how categories are formed and how categories are structured has become of recent interest.

In the 19th and 20th centuries, five prominent models came to be evaluated in the realm of categorization: 1) Necessary and sufficient conditions, 2) Cue validity, 3) Exemplar Theory, 4) Prototype Theory, and 5) Decision-Bound Theory. They take varying stances on important aspects of the stimuli and the decision process. The traditional stance of categorization was heavily influenced by the Aristotelian philosophical considerations of categorization, defining the world by necessary and sufficient conditions. While an excellent starting point for the philosophical concerns that underlie categorization and concepts, this method was descriptively inadequate both to describe real categories (what are the necessary and sufficient conditions for "bird" category membership, versus "dinosaur" membership?) and the psychological relationships or realities underlying categorization behavior.

Succeeding this highly introspective proposition were accounts based on the specific stimuli used during learning. One method of categorization that seems closely tied to the idea of necessary and sufficient cues was the use of cue validity. In this scheme, the categorization of a novel object depends on the learned probabilities of category membership with cue distributions (Beach, 1964). Upon exposure to different stimuli and their correct classification, the association between the cues within the stimuli (usually framed in a cue present vs. cue absent design) and the classification can guide discrimination. Whether classification is decided according to which is the most likely or probabilistically decided in proportion to relative likelihood of the outcomes is one tunable parameter within this model.

In order to account for recognition ability as well as categorization, exemplar theory was developed. In this theory, the weighted sum of psychological distance of a novel item to all previously experienced exemplars is used to guide behavior (Medin & Schaffer, 1978). The original modern version of this model, termed the context model, was discussed using binary-valued features (e.g., is-red, is-large, etc.), although there were earlier hints of this strategy using continuous values (see "average distance model" in Reed, 1972). The generalized version of the context model utilized psychological distances based off of multidimensional scaling, which allowed the scheme to be applied to a greater scope of tasks and conditions with rigor (Nosofsky, 1988). Exemplar theory has been shown to successfully account for many features in empirical data (Nosofsky & Johansen, 2000).

Models of categorization using only the prototype of the stimulus distributions also found empirical support. Instead of defining categories by a sum of all previously seen exemplars, the categories were defined according to a category prototype that was at an appropriate centroid of the distribution (Reed, 1972). For example, given one item and several category prototypes, the items assumed assignment could be determined by the relative distances to the prototypes (in the psychological space discussed above). Prototype theory imbued the prototype with many properties and a type of centrality that found echoes in empirical results (e.g., see brief review in Medin & Schaffer, 1978). For example, if all training stimuli are distortions of a single exemplar (i.e., prototype), responses to the never-before-seen prototype are more accurate than distance-matched controls (Reed, 1972).

A more recent approach to the problem built upon advances from the recognition and detection literature, resulting in the use of multidimensional stimuli that varied in a clearly perceptual space. Initially defined according to Gaussian perceptual noise around binary values, this method evolved into using distributions of continuously-valued stimuli. This method more readily maps onto schemes of behavior and has "ethological" validity to the sorts of categorizations made in everyday life. In categorizing a distant colleague as a faculty, graduate student or undergraduate, many more clearly perceptually parametric parameters can be extracted, such as franticness of motion and the degree of the freezing response at having been seen, than there are discrete identifiers. The generalization of formal methods from recognition and detection allowed for

analysis of categorization in a more rigorous fashion than previously. The use of simple parametric features were effective due to their well-studied nature and the benefited from knowledge in how the dimensions operate. Neural studies could pinpoint brain regions responsible for processing this exact data and single-cell studies with animal models could identify the tuning of these cells to exact dimensional values. Thus a rigorous analytic method for the categorization of continuous bi-dimensional perceptual stimuli was developed.

Categorization by Parametric Optimization

The use of Gaussian-distributed stimuli caused the categorization task to be a multidimensional generalization of Signal Detection Theory (SDT; also Statistical Decision Theory). In traditional SDT, the stimuli are generated from two unidimensional normal distributions (see Figure 1). One is the "noise" category distribution and the other the "signal," although this does not mean that this method is only appropriate or available for the problem of actually detecting signals in the presence of noise. In a common characterization of the problem in SDT, the observer gets a single sample and its value defined along the dimension of interest, and the observer has to declare whether the sample came from the signal distribution (category) or the noise distribution (category). This characterization maps well onto many decision making problems and frameworks, and is used widely in many fields, from communications (was that a signal from the emitter or just noise?) to security (are there illegal drugs in that luggage or is it just coffee?). The characterization lends itself to ready analysis, too. If the sample is more likely to have come from the noise distribution than the sample

distribution, then the optimal choice (i.e., strategy of highest accuracy) is to treat the sample as noise. If it is more likely to have been generated from the signal distribution, treat it as signal. Because of the unidimensional normal distributions, this decision making process comes down to identifying the singular value (threshold) that separates samples that are more likely to have been generated by the noise distribution and the samples that are more likely to have been generated by the signal distribution.

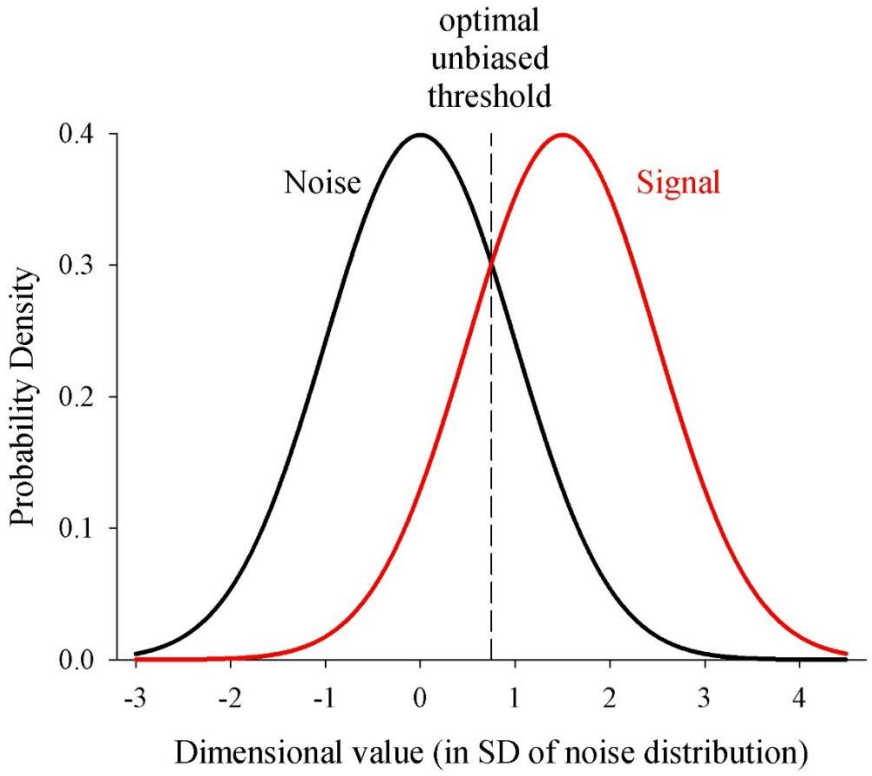

Figure 1. Depiction of Signal Detection Theory. Along the continuously-valued single dimension, the probability density is graphed for the noise distribution (black, left), and the signal distribution (red, right). The observer views the dimensional value, depicted here in terms of standard deviations from the noise distribution, and is required to assign the stimulus a label. The optimal unbiased threshold for labeling a dimensional value a Signal element versus a Noise element is also included, and reflects the point at which the likelihood of a sample coming from both distributions is equal.

In General Recognition Theory (GRT), which is the multidimensional extension of SDT, the same principles apply. Regarding the distribution of samples, instead of items that fall along a single dimension, there are multiple dimensions. For illustrative purposes, I will talk about two dimensions, but this can extend with appropriate caution to higher dimensions (Ashby & Townsend, 1986). The stimulus distributions are described by multivariate normal distributions, which are defined by a mean vector and a symmetric variance/covariance matrix. For a single multivariate normal distribution, there are three interesting classes of variance/covariance matrices: diagonal matrices that are multiples of the identity matrix, diagonal matrices generally, and nondiagonal but symmetric matrices (see Figure 2). In the case of a diagonal matrix that is a multiple of the identity, the variance in each dimension is equivalent, so if you were trace out lines of equal probability, the resulting contours would be circular. In the case of diagonal matrices that are not multiples of the identity matrix, the variances in each dimension are unequal and the covariance is zero so the contours of equal probability would be ovals, aligned with the axes. Finally, the contours of equal probability with only a symmetric variance/covariance matrix (i.e. covariances not equal to 0) would be rotated ovals (regardless of equality or inequality of the variance).

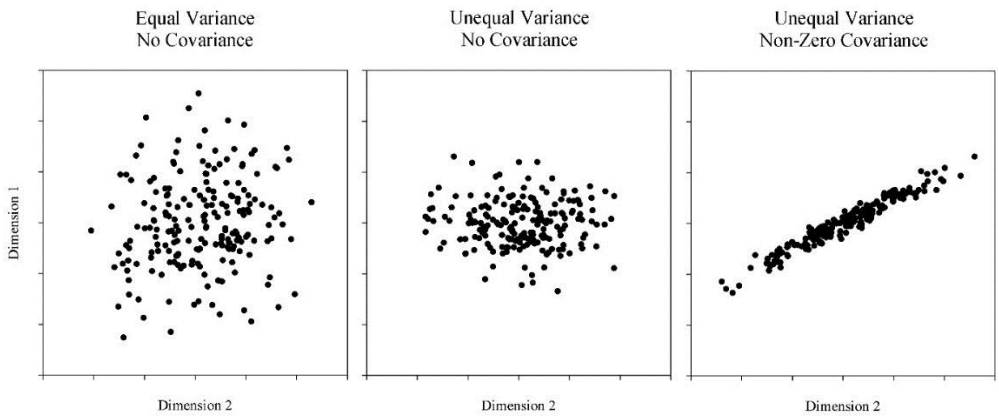

Figure 2. Different bivariate normal distributions. The three different panels depict the three interesting cases of bivariate normal distributions. The left panel depicts when the variance/covariance matrix is a multiple of the identity matrix. The 200 points are distributed with roughly circular lines of equal probability. The middle panel depicts when the variance/covariance matrix is a diagonal matrix, resulting in elliptical regions of equal probability. The right panel depicts a non-zero covariance, which appears as rotated ellipses of equal probability.

In GRT, there are two multivariate normal distributions that generate the samples, one which can ostensibly be called the noise distribution and another that can be considered the signal. Now they occupy different places on a 2D plane, like a sheet, and their likelihoods over the space can be plotted either as surface plots or as lines of equal probability. Again, for every point in the 2D space, the relative likelihood of a sample being generated at that point can be evaluated and the more likely model can be treated as the correct one. Whereas previously this corresponded to a singular threshold value where supra-threshold items could be labeled as being from distribution A and infra-threshold items labeled as from distribution B, the threshold is no longer a simple number and instead a curve of some sort called a discriminant. The actual values underlying the discriminant will follow a closed-form solution based on the means and covariances of the distributions being considered (Bishop, 2006, Section 4.2).

When the covariances are the same, the discriminant is linear, and when the covariances differ, the discriminant is quadratic (see Figure 3). Thus, in the first case, the discriminant line can be described from three parameters (some form of $0 = b_0 + b_1x + b_2y$) and in the second case the discriminant line can be described with up to six parameters (some form of $0 = b_0 + b_1x + b_2y + b_3xy + b_4x^2 + b_5y^2$). The optimal choice can be determined by the sign of the evaluation on the right hand of the equation. Note that with these multidimensional categorizations, there is no "supra-threshold" or "infra-threshold" concept; instead there is an A region and a B region, separated by a "rule." In general, these rules result from parametric classification schemes derived from generative classification problems. For the purposes of this manuscript, I will use the term rule to describe a curve through the stimulus space that divides the space into A and B regions.

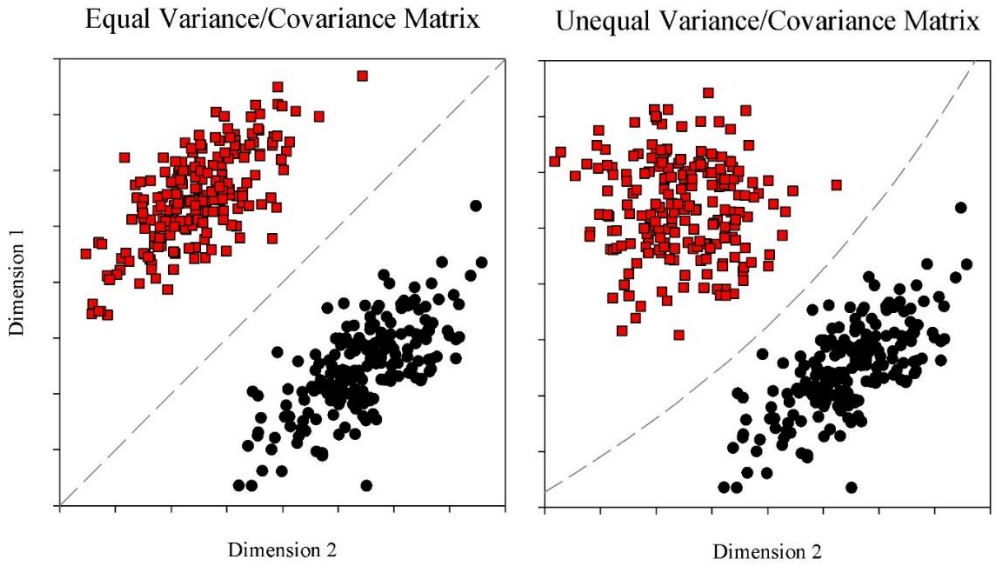

Figure 3. The optimal discriminant for two examples of a two category discrimination task. For the distributions on the left, the variance/covariance matrix used for each distribution is identical, whereas on the right, the distributions use different variance/covariance matrices. Note the optimal discriminant on the left is linear while that on the right is quadratic.

In discussing this optimal categorizing strategy, we have been dealing with behavior that would optimize accuracy. However, in psychological testing we know that not all observers follow optimal performance rules when probabilities are involved. In simple situations of 50/50 options, some animals in some conditions will choose to engage in only once choice option, a method that is called "maximizing." Other animals or in other situations, animals may choose to allocate responding among the responses in accordance to their relative likelihood of resulting in reinforcement, a method that is called "matching" (Herrnstein & Loveland, 1975). Thus, in this task (and most probabilistic situations), the optimal rule is to maximize because there is no additional information to draw on and the likelihood of a probabilistic response and a probabilistic situation both happening is far lower than of the probabilistic

situation alone. Some observers and some classifiers, however, may choose to "match" instead of "maximizing" and in this case, the question can be asked about what controls the gradated level of responding. There are two clear possible probabilistic methods when using these linear or quadratic rules. One method of non-optimal deciding is to base the decision on the distance to the decision bound. The items closer to the bound are also more likely to be from the other category than items distant from the bound. A second possibility is to use the value of the quadratic function, instead, as the distance to the bound is a mathematically more convoluted procedure and the value of the function behaves in similar ways. In one version of general recognition theory, maximizing is used in the decision process, but noise in the perception of the stimulus is used to generate probabilistic responding.

Optimal discrimination in the generative statistical scheme depicted creates a potentially curved rule which may govern behavior, and in limited circumstances (i.e., identical variance/covariance structures), the rule would become a line. The other distance-based method discussed, prototype theory, also posit a linear decision rule when given the alternative between two prototypes (Ashby, 1992a). This relationship can be shown to be one of super-setting; the GRT as described here is a more flexible categorization scheme that can also represent results generated via prototype theory.

Multiple Systems Models

Optimal discrimination rules have been suggested to exist in humans, although primarily as one of two separate mechanisms of categorization, one

explicit or rule-based and one implicit or non-rule based (Ashby, Alfonso-Reese, Turken, & Waldron, 1998). The rule-based system operates via optimal, GRT-like mechanisms with parametric classifiers, and the non-rule based system operates on associations in visual cortex. Evidence for this dichotomy has been found in different experimental schemes (Ashby & Maddox, 2005). In one scheme, identical categorization task structures using overlapping multivariate normal distributions were found to be affected by the stimulus dimensions used (Maddox & Ashby, 1993). When the dimensions allowed for ready comparison in a verbalizable fashion (i.e., both length dimensions of a rectangle), the participants were able to learn the categorization task more quickly than when they could not (i.e., one length dimension and one orientation dimension). Another important example is the demonstration of fast learning (e.g., less than 50 trials) for rule-based or unidimensional tasks (as in Figure 4, left) and slow learning of information-integration tasks (as in Figure 4, right; Smith, Beran, Crossley, Boomer, & Ashby, 2010). These figures demonstrate hypothetical category membership and boundaries from simple categorization tasks. In the rule-based (RB) condition, the stimuli can clearly be divided along a single dimension with a single threshold value, and the other dimension can be functionally ignored. In the information integration (II) condition, no such threshold exists in either dimension and the information in both dimensions is thus necessary at the same time to correctly categorize the stimulus.

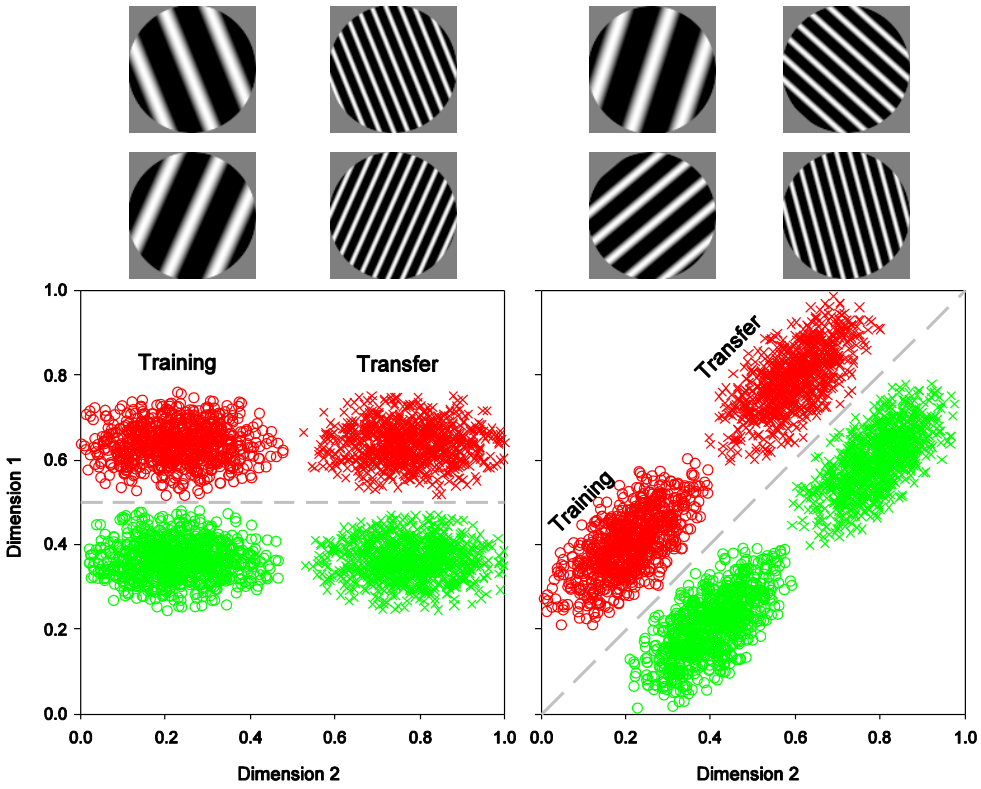

Figure 4: Distributions of the stimuli investigated in Experiment 1. The left graph depicts one of the Rule-Based training conditions, and the right graph depicts one of the Information Integration training conditions. Sample stimuli depict category means with "Dimension 1" corresponding to orientation and "Dimension 2" to spatial frequency. In these examples, the circles denote the training distributions, with the pigeons required to appropriately discriminate between the red and green distributions. The exes denote transfer test stimuli from Experiment 1, although only 72 of these points were tested. Correct category assignments for the transfer stimuli are based on the dashed gray line separating the distributions. Note that individual birds may have had these setups rotated by 90 degrees (RB) or 180 degrees (RB and II).

The investigation of categorization in animal models has also been a focus of interest. Two styles of comparative investigation are particularly salient. The first is the analysis of related species (i.e., non-human primates) and the other is that of different species (i.e., fish, amphibians, birds). Non-human primates placed in similar contexts show similar learning rate differences between the conditions (Smith et al., 2010). Pigeons, however, when tested with RB and II designs show no differences in the two conditions (Smith et al., 2011). This has been taken as

further evidence of multiple systems of learning in humans and non-human primates and evidence of a single learning system in pigeons. Because the two rates are equivalent in pigeons, the inference is that the pigeons only use associative mechanisms to learn the categorization task.

Another comparative investigation with humans and non-human primates has used analogical transfer to support the idea of a multiple systems categorization process. In this design, the traditional RB or II categorization task is developed, but then the observers are tested in novel regions of the stimulus space. This is called analogical transfer because the stimuli in the new region of space look starkly different than the stimuli in the original training region. In the RB condition, the observers are able to withstand the shift in the stimulus space, and in the II condition, the observers show no savings or benefit of learning (Casale, Roeder, & Ashby, 2012; Smith et al., 2015). This difference in savings has been suggested to be the effect of multiple systems and their context dependence.

Here are reports of two experiments with pigeons and one experiment with humans, examining the structure of generalization during "analogical transfer." In order to explain the performance observed, a new stimulus representation is proposed to underlie association-based learning. The ability of this representation to underlie learning is evaluated, and the possibility that the pigeons and humans use this method is evaluated.

Experiment 1

We trained pigeons on a two-alternative forced-choice (2AFC) categorization task to discriminate bi-dimensional sine-wave gratings. These stimuli have been well investigated in multiple species, especially when convolved with a Gaussian filter to produce Gabor patches (e.g., Jassik-Gerschenfeld & Hardy, 1979; Tappeiner et al., 2012). These stimuli in primates are designed to optimally activate a single patch of visual cortex and were used in previous investigations with humans, non-human primates, and pigeons (Smith et al., 2011). During training, the pigeons were only presented stimuli from one half of the total stimulus space. Half of the pigeons were trained using RB conditions and half of the pigeons were trained using II conditions.

During testing, the pigeons were presented with stimuli from the untrained half of the stimulus space. This technique should reveal what cognitive processes the pigeons use to discriminate the categories. If the pigeons employ two systems, one rule-based for rule-based training and one association-based for information-integration training, then we should see differential results during analogical transfer based on the training conditions. The pigeons with RB training should have perfect transfer, similar to humans, and the pigeons with II training should have little to no transfer. If the pigeons have a single, association-based learning system, as suggested by the previous Smith et al. (2011) findings, then neither group should show any analogical transfer. Finally, if the pigeons have a single, rule-based learning system and we assume that these dimensions are not

meaningful to them, allowing the RB and II tasks to be learned in the same way, then we should find perfect transfer in both groups.

Methods

Subjects

Eight male pigeons (*Columba livia*) were tested. The pigeons were housed and tested at 80-85% of their free-feeding weights, with ad libitum grit and water in their home cage, and they were experimentally naïve at the time of training. Prior to these experiments, they only received training to peck at a white signal on a display for food reinforcement. All animal procedures were reviewed and approved by Tufts University's Internal Animal Care and Use Committee.

Apparatus

Pigeons. A touchscreen (EZ-170-WAVE-USB) operant chamber was used to present video stimuli and record peck responses. Stimuli were displayed on an LCD computer monitor (NEC LCD 1525X; 1024x768, 60 Hz refresh rate) situated just behind the touchscreen. Mixed grain reward was delivered via a central food hopper positioned beneath the touchscreen. A houselight in the ceiling was constantly illuminated, except during timeouts.

Stimuli

The stimuli in these experiments were sine-wave gratings that varied in spatial frequency and orientation, designed after those used in (Smith et al., 2011). These stimuli were composed of a solid gray square with a circular aperture that contained the sine-wave grating (see Figure 4, top). Stimuli were generated using

ImageMagick (http://www.imagemagick.org). Each image was a 100 pixel × 100 pixel square.

The category distributions were defined using the dimensions of spatial frequency and orientation, and they were designed after those used in (Casale et al., 2012). These bivariate normal distributions were generated using MATLAB (MathWorks) with fixed mean and covariance parameters as described in Table 1, and the sampling was restricted such that the Mahalanobis distance for all points was less than 7.5. The resulting distributions are depicted in Figure 4. The dimensions mapped onto spatial frequency with a minimum of 0.297 peaks per image (i.e. normalized 0) up to a maximum of 12.31 peaks per image (i.e. normalized 1) and orientation with a minimum of 4.41° and a maximum of 173.31°, with 0° corresponding to horizontally oriented bars and positive angles corresponding to counter-clockwise rotation.

Distribution		μ_1	μ_2	σ_1^2	σ_2^2	cov_{xy}
RB Training	A	36.28	22.50	20.93	91.72	00.00
	B	63.72	22.50	20.93	91.72	00.00
RB Transfer	A	36.28	77.50	20.93	91.72	00.00
	B	63.72	77.50	20.93	91.72	00.00
II Training	A	20.85	40.26	56.32	56.32	36.71
	B	40.26	20.85	56.32	56.32	36.71
II Transfer	A	59.74	79.15	56.32	56.32	36.71
	B	79.15	59.74	56.32	56.32	36.71

Table 1. Distribution parameters for training and transfer distributions in Experiment 1. The values listed here indicate the means, variances, and covariance between the dimensions in the normalized (0 to 100) stimulus space. Note that training and transfer designations here are only accurate for a subset of the birds; for the remaining subjects, the data need to be rotated around the point (50, 50) by 90° or 180°.

Procedure

Pre-training. The pigeons were trained to peck at a centrally located, white, 2.5 cm *ready signal* prior to the start of this experiment. They were then trained to peck at a sample when it appeared in return for food on a fixed-ratio schedule. Each training trial used a randomly selected stimulus from either training distribution as the sample. The FR to this sample was slowly increased to accommodate the final variable-ratio schedule. After they were pecking reliably to the sample, we began training the choice key response. After completing the FR on a trial, a single red (RGB 255, 0, 0) or cyan (RGB 0, 255, 255) choice alternative positioned 275 pixels to either side of the sample appeared, and one peck to this alternative resulted in food. Once the pigeons were pecking reliably in all phases of the trial, discrimination training began.

Training. On every trial, a centrally-located, white, 2.5 cm *ready signal* appeared. When the pigeon pecked this signal, the signal was replaced with a sample stimulus. The sample stimulus was a randomly selected stimulus from the two categories such that the mean and variance parameters as estimated on a session-wise basis matched the values listed in Table 1. After pecking at the sample stimulus on a variable ratio schedule that was uniformly distributed between 13 and 15 pecks, choice alternatives appeared on both sides of the sample. The red and cyan choice alternatives corresponded to the category of the stimulus. A single peck at the red choice alternative indicated that the pigeon categorized the sample as a "red" category stimulus, and a single peck at the cyan choice alternative indicated that the pigeon categorized the sample as a "cyan" stimulus. Each alternative appeared equally often on either side of the display. Correct choices resulted in access to mixed grain (i.e. food reward) for 2.5 s (for one subject, this was increased to 4 s), and incorrect choices resulted in an 8-s timeout during which the houselight was also turned off. A 3-s inter-trial interval then followed, and then the ready signal would appear to allow the next trial to be initiated. A correction procedure was used from the beginning of training such that incorrect responses resulted in the trial being re-presented until the correct response was given. Only the first trial in this sequence was considered for accuracy metrics.

Four of the pigeons were trained in the RB condition and four in the Information Integration II condition. The distributions used are listed in Table 1. Half of the birds in each case were trained using the "lower" distributions, where

the features of interest occupy the lower portion of the total values used, and the other half were trained using the "higher" distributions. Training was considered completed when the pigeon achieved an accuracy of at least 80% for five sessions (non-consecutively).

Transfer. The pigeons were then given six sessions of testing with the appropriate transfer distributions (i.e. birds trained on "high" distributions were given transfer tests from the corresponding "low" distributions). For transfer tests, a subset of six stimuli from each transfer distribution was randomly selected each session for testing (72 total test trials, 36 for each category). All responses for these test trials resulted in food reward and no time out (i.e. non-differential reinforcement).

Results

Training: Seven of the eight pigeons learned the discrimination to criterion, requiring from 16 to 50 sessions to learn the task. The speed of this acquisition did not differ between the birds in the two training conditions. The pigeons in the II condition required 24, 25, 26, and 37 sessions to complete training, and the pigeons in the RB condition required 16, 26, and 50 sessions to complete training. The one pigeon who did not learn was in the RB condition and did not meet criterion despite 100 additional sessions of training. Altogether, the birds in the II condition averaged 28 sessions to criterion while the birds in the RB condition averaged 30.7 sessions, which is negligibly different ($t(5) = 0.47$, $p = .782$). Therefore, we consider this result a replication of the of the previous pigeon data.

Analogical Transfer: In the transfer to the new distributions, the pigeons' performance was systematically related to their training condition. As shown in Figure 5, the four II pigeons performed nearly below chance while the three RB pigeons performed above chance. Each pigeon's performance was evaluated using a binomial test with all 72 test trials, revealing that all three pigeons in the RB condition were significantly above chance (46, 46, and 58 trials correct; $p = .012$, $p = .012$, $p < .001$). The same analysis showed that three of the pigeons in the II condition were not performing significantly different than chance accuracy while the fourth pigeon was below chance (23 trials correct, $p = .001$).

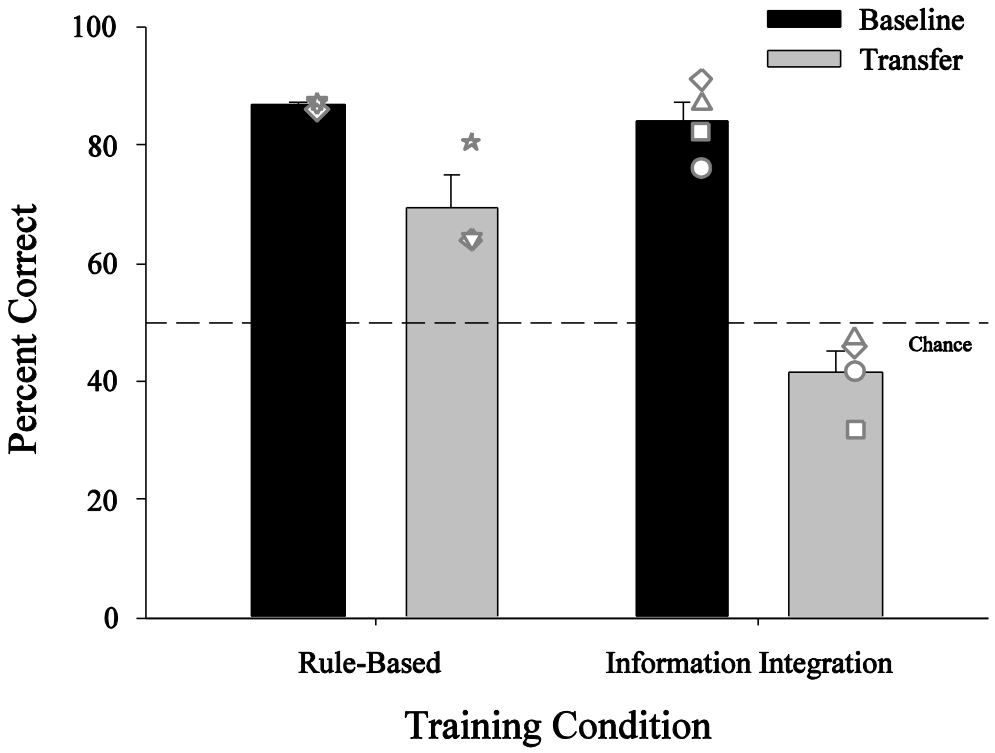

Figure 5: Analogical transfer performance from Experiment 1. Error bars depict standard error. Individual birds are depicted slightly offset from the bars using bird-unique symbols.

Discussion

During the test of analogical transfer, the pigeons exhibited similar transfer effects as a function of task learning as previously shown by humans. Like humans, the pigeons who learned the RB task extended their learning beyond the training portion of the stimulus space along the length of the irrelevant dimension. This resulted in greater-than-chance categorization accuracy with the novel distributions of stimuli. Like humans, the pigeons who were trained with an information-integration task were unable to extend their learning. They showed no discrimination during analogical transfer, being at chance statistically and all seemingly below-chance numerically (though non-significantly, $t(3) = 2.4$, $p = .09$). In previous investigations, the relative success with RB training has been suggested to be an example of analogical rule transfer.

Given these two patterns of transfer, partial transfer for the RB trained birds and no transfer for the II trained birds, this should be considered evidence for two systems of categorization in pigeons, as in primates. The human and non-human primate investigations used this pattern as support for two systems. However, the previous pigeon investigation and the similar equivalence of learning rate for RB and II in this experiment are still most consistent with a single system. How are these conflicting interpretations for the pigeons to be resolved?

Two possible methods of resolution are readily apparent. The first method is to assert that birds, similar to humans and non-humans primates, use a rule-based solution with RB task and a non-analytic solution for the II task. This

would allow linear transfer of the rule to novel regions of stimulus place. It would not easily explain why RB and II birds always learn at the same rate, however. Why is there not the same rule-based benefit in acquisition as seen with humans then? How are dimensions meaningful to one categorization system during analogical transfer but not during learning? A second resolution is that both RB and II task are learned by pigeons with same non-analytic association-based system. This would explain why the learning rates are so similar, but this would not account for why the two groups show different degrees of "analogical" transfer. How does a single learning system show no dimensional benefit in training, but then show it in the analogical transfer? One possibility is that the unidimensionality of the stimuli in the RB case permitted a greater degree of generalization than is permitted with the II task. To better understand to these alternatives, Experiment 2 more precisely evaluated how the pigeons categorized regions of the stimulus space.

Experiment 2

The previous experiment leaves open some questions about how the pigeons learned the discrimination and then how that training altered their responding to the untrained portion of the stimulus space. The scope of potential models that describe the pigeons' discrimination is large, encompassing rule-based and association-based models of categorization. Thus far, there is little data to differentiate between the various models, and what little data we have is concentrated in a two large, diffuse areas. In order to detangle the many possible

categorization schemes, we needed to expand and make more precise our knowledge of how the pigeons response to stimuli from multiple regions across the stimulus space. To accomplish this, we tested small bivariate normal clusters to evaluate regions of the space that we suspected would be helpful in understanding the pigeons' categorization of the displays. The specific cluster areas were chosen to elucidate the pigeons' overall response patterns with the different possible rule-based and associated strategies in mind. Thus, in this experiment, we tested small, focused clusters distributed throughout the stimulus space to examine the effectiveness and reliability of the pigeons' performance in order to decide among the multitude of categorization methods.

Methods

Participants and Apparatus

The seven successful pigeons from the previous experiment were used in this experiment, and the same apparatus was used. For part of this experiment, a second similar operant chamber was used.

Stimuli and Procedures

Stimulus values for the ten new clusters tested are listed in Table 2 and graphically depicted in Figure 6. The standard deviations were fixed to 3.0 in both directions with zero covariance for these clusters.

	Information Integration		Rule Based	
Distribution #	μ_1	μ_2	μ_1	μ_2
1	59.74	79.15	77.5	63.72
2	79.15	59.74	77.5	36.28
3	45.00	95.00	78.28	85.36
4	95.00	45.00	78.28	14.64
5	35.00	65.00	50.00	71.21
6	65.00	35.00	50.00	28.79
7	25.00	85.00	57.07	92.43
8	85.00	25.00	57.07	7.57
9	10.00	60.00	28.79	85.36
10	60.00	10.00	28.79	14.64

Table 2. Distribution parameters for the transfer clusters in Experiment 2. The values listed here indicate the means in the normalized stimulus space. Note that these are only representative for a subset of the pigeons; for the remaining subjects, the data need to be rotated around the point (50, 50) by 90° or 180°.

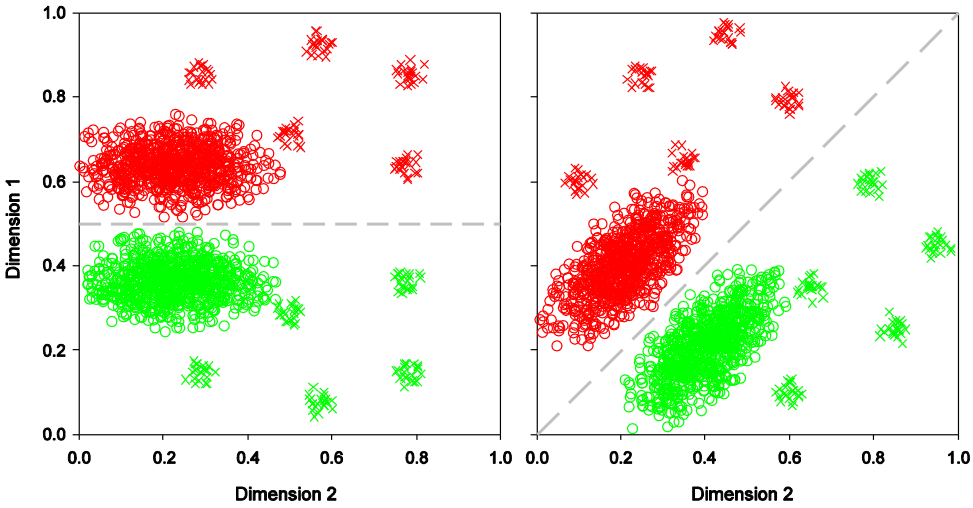

Figure 6: Distributions of the stimuli investigated in Experiment 2. The left graph depicts one of the Rule-Based conditions, and the right graph depicts one of the Information Integration conditions. As in Figure 4, the circles denote training stimuli, and the exes denote the transfer tests. Note that individual birds may have had these setups rotated by 90 degree (RB) or 180 degrees (RB and II), and that category assignment for transfer stimuli are based on the dashed gray line separating the distributions.

Each test session contained ten randomly inserted probe trials, one from each of the ten clusters. As before, these test trials were non-differentially reinforced. Ten test sessions were conducted, so that each pigeon received a total of 100 test trials, equally divided among the ten new clusters. For analysis, "correct" category assignments for these transfer clusters were determined according to the extension of the linear rule dividing the training categories.

Results

These data will be analyzed in two ways. First, we will discuss the overall accuracy of the pigeons, because that is a naturally meaningful metric and relates to our analysis from Experiment 1. Second, we will evaluate and explore the

patterns of their categorization as it relates to both the location of the training clusters and the pigeons' overall evaluation of stimuli from that cluster.

All of the pigeons demonstrated above-chance transfer of their trained discrimination to these new clusters. Figure 7 depicts categorization accuracy for both groups. Binomial tests of the pigeons' accuracy shows that all seven pigeons were significantly above chance (number of correct responses $\geq$ 61, *p*s <.02). While it may be surprising that the pigeons who previously did not show analogical extension were able to produce "correct" transfer here, the pattern of categorized and non-categorized clusters is most telling. The different training conditions resulted in different patterns of transfer. Figures 8 and 9 report the results of the transfer tests as they relate to the normalized stimulus space for the birds in the RB and II groups, respectively. The dashed gray line is the ideal extended linear rule that would divide the training categories, which was used to determine "correct" category assignments. In these displays, the different bird conditions have been normalized to facilitate inter-bird comparisons (i.e., all training conditions in the same half of stimulus space, all category assignments made to look the same). The transfer stimuli clusters are shown with the number of "A" or (normalized) "red" responses (out of 10) placed at the cluster center. The graphs were manually annotated to highlight where the pigeons made reliable categorization judgments (greater than six or less than four "red" responses) to better emphasize the patterns as they relate to the overall space.

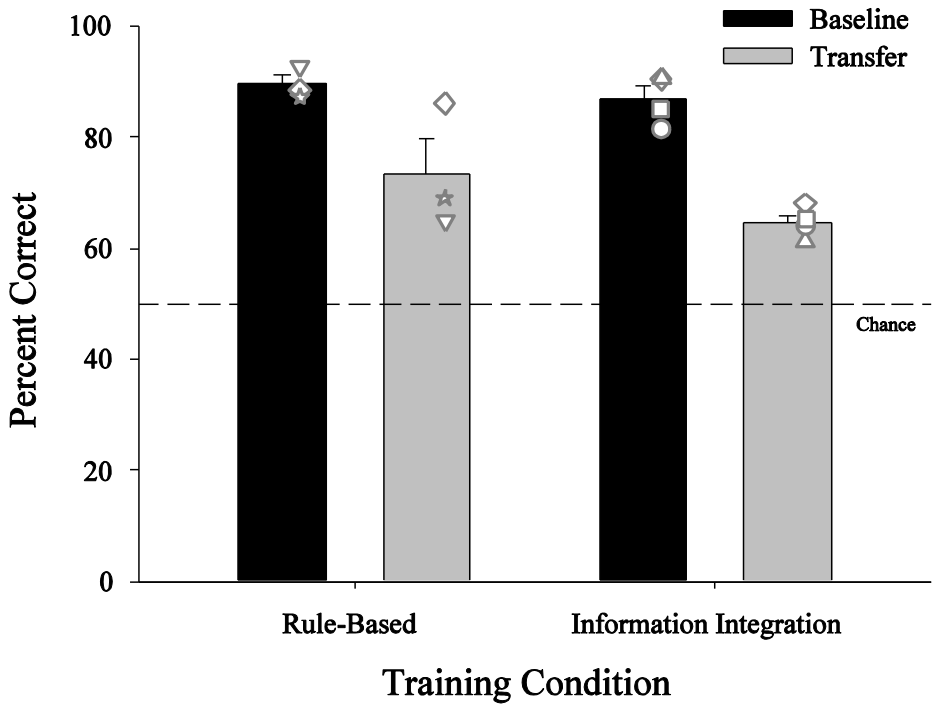

Figure 7: Analogical transfer accuracy from Experiment 2. Error bars depict standard error. Individual birds are depicted slightly offset from the bars using the same bird-unique symbols as in Figure 2.

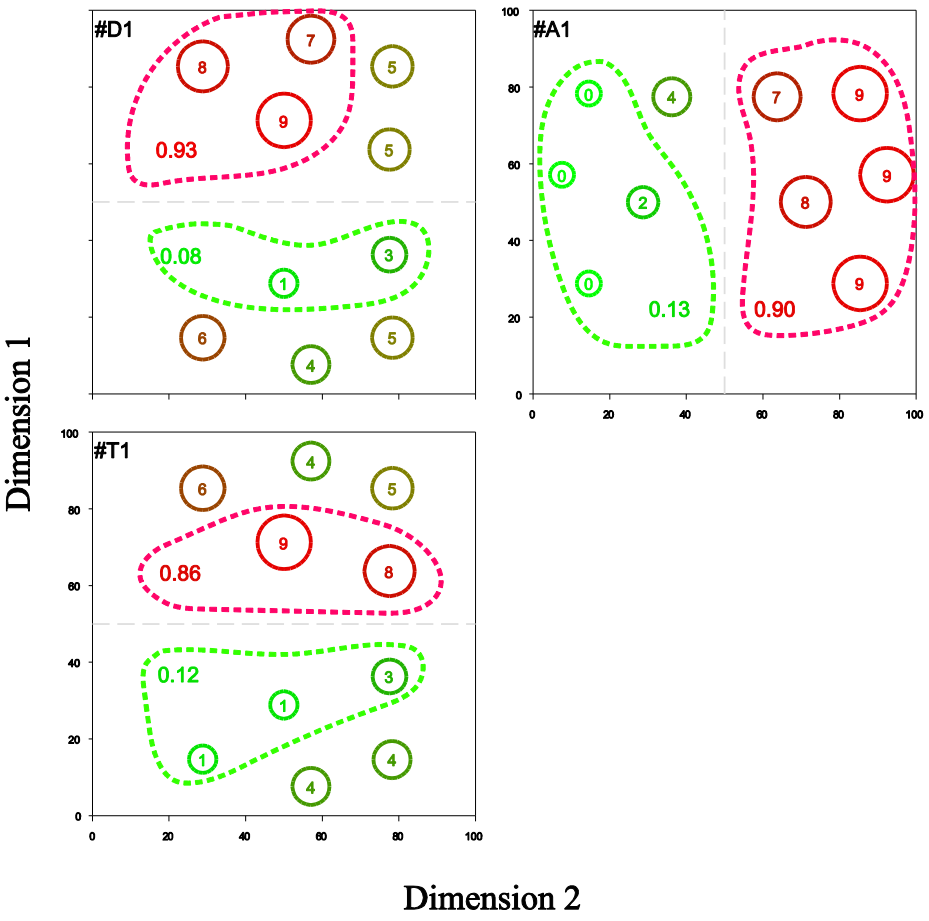

Figure 8: Analogical transfer performance for the birds in the Rule-Based condition from Experiment 2. Values are positioned where the clusters of interest are positioned. Integer digits (redundantly coded with color and surrounding circle size) indicate number of times out of 10 that response category "red" was selected for items from that cluster. Decimal fractions indicate the same but for the proportion of baseline trials that generated a response of category "red". The dashed gray lines indicate the space-dividing category line. The dashed red curve is a manually applied annotation of the figure to highlight the "red" category for each bird, and the dashed green curve is a manually applied annotation to highlight the "green" category for each bird. Note that some of responses and assignment have been rotated and/or flipped to provide a more understandable, uniform appearance to the task.

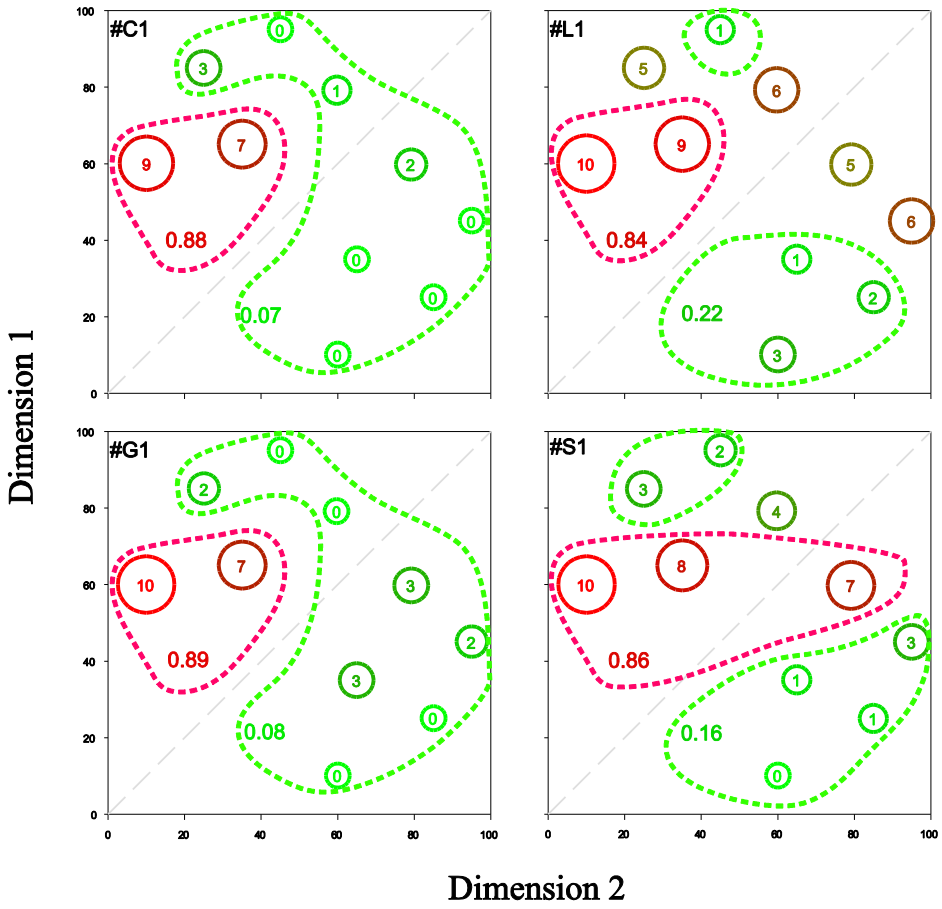

Figure 9: Analogical transfer performance for the birds in the Information Integration condition from Experiment 2. Values are positioned where the clusters of interest are positioned. Integer digits (redundantly coded with color and surrounding circle size) indicate number of times out of 10 that response category "red" was selected for items from that cluster. Decimal fractions indicate the same but for the proportion of baseline trials that generated a response of category "red". The dashed gray lines indicate the space-dividing category line. The dashed red curve is a manually applied annotation of the figure to highlight the "red" category for each bird, and the dashed green curve is a manually applied annotation to highlight the "green" category for each bird. Note that some of responses and assignment have been rotated and/or flipped to provide a more understandable, uniform appearance to the task.

The RB training in Figure 8 shows two prominent response patterns. Pigeon #A1 (top right; 86% overall accuracy) was the only pigeon successfully trained on the spatial frequency rule, and in this test of analogical transfer, his categorization performance seems to be clearly related to the dashed gray line that

represents the extended linear rule. If we consider the variability in responding in each of the clusters, his transfer data are clearly related to distance from the space-dividing category line. Transfer distributions distant from the line are relatively absolute in terms of responding, and those distributions closer to the line engender more mixed responding. The same is not true of pigeons #D1 (top right, 65% overall accuracy) and #T1 (bottom right, 69% overall accuracy), who had rule-based discriminations of orientation. While responding is still clearly related to the category line, the variability within transfer clusters seems unrelated to distance from the line, or possibly inversely related. Overall, these two pigeons' categorization seems more constrained to the training values than pigeon #A1.

The II training yielded a different set of results. These birds' categorization performance was seemingly unrelated to the dashed gray line that represents the extended linear rule. Here, again, we see two patterns of data. On the left of Figure 9, pigeons #C1 (68% overall accuracy) and #G1 (61% overall accuracy) both show responding that look fairly similar to having learned a single category's prototype. If a transfer item was sufficiently close to the learned "A" category, there was a fair amount of "A" selection. Accordingly, transfer items not in that region of space were primarily classified as "B." Pigeons #S1 (65% overall accuracy) and #L1 (64% overall accuracy) on Figure 9 right may show something similar, but the annotations emphasize the fact that there are specific clusters that challenge this simple story. For pigeon #L1 (top right), there is a clear band of ambiguity where A responding otherwise dominates, and yet at the

top of the graph is a cluster where nine of 10 responses were “B”, suggesting that beyond the trained “A” distribution and opposite the core “B” distribution, there was a high confidence that the stimuli were from category “B.” For pigeon #S1 (bottom right), the rightmost cluster in the annotated region is clearly more distant from the “A” training distribution than the clusters above the annotated red region, but the clusters above the region received primarily “B” responses. Some hints of this may also exist in the left two II pigeons, but for all four II pigeons, whatever the hypothesized rule is, categorization performance and “confidence” of responding is not related to the extended linear rule.

Discussion

The results revealed that all seven pigeons showed systematic behavior to the more widely-spaced novel transfer. Only one of these pigeons (#A1; in the rule-based condition) demonstrated transfer in a manner consistent with rule-based responding. The other two pigeons in the rule-based training condition demonstrated better performance for stimuli closer to the hypothetical rule-bound rather than improving as the stimuli were more discriminable (i.e., further from the rule-bound). In contrast to the analogical test in Experiment 1, all four of the pigeons in the II condition here showed systematic, and discriminative transfer to novel items. Furthermore, all four pigeons agreed in terms of how they categorized six of the ten transfer clusters. Five of these six agreed upon clusters were relatively close to the training distributions, but the sixth (topmost cluster in Figure 9) was on the other side of the stimulus space from its categorized training distribution.

While the results for the II pigeons appear to conflict with the results from Experiment 1, if we consider the two clusters in this experiment that are most similar to the transfer distributions from the previous experiment (top right clusters in Figure 9), the data are visibly consistent. In those distributions, the responding is either completely biased towards one stimulus or fairly non-discriminate with a slight bias towards a "reversal" of responding exactly as found in Experiment 1. This suggests that the failure to find transfer in Experiment 1 in the information-integration condition was a direct result of the location tested by specific distributions, and not due to a failure of the learned response to generalize to novel regions of the stimulus space. However, the birds' behaviors do not conform to a simple model of similarity based on distance in this stimulus space. In Figure 9, all four birds show patterns of transfer to items that are far away from the original location of the training stimuli. Given this information about how the pigeons categorize and how their transfer performance systematically varies over the stimulus space, we began to consider how the different models of categorization accord with the results.

Formal Methods

Overview

Investigations of categorization are regularly presented in light of specific mathematical models that employ different conceptual schemes. This mathematical approach allows for a more unbiased evaluation of the different categorization frameworks. In order to evaluate how well our data are predicted

by different analytic and non-analytic categorization schemes, we do the same. We considered two broad types of models. The first involves different analytic categorization schemes, while the second involves different non-analytic schemes. Following brief explanations of each, we simulate the experimental methods thus far with a novel representation in the association scheme. We then discuss model-fitting methods and then present and discuss the outcomes of this modeling with artificial and experimental data.

Analytic Overview

There are many analytic schemes available to model categorization (Ashby, 1992b), but we will focus on two broad classes of current interest: prototype learning and discriminant-based categorization (COVIS/general recognition theory). The two classes have some common underlying assumptions. They both assume that the animal (human, non-human primate, pigeon, rat, etc.) perceives the stimulus in some dimensionalized perceptual space. They also assume that every point in that perceptual space has some category membership(s) value that informs responding. These category memberships are determined according to a distance metric within the perceptual space that can usually depend on the stimulus features being evaluated (i.e., city-block distance or Euclidean distance as appropriate). Responding may be deterministic based on those evaluations of category membership(s), so that re-presenting a stimulus will result in the same response every time, or alternatively responses could be probabilistically emitted as a function of category membership(s) value. Thus, both can identify contours in the perceptual space where the decision is most

"uncertain" – minimal distance to the bound if deterministic or close to equiprobable if probabilistic – which are referred to as the category bounds.

The differences between prototype theory and discriminant theory are the source and form of these category bounds (Ashby, 1992b). In prototype theory, prototypes are generated by exemplars from within that category, although how they are learned is a matter of debate. The distance between the prototype and the sample is the controlling factor in categorization behavior, with some critical cutoff between 'A' (sample is in the prototype's category) and 'not A' (sample is not in the prototype's category). Alternatively, if there are two prototypes representing separate categories, a stimulus in perceptual space is categorized to belong to the category with the closer prototype. Other versions of the theory also posit multiple prototypes (forming a spectrum with exemplar theory), with decision bounds that function largely the same way.

Discriminant theory works with the stimulus distributions in perceptual space similar to General Recognition Theory. Here, the theory assumes that the stimuli are distributed according to multivariate normal distributions, so that at all points in the perceptual space, there is a possibility (however remote) that a stimulus came from either category. Categorization behavior in this scheme is controlled by the relative likelihoods that a stimulus came from the distribution of category A instead of category B. At many points in the perceptual space, the likelihood that a stimulus belongs to the category A distribution will equal the likelihood that it belongs to the category B distribution. These points generate iso-probability contours that can be described by simple discriminant functions.

Variations on sensitivity to distribution features (i.e. variances, covariance, means, skewness, etc.) in this space result in different iso-probability contours.

Analytic Categorization Formalization

To formalize the statements from the introduction to model fitting, both the prototype theory and the discriminant theory state that a stimulus (S) can be represented in a (veridical) stimulus space and an internal perceptual space. Without loss of generality, we use two dimensions: specific stimulus *i* will be designated $S_i = (x_{1i}, x_{2i})$. Its representation in perceptual space is a fair approximation of the veridical stimulus, plus some noise from sensors: $P(S_i) = P_i = (p_{1i} + \varepsilon_{1i}, p_{2i} + \varepsilon_{2i})$. Given research on generalization as well as examining neural decay functions, we will assume that the noise is distributed according to a Gaussian with zero mean and uncertain variance (Ghirlanda & Enquist, 2003). In this same perceptual space, there is a function that partitions in the space into distinct "A" or "B" regions, which we will designate F(S). Prototype theory also posits a prototype or prototypes $\mathbf{E}_A$ and/or $\mathbf{E}_B$ that represent the categories. If using a version of exemplar theory similar to A/Not-A categorization, the function F(S) is defined as $F(S) = D(P,\mathbf{E}) - C$. In this, C is a (possibly noisy) criterion, and D is a function that defines the distance between points in the perceptual space (e.g. city-block distance, Euclidean distance, as appropriate). Alternatively, if two categories are represented, then $F(S) = D(P,\mathbf{E}_A)/ D(P,\mathbf{E}_B) - 1$. In either situation, if F(S) evaluates to a negative value, response A is produced, and if F(S) evaluates to a positive value, response B is produced. In both instantiations, the curve produced at F(S) = 0 is the discriminant rule or decision bound described above.

In contrast, discriminant theory posits the existence of a series of discriminant functions or rules **R**. In most cases, **R** is comprised of a single rule that can be described as a polynomial combination of terms in the stimulus space: for example in two dimensional space, rule *i* may be $R_i = A*p_1^2 + B*p_1*p_2 + C*p_2^2 + D*p_1 + E*p_2 + F$. This rule describes a curve when its roots are evaluated, and the curve described by the roots of this function is the discriminant function that determines responding. Thus, if a stimulus is evaluated to be in the (e.g.) positive region of this function response, an A response would be generated, and if a stimulus is evaluated to be in the negative region, a B response would be generated. Again, this yields a function F(S) = D(P,**R**) – C, although here C is clearly 0 but needs to be indicated to allow for the noisy criterion. In the discriminant based theory, D is the appropriate distance function between the percept and the bound. Both methods can also allow for multiple exemplars or rules, which increase the complexity of the decision method, but those are omitted here. Finally, while the above methods describe a deterministic system, if a probabilistic method is used, then the likelihood of responding A is increased using an appropriate function (e.g. logistic) as the distance from the exemplar-based or discriminant-based bound increases (although this coincides with the various noise sources in the case of the discriminant theory).

Association Overview

The above two classes of categorization are analytic methods that posit a "rule," more than just the line through the space. The "rule" is a higher-level cognitive representation of a strategy. An alternative to these rule-based theories

is the relatively simple procedural learning method. One implementation of this is the procedural learning system of the SPEED model which uses a grid of neural network units to represent the perceptual space (Ashby, Ennis, & Spiering, 2007). When a stimulus is presented, a region of the units in that part of the perceptual space is activated according to a radial basis function (see Figure 10, left, for a visual depiction of this activation). This grid of units therefore represents a configural activation of features, similar to the configural representation used in other models of animal cognition (George & Pearce, 2012; Pearce, 2002). This entire grid of units is connected to a much smaller set of units (maybe even just one unit) whose activity level generates the category response. Over multiple presentations, as a result of feedback-based learning, the weights are adjusted so that ultimately the correct response is produced when the stimuli are presented.

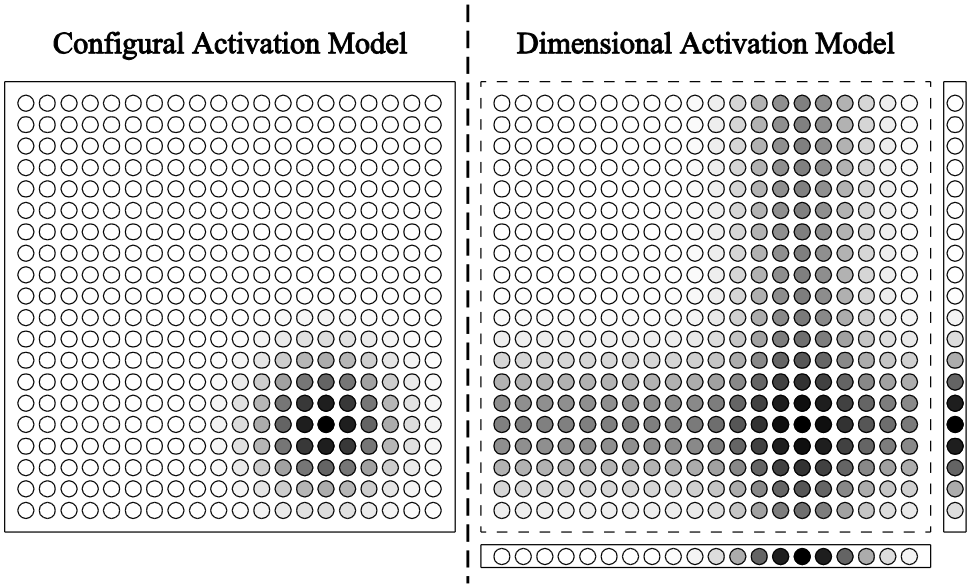

Figure 10: A depiction of stimulus representation for the configural activation model and the dimensional activation model, using a 20-unit based grid. Each unit is represented by a circle, with highly activated units filled with black, less activated units filled with grey, and inactivated units filled with white. The configural activation model uses a radial basis function to compute the activation of units, while the dimensional activation model uses a Gaussian decay. Both representations are depicting the activation from the same external stimulus. Note that the grid on the right with the dashed boundary is for comparison purposes only – it does not accurately reflect the stimulus representation. The dimensional activation model is best represented by the two lines of units to the right and below the grid on the right.

Finally, we considered an alternative version of this association model. This model uses a separate set of units *for each feature* in the stimulus to represent activation in just that dimension. Thus, instead of a grid of units, we end up with two lines of units, with each unit representing different values within each dimension (see Figure 10, right). When a stimulus is presented, it generates an activation in each dimension. The activation of these lines of units are then connected to a much smaller set of units (maybe even just one unit), and through feedback-based learning the weights are adjusted to generate the correct response. For clarity, we will refer to this as the *dimensional activation* model and treat it

separately from the model in SPEED, which we will refer to as the *configural activation* model. These names identify that the critical difference is the use of a configural representation in the SPEED model versus the dimensional representation described here. An illustration of the difference between the two models is shown in Figure 10. Note that while the units are displayed in the context of the grid of stimuli used in the configural activation model, this is only to emphasize the differences in activation patterns in the model and not to suggest that there are configural units become activated in this fashion. If these dimensional units representing the unbound activation within each dimension were later bound together for a different task, how their activation would pattern is an open question.

These association methods discussed do not precisely generate bounds in the same manner as the analytic methods described above. Nevertheless, if the decision layer has a single unit, then its activation level will reflect the relative likelihood of the distributions in the stimulus category distributions. In the case of the configural activation model, this reduces to the GRT optimal classifier bounds for regions near and between the sampling distributions. The critical difference between the configural activation model and GRT bounds relates to stimuli far from the training region (i.e., during analogical transfer). GRT posits that these distant points would be responded to according to the rule, but the configural activation model would treat them differently depending on how far from the training distribution the sample stimulus is. For regions near the training distributions, largely trained units will be activated and so the response by the

system will be systematic and correct. For regions further from the training distributions, the weights that connect those units to the decision layer were never trained or modulated to any serious degree, resulting in arbitrary (although potentially not-chance-level) responding. Thus, for the space between the two category distributions, the GRT model and the configural activation model would (asymptotically) yield similar results, but in the space outside the two category distributions, the configural activation model has less systematic performance because of untrained areas with non-differential responding.

The difference between the dimensional activation and the configural activation model is readily apparent in Figure 10, but understanding how the dimensional activation model changes the decision process during our analogical transfer testing is not as self-evident. The dimensional activation model and the configural activation model would most likely yield similar results for the region between the training distributions. However, important differences emerge with increased distance from the training distributions. Because of the dimensional activations, the within-dimension distance from the trained values would control responding. This means that outside of the training region, different regions of the space support different levels of performance, so the responses can be quite different from that proposed by the configural activation model. As long as a stimulus is similar in some of its dimensional values with a training stimulus, the dimensional activation model would predict systematic behavior. Thus, this dimensional activation model could readily produce some of the aberrations depicted in Experiment 2. The “distant” cluster that was readily and

systematically categorized by the pigeons was only distant from the training distributions along a single dimension while remaining relatively close in the second dimension. These two models would predict vastly different outcomes as a result: the configural activation model will likely falter due to the distance (i.e., predict chance-level behavior) while the dimensional activation model may predict systematic responding.

Association-Based Categorization Formalization

We start again with stimulus S_i, but now its representation in perceptual space is no longer provided by a simple formula. Instead, the perceptual representation of a stimulus is the activation it generates in the neural network units of the model. If we assume Euclidean distance for the distance metric and Gaussian decay for all units' sensitivity, the activity level for each of these activated units can be described using Gaussian distributions. In the configural activation model, each configural unit W_{ab} has optimal responding to dimension 1 value *a* and dimension 2 value *b*. Its activation in response to S_i would be computed by $A_{ab}(S_i) = \varphi(D(S_i, [a,b]), 0, \sigma_{ab}^2)$, where $\varphi(X, \mu, \sigma^2)$ is the probability density function of a Gaussian with mean μ and variance σ^2 evaluated at X. The use of a radial basis function instead of a normal distribution curve was also evaluated, and that showed no fundamental differences in its results. Contrastingly, in the dimensional activation model, each dimension unit U_a or U_b is sensitive only to the values within their dimension. Their activations for $S_i = (x_1, x_2)$ are given by $A_a(S_i) = \varphi(D(x_1, a), 0, \sigma_a^2)$ and $A_b(S_i) = \varphi(D(x_2, b), 0, \sigma_b^2)$; if we wanted to continue representing this in the grid space of the SPEED model

(i.e., as in Figure 10), this would provide us with something similar to $A_{ab}(S_i) = \varphi(D(x_1, a), 0, \sigma_a^2) + \varphi (D(x_2, b), 0, \sigma_b^2)$. The output of the activations from all units would map onto a hidden layer, and from there to a category decision unit. These units could be activated according to a cumulative distribution function of the standard normal Gaussian (Φ) or a logistic function. The evaluation of the final unit will correspond to the likelihood of emitting (e.g.) an "A" response.

Simulations 1

Simulations provide insights into how these models of categorization behave when their assumptions are tested rigorously and repeatedly. The analytic methods have been investigated and functionally simulated in the COVIS model previously. In order to better understand how the two different associative models of categorization will learn the task and generalize during these analogical transfer tasks, we simulated functional learning networks using MATLAB. The goal of this task was to determine whether a neural network with representational units like those described above could in fact solve the discrimination. Next, we will evaluate two critical questions: 1) are the learning rates of the associative methods different when presented with the rule-based training condition versus the information-integration training condition, and 2) what performance do the models yield during the transfer tests from these experiments? Finally, we will explore methods of determining which underlying representational model generated a set of data, and we will ultimately carry that forward to evaluate the behavioral data.

The association models have many components to them, organized into three layers. The first layer is an input layer, the organization of which is the fundamental difference between the two models proposed here and formalized above. In the dimensional activation model, this layer comprises two sets of units; and in the configural activation model, this layer comprises a grid of units. The second layer is termed the "hidden layer" because it is not exposed to the inputs nor is it an output. The size of this layer is proportional to the complexity of the neural network. As this layer grows, the neural network gains the ability to subdivide the input space into more chunks. A familiar demonstration of this is a neural network for categorization of binary valued bidimensional stimuli. A neural network with one unit in the hidden layer would not be able to solve a "disjunction" category, but a neural network with two units would. As the size of this hidden layer grows, the neural network is capable of discriminating more and more complex categories. Finally, the third layer is the output layer, used to indicate the models "choice" based on the input provided.

Every layer will have some number of neural network units, and these units will be connected to the next layer by some number of connections. A connection is a gain modifier on the signal strength from the input. By convention, input units become activated by some amount (i.e., 0 to 1), and then the connection modifies that activation as a multiplier. A common functional assumption is that the layers will be fully connected, so that all units in the input layer are connected to all units in the hidden layer, and all units in the hidden layer are connected to all units in the output layer. In the event that a connection

is not needed (e.g., an input can be ignored), then its connection value is simply set to zero, so that it completely "turns off" the attention to that unit. In this case, although the connection is still there computationally, it functionally does not exist.

A functional fully-connected neural network, one that has an input layer, hidden layer, and output layer that when provided with input produces sane and meaningful output, needs to be trained. All of the weights, the connections between the layers, need to be correctly set in order for the neural network to function. The Rescorla-Wagner model of learning, derived through experimental work, can be applied to neural networks to successfully train simple neural networks. Known to computer scientists as the delta learning rule, it modifies each connection in the "appropriate direction" proportional to "how wrong" the output is. This method can extend to more complex neural networks if an intermediate proxy for success can be determined for the hidden layer. This is one way to describe what happens during the method of backpropagation. Combined, the delta learning rule and backpropagation can be used to learn complex decision rules in neural networks.

This first simulation has two focuses. The first is a proof-of-concept test to determine whether or not this representational scheme has the ability to learn the II task. Given that the representation is dimensional, the likelihood that this model could learn the RB task is high, but the II task seems to require a coordination between the dimensions that is possibly lacking here. The second focus is to determine how the parameters of the model affect how the neural network learns.

Methods

To simulate the basic task, MATLAB 2015a was used to program a neural network with the three layers described above. The first layer, the input layer, was comprised of 2 sets of 300 units (600 units total). These 600 units were used to represent the input activation, but only a portion of them in the interior were used. By padding the units of interest, we prevented edge units from gaining special salience during the categorization process. Thus, whereas the data that we have examined previously was in the range of 0-100, the data for these neural network models is in the range of 150-250. The hidden layer comprised possibly many units with a logistic activation function. The activation of the hidden layer was then sent to a single output unit with a logistic activation function. When this output unit was higher than .5, it was considered a category "A" decision, and when it was below .5, it was considered a category "B" decision. The network was trained iteratively using the delta rule with backpropagation.

There are many parameters that affect the success of the neural network. The simplest one is the learning rate, which was fixed at 0.1 for all simulations. There are four critical parameters to also consider: the perceptual noise experienced by the artificial observer and the decay of activation in the representation of the stimulus. For the initial simulations, the perceptual noise in both dimensions is set to a value of 3 (i.e., a stimulus will on average be perceived at its true value, according to a normal distribution around the true value with a standard deviation of 3). The representational noise (i.e., the decay on the activation function) will be initially fixed at 15.

The method of training was systematic and very similar to how the pigeons were trained. First a "naïve" network was instantiated by setting all of the weights between the layers set to random values from -0.2 to 0.2. Then a set of 80 trials was generated using the distributions listed in the II condition of Table 1. For each trial, the stimulus sample was taken and perceptual noise was added to it. This perceived stimulus was then represented according to the Gaussian decay functions for each dimension. These activations were fed forward to the hidden layer and to the final decision node, and then the delta rule and backpropagation were used to update the weights based on the difference between the observed output and the true output. This process continued until five sessions with above 80% accuracy were observed.

Results

Proof of Concept Tests

The first important question to evaluate was whether or not the dimensional model developed here was capable of learning the basic task in both the II and RB conditions. The first model used 10 units in the hidden layer. As the model was learning the task, the outputs and state of the weights were observed to change from randomness to having a specific banding pattern. One such example of this is shown in Figure 11. Here, the model's performance on the "session" of data is shown to the left and the current state of its weights are displayed on the right. The weights on the right are scaled between yellow and blue, and demonstrate the state of the neural network. In the first panel, the units are fairly indiscriminately patterned. After approximately 40 sessions, the final form of the

neural network can be seen more readily. Hidden unit 3 seems to have the clearest pattern or highest loading, although units 4, 5, 6, and 9 show related patterns.

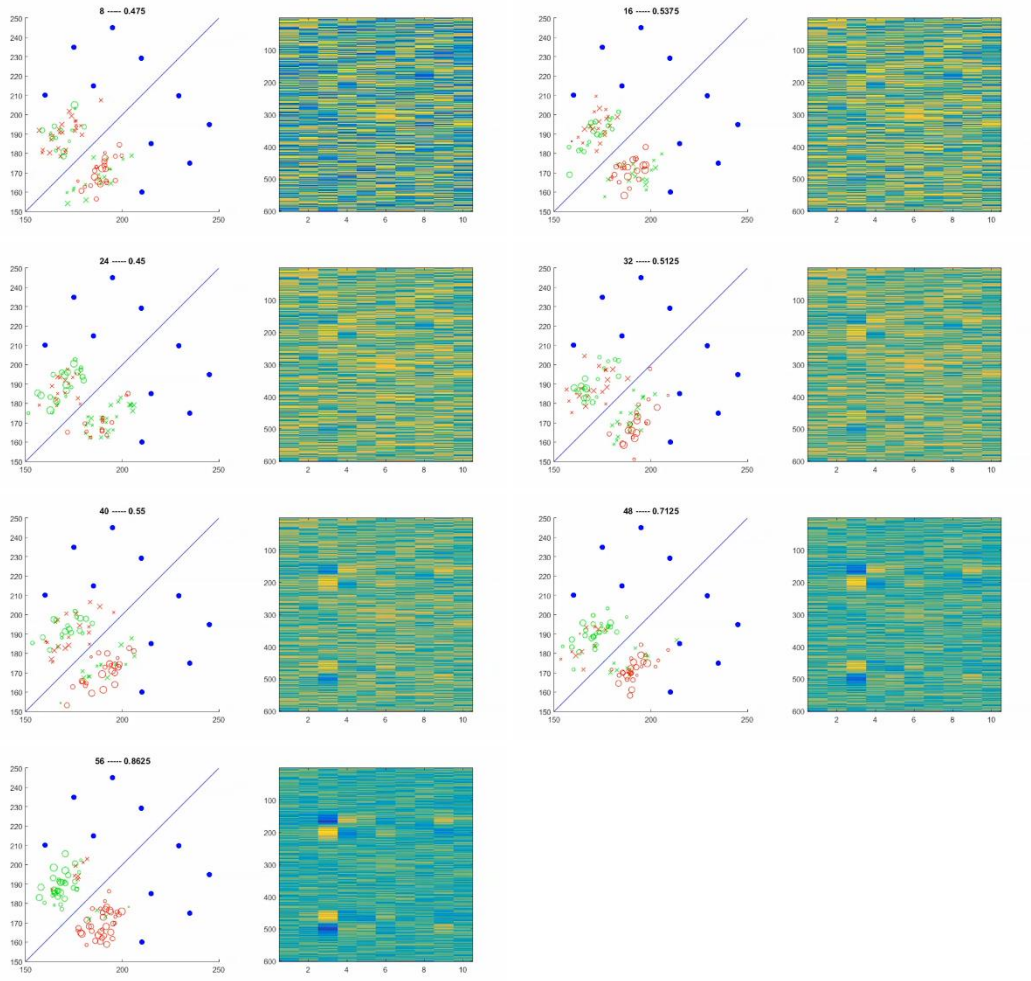

Figure 11. Neural network acquisition of the II categorization task with ten hidden units. The seven panels shown snapshots of learning at the different stages. Each plot shows the stimuli presented to the neural network as either red or green with X or O. The color indicates the category membership as determined by the neural network, O symbols indicate correct responses, and X symbols indicate incorrect responses. These plots are titled with the number of "80-trial sessions" exposed. On the right are graphic depictions of the weights from the input layer to the hidden layer. Rows 1 to 300 are for Dimension 1, and rows 301 to 600 are for Dimension 2. Blue indicates low or negative values and yellow indicates high or positive values.

That the dimensionally activated neural network can learn is an important first step. In order to understand how this network functions, we need to examine

how the hidden layer converts the stimulus representation to something that is decided upon. One way to do this is to discern what “concept” each of the useful hidden units extracts and then determine how those “concepts” are integrated into the final decision. Before conducting this endeavor, we thought to reduce the number of units to determine how many “concepts” in total were necessary. Figure 12 shows the final learned state of three different iterations of this process, first with five hidden units, then two, and finally only one hidden unit. That the neural network was able to learn this using a single hidden unit is revealing. A single hidden unit suggests that little to no functional value is added by the hidden layer, and that in fact the problem could be solved with a simple perceptron.

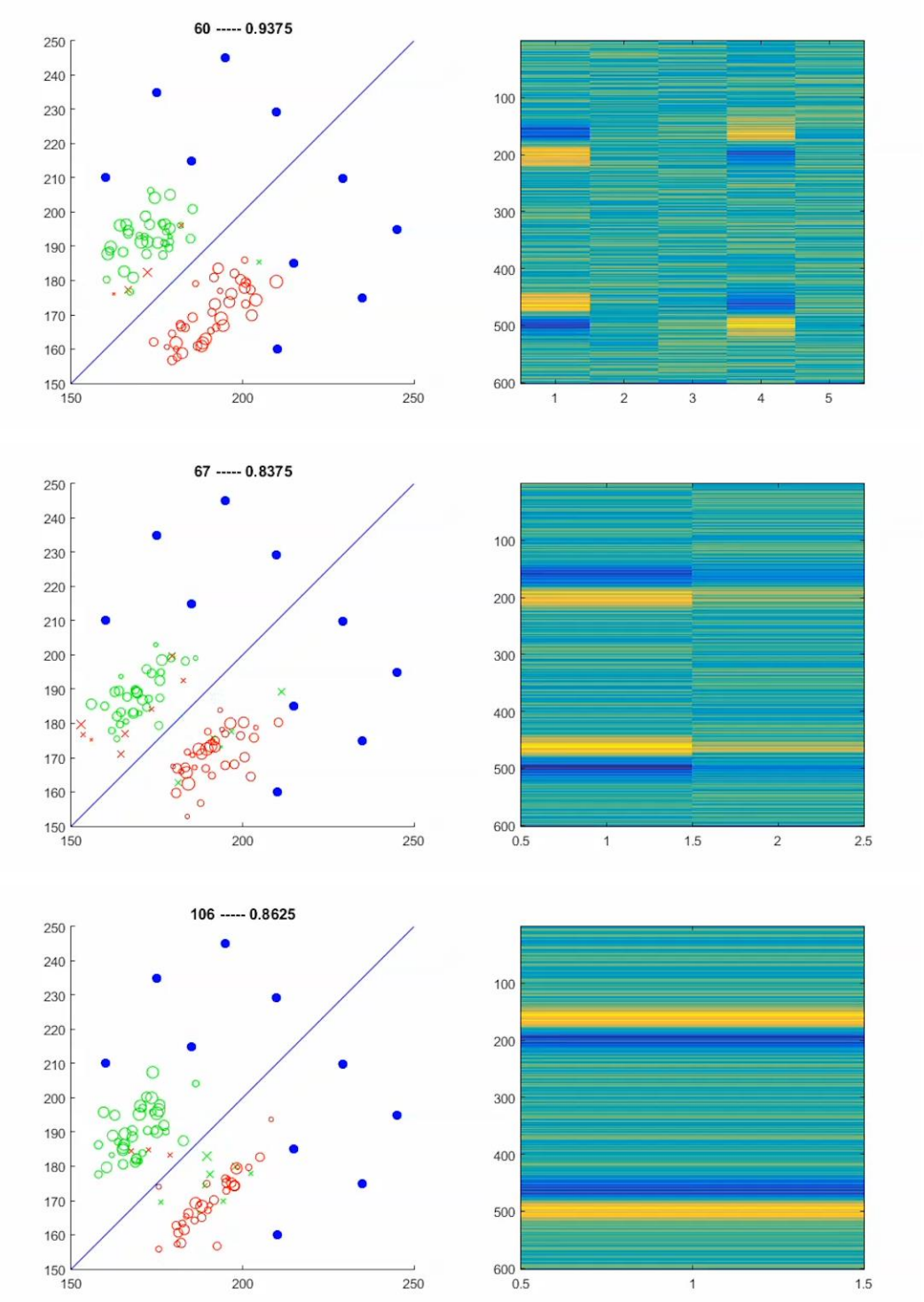

Figure 12. Final iterations for neural network simulations with reduced hidden units.

Parametric Evaluation

To better understand the ability of the model, the neural networks' dependence on the different parameters was evaluated. Two critical questions were the contribution of the perceptual and representational noise on the discrimination. Varying these sources of noise did not undermine the fundamental learning ability of the neural network, although it was delayed. As the representational noise increases, the breadth of the bands in the weight matrix increases. As the perceptual noise increases, the breadth of the bands may increase; more importantly, the neural network may stop demonstrating the ability to achieve criterion. As the perceptual noise increases, although the appropriate areas generally end up associated with the correct responses, the initial noise prevents all 80 trials from being responded to correctly. Thus, these two sources of noise clearly affect the learning and asymptotic level of performance that the network can reach.

Examining the weights from the input layer to the single-unit hidden layer enables us to understand how the neural network functions in a basic way. Figure 13 shows the average weights of the models for 1000 runs for one of the dimensions. The weights, represented by the dots, seem to follow a systematic pattern, basically flat, broadly distributed, and near zero up until 100, when they begin to increase to a peak before decreasing a possibly equivalent amount, after which it returns to near-zero values. A close examination of this function revealed similarity to what might be expect from peak shift curves from prothetic dimensions. In order to evaluate this possibility, a curve was fit to the weight values using three key parameters controlling two normal functions: two means

and one variance parameter (one additional parameter as included to fit the data at the right scale, but this does not change the shape of the function and is ignored). The best-fitting means were 171.2 and 189.9, which are both within 0.5 units of the transformed means of the two distributions when projected onto this dimension. The best fitting standard deviation used was 17.1, which was much higher than the standard deviation of the distributions themselves. Accounting additively for all sources of Gaussian distributed noise (distributional noise, perceptual noise, and representational noise), this is larger by a factor of 1.89. This factor is likely equivalent to that of the marginal distribution of the representational unit activation conditional on the perceptual value, which is a conditional marginal distribution dependent on the stimulus distribution. The data from the second dimension agree, with means of 171.6 and 189.4 and a standard deviation of 17.2. Given the relationship between these weights and the underlying stimulus distribution, it seems like these weights are systematically and predictably determined by the input.

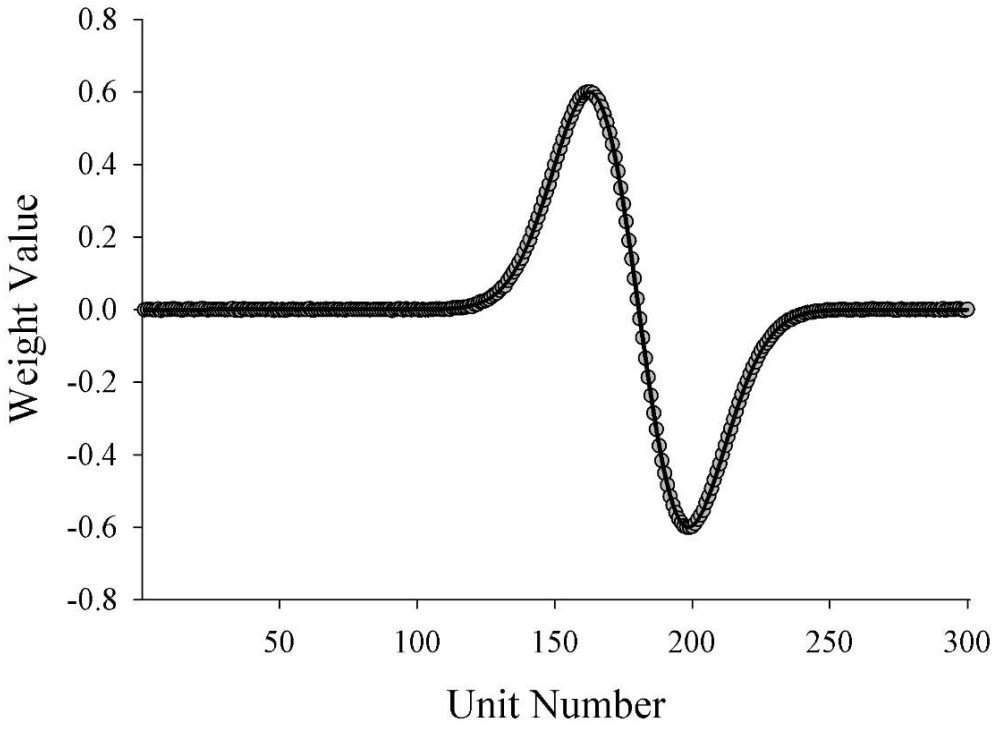

Figure 13. Average final weights for the 1000 neural net simulations. Dots represent average values, and the line is a manually approximated curve derived from normal distributions. Stimuli were presented as being in the 150-250 range (instead of 0 to 100).

Discussion

The neural network model with dimensional representations was able to successfully learn the basic discrimination. We demonstrated that the model will, on average, learn to weight the units in a manner consistent with the underlying stimulus distribution and its classification to the different responses. The use of a single hidden layer proved the network to be simple in nature. Further simulations without the hidden layer shows rapid learning as compared to when the hidden unit is also present. In order to understand how this learning and feature weighting may differ from a configural model of stimulus representation, we conducted

another set of simulations to directly compare the neural networks and task performance.

Simulations 2

Having determined that the dimensional representation of the stimuli can produce a model that discriminates, we wanted to evaluate how this novel neural network model would compare to the already established configural model used in COVIS and SPEED. In order to do this, we first simulated learning to the same criterial levels as the pigeons. We also thought to take this opportunity to evaluate how the models would behave in the different training and transfer conditions. Thus, we simulated the transfer results of the exact experiments conducted with pigeons above.

Methods

The same network architecture as used above was repeated here for the dimensional activation model. Additionally, the configural activation model was implemented by using a representation that utilized a grid instead of rows of units. To be comparable, the grid was 300 units on an edge (300 x 300 = 90,000 units total), with the target region of stimulus activation from units in the 150-250 range. The grid was activated on a trial using a Gaussian decay function to its optimal activation value, using Euclidean distance. The representational noise was set to be the square root of the sum of the dimensional noise components squared. The other parameters of the neural network remained the same. To make the comparison, 10,000 simulations of each neural network were conducted in both

the rule-based and information-integration tasks. In order to accomplish this, Tufts High-Performance Computing cluster was used with MATLAB 2012b.

Results

Both models learned both forms of the task, as expected. Figure 14 depicts distributions of the number of "sessions" required until criterion for the 10,000 simulations run. The configural activation model clearly results in faster learning, showing distributions that peak around 18 sessions to criterion. On average, the rule-based task was learned by the configural model in 21.22 (SD = 4.72) sessions, and the information integration task was learned in 21.22 (SD = 4.73) sessions. Clearly these two tasks are equivalent in the scope of this model. A two-sample Kolmogorov-Smirnov test for equality finds no difference between the distributions (D = 0.0085, p = .86). Visually, Figure 14 suggests that the dimensional activation model also appears to treat the two tasks equally, as their lines are right on top of each other. However, numerically the two are in fact different; rule-based 91.01 (SD = 24.55) sessions and information-integration 92.31 (SD=24.96). Furthermore, a two-sample Kolmogorov-Smirnov test for equality shows that in fact these distributions are different (D = 0.03, p < .001). However, finding this effect in biological systems would be incredibly difficult because the effect size is very small (Cohen's D = 0.04).

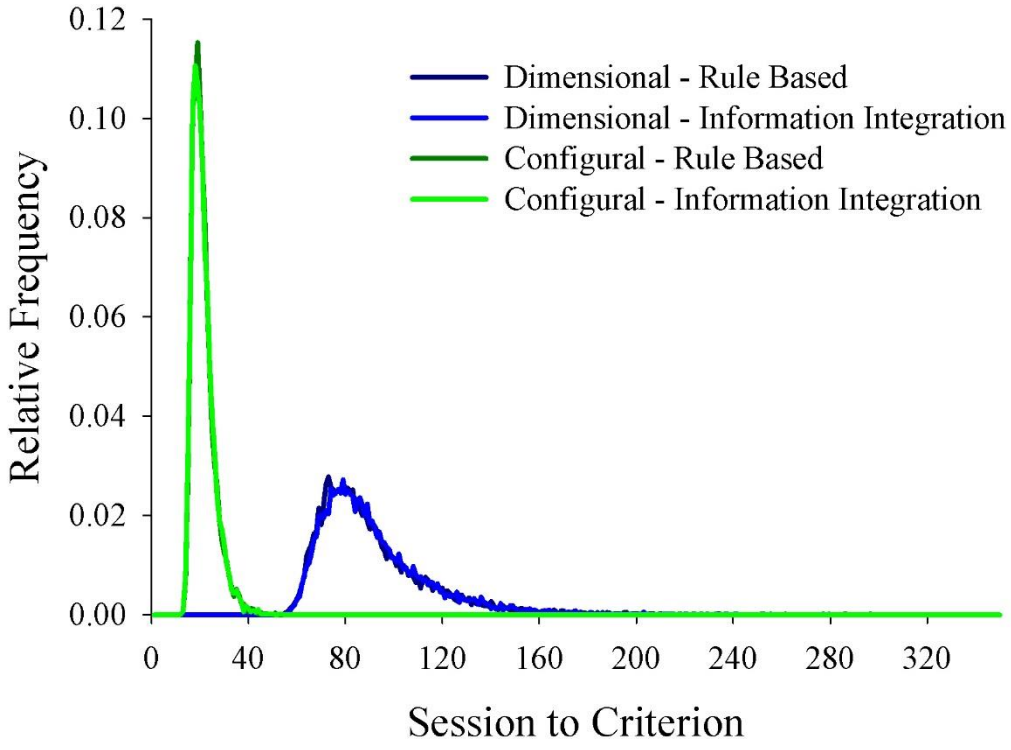

Figure 14. Distribution of learning times for the two representations in the two tasks. The blue lines depict the distributions for the dimensional activation model and the green liens depict the distributions for the configural activation model. The bright shades depict performance with the information integration task while the dark shades depict performance with the rule based task.

Next we examined how the different networks functioned when given the transfer tests as the pigeons were in Experiments 1 and 2. This data is reported in Figure 15. The first two bars in the panels of Figure 15 represent the two transfer distributions tested in Experiment 1. These are the distributions that are opposite the training distribution within the stimulus space. In the case of the configural model (bottom row), performance in both tasks is stable and (on average) above chance at 55%. The performance of the dimensional activation model, in contrast, depends on the task type. In the rule based task, the dimensional activation model

shows above-chance transfer at 60% accuracy, but in the information integration task, it performs below chance at 45%.

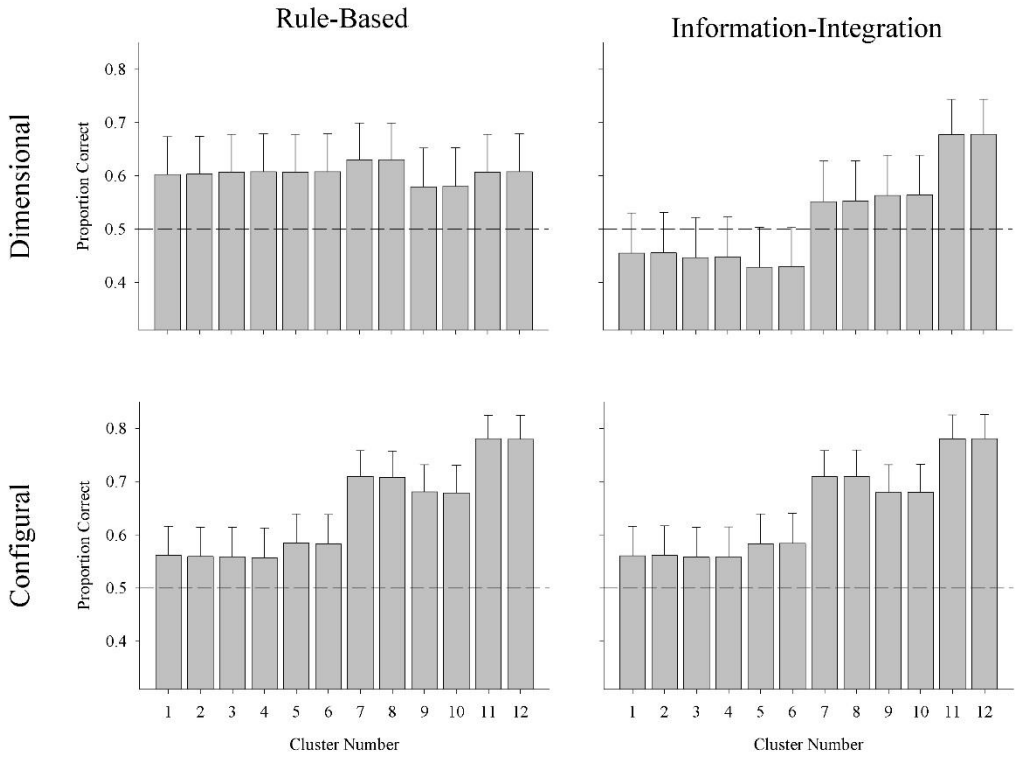

Figure 15. Model performance on transfer stimuli. The left panels show performance on RB tasks and the right panels show performance on II tasks. The top row shows the performance by the dimensional activation model, and the bottom row shows the performance by the configural activation model. Each bar depicts accuracy for a subset of transfer data; the first two are the two distributions from Experiment 1 and the remainder are the clusters from Experiment 2. Error bars indicate one standard deviation.

The transfer results for the clusters from Experiment 2 show a related pattern. The configural activation model tends to classify everything correctly, and there is virtually no difference between performance on the RB task or the II task. However, the dimensional activation model's performance is highly dependent on the task. In the RB task, the dimensional activation model performs

consistently above chance. In the II task, the individual clusters produce large variations in the classification performance. Clusters 1 through 4 (bars 3 through 6) are regularly below chance, and these are the clusters that were regularly misclassified by the pigeons as well. Clusters 5 through 8 are middlingly above chance, at about 55% accuracy, and clusters 9 and 10 are the highest performing clusters. These last two clusters are also the closest to the training displays.

Discussion

Here we examined how the dimensional activation model and the configural activation model would perform in the experiments conducted with the pigeons. The simulations demonstrated that in neither model would there be a detectable difference between the learning rates for the two tasks. We did, however, find differential transfer performance based on the training conditions and the model employed. The configural model was agnostic to the dimensions so dimensional transfer or non-dimensional transfer made no difference. The dimension activation model, however, finds that in the RB task, the transfer stimuli are still familiar because how they activate the input representational layer has not changed. In the II task, however, the stimuli are now activating units that previously accorded with the "wrong" category, thus yielding the incorrect categorization.

Since the pigeons' data mirrored these simulated transfer results, these simulations seem to suggest that the pigeons may be completing the task using a dimensional representation of the stimuli. Simulations 1 revealed some success in matching the weights from the neural network to an underlying normal

distribution suggesting that the weights, and therefore potentially the model's performance, can be determined from six normal distribution parameters. In order to more thoroughly evaluate the pigeons' discrimination and categorization method, we need to develop an effective method of comparison.

Simulations 3

The traditional approach to determining whether an observer is using one method or another is to determine the likelihood of observing the data under the different model schemes. The goal of this next section is to develop a method by which the simulated data can be evaluated for its adherence to a specific model. Because we will be interested in applying this to analytic categorization methods as well, this will involve generating models of prototype rule discrimination and GRT-rule discriminations. All such methods require positing a set of parameters that the observer is using to control the discrimination; in the case of the prototype model it would be the prototype stimulus, and in the case of the GRT model, it would be the coefficients controlling the rule bound. The first question we need to tackle is determining a method using a small set of parameters that describes the performance of the dimensional activation model. The results from Simulations 1 are instructive here by demonstrating the dependence on the model to the underlying normal distributions. Once we develop a method for the dimensional activation model, we will extend it to the configural activation model, and finally, we will simulate data using all four model methods and evaluate whether or not these methods can correctly identify the models used to generate the data.

Association-Based Categorization Reformulation

To model the performance of the model, we need to acknowledge the simple associations for the stimuli to each response category. The curve in Figure 14 not only represents the weight values, but also reflects the association of a particular stimulus value with category A (positive) or category B (negative). After the transformation through this hidden unit, where the information is weight as depicted in Figure 14, the process is fairly linear. Thus, the model ultimately seems to simply approximate this curve.

Because of the assumptions of our model and the functional simplicity of our distributions, we can model the neural network outcome without having to instantiate the hundreds of connections posited by the networks. With a single unit in the hidden layer, which is as functionally useful as no hidden layer at all, what the neural networks reduce to are simple association networks. Output activations in simple association networks should be proportional to the relative predictability of the input units. For the networks provided here, the predictability of a given unit depends on the relative activation by stimuli from one distribution versus the stimuli from the other distribution. We assume that both stimuli activate units using a Gaussian decay function. The categories and stimuli are distributed according to a normal (i.e., Gaussian) distribution. Given stimuli distributed normally and Gaussian decay functions, the overall activation by all the stimuli of a single category will also be Gaussian. The mean of the activation distribution will be the perceived mean of the stimuli (assuming an unbiased perception, equal to the true mean of the distribution). The variance of the

activation distribution will be composed of the true variance in the stimulus category as well as any added variance from the perceptual or decisional processes; however, without further careful experimentation it will not be possible to separate the processing variance perfectly from the distributional variance. Nevertheless, computationally identifying data generated by this model can be accomplished by determining these six values (four means, two variances) underlying the activation distributions (including the variance from perceptual and decisional processes). Finally, we included a scaling parameter for the final decision process so that the absolute activation levels could be attenuated. The result for a single stimulus in the dimensional activation model was given by the following equation $\Phi(s * [\varphi(D(x_1, \mu_{1A}), 0, \sigma_1^2) - \varphi(D(x_1, \mu_{1B}), 0, \sigma_1^2) + \varphi(D(x_2, \mu_{2A}), 0, \sigma_2^2) - \varphi(D(x_2, \mu_{2B}), 0, \sigma_2^2)])$, where Φ and φ are the c.d.f. and p.d.f (respectively) of the standard normal, D is the distance function, *s* is the scaling parameter, (x_1, x_2) is the stimulus values in dimensions 1 and 2, μ_{1A} and μ_{1B} are the means of the A and B distributions in the first dimension, μ_{2A} and μ_{2B} are the means of the A and B distributions in the second dimension, and σ_1 and σ_2 are the variances of the activation distributions for the first and second dimensions. Note that for the configural activation model, the function only changes by the joining ("binding") of the dimensional activations by using a multi-variate normal p.d.f., but the decisional c.d.f. is still unidimensional. To distinguish these formulations from the neural network instantiations of the same networks, we will refer to these as the extracted dimensional activation model and the extracted configural activation model.

Model Fitting Methods

Modern computing affords us the ability to determine which model is most likely to have generated the data. We will evaluate four total models, two analytic and two non-analytic. The analytic, rule-based models will consist of the prototype model and the general quadratic classifier. For proper comparison to the non-analytic model, we will use probabilistic versions of the models.

The prototype model and the general quadratic classifier have been dealt with fairly thoroughly in the literature (Ashby, 1992a) as well as in the previous section, and the association models and their parameters for fitting have been described above. Consequently, we will not expend too much time here to re-iterate their differences. Both classifiers use a distance function to evaluate a given stimulus. For our models, in both cases, the stimulus' distance value (to either the rule or the prototype) was subtracted from a threshold and divided by a scaling factor and the resulting value was mapped to the likelihood of the two responses using the inverse standard normal function. This yielded a probabilistic version of these analytic models, allowing us to compare their efficacy against the association based models using log likelihood, which we used as a step of evaluating model fit. Thus, for each response the pigeon made, we will evaluate the likelihood of seeing that response for each model.

All model fitting was conducted in MATLAB, using the optimization toolbox. Each categorization method was implemented as a separate function that generated a probability of seeing a given response value. These probabilities were compared against the pigeons' response data to compute the sum squared error,

which was used to generate log likelihood (LL) via the formula n*(SSE/n). While this typically only used in regression contexts when the error in the model is normally distributed (and not binomially distributed as in this case), using log-likelihood allows us to make comparisons among the models without making additional contentious and complex assumptions. In order to determine the optimal set of coefficient values for each of these models, we replicated this process at least 50,000 times for each problem using a grid that encompassed all likely values.

For each dataset and for each model, we determined the value of the Akaike Information Criterion (AIC), which is a common metric used to determine minimally-complex, best-fitting models. The AIC is defined as AIC = 2*k – 2 * LL, where *k* is the number of meaningful parameters that were fit and LL is the log likelihood. So for example, although the GQC has 6 parameters, there are only 5 degrees of freedom to specify a model and hence 5 free parameters for the AIC's *k* value. The *k* component of AIC therefore penalizes complexity, while the LL component of AIC penalizes poor fits to the data. The best AIC is the smallest (i.e., closest to negative infinity), with variations that can be difficult to judge as meaningful. The rule of thumb that has been widely adopted is that a difference of 2 is considered "meaningful" or "significant" (not statistically). AIC is reported for each model and animal. The best-found log likelihood was used in the AIC computation.

Simulation Methods

In order to discern how well this model can discriminate the various models, the other models under consideration needed to be generated. For this purpose, parameters for the other models were randomly generated until they produced data that conformed to an 80% accuracy criterion on 400 trials (i.e., the same as the previous 5-session criterion). In this fashion, datasets were generated using each of the general quadratic classifier, the prototype model, the analytic dimensional activation model, and the analytic configural activation model. For the purposes of this exercise, 100 datasets were generated for each of these models in both the II and RB conditions. Additionally, 100 datasets were generated from the RB- and II-train dimension- and configural-activation neural networks from the previous simulations. Each dataset was comprised of the same number of total baseline trials and transfer trials as the pigeons received in Experiments 1 and 2, combined.

Results

The simulation results are shown in Table 3, below. The model most likely to yield a false recognition when the true training is the dimensional or extracted dimensional model is the prototype model; in every other case the most likely model is the extracted dimensional model. In the case where the true model was the extracted configural model, this process was able to recover that model as the best fitting model. In every other case, the extracted dimensional model was found to be the best fitting model.

True\Fitted Model	Extracted Dimensional	Extracted Configural	GQC	Prototype
Dimensional	68.53%	0.00%	1.02%	30.46%
Extracted Dimensional	73.50%	0.00%	0.00%	26.50%
Configural	96.00%	0.00%	0.00%	4.00%
Extracted Configural	4.19%	95.29%	0.00%	0.52%
GQC	79.00%	4.50%	0.00%	16.50%
Prototype	99.50%	0.50%	0.00%	0.00%

Table 3. Formal model fitting results. Each cell reports the percentage of simulated datasets in the row that were fit by the model listed in the column.

Discussion

The results seem to suggest that the extracted dimensional model is able to account for much of the patterns of data found in the other generated datasets, including GQC and prototype models. This is likely the result of the fact without further optimization, the model fits for the traditional GQC and prototype models suffer. Alternatively, these mixed results may also indicate that the space in which the models are being searched are qualitatively different. The parameter space surrounding this problem needs further investigation in order to best determine why this version of the analysis failed. Notably, in the next section, it will be shown that the GQC is the best fitting model for some of the birds, therefore the failure in this section is likely the result of the lower number of iterations or further optimizations in this validation step. Furthermore, the critical investigation here is between the configural model and the dimensional model, and these results suggest that datasets generated from these two models are identifiable using this procedure. Therefore, we can apply this method to the data from the pigeons to determine what algorithm they use to complete these II and RB tasks.

Behavioral Model Fitting

Having established in the Formal Methods that the model fitting procedure can sufficiently recover the algorithm used to generate artificial data, we can analyze the pigeons' data to determine how they complete the discrimination task. Given the poor fits from the formal methods, the search scheme was intensified in order to fit this data, including 200 times as many random searches followed by a second optimization step. Applying this method for the 7 pigeons was feasible as compared to the 1200 simulated datasets from the previous section.

Methods

In order to determine the optimal set of coefficient values for each of the models, we randomly selected parameters 1,000,000 times for each problem using a grid that encompassed all likely values. We then took the best parameter sets and used MATLAB's Optimization Toolbox functions to minimize the error further. The final log likelihood was used in the AIC computation.

Model Results

The AIC values and the rank position of the four models are in Table 4 for the combined transfer results of Experiments 1 and 2. Examination of the table reveals that the dimensional activation model most consistently supported the best fit to the birds' overall results. It was the best fitting for three birds and the second best for the remaining four. It is the only model to do this well. The most serious competitor was the general quadratic classifier, which also best described three birds' results, but was second and third twice with the remaining birds. The

configural activation model was a poorer third, although the RB trained birds show some accordance with the configural activation model. The prototype model was always the worst model.

	Rule-Based			Information Integration			
	#A1	#D1	#T1	#C1	#G1	#L1	#S1
1st (best)	gqc -3182.08	conf. -3332.82	dim. -3149.32	gqc -3526.06	dim. -3320.54	dim. -2590.88	gqc -2876.45
2nd	dim. -3161.72	dim. -3301.19	conf. -3139.73	dim. -3524.71	gqc -3308.80	gqc -2588.26	dim. -2858.36
3rd	conf. -3154.19	gqc -3245.18	gqc -3131.71	conf. -3366.61	conf. -3283.58	conf. -2570.55	conf. -2842.10
4th	prot. -3125.86	prot. -3165.37	prot. -3008.84	prot. -3274.87	prot. -2986.51	prot. -2424.44	prot. -2667.60

Table 4. Model fitting results for the pigeon data from sessions containing transfer data in Experiments 1 and 2. The values indicate Akaike Information Criterion (AIC). Definitions of the models and the source of AIC are in the text. The model results are ranked from 1 (best) to 4 (worst). General Quadratic Classifier (gqc), Dimensional Activation (dim.), Configural Activation (conf.), Prototype (prot.)

The birds trained in the RB task showed wider disagreement regarding the "best" fitting model. Pigeons #A1, #D1, and #T1 showed best fits by the general quadratic classifier, configural activation model, and dimensional activation model, respectively. The second best was the dimensional activation model for pigeons #A1 and #D1, and the configural activation model for #T1. The pigeons trained with the II task show more consistency. For all four of these pigeons, the general quadratic classifier or the dimensional activation model was the best-fitting or second-best fitting model, often closely.

Discussion

Overall, the model that that best describes the pigeons' behavior on average was the dimensional activation model (average rank 1.57 of 4). Between the two non-analytic models, the dimensional activation model described the pigeons' behavior better than the configural activation model for five of the seven pigeons. As a result, this modeling suggests that the pigeons were most likely using the non-analytic method of dimensional activation. That said, the most serious competitor was the general quadratic classifier. This analytic model had an average rank 1.86 of 4. Thus, if the pigeons were using an analytic mode of processing, they were most likely using some form of general quadratic classifier.

These model fitting results support the hypothesized framework being developed in this manuscript; however, a close scrutiny of the methods requires an additional comment on the approximation of likelihood. The fit of the models was determined by inferring log likelihood from error. However, the models as developed here should provide precise probability values for every stimulus point. However, in several many cases, the likelihood of a given response is functionally zero (i.e., below the threshold for numerical accuracy for the computers involved). In this case, the computed log likelihood becomes negative infinity. However, there are many potential sources for error in these behavioral experiments, from observer fatigue effects to response errors. The focus of this investigation was the representation underlying performance, and so building a decision framework with representational value as one component was avoided by using the error approximation to likelihood. One alternative would be to create

mixture models involving guessing or biased guessing or other factors to account for the otherwise impossible data points. Another alternative would be to impose a lower limit at a previously determined value (i.e., 0.5%) instead of 0. This would weight each incorrect value at a specific number, possibly configurable by the researcher. Instead of arbitrarily determining this value or creating a framework, I adopt the error approximation to likelihood used in regression contexts.

These results raise an apparent incongruity. If the pigeons' representation of the stimuli is dimensional and there are potentially no configural units to permit that sort of joint representation, then how does learning proceed at the same rate? Would not the dimensional representations yield an advantage during learning for the RB condition above the II condition? The result from Simulations 2 suggest that the answer is "apparently not" because the performance gain is sufficiently subtle. Thus, assuming a dimensional representation instead of extant configural representation model comes at no cost; it succinctly captures the results in these experiments and does not contradict previous results (i.e. similar learning rates as evidenced across these 7 and the previous 17 pigeons; Smith et al., 2011). If there is no cost for the shift in the pigeon model, what about the procedural learning system in the human animal? Perhaps it, too, has a dimensional representation. We extended these test using human participants in the next experiment to evaluate this question.

Experiment 3

We now have two models of procedural learning, one which relies on configural activation and one which relies on dimensional activation. Comparing the two models of procedural learning, the previous experiment made clear that the pigeons use the dimensional representation. This then raises the question of whether the same representation is used during procedural learning in humans. The procedural learning component of the SPEED model, which was developed from the striatal pattern classifier (SPC), uses configural activation, but what is the source of this choice? Examining the original COVIS paper, the most logical conclusion is that researchers by default have a bias towards bound representations. Some of this, however, may have been supported by the knowledge of visual cortex cells that are both orientation and spatial frequency specific (e.g., De Valois, Albrecht, & Thorell, 1982; Issa, Trepel, & Stryker, 2000) along with a belief that as receptive fields increase, the nature of the information fundamentally does not change.

This leaves open the question of whether or not humans have dimensional activation during procedural learning or if it is configural as the SPEED model proposes. Some support for the use of dimensional procedural learning was suggested in previous work; Casale et al. (2012) found a negative correlation on II training and analogical transfer performance in Experiment 3. While his results were “extremely non-significant,” they were in the correct direction for the inversion that we observed and that the dimensional activation model predicts. The remainder of their data are hard to evaluate with respect to the scheme we

provide here because they trained responding in the novel region, making it more difficult to evaluate the underlying representation. By providing feedback about the correct response structure, a participants' categorization rule may change. We thought that our method here of testing probes during regular baseline training best resulted in continued responding without changing the underlying categorization process.

We thus tested 27 Tufts undergraduates in the II task and tested analogical transfer in a variety of locations to determine which model of categorization most veridically described their behavior. In order to best capture their performance, we did not provide feedback on those trials. While this differs from the procedure with pigeons, it seemed justifiable given that we could instruct participants to continue despite non-reinforcement (pigeons would not have accepted these instructions as happily). In order to prevent the no-feedback trials from altering performance, we first gave participants a partial-feedback phase so that we could evaluate any effect of not providing feedback. The data from the final blocks were then subjected to the same modeling methods described above.

Methods

Participants

We recruited 27 Tufts undergraduates from a common participant pool used by multiple introductory undergraduate Psychology courses.

Apparatus

Two testing stations were used for testing. Both stations had modern Dell PCs, running a custom built Visual Basic 6.0 program that presented stimuli on a monitor. Both stations had a Dell Dimension 745 machine, and one had a Dell 1907FPt monitor or while the other had a Dell 1903 FPt monitor. Participants provided input using the keyboard, and received feedback both on the computer display as well as through headphones connected to the computer.

Stimuli

The values and distributions tested with the human participants were largely the same as the II conditions tested with the pigeons. The only exception to this is an additional 11 clusters of tests on the border of the stimulus space. The values for these clusters are reported in Table 5, and their spatial distributions are displayed in Figure 16. These clusters were added to provide stronger disambiguation of the general quadratic classifier model and the dimensional activation model. Note that the cluster in the corner opposite the training distributions (cluster #6) does not have a clearly "correct" assignment.

Distribution #	μ_1	μ_2
1	5.00	95.00
2	23.00	95.00
3	41.00	95.00
4	59.00	95.00
5	77.00	95.00
6	95.00	95.00
7	95.00	5.00
8	95.00	23.00
9	95.00	41.00
10	95.00	59.00
11	95.00	77.00

Table 5. Distribution parameters for the new transfer clusters in Experiment 3. The values listed here indicate the means in the normalized stimulus space. Note that these are only representative for a subset of the human participants. For the remaining subjects, these distributions must be rotated around (50, 50) by 180° to be accurate.

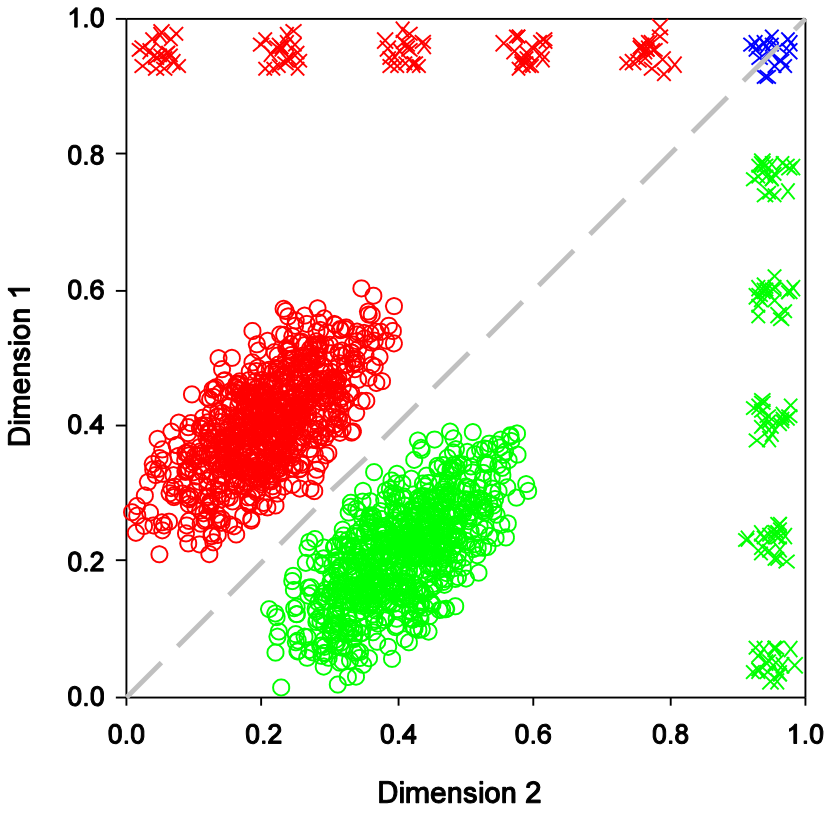

Figure 16: Distributions of the novel stimuli investigated in Experiment 3. In these examples, the circles are the training stimuli and the exes are the transfer stimuli. The red/green assignment denote correct category assignment based off the dashed linear separator in grey. The blue corner distribution does not have a clear a-priori correct assignment. Note that some participants received a 180° rotation from this setup.

Procedure

Participants were brought into the lab and seated in an anteroom to sign their informed consent. They were then led to the testing station where the program had instructions listed on the screen. The experimenter led the participant through the on-screen instruction and confirmed that they had no questions before continuing. The instructions were minimally informative regarding the purpose of the experiment. Participants were told that they will be making choices about images. When presented with a white circle at the center of the screen, pressing

spacebar would advance the trial, and then an image would appear. They were instructed to make a choice about the image by pressing the blue or yellow keys (indicated using electrical tape over the "Z" and "/" keys on the keyboard). They were told that the program would indicate a correct response by displaying a green square after the choice and playing a chime sound, and that it would indicate incorrect responses with a red square and a thud. They were also told that after some choices, the program would not provide feedback, which was normal, and they should just continue working through the experiment as normal. Finally, they were also told that the experiment may be difficult, but it is possible to make only correct choices.

The manner of training was fixed for all participants. First, they received the baseline (i.e. training) distributions for three 80-trial blocks. The next three 80-trial blocks used the baseline distributions, but every trial had a 50% chance of providing feedback. Finally, three test blocks finished the session. The three test blocks were comprised of 160 baseline trials plus the additional test trials. In "first distribution transfer," this was 80 trials of the untrained distribution (i.e., analogical transfer tests from Experiment 1). In the "second cluster transfer" (i.e., Experiment 2 transfer tests), 100 trials were added testing each of the ten cluster distributions that the pigeons received. In the "third cluster transfer" (i.e., Experiment 3 unique trials described above), 110 trials were added testing the 11 clusters around the border of the stimulus region. These three test blocks were provided in a randomized order for each participant.

Results

Two participants failed to learn the task (accuracy never greater than 60% in a single block) and one participant decreased in accuracy over the course of training. In addition to one participant who did not complete the experiment, these data have all been removed and are not depicted or considered further. For the other participants, their data are considered by first evaluating baseline accuracy, then transfer accuracy, and finally examining the model fits.

The remaining 23 participants learned the task quickly and were largely unaffected by partial feedback. Figure 17 depicts accuracy on only the baseline trials over the course of the session divided into the nine blocks. Participants reach good accuracy within the first block, attaining high levels tightly clustered around 90.2% (SE = 0.93%) by the third block. There was little to no decrement in the partial feedback blocks (paired samples t-test comparing blocks 3 and 4, $t(22) = .06$, $p = .95$). The same was not true when introducing the transfer items. There was a substantial drop in accuracy on the baseline trials from block six to block seven ($t(22) = 4.6$, $p < .001$, $d = .96$), despite there being no change to the baseline trials during this time. Throughout the testing block, there were no further changes in baseline accuracy (ts$(22) < 1.3$, ps $> .2$).

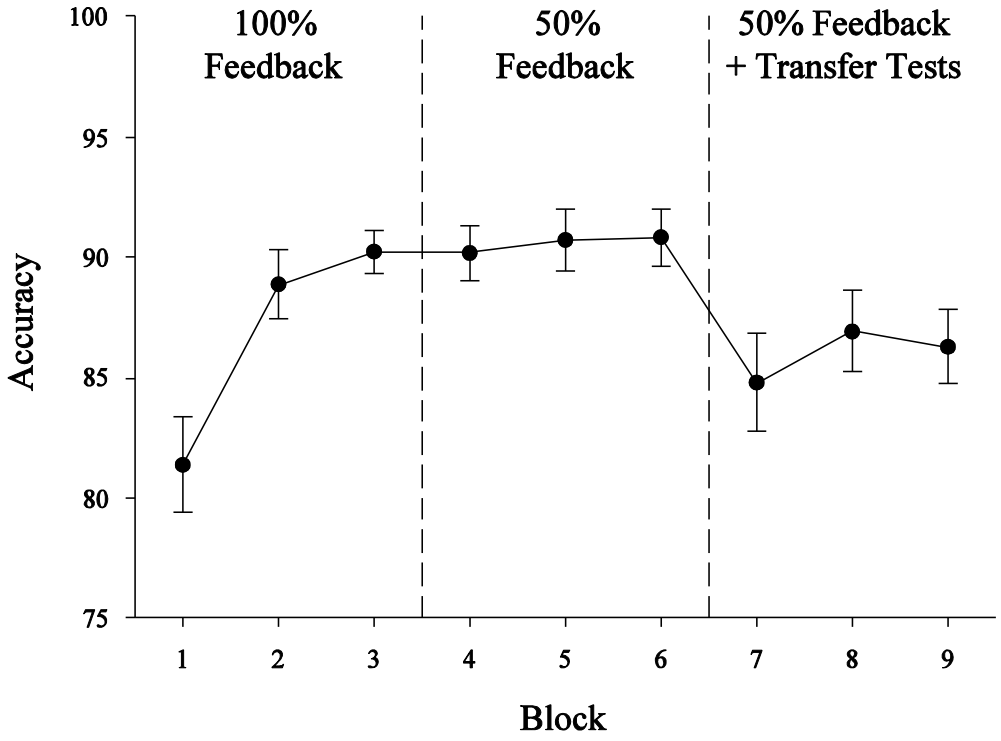

Figure 17: Baseline performance through the task for human participants in Experiment 3. The first six blocks were composed of 80 trials, and the last three blocks contained 160 baseline trials as well as 80 to 110 transfer trials. In the first three blocks, the participants received full feedback, and during the last six blocks, the participants received partial feedback. These boundaries are indicated via vertical dashed lines in the figure.

The accuracy during analogical transfer was very similar to that which was seen with the pigeons in the II training condition (Figure 18). The first distribution transfers were near or below chance (M = 47.8%, SE = 2.2%), despite high baseline accuracy (M = 85.4%, SE = 2.1%). Similar to previous investigations, transfer performance was negatively correlated with baseline accuracy ($r(23) = -.29$) although this was not significant ($p = .156$). Like the pigeons, the humans in this II training transferred at above-chance levels to the second cluster transfers ($t(22) = 7.0$, $p < .001$, $d = 1.5$). The data for third cluster transfer looks similar (the ambiguous "corner" distribution, cluster #6, has been

excluded). For these clusters, participants on average "correctly" identified them as belonging to one category or the other ($t = 4.6$, $p < .001$, $d = 0.96$).

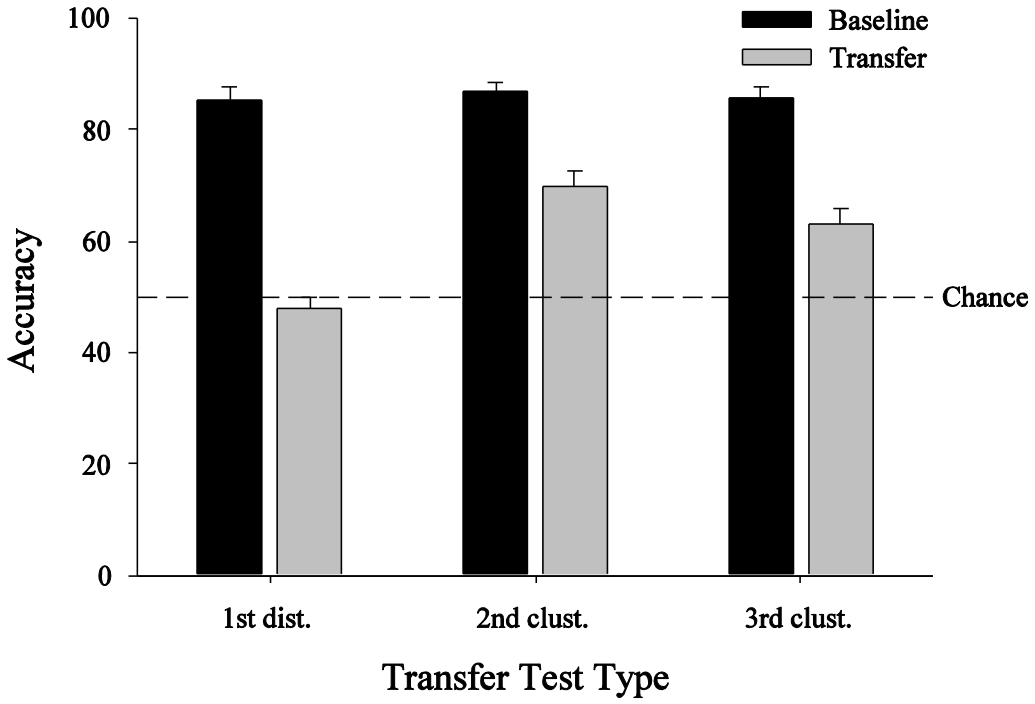

Figure 18: Accuracy on transfer tests for human participants in Experiment 3. The relevant baseline data is plotted alongside the transfer data. The tests are broken up into the Experiment 1 distribution transfers (1st dist.), Experiment 2 cluster transfers (2nd clust.), and Experiment 3 cluster transfers (3rd clust.). Note that for the last set of transfers, the ambiguous distribution was removed from the transfer data.

Again, for this dataset, the categorization behavior is better informed by model fitting than by examining accuracy values. We subjected the human data to the same model fitting analyses as the pigeons above. The results for this analysis are reported in Table 6. The model fitting data are reported according to how many participants agreed for a given model order, along with the average AIC values. The first three columns, which comprise over half the participants, contain model rankings in which the dimensional activation model was the best fitting model. In most of these cases, the second-best was the general quadratic

classifier. The fourth and fifth columns, which summarize the data for 8 additional participants, were cases where the best model was the general quadratic classifier, and in all of these cases, the dimensional activation model was second best. Finally, the last two columns report one participant each, first a participant for whom the prototype model was the best and dimensional activation second, and the other a participant for whom the configural activation model was best and the general quadratic classifier second.

# of participants	7	2	4	4	4	1	1
1st (best)	dim. -2013.73	dim. -1806.67	dim. -1807.67	gqc. -1990.00	gqc. -1812.68	prot. -2019.63	conf. -1224.94
2nd	gqc. -2004.92	gqc. -1796.21	prot. -1794.53	dim. -1979.27	dim. -1770.29	dim. -2013.77	gqc. -1204.44
3rd	prot. -1871.09	conf. -1501.18	gqc. -1756.92	prot. -1876.67	conf. -1596.82	gqc. -1958.09	dim. -1201.30
4th	conf. -1601.10	prot. -1398.98	conf. -1531.98	conf. -1581.06	prot. -1525.71	conf. -1663.59	prot. -1090.13

Table 6. Model fitting results for the human data from Experiment 3. The values indicate average Akaike Information Criterion (AIC). Definitions of the models and the source of AIC are in the text. The model results are ranked from 1 (best) to 4 (worst). General Quadratic Classifier (gqc), Dimensional Activation (dim.), Configural Activation (conf.), Prototype (prot.)

Discussion

This experiment confirmed that the procedural learning that is assumed to function in human Information Integration category learning does systematically alter an observer's responses to novel parts of the stimulus space. The original

transfer scheme used by Casale et al. (2012) and J.D. Smith et al. (2015) demonstrated below-chance transfer performance to those broad distributions, with a negative correlation between learning the original task and transfer accuracy. Our subsequent transfer tests supported good "accuracy," and the model fitting suggests that the general quadratic classifier (average rank 1.87 out of 4) or the dimensional activation model (average rank 1.48 out of 4) are most consistent with the data. Except for one subject, the dimensional activation model better described the participants' categorization performance compare to the configural activation model (average rank 3.61 out of 4). Given the large corpus of other experiments that eliminates the prospect of rule-based responding in learning the Information Integration task, these data suggest that the dimensional activation model proposed in this paper most likely underlies procedural learning in these classification tasks.

The model that best described the human data by far was the dimensional activation model, but again the most serious competitor was the general quadratic classifier. This raises the need to consider the viability of the GQC as a model of procedural learning. The original COVIS proposal identified that a small decision layer (one or two units) using a threshold on the value of the hyperbolic function (i.e., $Ax^2 + Bxy + Cy^2 + Dx + Ey + F$) would divide the stimulus space sufficiently well into three regions (note: this is not how the configural activation model in this paper was computed). Thus, the authors argue, the GQC function could easily be instantiated in a neural network and could therefore be the model controlling procedural learning. In this scheme, however, the authors

acknowledge that the GQC model becomes a description of performance without psychological reality. The same can be said of the GQC in the current investigation. Despite its ability to describe the data, the GQC model only reflects patterns in the data, but it does not explain the data or properly model the process generating the data.

One interesting feature of the data is how the human participants demonstrated a deficiency in their baseline accuracy once the transfer stimuli were introduced, a feature that is absent in the pigeon data. The causes for this drop are unclear. Given that the preceding 240 trials had included (on average) 120 trials without feedback, the critical difference must be the information gained or altered by seeing the transfer stimuli at all, not any methodological differences. One possibility is that during the training, the participants had learned and considered the training distributions' region as the entirety of the stimulus space. Thus, they had attuned their discrimination and expectation of variability to only the training distributions. Once the transfer stimuli were introduced, the participants learned that the stimuli can vary to a greater extent than they previously learned, and perhaps this altered the discriminability of the stimuli. This can also be thought of as a sort of priming effect, where the training trials all primed the same area of the dimensions and thus increased the items' discriminability. Once the transfer trials were introduced, the amount of priming for the baseline region was reduced and this reduced the discriminability of the original training stimuli. Normally, variation in perceptual noise across the stimulus space (and to a certain extent across the experiment) is ignored or

assumed to be relatively trivial. In order to fully account for these results, this assumption may need to be relaxed or removed. In our model fitting, we only analyzed transfer blocks, and since baseline performance in these blocks is stable, we assume the perceptual noise within the discrimination is also stable. Given that we only tested the information integration condition, it would be interesting to examine whether the participants would show the same drop in baseline performance during analogical transfer tests with a rule-based discrimination.

General Discussion

Using the presumably non-analytic pigeon model, we replicated the previously demonstrated benefit of rule-based discrimination during analogical transfer in humans and primates. Further, we also importantly demonstrated that the pigeons' "analogical transfer" was not the result of an abstract "rule" extending to the novel region of the stimulus space. Instead, we found that the successful transfer was supported by systematic associative generalization of the previously learned discrimination to novel regions. We found that a mathematical model of unbound, dimensional activation of the stimulus features well predicted and accounted for the pigeons' generalization to the "analogical transfer" stimuli. More intriguingly, we found that this same model could also account for "analogical transfer" in humans' procedural learning. Thus, it appears the representation of the stimuli during procedural learning is unbound and dimensional in nature and is likely an associative mechanism that operates across a wide variety of species.

The comparison between pigeons and humans on these tasks is somewhat striking. In the previous reports, primates showed strong or complete savings during analogical transfer when in the rule-based condition, and showed little-to-no transfer when conducting the same analogical transfer in the information-integration condition. A similar pattern of transfer emerged with the pigeons in the first experiment. That is, pigeons showed above-chance analogical transfer discrimination in the rule-based condition, but not in the information-integration condition. Unlike the primates, however, their performance was not perfect, and they showed a performance decrement with the novel stimuli. Nevertheless, their above-chance performance suggests that something from the dimensional training – and hence the dimensions themselves – was relevant and useful to these animals. The pigeons in the information integration condition, like humans, failed to transfer their learning to the novel region of "analogical" stimuli used in previous investigations. Additional testing across the entire stimulus space, however, revealed that the information integration condition could support systematic transfer of a dimensional nature. Such results contrast with theories of configural perceptual categorization by demonstrating that both pigeons' and humans' responses to the novel stimuli depended independently on the values of each component dimension.

Our results are not well accorded by previous computational models of procedural learning and categorization. The critical issue raised by our methods and results is that any successful model needs to divide the stimulus space into three distinct areas. Two of these areas are the two distributions of the initial

training, which comprise the basic categorical task. Most models can divide and associate these areas effectively. The third clearly associated area, however, is in an untrained region of the stimulus space, and critically, it is associated with the more distant training distribution. Previous rule-based models would primarily rely on extensions of straight line decision rules to divide the space into two response areas, but not three. Traditional configural-based associative models also do not suffice, because their associations are driven by simple Euclidean distance – so they cannot effectively create the third response area we found. Instead, we found that an associative model that used independent dimensions worked to best capture the pattern of data. In this scheme, the specific placement and organization of the two training distributions in the II condition generated this three-area pattern.

Overall the data and their associative account further confirm the idea that the pigeons use a singular categorization and learning system during both RB and II tasks. This is unlike humans who appear to employ two distinct systems in dealing with these conditions. There is ample evidence describing differences in human categorization mechanisms during procedural learning tasks (i.e., information-integration conditions) and matched rule-based dimensional discriminations. The II and RB training designs demonstrate differences in acquisition rate and efficacy (Ashby & Maddox, 1990; Smith et al., 2010), selective effects of task interference (Waldron & Ashby, 2001), selective effects of feedback availability (Maddox, Ashby, & Bohil, 2003), selective effects of feedback processing (Maddox, Ashby, Ing, & Pickering, 2004), and selective

effects of task shifts (Ashby, Ell, & Waldron, 2003), to name a few. That said, the current results with both the pigeons and humans call into question using evidence of differential "analogical" transfer as evidence for these separate mechanisms.

Not all of these distinctions have been tested within pigeons, although the present experiments represent an effort to continue to do so. What has been found is that pigeons failed to show acquisition differences between the two tasks (Smith et al., 2011). The acquisition data of Experiment 1 replicated the result of the previous investigations, showing equivalent rates of acquisition times in both RB and II conditions. One implication raised from that finding was that perhaps dimensions are irrelevant to pigeons (Smith et al., 2011). The analysis of the current results and supporting model from Experiments 2 and 3 suggest that this is not so. It appears that their single categorization system is sensitive to and functions on a dimensional basis. Further, the current results critically show that pigeons' singular system and the humans' procedural learning system operate very similarly in this manner.

With novel results such as these, one possibility that may be raised is that the results are unique to the experimental circumstances tested. In this case, perhaps the sine-wave gratings used here were critical to the dimensional, separable perception and therefore its resulting categorization. Perhaps the use of an alternative stimulus where the continuous-valued multivariate feature space is perceived as a configuration or a gestalt would be more likely to conform to the traditional configural model of categorization. To this end, the study of known bound dimensions such as the hue of a patch with its brightness or the perception

of a face may be illuminating (Farah, Wilson, Drain, & Tanaka, 1998; Garner & Felfoldy, 1970). However, examining these stimuli in an appropriately comparative scope may be difficult. Pigeons are known to attend more to local features than global features (Qadri & Cook, 2015), and the spatial aspects of faces seem critical to the configural perception of faces (Leder & Bruce, 2000). Determining a common configural perceptual stimulus will be important to further advance this study.

These developments require us to reconsider more broadly the implications for human categorization, which we can begin by considering the impact on the development of the COVIS/SPEED two-system models. We have concluded that the stimulus representation as it enters the procedural categorical learning system is dimensional, rather than configural. These dimensional stimulus features are not bound. How does this conclusion impact our understanding of the analytic, rule-based, "dimensional" system? Previously, the critical difference between the two systems was proposed to be that the analytic system is dimensional in nature and capable of decomposing the configural stimulus into its subcomponents. If the procedural system uses the unbound, dimensional activations to begin with, how could the analytic system then use a bound representation that it decomposes to ultimately produce faster and more robust discrimination?

The differences between the rule-based and procedural-learning systems can no longer be that one is dimensionally "aware" and the other is not. Perhaps the rule-based system can process rules more effectively in a hypothesis- or

model-testing fashion. In this case, the system may be able to consider and evaluate outcomes with respect to multiple hypotheses simultaneously, resulting in faster and more robust learning. Alternatively, the rule-based system may have access to different, more expansive dimensional representations than the association-based system. Perhaps, the activation of a value of 60° may simultaneously activate the representations "less than 75°," "less than 80°," etc., and "more than 45°," "more than 40°," etc. Utilizing such comparison relations could potentially allow for the sort of robust rule learning that underlies the analogical transfer in this task. Yet again, perhaps the rule-based system has the ability to use attention to change the salience of irrelevant dimensions to zero. By applying this sort of attentional hyper-modulation, rule-based discriminations are sped up and information-integration discriminations are not. These possibilities need to be considered in the scope of larger datasets and models.

These results suggest an interesting problem of the integrated processing of separable dimensions, which also bears on the problem of dimensional binding. The "binding problem" refers to (among other things) the algorithmic complexity introduced by having specific feature receptors that separately process visual input for the same object or region of space as well as the problem of preattentive featural responding in behavior (Treisman, 1996). With rapidly processed or separately processed information, how do the features later get related and bound into a single object? The question of integral and separable dimensions bears some similarity to the binding problem, but instead of questioning how separated features are re-processed together, it examines the problem of dimensions that

empirically seem to be bound together in their processing. For example, the hue of a color patch seems to be impossible to process without also processing its intensity, so a patch's hue and intensity are considered integral dimensions (Garner & Felfoldy, 1970). This integrality can be demonstrated by finding redundant benefits and/or interference effects when processing both dimensions or attempting to attend to only one of the dimensions. The stimuli used in the experiments here were intentionally composed of separable dimensions, line angle and spatial frequency. In some schemes, the fact that the dimensions are separable would imply that a bound representation would be unlikely to underlie categorization, but integrality can occur at several levels in the perception-action sequence (Ashby & Townsend, 1986). Previous work suggests that the design of the distributions in this task forces integration in representational or decisional processes using associative methods. However, these data reveal that the integrated decisional process does not require bound representational information.

Some developmental studies elucidate the possibilities in this domain. These studies use a Wisconsin Card Sorting Task methodology, where participants try to sort a deck into two piles with given exemplars as quickly as possible. An investigation of binding in 2 year old, 3 year old, and 5 year old children demonstrated that the processing of dimensions as integral or separable has been shown to have a developmental trajectory (Kemler & Smith, 1978), where bound, integral representations are initially used and unbound, separable representations are later developed or learned. This could have been one reason to believe that procedural learning may utilize a configural, bound representation,

because it appears ontogenetically earlier in humans than unbound processing. Perhaps testing in a similar age range could identify a developmental trajectory to the rule-based benefit. This was examined by Huang-Pollock, Maddox, and Karalunas (2011), although their participants were of an age past the development of unbound processing. In that study, the authors found that the children had difficulty inhibiting the explicit dimensional strategies and consequently suffered in the implicit information integration condition. Testing with younger children may be necessary to understand the relationship between binding or integrality/separability and the analytic benefit.

Since the COVIS/SPEED model is based on neurophysiology and tries to be a realistic model of cognitive processing, these conclusions inform us about neuroprocessing. In the SPEED model, a 100×100 grid of units creates a 10,000 unit representation that is bound and configural (Ashby et al., 2007). The same representation seems to be implied in the SPC model and the COVIS model (Ashby et al., 1998; Ashby & Waldron, 1999). The assumption that the stimulation or representation is bound with a decay to nearby units is a straightforward one to accept: activation in visual cortex has been suggested to be spatial frequency and orientation specific in a similar grid-like fashion (De Valois et al., 1982; Issa et al., 2000). Specific neurons in early visual cortex are thought to be optimally activated by a stimulus with specific properties – a stimulus at a specific point in visual space of a specific orientation and specific spatial frequency (Baker & Issa, 2005; Daugman, 1985). Thus, a neuron's singular activation represents a bound activation of the stimulus properties. However, this

bound representation can (and according to these results, does) become unbound again, not ensuring that all activation thereforward is bound (Treisman, 1996). The COVIS model originally posited the inferotemporal cortex and extrastriate visual areas as the input into the decision process that occurs in the tail of the caudate nucleus (Ashby et al., 1998), and thus, these results suggest that the representation of spatial frequency and orientation in these neural structures is dimensional and unbound.

Finally, the difference between rule-based and information-integration strategies could be attentional in nature. Attentional strategies in perceptual decision making would require an organism to possess neural structures with the ability to modulate incoming sensory signals to alter behavior. In the primate brain, the rule-based system has been attributed to utilize specifically this sort of attention hardware in the prefrontal cortex (PFC). In both mathematical models and neural imaging, PFC and related structures were shown to be relevant for the differences in rule-based and information-integration discriminations (Ashby et al., 1998; Ashby et al., 2007; Nomura et al., 2007). The procedural learning mechanisms are modeled according to the striatum activations and behavior (Ashby et al., 1998). These structures have both been repeatedly implicated and analyzed with respect to human visual categorization – but what about pigeons? The visual system of the pigeon is organized in a fundamentally different manner as compared to the primate visual system, making direct comparisons difficult as structural homology is not widespread (Cook, Qadri, & Keller, 2015). The laminar processing of early visual information in primates is instead nuclear

processing in pigeons, with the primary work seemingly occurring in the nucleus rotundus. Modulatory attentional structures have been difficult to identify, but there is some recent evidence that the nidopallium caudolaterale processing relates to human PFC processing (Lengersdorf, Pusch, Güntürkün, & Stüttgen, 2014). Thus, pigeons may have the capacity for similar two-system categorization processing of these stimuli, but these procedures may not tap into those cognitive strategies. Alternatively, the strength of modulation by the nidopallium may not rival the strength of modulation in the primate PFC. Further knowledge of the neural structures involved, in pigeons especially, may help identify the critical difference between the procedural learning system that is common to both pigeons and humans and the rule-based system that appears to be absent in pigeons.

References

Ashby, F. G. (1992a). Multidimensional models of categorization. In F. G. Ashby (Ed.), *Multidimensional models of perception and cognition* (pp. 449-483). Hillsdale, NJ, England: Lawrence Erlbaum Associates, Inc.

Ashby, F. G. (Ed.) (1992b). *Multidimensional models of perception and cognition*. Hillsdale, NJ, England: 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*(3), 442-481. doi:10.1037/0033-295X.105.3.442

Ashby, F. G., Ell, S. W., & Waldron, E. M. (2003). Procedural learning in perceptual categorization. *Memory & Cognition, 31*(7), 1114-1125. doi:10.3758/bf03196132

Ashby, F. G., Ennis, J. M., & Spiering, B. J. (2007). A neurobiological theory of automaticity in perceptual categorization. *Psychological Review, 114*(3), 632-656. doi:10.1037/0033-295X.114.3.632

Ashby, F. G., & Maddox, W. T. (1990). Integrating information from separable psychological dimensions. *Journal of Experimental Psychology: Human Perception and Performance, 16*(3), 598-612. doi:10.1037/0096-1523.16.3.598

Ashby, F. G., & Maddox, W. T. (2005). Human category learning. *Annual Review of Psychology, 56,* 149-178. doi:10.1146/annurev.psych.56.091103.070217

Ashby, F. G., & Townsend, J. T. (1986). Varieties of perceptual independence. *Psychological Review, 93*(2), 154-179. doi:10.1037/0033-295X.93.2.154

Ashby, F. G., & Waldron, E. M. (1999). On the nature of implicit categorization. *Psychonomic Bulletin & Review, 6*, 363-378. doi:10.3758/BF03210826

Baker, T. I., & Issa, N. P. (2005). Cortical Maps of Separable Tuning Properties Predict Population Responses to Complex Visual Stimuli. *Journal of Neurophysiology, 94*(1), 775-787. doi:10.1152/jn.01093.2004

Beach, L. R. (1964). Cue probabilism and inference behavior. *Psychological Monographs: General and Applied, 78*(5-6), 1-20. doi:10.1037/h0093853

Bishop, C. M. (2006). *Pattern recognition and machine learning*: Springer-Verlag New York, Inc.

Blackwell, H. R., & Schlosberg, H. (1943). Octave generalization, pitch discrimination, and loudness thresholds in the white rat. *Journal of Experimental Psychology, 33*(5), 407-419. doi:10.1037/h0057863

Blough, D. S. (1961). The shape of some wavelength generalization gradients. *Journal of the Experimental Analysis of Behavior 4 1961, 31-40 Journl of the Experimental Analysis of Behavior, US.*

Casale, M. B., Roeder, J. L., & Ashby, F. G. (2012). Analogical transfer in perceptual categorization. *Memory & Cognition, 40*(3), 434-449. doi:10.3758/s13421-011-0154-4

Cook, R. G., Qadri, M. A. J., & Keller, A. M. (2015). The analysis of visual cognition in birds: Implications for evolution, mechanism, and representation. *Psychology of Learning and Motivation, 63*, 173-210. doi:10.1016/bs.plm.2015.03.002

Daugman, J. G. (1985). Uncertainty relation for resolution in space, spatial frequency, and orientation optimized by two-dimensional visual cortical filters. *Journal of the Optical Society of America A, 2*(7), 1160-1169. doi:10.1364/JOSAA.2.001160

De Valois, R. L., Albrecht, D. G., & Thorell, L. G. (1982). Spatial frequency selectivity of cells in macaque visual cortex. *Vision Research, 22*(5), 545-559. doi:10.1016/0042-6989(82)90113-4

Farah, M. J., Wilson, K. D., Drain, M., & Tanaka, J. N. (1998). What is "special" about face perception? *Psychological Review, 105*(3), 482-498. doi:10.1037/0033-295X.105.3.482

Ganz, L., & Riesen, A. H. (1962). Stimulus generalization to hue in the darkreared macaque. *Journal of Comparative and Physiological Psychology, 55*(1), 92-99. doi:10.1037/h0044987

Garner, W. R., & Felfoldy, G. L. (1970). Integrality of stimulus dimensions in various types of information processing. *Cognitive Psychology, 1*(3), 225-241. doi:10.1016/0010-0285(70)90016-2

George, D. N., & Pearce, J. M. (2012). A configural theory of attention and associative learning. *Learning & Behavior, 40*(3), 241-254. doi:10.3758/s13420-012-0078-2

Ghirlanda, S., & Enquist, M. (2003). A century of generalization. *Animal Behaviour, 66*(1), 15-36. doi:10.1006/anbe.2003.2174

Herrnstein, R. J., & Loveland, D. H. (1975). Maximizing and matching on concurrent ratio schedules. *Journal of the Experimental Analysis of Behavior, 24*(1), 107-116. doi:10.1901/jeab.1975.24-107

Hovland, C. I. (1937). The generalization of conditioned responses; I. The sensory generalization of conditioned responses with varying frequencies of tone. *Journal of General Psychology, 17*, 125-148. doi:10.1080/00221309.1937.9917977

Huang-Pollock, C. L., Maddox, W. T., & Karalunas, S. L. (2011). Development of implicit and explicit category learning. *Journal of Experimental Child Psychology, 109*(3), 321-335. doi:10.1016/j.jecp.2011.02.002

Issa, N. P., Trepel, C., & Stryker, M. P. (2000). Spatial Frequency Maps in Cat Visual Cortex. *The Journal of Neuroscience, 20*(22), 8504-8514.

Jassik-Gerschenfeld, D., & Hardy, O. (1979). Single-neuron responses to moving sine-wave gratings in the pigeon optic tectum. *Vision Research, 19*, 993-999.

Kemler, D. G., & Smith, L. B. (1978). Is there a developmental trend from integrality to separability in perception? *Journal of Experimental Child Psychology, 26*(3), 498-507. doi:10.1016/0022-0965(78)90128-5

Leder, H., & Bruce, V. (2000). When inverted faces are recognized: The role of configural information in face recognition. *The Quarterly Journal of*

Experimental Psychology Section A, 53(2), 513-536. doi:10.1080/713755889

Lengersdorf, D., Pusch, R., Güntürkün, O., & Stüttgen, M. C. (2014). Neurons in the pigeon nidopallium caudolaterale signal the selection and execution of perceptual decisions. *European Journal of Neuroscience, 40*(9), 3316-3327. doi:10.1111/ejn.12698

Lumsden, E. A. (1977). Generalization of an operant response to photographs and drawings/silhouettes of a three-dimensional object at various orientations. *Bulletin of the Psychonomic Society, 10*(5), 405-407.

Maddox, W. T., & Ashby, F. G. (1993). Comparing decision bound and exemplar models of categorization. *Perception & Psychophysics, 53*(1), 49-70. doi:10.3758/bf03211715

Maddox, W. T., Ashby, F. G., & Bohil, C. J. (2003). Delayed feedback effects on rule-based and information-integration category learning. *Journal of Experimental Psychology: Learning, Memory, and Cognition, 29*, 650-662. doi:10.1037/0278-7393.29.4.650

Maddox, W. T., Ashby, F. G., Ing, A. D., & Pickering, A. D. (2004). Disrupting feedback processing interferes with rule-based but not information-integration category learning. *Memory & Cognition, 32*(4), 582-591. doi:10.3758/bf03195849

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

Nomura, E., Maddox, W., Filoteo, J., Ing, A., Gitelman, D., Parrish, T., . . . Reber, P. (2007). Neural Correlates of Rule-Based and Information-Integration Visual Category Learning. *Cerebral Cortex, 17*(1), 37-43. doi:10.1093/cercor/bhj122

Nosofsky, R. M. (1988). Exemplar-Based Accounts of Relations between Classification, Recognition, and Typicality. *Journal of Experimental Psychology: Learning Memory and Cognition, 14*, 700-708.

Nosofsky, R. M., & Johansen, M. K. (2000). Exemplar-based accounts of "multiple-system" phenomena in perceptual categorization. *Psychonomic Bulletin & Review, 7*(3), 375-402.

Pearce, J. M. (2002). Evaluation and development of a connectionist theory of configural learning. *Animal Learning & Behavior, 30*(2), 73-95. doi:10.3758/BF03192911

Qadri, M. A. J., & Cook, R. G. (2015). Experimental divergences in the visual cognition of birds and mammals. *Comparative Cognition & Behavior Reviews, 10*, 73-105. doi:10.3819/ccbr.2015.100004

Reed, S. K. (1972). Pattern recognition and categorization. *Cognitive Psychology, 3*, 382-407.

Shepard, R. (1987). Toward a universal law of generalization for psychological science. *Science, 237*(4820), 1317-1323. doi:10.1126/science.3629243

Smith, J. D., Ashby, F. G., Berg, M. E., Murphy, M. S., Spiering, B., Cook, R. G., & Grace, R. C. (2011). Pigeons' categorization may be exclusively

nonanalytic. *Psychonomic Bulletin & Review, 18*(2), 414-421. doi:10.3758/s13423-010-0047-8

Smith, J. D., Beran, M. J., Crossley, M. J., Boomer, J., & Ashby, F. G. (2010). Implicit and explicit category learning by macaques (*Macaca mulatta*) and humans (*Homo sapiens*). *Journal of Experimental Psychology: Animal Behavior Processes, 36*(1), 54.

Smith, J. D., Zakrzewski, A. C., Johnston, J. J. R., Roeder, J. L., Boomer, J., Ashby, F. G., & Church, B. A. (2015). Generalization of category knowledge and dimensional categorization in humans (Homo sapiens) and nonhuman primates (Macaca mulatta). *Journal of Experimental Psychology: Animal Learning and Cognition, 41*(4), 322-335. doi:10.1037/xan0000071

Tappeiner, C., Gerber, S., Enzmann, V., Balmer, J., Jazwinska, A., & Tschopp, M. (2012). Visual acuity and contrast sensitivity of adult zebrafish. *Frontiers in Zoology, 9*(1), 1-6. doi:10.1186/1742-9994-9-10

Treisman, A. (1996). The binding problem. *Current Opinion in Neurobiology, 6*(2), 171-178. doi:10.1016/S0959-4388(96)80070-5

Waldron, E. M., & Ashby, F. G. (2001). The effects of concurrent task interference on category learning: Evidence for multiple category learning systems. *Psychonomic Bulletin & Review, 8,* 168-176. doi:10.3758/BF03196154