Selective and distributed attention in human and pigeon category learning

Leyre Castro[1], **Olivera Savic**[2], **Victor Navarro**[1], **Vladimir M. Sloutsky**[2], **Edward A. Wasserman**[1]

[1]The University of Iowa

[2]Ohio State University

Abstract

Attention to relevant stimulus features in a categorization task helps to optimize performance. However, the relationship between attention and categorization is not fully understood. For example, even when human adults and young children exhibit comparable categorization behavior, adults tend to attend selectively during learning, whereas young children tend to attend diffusely (Deng & Sloutsky, 2016). Here, we used a comparative approach to investigate the link between attention and categorization in two different species. Given the noteworthy categorization ability of avian species, we compared the attentional profiles of pigeons and human adults. We gave human adults (Experiment 1) and pigeons (Experiment 2) a categorization task that could be learned on the basis of either one deterministic feature (encouraging selective attention) or multiple probabilistic features (encouraging distributed attention). Both humans and pigeons relied on the deterministic feature to categorize the stimuli, albeit humans did so to a much greater degree. Furthermore, computational modeling revealed that most of the adults exhibited maximal selectivity, whereas pigeons tended to distribute their attention among several features. Our findings indicate that human adults focus their attention on deterministic information and filter less predictive information, but pigeons do not. Implications for the underlying brain mechanisms of attention and categorization are discussed.

Keywords

Categorization; Attention; Cognitive Development; Animal Cognition

---

If it has fur, then it must be a mammal; if it has feathers, then it must be a bird. This kind of reasoning is typical of our daily inductive inferences. Indeed, in order to classify numerous

---

objects and events, an organism must perceive and attend to those features that are common to exemplars of one category and that distinguish this category from all others. Thus, it seems logical to say that, when humans and nonhuman animals learn to categorize diverse stimuli, it is advantageous for them to focus on those features that are relevant to mastering the task.

Attention figures prominently in many models of human categorization, including exemplar models (Kruschke, 1992; Nosofsky, 1986, 1992), prototype models (e.g., Smith & Minda, 1998), and clustering models (e.g., Love, Medin, & Gureckis, 2004), as well as in various animal associative learning models (e.g., George & Pearce, 2012; Mackintosh, 1965, 1975). According to these accounts, attention is pliable; it tends to be distributed along multiple stimulus dimensions at the beginning of learning, but it converges on the most relevant dimensions as learning proceeds. In addition, deploying attention is assumed to help optimize performance, particularly when categories can be distinguished on the basis of only a few dimensions. For example, to discriminate between squirrels and chipmunks, one should shift attention to stripes (diagnostic feature) and away from the tail or fur (nondiagnostic features).

Although it is entirely reasonable for organisms to focus their attention on the stimulus feature(s) conveying information that is relevant to solving a categorization task, it is important to appreciate that a category discrimination can also be accomplished by perceiving the overall similarity or family resemblance of the exemplars in each category, so that attention may become more widely distributed among multiple features. It has often been observed that, under explicit, intentional classification and learning conditions, healthy human adults tend to use a single *deterministic* dimension; however, they tend to rely on multiple *probabilistic* dimensions that contribute to overall exemplar similarity under implicit learning conditions (e.g., Kemler Nelson, 1984; Love, 2002; Waldron & Ashby, 2001).

Learning under those different circumstances may involve separate mechanisms, as suggested by COVIS, the categorization theory proposed by Ashby and his colleagues (Ashby et al., 1998; Ashby & Valentin, 2005; Ashby & Waldron, 1999). According to COVIS, category learning may be accomplished by two different systems: 1) a frontal-based explicit system that uses language and logical reasoning, and allows the organism to learn relatively quickly, and 2) a basal ganglia-mediated implicit system that involves procedural learning, and results in learning taking place slowly, in an incremental fashion, and being highly dependent on immediate feedback. Correspondingly, these two systems involve different types of attention. Attention is hypothesis driven under the explicit mechanism, whereas attention is stimulus-driven by the contingencies of reinforcement under the implicit mechanism; that is, stimulus features or dimensions are differentially weighted based on their capacity to predict the correct category. As it has become the custom in the human cognition literature, we will reserve the term *selective attention* for the top-down, hypothesis-driven type of attention.

Early in human development, categories can be learned without engaging selective attention (e.g., Best, Yim, & Sloutsky, 2013). Infants can learn the statistical co-occurrence of several

features within the exemplars of different categories; thus, infants' attention tends to be distributed rather than focused on specific diagnostic features. In a later study, Deng and Sloutsky (2016) gave 4-year-olds, 7-year-olds, and adults a category learning task in which there was a single rule-like deterministic feature that perfectly predicted category membership accompanied by multiple probabilistic features that only probabilistically predicted category membership (a paradigm developed by Kemler Nelson, 1984). After training, participants' categorization behavior and memory were tested in order to identify which features controlled participants' categorization choices and how well these features were remembered. When the instructions directed participants' attention to both the deterministic and probabilistic features (Experiment 1), adults and 7-year-olds tended to rely on the deterministic feature, whereas 4-year-olds relied on the probabilistic features. The 4-year-olds could and did rely on the deterministic feature when the instructions directed them to it (Experiment 2), just as did the 7-year-olds and adults. Yet, even when the categorization choices of children and adults were based on the deterministic feature, their memories were strikingly different. Consistent with their engagement of selective attention, the 7-year-olds and adults exhibited robust memory for the deterministic feature, but not for the probabilistic features; in contrast, the 4-year-olds remembered all of the features equally well, at odds with the idea of selective attention. Thus, older children's and adults' memory pointed to selective attention during category learning, whereas younger children's memory suggested more distributed attention.

Following Ashby and colleagues' theoretical proposal (Ashby et al., 1998; see also, Cincotta & Seger, 2007; Sloutsky, 2010), one possible explanation for this discrepancy between very young children and adults is that selective attention requires the active involvement of brain structures mediating what is called *executive function:* specifically, the prefrontal cortex (PFC), which is immature early in development. From this perspective, selective attention in category learning requires the participation of a fully developed and functional PFC, whereas similarity-based categorization can be accomplished by more primitive brain regions, such as the inferotemporal cortex and the basal ganglia. If a fully developed and functional human PFC is required to exhibit selective attention, then animals that have a less well-developed PFC or that do not have this structure at all may show an attentional pattern more similar to that of young children who have an immature PFC.

Given these premises, Couchman, Coutinho, and Smith (2010) explored whether humans and monkeys (*Macaca mulatta*) would rely on a single deterministic predictor or on several probabilistic predictors when learning to discriminate two artificial categories. Because monkeys have proportionally smaller frontal cortices than humans (Semendeferi, Lu, & Schenker, 2002), it was suspected that they might not show the same attention capacities as do humans. Human adults were trained under either explicit or implicit learning conditions, with the explicit condition being more directly comparable to the task given to monkeys. The prediction was that, as in earlier studies (e.g., Kemler Nelson, 1984), human adults would strongly rely on the deterministic predictor in the explicit condition, but strongly rely on the probabilistic predictors in the implicit condition. However, only 58% of the human participants relied on the single perfect predictor in the explicit condition (36% relied on the probabilistic features in the implicit condition). Two monkeys were given the same basic explicit task in a pair of experiments. The first experiment suffered from low categorization

accuracy during the testing phase. The second experiment, with the same two monkeys, required several procedural modifications to improve their testing accuracy; its results suggested that the monkeys were distributing their attention among all of the features in the categorization stimuli. It seems safe to conclude from this study that humans were inclined to use to the deterministic information, whereas monkeys were more inclined to use the probabilistic information, although the results of the study were not straightforward.

A subsequent study addressing the same issue with pigeons was conducted by Nicholls, Ryan, Bryant, and Lea (2011). Pigeons do not have a PFC, so their tendency to distribute attention and to rely on all of the available features when learning to categorize complex stimuli may be even more clearcut. Indeed, some researchers have contended that pigeons lack the capacity for rule formation and selective attention (Smith et al., 2012); if so, then pigeons' excellent categorization performance may actually be based on their recognition of the overall similarity among the trials belonging to the same category rather than on their deployment of selective attention to the most diagnostic information. In Nicholls et al. (2011, Experiment 2), pigeons were shown artificial categories in which each exemplar was created from four spatially separated features: one was a perfect predictor, whereas the other three were probabilistic predictors (using only one feature could yield 75% accuracy; using all three features was required to reach 100% accuracy). When, in testing, the perfect and probabilistic predictors were put into conflict, most pigeons relied on the perfect predictor to classify the stimuli. Interestingly, those pigeons that did not, also focused on one feature, just not the perfect predictor. So, overall, pigeons' categorization behavior was controlled by only one feature (see also Lea & Wills, 2008, Lea et al., 2009, and Wills et al., 2009, for further results and discussion).

Categorization controlled by a single feature suggests selective attention, because that single feature must be preferentially processed amid all of the available features. More explicit evidence implicating attention to specific features in pigeons' categorization was provided Castro and Wasserman (2014). In that study, pigeons were trained to classify stimuli from two different artificial categories, in which the exemplars contained both relevant (perfect predictors) and irrelevant features. Because tracking of peck location—similar to human eye tracking (e.g., Rehder & Hoffman, 2005)—is a promising proxy for measuring pigeons' allocation of visual attention, Castro and Wasserman required their pigeons to peck anywhere at the category exemplars when they were presented on a computer screen (see also Dittrich, Rose, Buschmann, Bourdonnais, & Güntürkün, 2010). The authors found that, as pigeons' categorization accuracy progressively rose, so too did their pecks to the relevant category features; conversely, pigeons' pecks to the irrelevant category features progressively fell. In short, as pigeons were learning to categorize the stimuli, they also seemed to learn to attend preferentially to the relevant stimulus features (see also Castro & Wasserman, 2016a, 2017).

Although pigeons do not have a PFC—and their pallium is nucleated and lacks the distinctive laminar organization observed in the mammalian cortex (Jarvis et al., 2005) — there are noteworthy parallels between the avian and the mammalian forebrains at the connectivity level. For example, birds' forebrains are both modular, small-world networks with a connective core of hub nodes that includes structures (e.g., the nidopallium

caudolaterale) similar to the mammalian PFC. These hub nodes are centrally located and richly connected (Shanahan, Bingman, Shimizu, Wild, & Güntürkün, 2013). It is thus conceivable that pigeons' brain characteristics are sufficient to allow them to solve categorization tasks akin to those solved by human adults (see Lazareva & Wasserman, 2010, for a comprehensive review).

Considering the prior empirical findings and theoretical analyses, we are left with two related, but unanswered questions. First, do animals lacking a mature mammalian PFC learn categorization tasks by selectively attending to the most predictive information (e.g., Castro & Wasserman, 2014; Nicholls et al., 2011) or do they tend to distribute their attention in the process of category learning so that their usage of the most predictive features is merely the result of those features acquiring high associative strength because they are strong predictors of reinforcement (e.g., Ashby et al., 1998; Couchman et al., 2010; Smith et al., 2012)? Second, to what extent is the role of attention in category learning similar in animals and humans?

To answer these questions, we deployed a category learning paradigm similar to that of Deng and Sloutsky (2016) with both human adults (Experiment 1) and pigeons (Experiment 2), to better understand how these different species attend to and process the available information for solving a categorization task. We also used computational modeling to determine people's and pigeons' attentional profiles in categorization testing. We expected that human adults would be prone to optimize attention and, thus, to focus on the deterministic feature of the category exemplars. At greater issue was pigeons' attentional performance. Would they too optimize their attention and focus on the deterministic feature? Or would they attend more diffusely, relying as well on the probabilistic features, in line with the behavior of very young children?

2. Experiment 1

In Experiment 1, human participants had to learn to categorize exemplars belonging to two categories. Each category had a prototype that was completely different in seven features from the prototype of the other category. The prototype itself was *never* presented in training, but the training exemplars highly matched the prototype (see Figure 1). All of the training exemplars contained one deterministic feature that perfectly distinguished the two categories (e.g., the circle in the center of the exemplars in Figure 1); in addition, the exemplars contained four probabilistic features that were consistent with the corresponding prototype plus two more features that were consistent with the opposite prototype (e.g., Kemler Nelson, 1984). Thus, participants could use either the deterministic feature or the probabilistic features to learn the category discrimination. A minimum of three probabilistic features was necessary to reach accuracy as high as accuracy using the deterministic feature; accuracy could not exceed 66% if only one probabilistic feature was being used.

Testing trials allowed us to determine the participants' categorization strategy. Critically, we included Incongruent trials, which had the deterministic feature of one category, but most of the probabilistic features that were consistent with the *opposite* category. These Incongruent trials permitted us to determine whether participants relied on the overall similarity of the

exemplars (if they were to choose the category that was consistent with the majority of the probabilistic features) or on the deterministic rule (if they were to choose the category that was consistent with the deterministic feature).

In testing, participants were also presented with the prototype of each category in order to assess whether accuracy would increase compared to accuracy with the training exemplars, which always contained four probabilistic features from the training category plus two probabilistic features from the opposite category. If participants were focused on just the deterministic feature, then no difference in accuracy between the prototype and training trials should be observed. But, if the probabilistic features also commanded participants' attention, then higher accuracy for the prototype would be expected, because all of the features in the prototype came from the same category.

The remainder of the testing trials involved replacing one or more of the trained features with a novel feature. We included Only-P trials, which had four of the probabilistic features of one category, two of the probabilistic features of the other category, and a novel feature replacing the trained deterministic feature; these Only-P trials allowed us to assess whether participants could rely on the probabilistic features alone when the deterministic feature was not available. There were also Only-D trials, which had the deterministic feature from one category and six novel features replacing the probabilistic features; these Only-D trials allowed us to determine whether participants could rely on the deterministic feature alone when none of the probabilistic features were available. Finally, we included One-New-P trials, in which a novel feature replaced one of the four probabilistic features of the correct category; because these One-New-P trials lacked one of the training features, just as the New-D trials, they allowed us to see whether a possible decrement in accuracy on New-D trials was due to a strong reliance on the deterministic feature or simply to the replacement of one trained feature.

Examples of testing trials can be seen in Figure 1. Note that the white circles are the novel features replacing the features presented during training.

2. 1. Method

2.1.1. Participants—Participants were 56 human adults. All were undergraduate students at The University of Iowa who received course credit for their participation. Participants provided their informed consent prior to beginning all experimental procedures. The study was conducted in accordance with the Declaration of Helsinki and all procedures were approved by the Institutional Review Board at The University of Iowa.

2.1.2. Stimuli

2.1.2.1. Training stimuli.: A total of 14 colored circles of 2.7 cm of diameter were used to create the different category training exemplars. These colors' wavelengths were all discriminable by humans and have been shown to be discriminable by pigeons as well (Emmerton & Delius, 1980; Palacios & Varela, 1992).

Participants were trained with two categories, A and B. The category structure of the training stimuli is detailed in Table 1 (top rows). Each of the category exemplars contained

seven features forming a circular shape: one feature was placed in the center and the other six surrounded it, separated from one another but connected by a white line (see Figure 1). Each image occupied a 10×10 cm square area. Each category had a prototype that differed in all seven colors from the prototype of the other category; the prototype was *never* presented in training. During training, all exemplars contained one deterministic feature that perfectly distinguished the two categories (e.g., the circle in the center of the exemplars in Figure 1); in addition, the exemplars contained four probabilistic features that were consistent with the corresponding prototype and two more features that were consistent with the opposite prototype. Each colored circle was always presented in the same location (for a given participant), so that the total number of unique training exemplars was 15 for Category A and 15 for Category B.

In order to control for the possibility that some color combinations were more discriminate than others, we created two sets of stimuli, so that each deterministic feature was presented with either Set 1 or Set 2 of probabilistic features; half of the participants were trained with one combination, whereas the other half was trained with the other combination. In addition, to control for a possible preference to attend to some spatial locations, the deterministic feature was placed in the center for 28 participants, and in one of four side locations (top left, top right, bottom left, bottom right) for the other 28 participants.

2.1.2.2.   Testing Stimuli.: In testing, participants were presented with six types of trials: Training, Incongruent, Only-P, Only-D, One-New-P, and Prototype. Examples of the testing stimuli are shown in Figure 1 and their category structure is detailed in Table 1. Training trials were used to assess learning of the trained categories. As explained earlier, Incongruent trials had the deterministic feature of one category, but most of the probabilistic features were consistent with the *opposite* category, so they allowed us to determine whether participants relied on the overall similarity of the exemplars or on the deterministic feature. Only-P trials had the probabilistic features of one category, two of the probabilistic features of the other category, and a novel white feature replacing the trained deterministic feature, so they allowed us to assess whether participants could rely on the probabilistic features when the deterministic feature was not available. Only-D trials had the deterministic feature from one category and all new white features replacing the probabilistic features, so they allowed us to determine whether participants could rely on the deterministic feature when none of the probabilistic features were available. In One-New-P trials, one of the probabilistic features of the correct category was replaced by a novel white feature to see if the replacement of just one relevant feature (as in Only-P trials) would result in a performance decrement. Finally, in Prototype trials, all of the probabilistic features belonged to the correct category, so they allowed us to see if the gain in positive probabilistic features and/or the loss of probabilistic features from the incorrect category would result in a performance increment.

2.1.3.   Procedure—In order to make the human experiment as similar as possible to the pigeon experiment, we minimized the verbal instructions that we gave to the participants. Each of the participants was seated in front of a computer and told that they would be observing a series of images and attempting to learn the correct response for each of them.

For each image, participants had to use the mouse to choose one response button located to the left or to the right of the image. The report buttons were 2.3×6 cm rectangles filled with two distinctive black-and-white patterns. Participants had to select one of the two report buttons, depending on the category presented. If their choice was correct, then they would hear a pleasant tone and move to the next trial; if their choice was incorrect, then they would hear an unpleasant buzz and the same image would again be presented until the correct response was made. No information was provided that could have directed the participant toward any particular aspect of the images. These are the full instructions given, in written form, to the human participants:

You will be observing a series of images and attempting to learn which response button is correct for each of them. These images will appear repeatedly, over a total of approximately 510 trials. First, you will see a white square with a black cross in the center of the screen; you need to click on it to start every trial. Then, an image will appear on the center of the screen. Click on it and the choice response buttons will appear to the left and right of the stimulus. One of the buttons will be correct, the other one will be incorrect. You have to learn to choose the correct one. At the beginning you will have to guess, but auditory feedback will indicate you whether or not your choice is correct. If you choose the correct button, then you will hear a pleasant tone and you will be moved on to the next trial. If you choose the incorrect button, then you will hear a buzz, and the stimulus will appear again until you choose the correct button. Your goal is to accurately choose the correct button as many times as possible.

Once the participants indicated their understanding of the procedure, the experimenter started the program. The program to run the experiment was developed in MatLab® with Psychtoolbox-3 extensions (Brainard, 1997; Pelli, 1997; http://psychtoolbox.org/).

All of the participants received a fixed number of training trials, with the order of presentation randomized for each participant. A total of 360[1] training trials were scheduled; each particular exemplar was presented a total of 12 times. After training was completed, the testing phase started; it continued without a noticeable change, but a total of 88 testing trials were randomly interspersed among 60 more training trials. Differential feedback continued for training trials, so that a pleasant sound was presented if the response was correct and an unpleasant buzz was presented followed by repetition of the trial if the response was incorrect. For testing trials, no differential feedback was given; participants always heard the pleasant sound and advanced to the next trial regardless of their responses. Unique testing trials (Prototype and Only-D trials) were presented eight times each, whereas testing trials that included multiple variations (Incongruent, Only-P, and One-New-P trials) were presented 24 times each, for later analysis of control by each individual feature (see section 4. Computational Modeling).

2.1.4. Data analysis—The data, here and in Experiment 2, were subjected to logit mixed-effects analyses (a generalization of logistic regression; see Jaeger, 2008). Mixed-

---

[1]A prior unpublished experiment in which we used a shorter training phase, 120 trials, resulted in a large proportion of participants not learning the task. We decided to have an extended training phase in order to maximize the number of participants learning the task.

effects models are especially well-suited to analyze repeated-measures data. Mixed-effects models add random effects (random intercepts and random slopes specific to the subjects taking part in an experiment) to the fixed effects (the independent variables familiar in traditional analyses). Thus, mixed-effects models allow one to take into account each subject's variability by computing a random intercept and/or a random slope for each subject and thereby ensure the best estimates of the fixed effects. To select an appropriate random-effects structure (only random intercepts or random slopes as well), we compared models with the same fixed-effects structure and varying complexity in their random-effects structure using the log likelihood ratio test (Wagenmakers & Farrell, 2004). All analyses were conducted using the lme4 version 1.1-21 (Bates, Maechler, Bolker, & Walker, 2015) package of R, version 3.3.2 (R Development Core Team, 2016).

2. 2. Results and Discussion

We chose an inclusion criterion of 85% correct to each of the two categories over the last 30 trials of the training phase; participants failing to meet this criterion were eliminated from later analyses. Of the 56 participants, 52 met this criterion. The 4 participants who did not learn had been presented with the deterministic feature on the left or right side locations of the category exemplars. The accuracy of these participants was at chance in the last block of 30 trials ($M = 54.16\%$, $SE = 4.56$), so we eliminated their data from subsequent analyses. The accuracy of those participants who learned was very high in the last block of training ($M = 99.48\%$, $SE = 0.18$), and there were no differences between participants trained with the deterministic feature in the center location ($M = 99.76\%$, $SE = 0.17$) and participants trained with the deterministic feature in the side locations ($M = 99.16\%$, $SE = 0.33$). In Supplemental Material, we separately present all of the data and all of the analyses for participants trained with the deterministic feature in the center location and participants trained with the deterministic feature in the side locations.

Categorization Choice.—All of the participants' responses to the training and testing trials presented during the testing phase are depicted in Figure 2 (top left). Accuracy was very high (over 95%) for all of the testing trials except for Only-P trials (around 50%, the chance level), the only testing trials in which the deterministic feature was not available. Thus, it seems that people relied almost entirely on the deterministic feature to make their choices. When the deterministic feature was absent, they failed to make the correct choice.

We first evaluated whether the counterbalancing of the location of the deterministic feature had any effect on the participants' testing performance. A logistic model was fit with the location of the deterministic feature (Center vs. Side) as the fixed effect (treatment coded, with center as the reference condition). This analysis did not yield differences in accuracy between the center and side locations, $B = -0.07$, $SE = 0.08$, $Z = -0.96$, $p = .34$; $M = 91.5$, 95% CI_M [90.7, 92.4] and $M = 90.9$, 95% CI_M [89.9, 91.8], respectively, so we dropped this factor from all subsequent analyses.

Next, a logistic mixed-effects model was fit with type of test trial (Training, Prototype, Incongruent, Only-D, Only-P, One-New-P) as the fixed effect (treatment coded, with training trial as the reference condition). A log-likelihood ratio test indicated that the best

fitting random-effects structure included a random intercept for participant as well as random participant slopes for test type. Critically, there were no statistical differences among the Training, Prototype, Incongruent, Only-D, and One-New-P trials; on all of those trials, the deterministic feature was available. Indeed, when the deterministic and probabilistic features were put into conflict on Incongruent trials, participants robustly relied on the deterministic feature, $M = 99.8$, 95% CI_M [99.6, 100] (for scoring purposes, on Incongruent trials we considered correct those choices that were consistent with the category to which the deterministic feature belonged). Only on those trials in which the deterministic feature was absent, Only-P trials, did accuracy drop to chance level, $M = 49.7$, 95% CI_M [46.9, 52.4] and was choice accuracy statistically different from Training trials, $B = -5.41$, $SE = 0.27$, $Z = -19.79$, $p < .001$, $d = 4.25$.

This pattern of performance reveals that human participants seem to have relied entirely on the deterministic feature when they made their categorization choices to the novel testing trials. Indeed, when only the probabilistic features were available, participants' performance fell to chance level, suggesting that they might not have learned anything about any of the probabilistic features.

Reaction time.—We also analyzed reaction time (RT) to perform the categorization choice during the testing phase. In order to normalize the distributions, the RTs were subjected to log-transformation before statistical analyses (see Ratcliff 1993, for several different methods to deal with non-normal reaction time distributions and reaction time outliers). Participants' RTs to the training and testing trials presented during the testing phase are depicted in Figure 2 (top right). RTs were very similar to all testing stimuli, except for Only-D and Only P trials, that yielded longer RTs.

A linear mixed-effects model was fit with type of test trial (Training, Prototype, Incongruent, Only-D, Only-P, One-New-P) as the fixed effect (treatment coded, with training trial as the reference condition). A log-likelihood ratio test indicated that the best fitting random-effects structure included a random intercept for participant. Critically, there were no statistical differences in RTs among the Training, Prototype, Incongruent, and One-New-P trials. But, on those trials in which the deterministic feature was absent, Only-P trials, participants were slower, $M = 590$ ms, 95% CI_M [572, 609], than on Training trials, $M = 508$ ms, 95% CI_M [501, 515], $B = 0.149$, $SE = 0.012$, $t = 11.95$, $p < .001$, $d = 0.31$. Participants were also slower when only the deterministic feature was present, on Only-D trials, $M = 603$ ms, 95% CI_M [577, 631], compared to Training trials, $B = 0.172$, $SE = 0.019$, $t = 8.85$, $p < .001$, $d = 0.40$.

It may have been that, when the deterministic feature was absent on Only-P trials, participants were uncertain as to which response was correct, consistent with their accuracy falling to chance level. However, such uncertainty is unlikely to have been the reason for participants' slowness to respond on Only-D trials, given that their accuracy was 96%. Perhaps because Only-D trials displayed only one feature, and all of the others were novel white circles, their appearance may have surprised the participants and, consequently, slowed their choice responses.

3.   Experiment 2

Next, we studied the categorization response patterns of a very different species. Pigeons had to learn to categorize exemplars belonging to the same two categories as in Experiment 1. The experimental design and category structure for training and testing exemplars were the same as in Experiment 1. The only disparities were related to the different regimens required to study both species.

3. 1.   Method

3.1.1.   Subjects—The subjects were 16 homing pigeons (*Columba livia*) maintained at 85% of their free-feeding weights by controlled daily feedings. The pigeons had served in unrelated studies prior to the present project. Pigeons' ages ranged from 2 to 11 years. All procedures were approved by the Institutional Animal Care and Use Committee at The University of Iowa.

3.1.2.   Apparatus—The experiment used four $36 \times 36 \times 41$ cm operant conditioning chambers detailed by Gibson, Wasserman, Frei, and Miller (2004). The chambers were located in a dark room with continuous white noise. Each chamber was equipped with a 15-in LCD monitor located behind an AccuTouch® resistive touchscreen (Elo TouchSystems, Fremont, CA). The portion of the screen that was viewable by the pigeons was 28.5 cm $\times$ 17.0 cm. Pecks to the touchscreen were processed by a serial controller board outside the box. A rotary dispenser delivered 45-mg pigeon pellets through a vinyl tube into a food cup located in the center of the rear wall opposite the touchscreen. Illumination during the experimental sessions was provided by a houselight mounted on the upper rear wall of the chamber. The pellet dispenser and houselight were controlled by a digital I/O interface board. Each chamber was controlled by its own Apple® iMac® computer. Just as in Experiment 1, the program to run the experiment was developed in MatLab® with Psychtoolbox-3 extensions (Brainard, 1997; Pelli, 1997).

3.1.3.   Stimuli—The stimuli were the same as in Experiment 1. As detailed above, each deterministic feature was shown with either Set 1 or Set 2 of probabilistic features; half of the pigeons were trained with one combination, whereas the other half was trained with the other combination. In addition, the deterministic feature was placed in the center for half of the birds and in one of the four side locations (top left, top right, bottom left, bottom right) for the other half of the birds.

3.1.4.   Procedure

3.1.4.1.   Training.: Daily training sessions comprised 120 trials: half presented Category A exemplars and half presented Category B exemplars, in a random fashion. At the beginning of a trial, the pigeons were presented with a start stimulus, a white square (3×3 cm) in the center of the computer screen. After one peck anywhere on this white square, one category exemplar was displayed in the center of the screen. The pigeons had to satisfy an observing response requirement (gradually increased from 2 to a maximum of 20 pecks, on a daily basis). This requirement was adjusted to the performance of each pigeon. If the bird was consistently pecking, but not meeting the discrimination criterion (see ahead) in a timely

fashion, then the number of pecks was raised to increase the cost of making an incorrect response.

On completion of the observing response requirement (number of pecks at the stimulus image), two report buttons appeared 4.5 cm to the left and right of the category exemplar. The report buttons were 2.3×6 cm rectangles filled with two distinctive black-and-white patterns. The pigeons had to select one of the two report buttons, depending on the category presented. If the choice was correct, then food reinforcement was delivered and the intertrial interval (ITI) ensued; the ITI randomly ranged from 6 to 10 s. If the choice was incorrect, then food was not delivered, the houselight was darkened, and a correction trial was scheduled. Correction trials were given until the correct response was made. No data were analyzed from correction trials.

We trained the birds until they reached an accuracy level of 85% for each of the categories on 2 consecutive days, to ensure that categorization performance had reached a high and stable level. Then, we started the testing phase.

3.1.4.2. Testing.: Each testing session began with 12 warm-up training trials. The next 128 trials comprised 108 training trials plus 20 randomly interspersed testing trials (four of each type: Prototype, Incongruent, Only-P, Only-D, and One-New-P trials). A total of 12 testing sessions were given. On training trials, only the correct response was reinforced; incorrect responses were followed by correction trials (differential reinforcement). On testing trials, any choice response was reinforced (nondifferential reinforcement); food was given regardless of the pigeons' choice responses, so that testing could proceed without explicitly teaching the birds the correct responses to the testing exemplars. No correction trials were given on testing trials.

3. 2. Results and Discussion

Of the 16 pigeons, 15 reached the learning criterion (85% correct on each category on 2 consecutive days) in an average of 16 days ($SD = 11.24$). The fastest bird reached criterion in 7 days, whereas the slowest bird took 36 days. After 52 days of training, the sole remaining pigeon (Bird 83Y) fell slightly short of criterion with an accuracy level of 79% on the last 2 days; none of the results of the subsequent analyses changed due to the presence or absence of the data of this bird, so we decided to include its data in all subsequent analyses.

Categorization Choice.—The pigeons' responses to the training and testing trials during the testing phase are shown in Figure 2 (bottom). Accuracy was very high for Training, Prototype, and One-New-P trials. On Incongruent trials, in which the deterministic and probabilistic features were put into conflict, the deterministic feature seemed to exert greater control than the probabilistic features. Consistent with those choices, when only the deterministic feature was available on Only-D trials, pigeons' accuracy was high, although lower than their accuracy on Training trials. And, although accuracy was not very high, it seems that pigeons could still solve the categorization task on Only-P trials, when only the probabilistic features were available.

Just as for people, we first evaluated whether the counterbalancing of the location of the deterministic feature had any effect on the pigeons' testing performance. A logistic model was fit with the location of the deterministic feature (Center vs. Side) as the fixed effect (treatment coded, with center as the reference condition). This analysis did not yield differences in accuracy between the center and side locations, $B = -0.22$, $SE = 0.21$, $Z = -1.01$, $p = .31$; $M = 91.1$, 95% CI_M [90.7, 91.6] and $M = 90.5$, 95% CI_M [90.0, 91.1], respectively; so, we dropped this factor from all subsequent analyses.

Next, a logistic mixed-effects model was fit with type of test trial (Training, Prototype, Incongruent, Only-D, Only-P, One-New-P) as the fixed effect (treatment coded, with training trial as the reference condition). A log-likelihood ratio test indicated that the best fitting random effects structure included a random intercept for bird, as well as random bird slopes for test type. Accuracy on Training trials was very high, $M = 92.8$, 95% CI_M [90.5, 95.1], yet it was even higher on Prototype trials, $M = 99.5$, 95% CI_M [98.9, 99.9], suggesting that seeing all of the probabilistic features from the correct category did improve pigeons' categorization accuracy. This improvement was statistically significant, $B = 2.64$, $SE = 0.68$, $Z = 3.86$, $p < .001$, $d = 0.61$.

When the deterministic and probabilistic features were put into conflict on Incongruent trials, pigeons tended to rely on the deterministic feature, $M = 76.3$, 95% CI_M [67.0, 85.6]. However, the decrement in pigeons' performance compared to Training trials was large and statistically significant, $B = -1.41$, $SE = 0.14$, $Z = -9.69$, $p < .001$, $d = 1.53$. Pigeons' accuracy was still robust on Only-D trials, when only the deterministic feature was available, $M = 70.2$, 95% CI_M [62.1, 78.3], but lower than their accuracy on Training trials, $B = -1.79$, $SE = 0.19$, $Z = -9.37$, $p < .001$, $d = 2.09$.

When only the probabilistic features were available, pigeons' accuracy dropped even more, $M = 63.3$, 95% CI_M [58.2, 68.4] compared to Training trials, $B = -2.21$, $SE = 0.21$, $Z = -10.44$, $p < .001$, $d = 2.72$. Still, pigeons' accuracy was significantly higher than the 50% chance level, $t(15) = 5.52$, $p < .001$, $d = 1.38$. So, it seems that the pigeons could also use some of the probabilistic features to perform the task (probably just a subset of them because, had they been able to use all of the probabilistic features, their accuracy would have been higher). When on One-New-P trials only one of the probabilistic features was removed, pigeons' accuracy was also very high, $M = 87.6$, 95% CI_M [83.5, 91.7], although there was a small performance decrement compared to Training trials, $B = -0.69$, $SE = 0.15$, $Z = -4.54$, $p < .001$, $d = 0.48$.

Thus, when the deterministic and probabilistic features were put into conflict on Incongruent trials, pigeons tended to rely on the deterministic feature more than on the probabilistic features, most likely because they had learned the perfect predictive value of the deterministic feature. However, pigeons' decrease in accuracy on Only-D and One-New-P trials, along with their improvement on Prototype trials and their higher than chance accuracy on Only-P trials, all suggest that at least some of the probabilistic features were having a measurable impact on their performance as well.

Reaction time.—We also analyzed, in the pigeons, reaction time (RT) to perform the categorization choice during the testing phase. As was the case for the human participants, the pigeons' RTs were subjected to log-transformation before statistical analyses in order to normalize the distributions. As can be seen in Figure 2 (bottom right), pigeons' RT to the training and all testing trials presented during the testing phase was almost exactly the same.

A linear mixed-effects model was fit with type of test trial (Training, Prototype, Incongruent, Only-D, Only-P, One-New-P) as the fixed effect (treatment coded, with training trial as the reference condition). A log-likelihood ratio test indicated that the best fitting random effects structure included a random intercept for subject. There were no statistical differences in RT among any of the stimuli. Thus, pigeons' time to choose the response buttons did not vary depending on the type of stimuli, and no differential processing can be inferred from this measure.

In sum, when given the same task, humans and pigeons exhibited both similar (greatest reliance on the deterministic feature compared to the other features), but nonetheless disparate patterns of categorization behavior (pigeons seemed to have learned about the probabilistic cues, but humans did not). These results suggest that the two species might have been attending to different features of the stimuli to different degrees. In order to gain a clearer understanding of humans' and pigeons' attention and categorization behavior, we next used a modeling approach to determine their attentional profiles during categorization testing.

4. Computational Modeling

We sought to determine to what extent people's and pigeons' attention was selective and focused on a single feature (presumably the deterministic feature) or distributed across some or all of the features in the stimuli. To do so, we modelled both species' categorization choices during testing to infer utilization scores for each feature (Macho, 1997).

A suitable modeling tool to better understand humans' and pigeons' performance is Nosofsky's (1986, 1988, 1992) Generalized Context Model (GCM). A core assumption of GCM is that organisms represent categories by storing individual training exemplars in memory; later classification—of novel testing exemplars, for example—is based on similarity comparisons between the novel exemplars and the stored exemplars. Thus, according to GCM, in the case of two mutually exclusive categories, A and B, the probability that Stimulus S_i is classified in Category C_A, that is, $P(R_A|S_i)$, is given by the following equation:

$$P(R_A|S_i) = \frac{b_A \sum_{a \in C_A} sim_{ia}}{b_A \sum_{a \in C_A} sim_{ia} + (1 - b_A) \sum_{b \in C_B} sim_{ib}}$$

where b_A ($0 \leq b_A \leq 1$) is Category A response bias and sim_{ia} and sim_{ib} are similarities between a given exemplar i and exemplars belonging to Categories A and B, respectively. Similarity between items is an exponential decay function of psychological distance d derived from this equation:

$$sim_{ia} = e^{-cd_{ia}}$$

where c ($0 \leq c \leq \infty$) is a sensitivity parameter reflecting overall discriminability in the psychological space, with larger values representing greater discriminability. Psychological distance d is calculated according to the following equation:

$$d_{ia} = \left[\sum_{m=1}^{M} w_m |x_{im} - x_{am}|^r \right]^{1/r}$$

where x_{im} is the psychological value of exemplar i on dimension m. The value of the parameter r typically takes values of 1 or 2, for separable or integral dimensions, respectively (Shepard, 1964); given the characteristics of our stimuli, we used an $r = 1$, that results in the city-block metric as the distance between exemplars in multidimensional space. Most critically, w_m ($0 \leq w_m \leq 1$) is the attentional weight given to a dimension or feature m. These attentional weights are free parameters and are interpreted as reflecting the attention that is allocated to each dimension during categorization (Nosofsky, 1986; Viken, Treat, Nosofsky, McFall, & Palmeri, 2002). Attentional weights are of critical importance because they change as a result of categorization training (Nosofsky, 1986); therefore, they reflect the amount of attention allocated to a particular feature. Because a limited-capacity system is assumed, the sum of all of the attentional weights equals 1; so, the more attention is paid to a particular feature, the less attention is paid to the other features.

We used GCM to estimate the attentional weights that best accounted for the responses of both humans and pigeons on the categorization testing trials. In order to do so, we assumed that all of the exemplars presented during training were in GCM's memory. In order to get a close assessment of the attentional weights, we fixed the sensitivity parameter ($c = 3$; which is often estimated from data) and adopted a city-block metric ($r = 1$). By means of numerical optimization, we found the attentional weights that minimized the sum of squared error (SSE) between GCM's and a given subject's responses to the stimuli presented during testing.

Once we obtained the vector of attentional weights for each subject, we quantified the attentional profile of each subject by calculating the entropy of their vector of attentional weights. Given that we were interested in determining whether our subjects focused on one feature or distributed their attention over some or all of them, we considered that *entropy*, a measure of variety or diversity provided by information theory (Shannon & Weaver, 1949), would be a good candidate for this purpose. Entropy measures the amount of informational diversity by computing a weighted average of the number of bits of information that, in our case, each of the features in a stimulus provides. When only one feature carries all of the information (e.g., $w = \{1, 0, 0, 0, 0, 0, 0\}$), there is no informational diversity, so entropy is 0. When all of the features carry some amount of information, informational diversity is larger; informational diversity, and entropy therefore, will be maximal when all of the features carry equal amounts of information (e.g., $w = \{.14, .14, .14, .14, .14, .14, .14\}$).

Therefore, we used the following equation to calculate the entropy of the vectors of attentional weights for each subject:

$$H_w = -\sum_m^M \log(w_m) w_m$$

where H_w is the entropy of the vector of attentional weights w, and w_m is the attentional weight of each feature m. Once entropy for each subject was obtained, it was then normalized based on the maximum possible entropy given the length of the vector of weights (uniform distribution of seven weights). Thus, normalized entropy was bound between 0 and 1, with 0 representing maximal selectivity (when the attentional weight for one of the features equals 1, whereas the remaining weights for all of the other features equals 0), and with 1 representing minimal selectivity or maximal distribution of attention (when all of the attentional weights are equal).

The entropies of the attentional profiles resulting from the GCM fits are shown in Figure 3. The violin plots depict the density distributions of normalized entropy for both species. Critically, of the human participants, 90% showed maximal selectivity (entropy = 0); that is, they attended only to the deterministic feature. The remaining 10% of the human participants showed varying levels of distributed attention (normalized entropy values ranged from 0.26 to 0.60). Table 2 details the attentional weights of the individual features for these 5 human participants (the attentional weights of the 47 participants exhibiting maximal selectivity were always 1 for the deterministic feature and 0 for all of the other features). Note that, the lower the entropy score, the higher the attentional weight was to one single feature. Conversely, the higher the entropy score, the more distributed the attentional weights were among a larger number of features. As can be seen in Table 2, the attentional weights for the deterministic feature were very high for these 5 participants; but, still, they were paying some measurable attention (weights equal to or greater than 0.05) to one, two, or three of the other features as well.

In contrast, Table 2 shows that all of the pigeons displayed distributed attention. Even when the pigeons had evidenced strong reliance on the deterministic feature to solve the categorization task (when the probabilistic and deterministic features were put into conflict on Incongruent trials, they preferentially, $M = 76\%$, chose the category predicted by the deterministic feature), their attention was not completely focused on a single feature; instead, attention was distributed across multiple features of the category exemplars. Pigeons did show varied levels of distributed attention (normalized entropy values ranged from 0.23 to 0.89); but, critically, no pigeon exhibited the maximal selectivity evidenced by the vast majority of human participants. The attentional weights for the deterministic feature were high for several pigeons (greater than 0.5 for 7 out of the 16 pigeons) and 13 of our 16 birds exhibited strongest control by the deterministic feature. Yet, all of the pigeons paid measurable attention to one, two, three, four, or five of the probabilistic features as well. Thus, we conclude that pigeons' attention was largely distributed.

In summary, the analyses of attentional weights disclosed that most humans completely focused their attention on the single deterministic feature. Pigeons, on the other hand, largely distributed their attention and none of them entirely focused on the deterministic feature.

5. General Discussion

In the reported experiments, human adults and pigeons mastered (terminal accuracy surpassing 90% correct) a categorization task that could be solved by selectively attending to a single deterministic feature or by distributing attention across multiple probabilistic features. When only the deterministic feature was available, both human adults' and pigeons' accuracy was well above chance (see Only-D trials in Figure 2), but humans' accuracy was much higher and did not suffer any decrement compared to training performance. In addition, when both deterministic and probabilistic features were in conflict, both species relied on the deterministic feature, but humans did so to a greater larger extent (see Incongruent trials in Figure 2). Subsequent modeling of subjects' attentional weights revealed a striking disparity between species. Whereas most of the humans' attention exhibited maximal selectivity (complete focus on the deterministic feature), all of the pigeons' attention was distributed; none of the pigeons exhibited maximal selectivity. Thus, it seems that pigeons' preferential use of the deterministic feature was due to their learning the statistical contingencies of the task rather than to their exhibiting selective attention. Because the deterministic feature was the best predictor of reinforcement, it acquired the highest associative strength and became the feature predominately, but not exclusively controlling the pigeons' behavior, as most associative learning theories would expect (e.g., Rescorla & Wagner, 1972; Mackintosh, 1975).

Our human participants' results agree with those prior findings showing that, when categorizing various stimuli, humans tend to focus their attention on a single stimulus feature or dimension and to deploy unidimensional rules or strategies (e.g., Ahn & Medin, 1992; Ashby & Ell, 2001; Kemler Nelson, 1984; Nosofsky, Palmeri, & McKinley, 1994; Regehr & Brooks, 1995). Pigeons may also use single features or dimensions to categorize stimuli (e.g., Castro & Wasserman, 2014, 2016a, 2017; Lea & Wills, 2008, Nicholls et al., 2011), but it seems more likely that this results from those features acquiring strong associative strength—given that they are the best predictors of the outcome—rather than because of pigeons deploying selective attention.

Role of length training and the interplay between accuracy and attention

Our human participants were trained for 360 trials in a single session, whereas our pigeons were trained for an average of 1,920 trials over an average of 16 daily sessions. Perhaps the longer training given pigeons allowed them to learn, not only the greater diagnostic value of the deterministic feature, but also the diagnostic value of several of the probabilistic features. A longer training phase might have provided humans with the opportunity to learn about the diagnostic value of the probabilistic features as well. We do not have any empirical evidence supporting that notion. We did find that some of our participants took as few as 4 or 6 trials to reach 85% correct, whereas other took as many as 320 or 340 trials to reach that level.

However, the number of trials to reach criterion did not correlate with participants' testing performance or with their distribution of attentional weights.

Theoretical accounts of categorization and learning that explicitly consider the role of attention (George & Pearce, 2012; Kruschke, 1992; Love, Medin, & Gureckis, 2004; Mackintosh, 1965, 1975; Nosofsky, 1986, 1992) all propose that attention tends to be distributed along multiple stimulus dimensions at the beginning of training, but that attention converges on the most relevant dimensions as learning proceeds. Our own observations support these accounts.

Most of our present human participants relied completely on the deterministic feature by the end of training and their attentional weights revealed maximal selectivity. However, in a prior unpublished experiment, after a training phase of only 120 trials, some 30% of the participants had not yet reached an even more lenient learning criterion of 75% correct; that is, they had not learned to solve the categorization task based on either the deterministic or the probabilistic features. More interestingly for this discussion, of those participants who learned the task, just 60% showed maximal selectivity, whereas 40% showed varying levels of distributed attention (normalized entropy values ranged from 0.22 to 0.74). So, it is conceivable that the 40% of learners who were distributing their attention might have become maximal-selectivity participants if the training phase had been longer. This possibilty merits further study.

Other theoretical accounts that consider attention to be goal- or hypothesis-driven assume that, once the learner's hypothesis proves to be correct, attention will fully focus on that relevant information and accuracy will assume its maximal level (e.g., Ashby et al., 1998). However, the story may not be that simple.

Rehder and Hoffman (2005) tracked human participants' eye movements while they were solving categorization tasks in which some elements were relevant to correctly classifying the category exemplars, whereas other elements were not. They found that, as learning progressed, participants' allocation of attention gradually shifted toward the relevant features of the stimuli. Curiously, but importantly, Rehder and Hoffman (2005) found that people's eye movements toward the relevant elements of the category stimuli tended to *follow* rather than to precede improvements in task accuracy; indeed, their correct responses were already very high *before* people fully deployed their attention to the relevant features of the category stimuli (see also Blair, Watson, & Meier, 2009). It appears that accuracy needs to reach a relatively high threshold for attention to hone in on the most relevant information. That is, the likelihood of correct categorization after presentation of a specific stimulus feature may have to be sufficiently high to influence the amount of attention that will later be allocated to that feature; the amount of attention allocated to that feature will, in turn, influence the likelihood of correct categorization after its presentation.

Neural substrates of selective attention

Although we did not include children in the current study, 4-year olds showed broadly distributed attention in a very similar categorization task (Deng & Sloutsky, 2016). Even when they were told to rely on the deterministic feature of the category exemplars during

learning (Experiment 2), later testing disclosed that the 4-year olds remembered the deterministic and probabilistic features equally well.

These phylogenetic and ontogenetic similarities (pigeons and very young children seem to attend diffusely) along with the disparities observed (only human adults showed maximal selectivity) prompted us to seek a reason for these different attentional profiles. One possibility involves the neural substrates participating in selective attention. The mammalian PFC is considered to be critical for executive functions, such as cognitive flexibility, working memory, and attentional control (e.g., Miller & Cohen, 2001; O'Reilly, 2006). The structural attributes of the PFC and its anatomical connections undergo a protracted maturational process during humans' first years of life; indeed, the PFC is immature early in development and, during the first 6 years, it expands more than twice as much as do other cortical regions (Hill et al., 2010). Because the PFC is not fully developed in very young children, their executive function capacities—attention among them—may greatly differ from the functional capacities of adults.

Mammals' and birds' evolutionary lines separated approximately 300 million years ago; hence, the anatomical organization of their forebrains differs considerably. Among the disparities, birds do not have a brain structure homologous to the PFC, although nonhomologous structures may perform similar functions (Briscoe & Ragsdale, 2018). Indeed, the avian nidopallium caudolaterale (NCL) —a multimodal telencephalic region located in the posterior forebrain—and the mammalian PFC share several physiological and functional attributes (Shanahan et al., 2013); thus, the NCL has been conjectured to be the avian neural structure most analogous to the mammalian PFC (Divac, Mogensen, & Björklund, 1985; Güntürkün, 2005, 2012). Still, there is relatively little research exploring executive control functions in avian species (Castro & Wasserman, 2016b; Rose & Colombo, 2005; Nieder, 2017). Hence, we do not know how the differences between the NCL and the PFC—for example, each of these structures receives projections from different regions of the thalamus (Waldmann & Güntürkün, 1993)—may affect the deployment of attention by birds, in general, and by pigeons, in particular.

We should also acknowledge that selective attention may not only involve *focusing* on relevant task information; it may also require *filtering* out irrelevant information, so that the less predictive and irrelevant stimulus dimensions or features are ignored and left out of the learning process (e.g., Corbetta & Shulman, 2002; Gulbinaite et al., 2014; Lennert & Martinez-Trujillo, 2011). Here, we found that pigeons did not focus solely on the most relevant or predictive information, but they also attended to, or were distracted by, other, albeit less diagnostic information. In dramatic contrast, 90% of our human participants focused solely on the most predictive information and ignored everything else.

Thus, it may be that the mature mammalian PFC is necessary to suppress attention to irrelevant or less reliable environmental information. Such suppression or filtering of irrelevant or less diagnostic information may not be within the capabilities of the avian NCL or other forebrain structures. Without strongly developed focusing *and* filtering mechanisms, pigeons may lack the concerted abilities to sustain attention to critical information *and* to

suppress attention to irrelevant information. The same may also be the case for young children.

Role of formal education and prior experience

It is also possible that language, formal education, and prior experiences determine how different subjects approach our categorization tasks. Human adults have extensive experience with verbal rules to solve problems and organize facts and knowledge based on deterministic information. Due to these experiences, adults might be more likely to approach a novel categorization task with a default strategy to find deterministic features. This may not be the case for preschool children and nonverbal animals. Therefore, human adults may be more likely to try to find deterministic information, whereas animals and young children may be more likely to engage with other various aspects of the environment.

This experiential explanation may help to better understand our findings, but it may not be entirely disconnected from the neural substrates explanation. It is commonly understood that structural maturation of the PFC drives cognitive development, because the functional developmental course of executive functions (selective attention among them) closely follows the maturational course of the PFC (Amso, Haas, McShane, & Badre, 2014; Davidson, Amso, Anderson, & Diamond, 2006; Wendelken, Munakata, Baym, Souza, & Bunge, 2012; Zelazo et al., 2003). It may well be, as Werchan and Amso (2017) suggest, that adaptation, and not maturation, is the process that best describes the developmental changes in the PFC.

It appears that the rate of development of the PFC is not fixed, but can be impacted by a variety of experiences, and that executive functions reflect changes in whole brain connectivity above and beyond simple PFC structural maturation. According to this ecological account, PFC development may reflect adaptation to different purposes relevant to the individual in their specific environment across their life span. As the environment begins to require children to exercise abilities related to executive function—such as when children begin formal schooling—the PFC will develop appropriately to meet the demands of the new situations. So, maturation of the PFC may allow for selective attention, as long as the environment requires this selective attention to be engaged (see Werchan & Amso, 2017, for a very insightful approach to the interaction between experience and PFC development). Thus, the emphasis that formal education places on deterministic information and rule-based learning may shape the PFC's functionality; this, in turn, may favor deterministic and unidimensional learning strategies in older children and adults, compared to younger children and animals.

Conclusions

In a categorization task that could be solved by selectively attending to a single deterministic feature or by distributing attention across multiple probabilistic features, both human adults and pigeons learned to rely predominately on the deterministic feature, the most predictive feature. However, computational modelling revealed a wide range of attentional profiles in both species; the vast majority of humans, but no pigeons exhibited maximal selective attention to the deterministic feature. These findings suggest that the interplay between

attention and categorization in humans and animals differs considerably. Elucidating the roles of focusing attention on relevant information and suppressing attention to irrelevant information should greatly advance our understanding of how different brain structures and mechanisms participate in category learning. Our experimental strategy and computational methods open the door to fresh possibilities for research in cognitive development and comparative cognition to illuminate that interplay.

Supplementary Material

Refer to Web version on PubMed Central for supplementary material.

Acknowledgments

References

Ahn W, & Medin DL (1992). A two-stage model of category construction. Cognitive Science, 16, 81–121. doi:10.1207/s15516709cog1601_3

Amso D, Haas S, McShane L, & Badre D (2014). Working memory updating and the development of rule-guided behavior. Cognition, 133, 201–210. 10.1016/j.cognition.2014.06.01 [PubMed: 25044248]

Ashby FG, Alfonso-Reese LA, Turken AU, & Waldron EM (1998). A neuropsychological theory of multiple systems in category learning. Psychological Review, 105, 442–481. [PubMed: 9697427]

Ashby FG, & Ell SW (2001). The neurobiology of human category learning. Trends in Cognitive Sciences, 5, 204–210. doi:10.1016/s1364-6613(00)01624-7 [PubMed: 11323265]

Ashby FG, & Valentin VV (2005). Multiple systems of perceptual category learning: Theory and cognitive tests In Cohen H & Lefebvre C (Eds.), Handbook of categorization in cognitive science (pp. 547–572). New York: Elsevier.

Ashby FG, & Waldron EM (1999). On the nature of implicit categorization. Psychonomic Bulletin & Review, 6, 363–378. doi:10.3758/bf03210826 [PubMed: 12198775]

Bates D, Maechler M, Bolker B, & Walker S (2015). Ime4: Linear Mixed-Effects Models Using Eigen and S4. R package version 1.1-10, http://CRAN.R-project.org/package=Ime4

Best CA, Yim H, & Sloutsky VM (2013). The cost of selective attention in category learning: Developmental differences between adults and infants. Journal of Experimental Child Psychology, 116, 105–199. [PubMed: 23773914]

Blair M, Watson MR, & Meier KM (2009). Errors, efficiency, and the interplay between attention and category learning. Cognition, 112, 330–336. doi:10.1016/j.cognition.2009.04.008 [PubMed: 19481733]

Brainard DH (1997). The psychophysics toolbox. Spatial Vision, 10, 433–436. [PubMed: 9176952]

Briscoe SD, & Ragsdale CW (2018). Homology, neocortex, and the evolution of developmental mechanisms. Science, 362, 190–193. doi:10.1126/science.aau3711 [PubMed: 30309947]

Castro L, & Wasserman EA (2014). Pigeons' tracking of relevant attributes in categorization learning. Journal of Experimental Psychology: Animal Learning and Cognition, 40, 195–211. doi:10.1037/xan0000022 [PubMed: 24893218]

Castro L, & Wasserman EA (2016a). Attentional shifts in categorization learning: Perseveration but not learned irrelevance. Behavioural Processes, 123, 63–73. doi:10.1016/j.beproc.2015.11.001 [PubMed: 26548717]

Castro L, & Wasserman EA (2016b). Executive control and task switching in pigeons. Cognition, 146, 121–135. 10.1016/j.cognition.2015.07.014. [PubMed: 26407340]

Castro L, & Wasserman EA (2017). Feature predictiveness and selective attention in pigeons' categorization learning. Journal of Experimental Psychology: Animal Learning and Cognition, 43, 231–242. doi:10.1037/xan0000146 [PubMed: 29120213]

Cincotta CM, & Seger CA (2007). Dissociation between striatal regions while learning to categorize via feedback and via observation. Journal of Cognitive Neuroscience, 19, 249–265. doi:10.1162/jocn.2007.19.2.249 [PubMed: 17280514]

Corbetta M, & Shulman GL (2002). Control of goal-directed and stimulus-driven attention in the brain. Nature Reviews Neuroscience, 3, 201–215. doi:10.1038/nrn755 [PubMed: 11994752]

Daneman M, & Carpenter PA (1980). Individual differences in working memory and reading. Journal of Verbal Learning & Verbal Behavior, 19, 450–466. doi:10.1016/S0022-5371(80)90312-6

Davidson MC, Amso D, Anderson LC, & Diamond A (2006). Development of cognitive control and executive functions from 4 to 13 years: Evidence from manipulations of memory, inhibition, and task switching. Neuropsychologia, 44, 2037–2078. doi:10.1016/j.neuropsychologia.2006.02.006 [PubMed: 16580701]

Deng W, & Sloutsky VM (2016). Selective attention, diffused attention, and the development of categorization. Cognitive Psychology, 91, 24–62. [PubMed: 27721103]

Dittrich L, Rose J, Buschmann J-U, Bourdonnais M, & Güntürkün O (2010). Peck tracking: a method for localizing critical features within complex pictures for pigeons. Animal Cognition, 13, 133–143. [PubMed: 19557444]

Divac I, Mogensen J, & Björklund A (1985). The prefrontal "cortex" in the pigeon. Biochemical evidence. Brain Research, 332, 365–368. doi:10.1016/0006-8993(85)90606-7 [PubMed: 3995275]

Emmerton J, & Delius JD (1980). Wavelength discrimination in the 'visible' and ultraviolet spectrum by pigeons. Journal of Comparative Physiology A, 141, 47–52. doi:10.1007/bf00611877

George DN, & Pearce JM (2012). A configural theory of attention and associative learning. Learning & Behavior, 40, 241–254. [PubMed: 22926999]

Gibson BM, Wasserman EA, Frei L, & Miller K (2004). Recent advances in operant conditioning technology: A versatile and affordable computerized touch screen system. Behavior Research Methods, Instruments and Computers, 36, 355–362.

Gibson B, Wasserman E, & Luck SJ (2011). Qualitative similarities in the visual short-term memory of pigeons and people. Psychonomic Bulletin & Review, 18, 979–984. doi:10.3758/s13423-011-0132-7 [PubMed: 21748417]

Gulbinaite R, Johnson A, de Jong R, Morey CC, & van Rijn H (2014). Dissociable mechanisms underlying individual differences in visual working memory capacity. Neuroimage, 99, 197–206. doi:10.1016/j.neuroimage.2014.05.060 [PubMed: 24878830]

Güntürkün O (2005). The avian 'prefrontal cortex' and cognition. Current Opinion in Neurobiology, 15, 686–693. [PubMed: 16263260]

Güntürkün O (2012). The convergent evolution of neural substrates for cognition. Psychological Research, 76, 212–219. doi:10.1007/s00426-011-0377-9 [PubMed: 21881941]

Hill J, Inder T, Neil J, Dierker D, Harwell J, & Van Essen D (2010). Similar patterns of cortical expansion during human development and evolution. Proceedings of the National Academy of Sciences of the United States of America, 107, 13135–13140. doi:10.1073/pnas.1001229107 [PubMed: 20624964]

Jaeger TF (2008). Categorical data analysis: Away from ANOVAs (transformation or not) and towards logit mixed models. Journal of Memory and Language, 59, 434–446. doi:10.1016/j.jml.2007.11.007 [PubMed: 19884961]

Jarvis ED, Güntürkün O, Bruce L, Csillag A, Karten H, Kuenzel W, … Butler AB (2005). Avian brains and a new understanding of vertebrate brain evolution. Nature Reviews Neuroscience, 6, 151–159. doi:10.1038/nrn1606 [PubMed: 15685220]

Kemler Nelson DG (1984). The effect of intention on what concepts are acquired. Journal of Verbal Learning and Verbal Behavior, 23, 734–759. doi: 10.1016/s0022-5371(84)90442-0

Kruschke JK (1992). ALCOVE: An exemplar-based connectionist model of category learning. Psychological Review, 99, 22–44. [PubMed: 1546117]

Lazareva OF, & Wasserman EA (2010). Category learning and concept learning in birds In Mareschal D, Quinn PC, & Lea SEG (Eds.), The making of human concepts. Oxford: Oxford University Press.

Lea SEG, & Wills AJ (2008). Use of multiple dimensions in learned discriminations. Comparative Cognition and Behavior Reviews, 3, 115–133.

Lea SEG, Wills AJ, Leaver LA, Ryan CME. Bryant CML, & Millar L (2009). A Comparative Analysis of the Categorization of Multidimensional Stimuli: II. Strategic Information Search in Humans (Homo sapiens) but Not in Pigeons (Columba livia). Journal of Comparative Psychology, 123, 406–420. doi: 10.1037/a0016851 [PubMed: 19929109]

Lennert T, & Martinez-Trujillo J (2011). Strength of Response Suppression to Distracter Stimuli Determines Attentional-Filtering Performance in Primate Prefrontal Neurons. Neuron, 70, 141–152. doi:10.1016/j.neuron.2011.04.011 [PubMed: 21482363]

Love BC (2002). Comparing supervised and unsupervised category learning. Psychonomic Bulletin and Review, 9, 829–835. [PubMed: 12613690]

Love BC, Medin DL, & Gureckis TM (2004). SUSTAIN: A network model of category learning. Psychological Review, 111, 309–332. doi:10.1037/0033-295x.111.2.309 [PubMed: 15065912]

Macho S (1997). Effect of relevance shifts in category acquisition: A test of neural networks. Journal of Experimental Psychology: Learning, Memory, and Cognition, 23, 30–53. doi: 10.1037/0278-7393.23.1.30

Mackintosh NJ (1965). Selective attention in animal discrimination learning. Psychological Bulletin, 64, 124–150. [PubMed: 14320077]

Mackintosh NJ (1975). A theory of attention: Variations in the associability of stimuli with reinforcement. Psychological Review, 82, 276–298. doi:10.1037/h0076778

Miller EK, & Cohen JD (2001). An integrative theory of prefrontal cortex function. Annual Review of Neuroscience, 24, 167–202. doi: 10.1146/annurev.neuro.24.1.167

Nicholls E, Ryan CME, Bryant CML, & Lea SEG (2011). Labeling and family resemblance in the discrimination of polymorphous categories by pigeons. Animal Cognition, 14, 21–34. [PubMed: 20652343]

Nieder A (2017). Inside the corvid brain—probing the physiology of cognition in crows. Current Opinion in Behavioral Sciences, 16, 8–14. doi:10.1016/j.cobeha.2017.02.005

Nosofsky RM (1986). Attention, similarity, and the identification–categorization relationship. Journal of Experimental Psychology: General, 115, 39–57. doi: 10.1037/0096-3445.115.1.39 [PubMed: 2937873]

Nosofsky RM (1988). Exemplar-based accounts of relations between classification, recognition, and typicality. Journal of Experimental Psychology: Learning, Memory, and Cognition, 14, 700–708.

Nosofsky RM (1992). Similarity scaling and cognitive process models. Annual Review of Psychology, 43, 25–53.

Nosofsky RM, Palmeri TJ, & McKinley SC (1994). Rule-plus-exception model of classification learning. Psychological Review, 101, 53–79. doi:10.1037/0033-295x.101.1.53 [PubMed: 8121960]

O'Reilly RC (2006). Biologically based computational models of high level cognition. Science, 314, 91–94. 10.1126/science.1127242 [PubMed: 17023651]

Palacios AG, & Varela FJ (1992). Color mixing in the pigeon (Columba iivia) II: A psychophysical determination in the middle, short and near-UV wavelength range. Vision Research, 32, 1947–1953. doi:10.1016/0042-6989(92)90054-m [PubMed: 1287991]

Pelli DG (1997). The VideoToolbox software for visual psychophysics: Transforming numbers into movies. Spatial Vision, 10, 437–442. [PubMed: 9176953]

R Development Core Team (2016). R: A language and environment for statistical computing. R Foundation for Statistical Computing, Vienna, Austria Retrieved from https://www.R-project.org.

Regehr G, & Brooks LR (1995). Category organization in free classification: The organizing effect of an array of stimuli. Journal of Experimental Psychology: Learning, Memory, and Cognition, 21, 347–363. doi:10.1037/0278-7393.21.2.347

Rehder B, & Hoffman AB (2005). Eyetracking and selective attention in category learning. Cognitive Psychology, 51, 1–41. [PubMed: 16039934]

Rescorla RA, & Wagner AR (1972). A theory of Pavlovian conditioning: Variations in the effectiveness of reinforcement and nonreinforcement In Black AH & Prokasy WF (Eds.), Classical conditioning II: Current research and theory (pp. 64–99). New York: Appleton-Century-Crofts.

Riley DA, & Leith CR (1976). Multidimensional psychophysics and selective attention in animals. Psychological Bulletin, 83, 138–160. doi: 10.1037/0033-2909.83.1.138

Rose J, & Colombo M (2005). Neural correlates of executive control in the avian brain. PLoS Biology, 3, e190. [PubMed: 15941358]

Semendeferi K, Lu A, & Schenker N (2002). Humans and great apes share a large frontal cortex. Nature Neuroscience, 5, 272–276. [PubMed: 11850633]

Shanahan M, Bingman VP, Shimizu T, Wild M, & Güntürkün O (2013). Large-scale network organization in the avian forebrain: a connectivity matrix and theoretical analysis. Frontiers in Computational Neuroscience, 7. doi:10.3389/fncom.2013.00089

Shannon CE, & Weaver W (1949). The mathematical theory of communication. Urbana: University of Illinois Press.

Sloutsky VM (2010). From perceptual categories to concepts: What develops? Cognitive Science, 34, 1244–1286. [PubMed: 21116483]

Smith JD, Berg ME, Cook RG, Murphy MS, Crossley MJ, Boomer J, Spiering B, Beran MJ, Church BA, Ashby FG, & Grace RC (2012). Implicit and explicit categorization: A tale of four species. Neuroscience & Biobehavioral Reviews, 36, 2355–2369. doi:10.1016/j.neubiorev.2012.09.003 [PubMed: 22981878]

Smith JD, & Minda JP (1998). Prototypes in the mist: The early epochs of category learning. Journal of Experimental Psychology: Learning, Memory, and Cognition, 24, 1411–1430.

Unsworth N, & Engle RW (2007). The nature of individual differences in working memory capacity: Active maintenance in primary memory and controlled search from secondary memory. Psychological Review, 114, 104–132. doi:10.1037/0033-295X.114.1.104 [PubMed: 17227183]

Viken RJ, Treat TA, Nosofsky RM, McFall RM, & Palmeri TJ (2002). Modeling individual differences in perceptual and attentional processes related to bulimic symptoms. Journal of Abnormal Psychology, 111, 598–609. doi: 10.1037/0021-843x.111.4.598 [PubMed: 12428773]

Wagenmakers E-J, & Farrell S (2004). AIC model selection using Akaike weights. Psychonomic Bulletin & Review, 11, 192–196. doi:10.3758/bf03206482 [PubMed: 15117008]

Waldmann C, & Güntürkün O (1993). The dopaminergic innervation of the pigeon caudolateral forebrain: immunocytochemical evidence for a "prefrontal cortex" in birds? Brain Research, 600, 225–234. doi:10.1016/0006-8993(93)91377-5 [PubMed: 8435748]

Waldron EM, & Ashby FG (2001). The effects of concurrent task interference on category learning: Evidence for multiple category learning systems. Psychonomic Bulletin & Review, 8, 168–176. [PubMed: 11340863]

Wendelken C, Munakata Y, Baym C, Souza M, & Bunge SA (2012). Flexible rule use: Common neural substrates in children and adults. Developmental Cognitive Neuroscience, 2, 329–339. 10.1016/j.dcn.2012.02.001 [PubMed: 22669034]

Werchan DM, & Amso D (2017). A novel ecological account of prefrontal cortex functional development. Psychological Review, 124, 720–739. doi:10.1037/rev0000078 [PubMed: 29106267]

Wills AJ, Lea SEG, Leaver LA, Osthaus B, Ryan CME, Suret MB, Bryant CML, Chapman SJA, & Millar L (2009). A comparative analysis of the categorization of multidimensional stimuli: I. Unidimensional classification does not necessarily imply analytic processing; evidence from pigeons (*Columba livia*), squirrels (*Sciurus carolinensis*), and humans (*Homo sapiens*). Journal of Comparative Psychology, 123, 391–405. doi:10.1037/a0016216 [PubMed: 19929108]

Zelazo PD, Müller U, Frye D, Marcovitch S, Argitis G, Boseovski J, … Sutherland A (2003). The development of executive function in early childhood. Monographs of the Society for Research in Child Development, 68, vii–137.
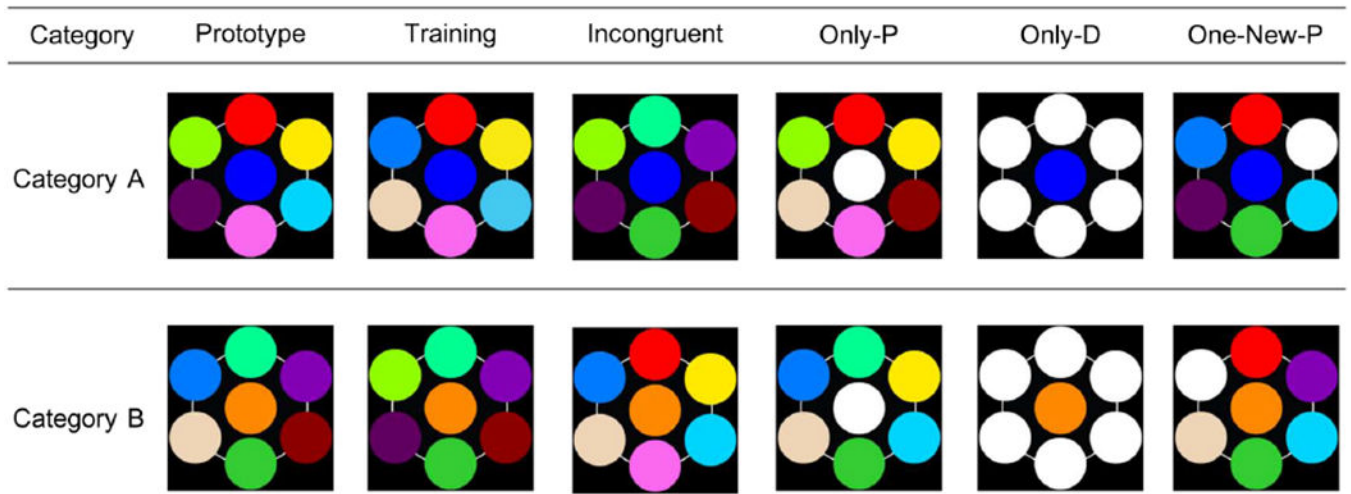
Figure 1.

Examples of the stimuli used in Experiments 1 and 2. Each row depicts trials within a category. Training trials were presented in both training and testing. Prototype, Incongruent, Only-P, Only-D, and One-New-P trials were presented only in testing. White circles are novel features that were not presented during training.

Figure 2.
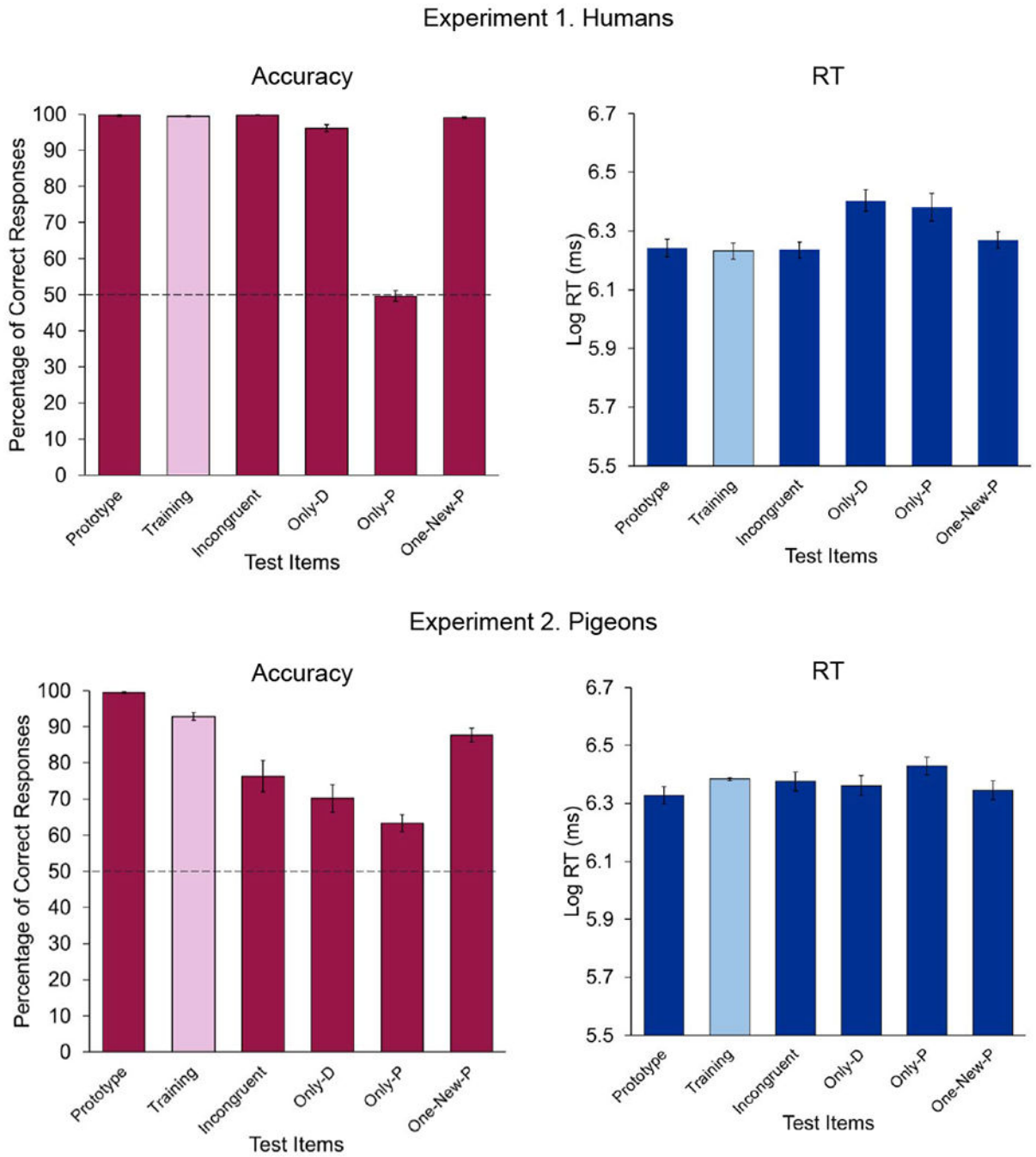Mean percentage of correct responses (left) and reaction times (right) to the different types of testing trials in Experiment 1, with humans (top), and in Experiment 2, with pigeons (bottom). In the case of Incongruent trials, responses based on the deterministic feature were considered to be correct. The dashed line, at 50%, represents the chance level. Error bars indicate the standard error of the means.
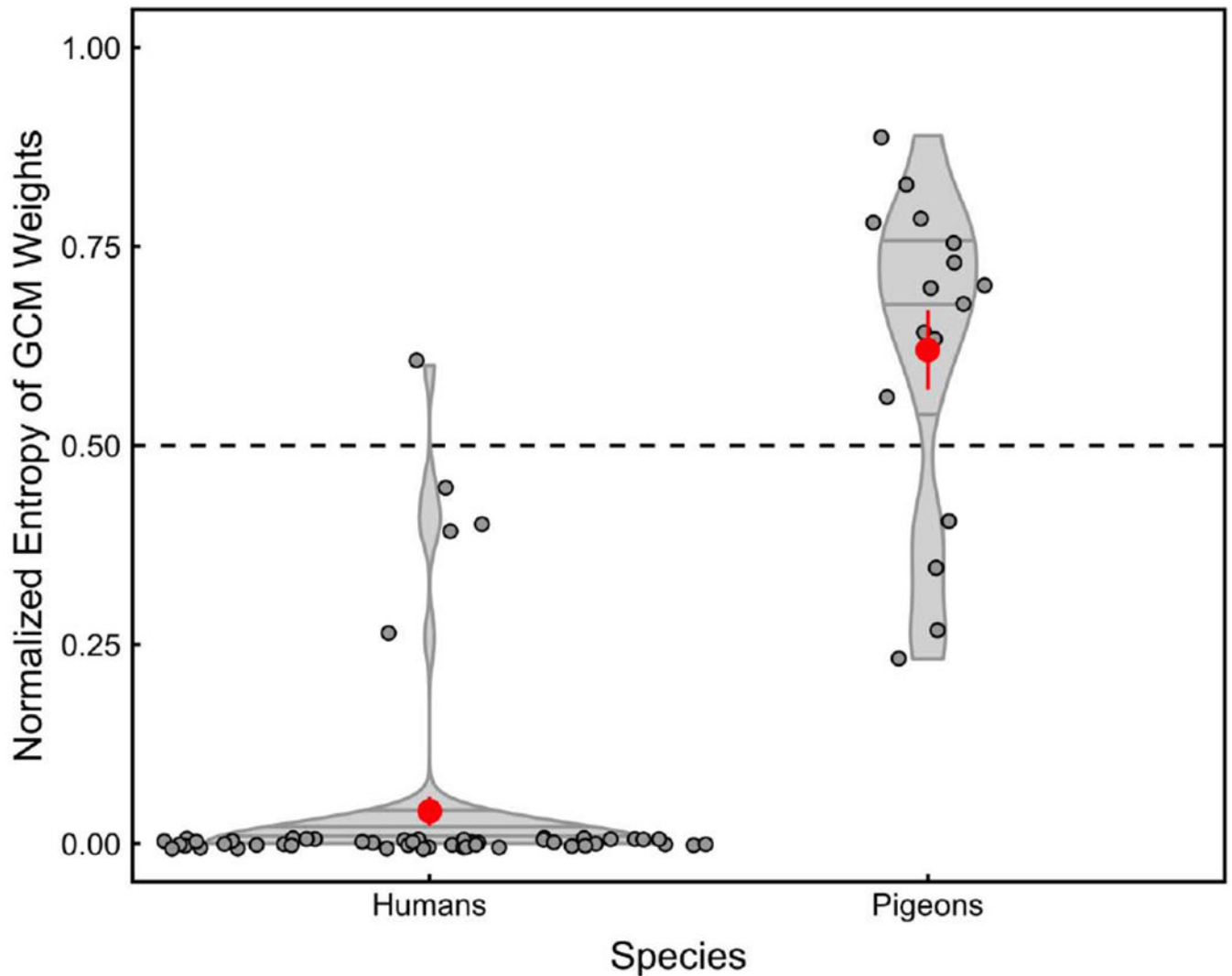
Figure 3.
Normalized entropy of GCM's best-fitting attentional weights for each individual subject. The violin plots depict the density distributions of normalized entropy for both species when the deterministic feature was presented either in the center or on the side of the category exemplar. A value of 0 represents maximal selectivity (when the attentional weight for one of the features equals 1 and the remaining weights for all other features equals 0), whereas a value of 1 represents minimal selectivity or maximal distribution of attention (when all weights are equal). The red points indicate the mean of each of the distributions. The dashed line, at 0.50, represents the middle value and is simply included as a reference.

Table 1.

Category structure used in Experiments 1 and 2.

	Category A Features							Category B Features						
	P_1	P_2	P_3	P_4	P_5	P_6	D	P_1	P_2	P_3	P_4	P_5	P_6	D
Training items														
	1	1	0	0	0	0	0	0	0	1	1	1	1	1
	0	1	1	0	0	0	0	1	0	0	1	1	1	1
	0	0	1	1	0	0	0	1	1	0	0	1	1	1
	0	0	0	1	1	0	0	1	1	1	0	0	1	1
	0	0	0	0	1	1	0	1	1	1	1	0	0	1
Testing items														
Prototype	0	0	0	0	0	0	0	1	1	1	1	1	1	1
Incongruent	1	1	1	1	0	0	0	0	0	0	0	1	1	1
Only-P	0	0	1	1	0	0	N	1	1	0	0	1	1	N
Only-D	N	N	N	N	N	N	0	N	N	N	N	N	N	1
One-New-P	N	0	0	0	1	1	0	N	1	1	1	0	0	1

Note. The value 0 = any of seven features identical to the prototype of the category A. The value 1 = any of seven features identical to the prototype of the category B. The value N = new feature; it was always a white circle. P = probabilistic feature; D = deterministic feature.

Table 2.

Attentional weights to individual features for all subjects that did *not* exhibit maximal selectivity

		Entropy	Features Rank Ordered						
			1st	2nd	3rd	4th	5th	6th	7th
Human	DPC_19	**0.260**	0.884*	0.074	0.011	0.009	0.009	0.007	0.006
	DPC_16	**0.396**	0.716*	0.239	0.030	0.009	0.003	0.002	0.002
	DPS2_5	**0.409**	0.750*	0.144	0.091	0.008	0.005	0.001	0.001
	DPS3_2	**0.449**	0.768*	0.098	0.064	0.036	0.022	0.007	0.005
	DPS3_5	**0.602**	0.682*	0.089	0.069	0.060	0.038	0.032	0.029
Pigeon	62B	**0.232**	0.891 *	0.075	0.019	0.005	0.005	0.004	0.000
	75B	**0.274**	0.853 *	0.115	0.013	0.010	0.007	0.001	0.001
	14W	**0.351**	0.806 *	0.122	0.008	0.007	0.054	0.001	0.003
	27Y	**0.408**	0.793 *	0.090	0.048	0.044	0.019	0.004	0.002
	76B	**0.562**	0.597 *	0.213	0.123	0.063	0.002	0.001	0.000
	66Y	**0.628**	0.552 *	0.197	0.144	0.088	0.019	0.001	0.000
	3W	**0.636**	0.575 *	0.220	0.114	0.030	0.028	0.020	0.012
	29B	**0.672**	0.417 *	0.325	0.181	0.048	0.021	0.008	0.000
	57Y	**0.699**	0.460 *	0.221	0.179	0.104	0.034	0.001	0.000
	83Y	**0.704**	0.540	0.180	0.111	0.088 *	0.054	0.019	0.008
	44W	**0.726**	0.513 *	0.167	0.141	0.108	0.041	0.021	0.009
	38B	**0.774**	0.438 *	0.149	0.149	0.143	0.108	0.013	0.000
	8R	**0.759**	0.479 *	0.151	0.139	0.105	0.104	0.018	0.004
	71R	**0.783**	0.322 *	0.267	0.182	0.143	0.085	0.001	0.000
	33W	**0.822**	0.235	0.225	0.205 *	0.169	0.165	0.000	0.000
	43R	**0.890**	0.285	0.228	0.167	0.151 *	0.089	0.067	0.013

Note: Attentional weights for each of the features in the category exemplars for the 5 human participants who did not show maximal selectivity (that is, $w = 1$ for the deterministic feature) and all 16 pigeons. The normalized entropy of their attentional weights appears in bold. The attentional weights are rank ordered, so that the first feature is always the one with the largest attentional weight. The asterisk (and red color) indicates the attentional weight for the deterministic feature. Note that, the lower the entropy score, the higher the attentional weight to a single feature; conversely, the higher the entropy score, the more distributed the attentional weights are among a greater number of features.