Anim Cogn (1998) 1 : 123–133

ORIGINAL ARTICLE

Joël Fagot · John K. Kruschke · Delphine Dépy
Jacques Vauclair

Associative learning in baboons (*Papio papio*) and humans (*Homo sapiens*): species differences in learned attention to visual features

Received: 30 May 1998 / Accepted after revision: 13 September 1998

Abstract We examined attention shifting in baboons and humans during the learning of visual categories. Within a conditional matching-to-sample task, participants of the two species sequentially learned two two-feature categories which shared a common feature. Results showed that humans encoded both features of the initially learned category, but predominantly only the distinctive feature of the subsequently learned category. Although baboons initially encoded both features of the first category, they ultimately retained only the distinctive features of each category. Empirical data from the two species were analyzed with the 1996 ADIT connectionist model of Kruschke. ADIT fits the baboon data when the attentional shift rate is zero, and the human data when the attentional shift rate is not zero. These empirical and modeling results suggest species differences in learned attention to visual features.

Key words Attention · Categorization · Primate · Baboon · Human

Introduction

This study investigates attention during learning from a comparative perspective. The term "attention" has been used to refer to many different attentional processes. Some research has emphasized *spatially* selective attention. Witte et al. (1996), for instance, investigated whether the location of a pre-cue stimulus relative to a target affected response times of rhesus monkeys in a target detection paradigm. Other studies have inferred *dimensionally* selective attentional mechanisms from discriminative performance in tasks involving complex stimuli. An example of these studies comes from research on conceptual discrimination by monkeys (e.g., D'Amato and van Sant 1988) and pigeons (e.g., Cerella 1979), which suggests that discrimination performance relies on an attentive analysis of stimulus features such as color, rather than on configurations of features. We consider here a different but related aspect of attention, namely, the ability of attention to *rapidly shift* away from one stimulus dimension to another, contingent upon prior associative learning.

The idea that attention to dimensions can shift during learning has a long history in the animal learning literature (see, for example, the historical summary provided in Chapter 1 of Sutherland and Mackintosh 1971). Rescorla and Wagner (1972), in their classic model of associative learning, suggested that different cues can have different associabilities, or attention strengths. Sutherland and Mackintosh (1971) and Mackintosh (1975) presented formal models of how attention to cues can change during learning. In these models, the attention allocated to a cue determines the cue's associability, such that a cue to which more attention is paid is more easily associated with other stimuli.

Many phenomena in human and animal learning have been explained by the idea that attention to cues can be changed by learning. One such phenomenon occurs when the cue-to-outcome correspondence shifts at some point during the course of training. Under conditions where the shifted correspondence has the same relevant, diagnostic, or valid cues as the initial correspondence, learning the shifted correspondence is relatively fast. Under other conditions, where the relevant cues for the shifted correspondence differ from those for the initial correspondence, learning the shifted correspondence is relatively slow (e.g., Kruschke 1996b). This difference can be explained by assuming that the subject learns to attend to the relevant cue(s) in the initial phase of training, and this learned attention perseverates into the shifted phase of training. When the same cue(s) continue to be relevant, this attentional perseveration is advantageous. When different cues are relevant in the shifted phase, the attentional perseveration is detrimental.

J. Fagot (✉) · D. Dépy · J. Vauclair
CNRS, CRNC, 31 chemin Joseph-Aiguier,
F-13402 Marseille, Cedex 20, France
e-mail: fagot@lnf.cnrs-mrs.fr

J. K. Kruschke
Department of Psychology, Indiana University,
Bloomington, IN 47405, USA

The ADIT model (Kruschke 1996a), which we use in this article to address attentional learning in humans and baboons, is closely related to the model of Mackintosh (1975) (Kruschke 1997). Mackintosh (1969) presented research comparing attentional learning across different non-human species. Our present research is the first to directly compare attentional learning abilities of humans and monkeys, using an experimental paradigm that dramatically highlights attentional abilities in humans.

Attentional shifts during learning by humans

In a recent experiment, Kruschke (1996a, experiment 2) presented people with a medical diagnosis task in which a learner had to diagnose a hypothetical patient as having one of several possible fictitious diseases. The basic design involved distinguishing a disease shown early in training (referred to as "E") from a disease shown later in training ("L"). Each disease had two symptoms. One of the two symptoms was shared by diseases E and L. It is labeled "I" for "imperfect predictor". The other symptom was a perfect predictor of the disease. It is labeled "PE" (for "perfect predictor" of E) for disease E, and PL for disease L.

Figure 1 shows the abstract design of this experiment (Kruschke 1996a). In an early training phase, subjects were asked to learn only disease E. In a later training phase, diseases E and L were intermixed across trials with equal frequencies. After this two-phase training, participants were asked to diagnose patients who had novel combinations of the possible symptoms, such as the symptom I alone, or symptoms PE+PL presented simultaneously.

How should a rational learner respond to these novel symptom combinations? Consider symptom I presented alone. During the entire course of training, it appeared twice with disease E for each time it appeared with disease L. Therefore a rational response would be to choose disease E. Consider now the symptom combination PE+PL. During the entire course of training, symptom PE appeared with disease E twice for each time that symptom PL appeared with disease L. A rational response to PE+PL would therefore favor E again.

Results showed that when tested with symptom I alone, people diagnosed it as the early disease E. By contrast, when presented with the conflicting PE+PL symptoms, people chose the later disease L. According to Kruschke (1996a), these results can be explained as follows. Because disease E is learned first, its two symptoms (I and PE) are each associated with the early disease. When subsequently learning disease L, the shared symptom I is already associated with E, which conflicts with the correct response, L. To avoid this conflict, people shift their attention away from I to PL, thereby encoding disease L predominantly by its distinctive symptom PL, and not by its shared symptom I. Consequently, when presented with symptom I alone during the test phase, people tend to respond with disease E. When tested with combination PE+PL, people diagnose it as the later disease, because PL is the strongly encoded perfect predictor of disease L, but PE is only half of the two symptoms needed to predict disease E.

This explanation suggests that people's response to symptom combination PE+PL is not irrational, but is instead a side effect of a highly adaptive mechanism for selective attention during learning. When learning the later disease, people dramatically reduce interference with their previously learned associations by selectively attending to the distinctive symptom PL, and by selectively ignoring the conflicting symptom I. This shift of attention protects and preserves previously acquired knowledge, and the shift simultaneously enhances rapid acquisition of new associations.

The principle of rapidly shifting attention in learning was formalized by Kruschke (1996a) in a connectionist model, called ADIT (attention to distinctive input). ADIT accurately fit data from several experiments, providing additional evidence that the theory of rapid attention shifts has merit.

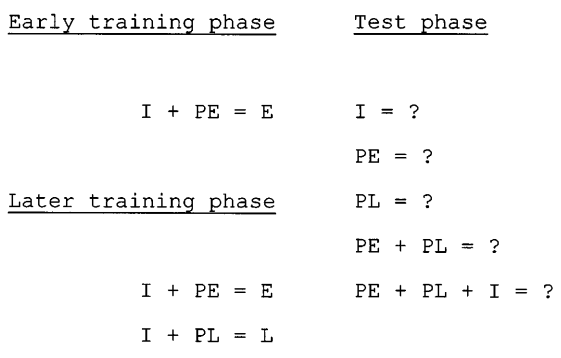

Fig. 1 Abstract design of the experiment from Kruschke (1996, experiment 2). *E* and *L* refer to the earlier and later categories to be learned, respectively. *I*, *PE* and *PL* refer to the stimulus features (i.e., symptoms: *I* imperfect predictor of the diseases, *PE* perfect predictor of the earlier disease, *PL* perfect predictor of the later disease)

Goal of this research

It is important to study learned attention for two reasons. First, learned attention arguably accounts for many phenomena in animal and human learning. Second, rapid shifts of attention are a computationally efficient means of reducing interference between previously learned associations and novel associations. Given the importance of learned attention, the rationale of the current study was to test baboons with an analogue, in the visual domain, of the disease diagnosis task, in order to assess the ability of these animals to shift attention while sparing previous category knowledge. Experimental results will suggest that baboons do not shift attention comparably to humans. In a second part of this paper, data from each species will be fit with the ADIT model. This model formalizes the notion of rapidly shifting attention, and makes detailed

quantitative predictions. In line with empirical data, the model fits will also suggest that baboons do not shift attention as humans do.

Experiment: evidence for species differences in attention shifts

Monkeys were tested with a conditional matching-to-sample task in which the visual forms to be categorized were presented on a monitor screen. The experimental procedure involved two training phases followed by a testing phase. During the training phases, monkeys sequentially learned two categories of two-feature visual objects, one after the other. In the testing phase, they were presented with various combinations of the features of the initially and subsequently learned categories. For comparative purposes, and as our visual task had never been presented before to humans, people were also tested in the same conditions as baboons.

Method

Participants

Two 6-year-old wild-born baboons (*Papio papio*), referred to as B04 and B06, were selected from a social group of 12 animals in the animal facilities at the C.N.R.S., Marseille, France. The two baboons were chosen as subjects in this experiment because they were already trained in various visual discrimination tasks involving the setup and the procedure employed in the present research (e.g., Fagot and Deruelle 1997; Vauclair et al. 1993). Animals were not deprived of food, but obtained their daily food ration at the end of the day, after completion of daily training and testing. Nine 22- to 26-year-old psychology students also participated for monetary payment. The human participants were not informed of the purpose of the experiment.

Apparatus

The apparatus comprised a 14-inch (35-cm) color monitor on which visual stimuli were displayed, and an analogue joystick controlling a cursor on the monitor screen. When baboons were tested, the set-up comprised an experimental cage ($68 \times 50 \times 72$cm) facing the joystick and the monitor. This cage was fitted with a view port, two hand ports for joystick manipulation, and a food dispenser for delivering 190-mg banana-flavored pellets. The testing apparatus for humans was identical to that of the baboons, except that the monitor and the joystick were placed on a table. For both humans and baboons, the viewing distance was 50 cm.

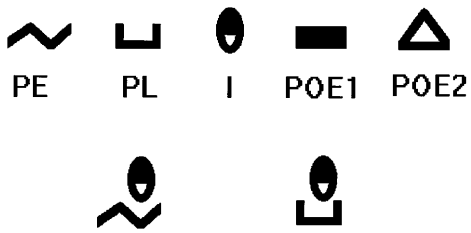

Fig. 2 **Upper row** Elementary features employed to construct the stimulus set, and corresponding labels used in specifying the abstract design. **Lower row** Example of two-feature stimulus used during training phase 1 *(left)* or training phase 2 *(right)*

Stimuli

Stimuli were composed of either one or two of the geometric features shown in the top part of Fig. 2. Each of these elementary features was composed of 510 yellow pixels displayed on a black background. For stimuli composed of more than one feature, such as those depicted in the bottom part of Fig. 2, the location of the features relative to each other was varied across trials, with the constraint that they did not overlap and they occupied a maximum of 3° of visual angle. For instance, in training phase 1, the oval form could be displayed either above or below the zigzag line.

Experimental design

Figure 3 illustrates the abstract structure of the categories in the two training phases. Consider first the structure of the categories in training phase 2. Subjects had to learn that stimuli composed of features I+PE required response E, and stimuli composed of features I+PL required response L. The feature I is an imperfect predictor of the two outcomes, whereas feature PE is a perfect predictor of outcome E, and feature PL is a perfect predictor of outcome L. Outcome E is denoted as such because it is also part of early training, in phase 1, whereas outcome L occurs only in later training, in phase 2. This much of the design is identical to the design of experiment 2 of Kruschke (1996a) (summarized in Fig. 1).

Fig. 3 Experimental design adopted during the two training phases. *E*, *OE* and *L* refer to the correct response categories. *I*, *PO1*, *PO2*, *PE* and *PL* refer to the stimulus features depicted in Fig. 2

Training phase 1

I+PE = E

I+POE^1 = OE

PE+POE^2 = OE

Training phase 2

I+PE = E

I+PL = L

Our account of the effect seen in humans relies on the assumption that in phase 1 subjects learn about both features, I and PE, of category E. Previous research (Fagot and Deruelle 1997) suggests that baboons will select local features of a compound stimulus, when this selection supports a viable solution. To prevent this from happening in this experiment, another category, labeled OE, was included in phase 1, which had instances consisting of features I+POE1, or of PE+POE2. With this structure, only the conjunction of features I+PE accurately predicted outcome E in the early training phase.

General procedure

The method of testing was based on the conditional matching-to-sample paradigm. At the beginning of each trial, a cursor appeared in the center of the monitor, along with a 0.5 × 0.5 cm "start" stimulus, 1.5 cm above or below the cursor. Participants then had to manipulate the joystick so as to place the cursor on the "start" stimulus to initiate the trial. Once accomplished, a sample stimulus appeared on either the left or the right side of the screen, along with two response squares of different colors, 4 cm above and below the cursor. Subjects then had to move the cursor to the color square designating the category to which the sample belonged. For instance, in phase 2 of the training, participants had to select the green response square when the sample belonged to E, or the white response square when the sample belonged to L. The location of response colors, either above or below the center, was random on each trial. For monkeys, correct responses were reinforced with food, and incorrect responses resulted in a time delay of 5 s before the next trial began. For humans, the outcome of each trial was indicated by the French words "vrai" or "faux" (correct or wrong) appearing for 250 ms on the screen.

The first training phase was designed to teach the participants categories E and OE (see Fig. 3). On each trial, a training stimulus was shown on the monitor, along with a blue and a green response square. To respond correctly, subjects were required to select the green square if it belonged to E, and the blue square if the sample belonged to OE. Because E stimuli shared one feature with all the OE stimuli, an accurate discrimination required participants to take into account the identity of both elemental features of the sample. In phase 1, training blocks contained 48 trials for humans and 96 trials for baboons. Baboons received one to three blocks of training per day, but humans received all training and testing blocks is immediate succession. Within each block, the OE stimuli were presented in 50% of the trials (25% for each type of OE stimulus, see Fig. 3), and the E stimuli were displayed in the remaining 50% of the trials. The order of stimulus presentation was randomly selected prior to each block, with the constraint that no more than three consecutive trials were from the same category.

Prior to training phase 1, human subjects were told how to initiate the trials and to manipulate the joystick.

Fig. 4 Stimuli used during testing, with corresponding abstract labels

They were never told, however, what the matching rule was, but were asked to discover it by themselves. Training phase 1 continued until the participants attained at least 80% correct responses for all three feature combinations.

The second training phase involved a discrimination between E and L categories (Fig. 3). All the procedural details of phase 2 were identical to those of training phase 1. The two categories were presented with equal frequencies. Moreover, participants had to select a green response square when the sample belonged to E, or a white response square when the sample was from L. The instruction given to human participants emphasized only the need to discover the matching rule.

The testing phase consisted of series of trials in which the sample form could be either an instance from phase 2 of training or any of the test stimuli shown in Fig. 4. Test stimuli were either features I, PE or PL presented alone, or the feature combinations PE and PL, or PE and PL and I. The response squares in the test phase were always green and white, corresponding to the E and L categories from phase 2. As there was no correct response for the five test stimuli, test trials with these stimuli never gave rise to feedback for humans. However, because an absence of feedback might have been considered as a negative reinforcement by baboons (since they had previously been rewarded in every correct trial), test trials involving the transfer stimuli were, for baboons, randomly reinforced at a 50% rate.

Each monkey received 20 test blocks of 106 trials. Each block comprised 48 E trials, 48 L trials, and 20 transfer trials involving the five test stimuli (see Fig. 4). People received only five sessions of 88 trials each, in order to keep a sustained level of motivation. Their sessions consisted of 24 E and 24 L trials intermixed with 40 test trials (eight trials per test stimulus). Moreover, prior to the test, human participants were instructed that they would receive trials, similar to those of phase 2, intermixed with some trials of a novel type, for which no feedback was given. For the novel trials, they were asked to give their best educated guess.

Results

Training

One baboon (B04) needed 9696 trials (101 blocks) to reach the training criterion in training phase 1. The other subject (B06) needed 9408 trials (98 blocks). Humans were much faster and needed 219 trials on average (SD =

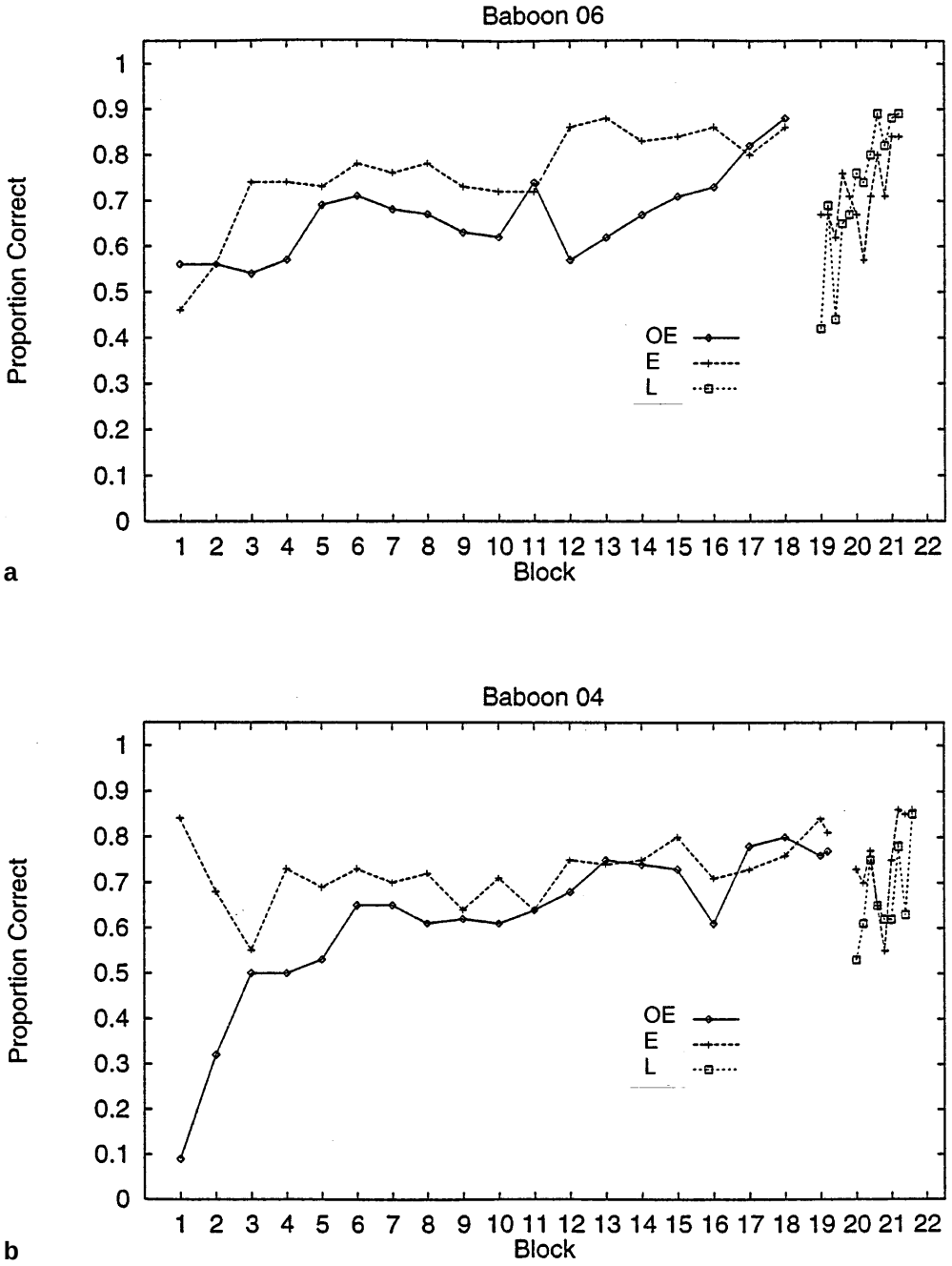

Fig. 5 Proportion of correct trials for **a** B06 and **b** B04, and for each stimulus category used in training. Training trials were blocked by 500 trials in training phase 1 and 50 trials in training phase 2. Thus, the *Block* on the abscissa of these graphs is different from the training and testing blocks definied in the main text

148) to reach the training criterion in phase 1. For training phase 2, criterion was reached in 864 and 1152 trials (9 and 12 blocks) for subjects B04 and B06, respectively. Humans all reached the training criterion in phase 2 within the first block of 96 trials.

Individual learning curves for training phases 1 and 2 are reported in Fig. 5a for B06 and in Fig. 5b for B04. In phase 1, B06 learned category E before category OE. This relative difficulty of OE might be accounted for by its complexity, in that it comprises two types of stimuli (Fig. 2).

In the initial 50 trials of phase 2, baboon B06 did not perform significantly above chance for category L, $\chi^2_{(1,\, n\,=\,50)} = 1.28$, $P > 0.10$, but performance on E remained above chance, $\chi^2_{(1,\, n\,=\,50)} = 6.48$, $P < 0.02$. Hence, this baboon did remember category E during the initial training sessions of phase 2. Results for B04 were very similar to those of B06, except that during the early sessions of phase 1 B04 showed a bias for response E, which resulted in a high proportion correct on E but a low proportion correct on OE. Like B06, however, B04 showed a significant tendency to respond correctly during the initial E trials of phase 2, $\chi^2_{(1,\, n\,=\,50)} = 11.52$, $P < 0.001$, and showed no significant bias for the initial 50 L trials of phase 2, $\chi^2_{(1,\, n\,=\,50)} = 0.08$, $P > 0.10$.

Table 1 Percentage of E responses for stimuli in the test phase, for baboons, humans, and best-fitting ADIT (attention to distinctive input) models; *No shift* best fit of ADIT when attention shifting is set to zero, but an additional parameter allowing different learning rates for present and absent categories is used; *Shift* best fit of ADIT with attention shifting, * significant ($P \leq 0.05$) response biases inferred from a two-tailed chi-square test, $^{+}$ significant ($P \leq 0.05$) response bias at the group level inferred from a two-tailed *t*-test comparing the mean number of E responses for the group to random 50-50 choice, *ns* no significant response bias). No statistical analyses were conducted on baboons' mean data, because of the limited number of subjects; results for each individual *(B04, B06)* are given separately

Stimulus	Baboons				Humans	
	B04	B06	Mean	No shift	Mean	Shift
I+PE	83.6*	88.9*	86.2	85.3	99.5^{+}	99.7
I+PL	19.6*	11.3*	15.5	15.3	1.3^{+}	1.2
PE	82.5*	92.5*	87.5	86.9	99.2^{+}	99.3
PL	20.0*	10.5*	15.3	13.8	0.0^{+}	0.2
I	40.0^{ns}	57.5^{ns}	48.8	48.2	78.3^{+}	79.6
I+PL	47.5^{ns}	55.0^{ns}	51.2	50.7	23.3^{+}	24.4
I+PE+PL	52.5^{ns}	37.5^{ns}	45.0	49.4	53.9	52.3

Testing

Table 1 reports the response proportions for individual baboons in the test phase, along with response proportions at the group level for humans. The two monkeys behaved in a very similar way during testing. First, both B04 and B06 showed a significant response preference for E when presented with PE, and for L when presented with PL. Because PE and PL corresponded to the perfect predictor of E and L, respectively, this finding shows that monkeys have encoded associations between the distinctive features and the corresponding categories they predict. Second, for both B04 and B06, there was no significant response bias for the imperfect predictor I. Finally, these two baboons exhibited random choices when presented with either PE+PL or I+PE+PL.

Not unlike baboons, human participants selected E and L when presented with PE and PL, respectively (see Table 1). They also showed no significant response bias for I+PL+PE, which comprised the imperfect and both perfect predictors of each category. However, humans behaved very differently from baboons when presented with feature I or with feature combination PE+PL. Indeed, for feature I there was a significant preference for response E instead of the random choice exhibited by baboons. For feature combination PE+PL, people showed a significant preference for response L instead of the ambivalence shown by the baboons.

The pattern of responses shown by human subjects in this experiment is consistent with the results of the experiment of Kruschke (1996a), despite the changes in stimuli and apparatus. Thus, the results observed by Kruschke (1996 a) cannot be due to the use of symptoms and diseases, with their implied direction of causality from diseases to symptoms, nor can the original results be due to the use of explicitly verbal material for stimuli and responses.

Discussion

This research investigated attention to visual features in animals and humans. For that purpose, baboons were tested with an analogue, in the visual domain, of the fictitious medical diagnosis task presented to humans by Kruschke (1996 a, experiment 2). Human participants were also tested under the same conditions as baboons. Our discussion will focus on attentional mechanisms involved in phased learning.

Humans and baboons both showed retention of category E while learning category L in phase 2. Humans and baboons also both showed a preference for response E when tested with feature PE, and they both showed a preference for response L when tested with feature PL. Humans and baboons also both showed no differential preference for either response when presented with feature combination I+PE+PL.

The two species differed quite dramatically, however, in their responses to test stimuli I and I+PL. Whereas the baboons showed no preference for either response option, humans strongly preferred response E for stimulus I, and response L for stimulus I+PL.

Results for the humans are consistent with those of Kruschke (1996 a). As described in the Introduction, these results can be explained by assuming that learners associated both features I and PE with the early learned category E, but shifted attention to the distinctive feature PL of the later learned category L. Why did baboons behave differently from humans in this experiment?

It might be argued that, in baboons only, associations related to the imperfect predictor acquired in phase 1 were forgotten in phase 2, due to a limited memory capacity. This hypothesis is unlikely for at least two reasons. First, as reported in Fig. 5, both baboons presented a significant propensity for correctly responding E during the initial 50 E trials of phase 2. Second, in phase 2, E trials which had already been presented in phase 1 were intermixed with 50% of L trials. Thus, the experimental design favored the retention of associations linked to E during acquisition of category L.

Another possible hypothesis is that people may have encoded E by the conjunction of its two elemental features, I and PE, whereas baboons may have encoded E merely as anything that does not contain the distinctive features of OE, namely POE1 and POE2. According to this hypothesis, the baboons responded OE if the sample contained either a POE1 or POE2, and responded E if the sample contained neither of these two features. This hypothesis predicts that, during the initial presentations of L in training phase 2, these I+PL stimuli should have been responded to as E, because they contained neither POE1 nor POE2. However, when the initial 50 presentations of L are considered, the two baboons responded E in 52% (B04) and 42% (B06) of the trials, which indicates ran-

dom choices (chi-square test, all *P*s > 0.1) rather than a preference for E. This latter result shows that, at least at the end of phase 1, the two baboons relied on the encoding of both features of E to respond correctly.

In brief, the results demonstrate that baboons encoded the two elementary features of E at the end of training phase 1 and, moreover, remembered category E in the initial training sessions of phase 2. Whereas the two elementary features of E were encoded during phase 1 and early in phase 2, only the distinctive feature of this category was coded at the end of phase 2, as demonstrated by the transfer tests. This phenomenon is probably not to be explained by limitations in memory, but rather by acquired changes in the association strength between the imperfect predictor I and response E. We propose that, during phase 2, the previously learned association between the imperfect predictor I and category E progressively diminished, due to the frequent co-occurrence of feature I with outcome L. Unlike humans, the baboons did not shift attention away from feature I when it appeared with outcome L. To compensate for this loss of association between I and E, the association strength between the perfect predictor PE and category E grew to be approximately equal to the association strength between the perfect predictor PL and the category L, leading to random choices when the two perfect predictors conflicted in test item PE+PL. In other words, rather than shift attention to protect associations learned in early training and to prevent interference with new learning, the baboons "overwrote" the earlier learned associations with new associations.

Modeling: additional evidence for species differences in attention shifts

The experimental results showed a dramatic difference between the generalization behavior of baboons and humans. In this section, we demonstrate that the ADIT model, which incorporates attention shifts, can fit the human data, and can also fit the baboon data when its attentional shift rate is set to zero. The modeling thereby supports the claim that a critical difference between the baboons and the humans is rapid attention shifting.

The Rescorla-Wagner (RW) model (Rescorla and Wagner 1972) is the best known model of associative learning in animals (cf. Miller et al. 1995; Siegel and Allan 1996), and it has also been successfully applied to some aspects of human category learning (e.g., Gluck and Bower 1988). The RW model associates cues with responses (or with unconditioned stimuli) such that the amount of associative change is proportional to the discrepancy between the correct response and the predicted outcome. In this way, the changes in associative weights are error-driven. In the basic RW model, any cue that is present in the stimulus will have its associative weights changed whenever there is an erroneous prediction. There is no mechanism of shifting attention whereby the different cues can participate more or less in accounting for the error. For example, if a cue is present in two different trials with different outcomes, then the cue must be attended to in both trials, and associative learning must accommodate the conflicting outcomes. In the particular experimental design used here (see Fig. 1), this property of the RW model implies that the cue, I, shared by stimuli I+PE and I+PL, will eventually be associated about equally with the two response categories, rather than unequally as demanded by the human data.

The configural connectionist model of Pearce (1994) also does not incorporate selective attention and also cannot accommodate the human data. In the Pearce model, there are configural nodes that mediate the mapping from input cues to output responses. Importantly, these configural nodes are recruited so that they exactly copy all the cues presented on a given trial, with no differential selective emphasis of individual cues within a configuration. The Pearce model does have a form of attentional capacity constraint expressed as activation normalization, but the model does not have any selective attention shifting. The Pearce configural model cannot, therefore, exhibit the preferences shown by humans for stimuli I and PE+PL, but it does predict the pattern of results shown by baboons.

Kruschke (1996 a) reviewed several other models of human categorization that have attempted to account for the type of phenomenon addressed here. These models included the component cue model of Gluck and Bower (1988), the attentional connectionist model of Shanks (1992), the context model of Medin and Schaffer (1978), the generalized context model of Nosofsky (1986), the ALCOVE model of Kruschke (1992), and the rational model of Anderson (1991). All of these models failed because they do not implement stimulus-specific attention shifts.

An extension of the RW model that includes shifting attention has been shown to address effects of phased training and base rates on category learning in humans (Kruschke 1996a). The model, called ADIT, formalizes two simple ideas: People learn about what they attend to, and they attend to cues that minimize interference with prior associations. ADIT builds associations between cues and outcomes in two corresponding steps on every trial: First, attention to cues is shifted on a trial-by-trial basis, so as to reduce error in prediction. In this step, attention is shifted to reduce interference with prior associations, i.e., attention is directed away from cues that are already associated with conflicting outcomes, and toward other cues that are available for associating with the current outcome. Second, once attention has shifted, associative weights are adjusted to further reduce any remaining error between the prediction based on the attended-to cues and actual outcome. This second step is identical to the basic RW model, using only attended-to cues instead of all presented cues. Thus, ADIT reduces to a form of the RW model when the rate of attention shifting is fixed at zero. After providing a detailed description of the model, we

will show that ADIT can fit both the baboon behavior and the human behavior, and the fit to baboon behavior is best when attention shifting is set to zero.

Formal description of ADIT

ADIT can be construed as a connectionist network in which each input node represents a cue and each output node represents a categorical response. To model this experiment, the network has five input nodes, corresponding to the five cues, and three output nodes, corresponding to the three response categories. When the *i*th cue is present, input node *i* has activation $a_i = 1$, otherwise it has activation zero. When the *k*th category is the correct response, the *k*th output node receives a "teacher" value of $t_k = 1$, otherwise it has teacher value zero[1]

Each input node is gated by an individual attention strength, which simply amplifies or attenuates the input node activation via multiplication. When a stimulus is presented, all activated nodes are initially attended to equally. There is a capacity constraint on the total attention, so that as more cues are presented, less attention can be allocated to each one. If the stimulus has *N* cues, then the initial attention given to each cue is $\alpha = 1/N^{(1/\eta)}$, where $\eta > 0$ and is a freely estimated parameter that expresses the attentional capacity. This causes the normed total attention, $(\Sigma_i \alpha_i^{\eta})^{1/\eta}$, to be equal to 1. If η is small, then increasing the number of cues reduces the attention per cue. If η is large, then increasing the number of cues has relatively little effect on the attention per cue.

Between input node *i* and output node *k* is a connection with associative weight w_{ki}. The associative weights are initialized at zero, but gradually learn according to the delta-rule described later. When a stimulus is presented, the input nodes are activated, attention is allocated to the activated nodes, and then activation spreads to the output nodes via the weighted connections, with output activation given by:

$$a_k = \Sigma_i w_{ki} \alpha_i a_i$$

The output node activations are converted to choice probabilities so that the probability of response *K* is given by:

$$p(K) = \exp(\phi a_K)/\Sigma_k \exp(\phi a_k)$$

where ϕ is a freely estimated constant that determines the "decisiveness" of the network. A large ϕ causes the most activated output node to garner a large choice probability at the expense of other partially activated response nodes. A small ϕ causes even slightly activated responses to be chosen with probability nearly as large as highly activated responses.[2]

After the network makes its initial prediction, the teacher values are delivered to the output nodes, just as feedback is delivered to experimental subjects. The attention strengths are then redistributed to reduce the error, *E*, across the output nodes, which is measured as:

$$E = 0.5\Sigma_k (t_k - a_k)^2$$

The primary goal of learning is error reduction, and the first reaction to the error is a rapid shift of attention away from cues that cause error and toward cues that reduce error. In ADIT, this shift of attention is accomplished by gradient descent on the error with respect to the attention strengths, which yields the following equation for shifting attention:

$$\Delta\alpha_i = \sigma\Sigma_k (t_k - a_k) w_{ki} a_i$$

where σ is a freely estimated constant of proportionality, called the attention shift rate. If the shift causes an attention strength to take on a negative value, then the value is set to zero because negative values might not have a clear psychological interpretation. After the attention strengths are shifted, they are renormalized by dividing each by:

$$(\Sigma_i \alpha_i^{\eta})^{1/\eta}$$

This renormalization respects the capacity constraint on attention.

With the model now having determined which cues to attend to, the cue activations are propagated again to the output nodes and a new error is computed. The association weights are then adjusted to reduce this remaining error, with the change in weights given by:

$$\Delta w_{ki} = \lambda (t_k - a_k) \alpha_i a_i$$

where λ is a freely estimated constant of proportionality, called the weight learning rate. This learning rule is formally equivalent to the learning rule in the original RW model, except that Rescorla and Wagner proposed that the learning rate for reinforced trials could be greater than the learning rate for non-reinforced trials. In the present formalism, this means that the learning rate takes on one value, λ, for output nodes with $t_k = 1$, and a different value $\beta \leq \lambda$, for output nodes with $t_k = 0$.

In summary, when stimulus cues are presented to ADIT, they are initially all attended to equally, and activation propagates from the cues, weighted by attention, to the category nodes. Corrective feedback is then supplied, and attention rapidly shifts away from cues that cause error and toward cues that reduce error. This shifting of attention constitutes reduction of interference between prior knowledge and new learning, because attention is shifted

[1] The "shift" and "no-shift" versions of ADIT implemented here actually used "humble" teachers, for which an activation more extreme than the zero or one teacher value does not generate an error (see Kruschke 1996a). Humble teachers turn out to have little influence on the model fits; in fact, they improve the fits of the two versions of ADIT slightly but not significantly

[2] The original version of ADIT also mixed category base rates with the output probabilities. This is not done here, primarily because the original base rate learning mechanism was ad hoc and is inappropriate for the present application. Moreover, the models fit the data here without considering base rates

away from exactly those cues that cause interference with prior knowledge. Any remaining error is used to drive changes in the associative weights. The no-shift version of this model, which is equivalent to the RW model, is simply the special case in which the attention shift rate is fixed at zero, and the associative weight learning rate for absent categories is allowed to differ from the associative weight learning rate for present categories.

The "shift" version of ADIT has four freely estimated parameters: the learning rate for the association weights, the decisiveness constant for mapping output activations to response probabilities, the attentional capacity constant, and the attention shift rate. The no-shift version of ADIT also has four freely estimated parameters: the attention shift rate is fixed at zero but the associative learning rate for absent categories is free.

Measure of fit

The model was fit to the empirical response frequencies for each test-phase stimulus. The fit of the predictions to the data was measured using a log-likelihood statistic, $G^2 = 2\Sigma_i f_i \ln(f_i/m_i)$ where f_i is the empirical response frequency in cell i, m_i is the predicted response frequency in cell i, and the sum is over all cells in the response frequency table (Wickens 1989).

Parameter value search method

Best-fitting parameter values were found using simulated annealing (e.g., Goffe et al. 1994). Simulated annealing randomly checks many thousands of parameter value combinations, and gradually increases the density of search in regions of parameter space with the best fit. We can be fairly confident that the resulting best fits are very nearly the global optimum in the searched range of parameter values. As an additional check, gradient-descent search was also performed from a number of different starting values.

Fit results

Best fit of ADIT to the baboon data

The two baboons' mean response frequencies in the test phase (see Table 1) were fit. For each of the seven stimuli in the test phase, there were two possible responses, E or L, thereby yielding 14 response frequencies to be fit. As the total number of responses for a given stimulus was fixed by the experimental design, the data included seven independent frequencies.

The model was trained on 2388 blocks of four trials (9552 trials) in phase 1, and 252 blocks of four trials in phase 2 (1008 trials), which represents the approximate mean training time of the two baboons. Phase 3 contained 20 blocks of 106 trials (2120 trials), with each block having 24 occurrences of stimulus I+PE reinforced with response E, 24 occurrences of stimulus I+PE reinforced with response L, 24 occurrences of stimulus I+PL reinforced with response E, 24 occurrences of stimulus I+PL reinforced with response L, one occurrence of each of the five test stimuli reinforced with response E, and one occurrence of each of the five test stimuli reinforced with response L. The results of two random training sequences were averaged to obtain the model predictions.

It turned out that an excellent fit to the baboon data could be obtained without use of a separate learning rate for absent categories (that is, $\beta = \lambda$). Moreover, the attentional capacity constraint was also unneeded ($\eta = 100.0$, the maximum permitted value). The best fit was obtained with $\phi = 1.22$ and $\lambda = 0.637$, which yielded $G^2(df = 5) = 0.88$, an excellent fit. The predictions of the "no-shift" version of ADIT are shown in Table 1. The model shows indifferent responses to stimuli I and PE+PL, just as exhibited by the baboons (but not by the humans).

Best fits to the human data

The nine humans' mean response frequencies in the test phase (see Table 1) were fit. Just as with the baboon data, for each of the seven stimuli in the test phase there were two possible responses, E or L, thereby yielding 14 response frequencies to be modeled. As the total number of responses for a given stimulus was fixed by the experimental design, the data included only seven independent frequencies. The two version of ADIT ("shift" and "no-shift") each have four estimated parameters, leaving three degrees of freedom. The models were trained for 55 blocks of four trials (220 trials) in phase 1, and 24 blocks of four trials (96 trials) in phase 2, representing the approximate mean training time of the human participants.

The best fit of the "no-shift" version of the model to the mean human data is quite poor, with best-fitting parameter values of $\phi = 6.44$, $\lambda = 0.0913$, and $\eta = 0.875$, and $\beta = 0.594$, which yielded $G^2(df = 3) = 225.4$. Thus the no-shift version of the model can be rejected with extremely high confidence. The model predicts indifferent responses to stimuli I and PE+PL, contrary to the human data.

The fit of the "shift" version of ADIT to the human data is excellent. The best fit, with $\phi = 6.65$, $\lambda = 0.0337$, $\sigma = 0.213$ and $\eta = 6.69$, yielded $G^2(df = 3) = 2.78$. Predictions are shown in Table 1, where it can be seen that responses for stimuli I and PE+PL, in particular, are very close to the empirical values.

ADIT achieves this fit by shifting attention away from the shared cue I when learning category L. By the end of training, the associative weight from the shared cue to category E remains much higher than the associative weight from the shared cue to category L. The associative weights from the distinctive features, PE and PL, are also noticeably asymmetric.

Model fit summary

The baboon data can be accurately fit without any attention shifts. The human data, on the other hand, can be accurately fit by ADIT only with attention shifts. When attention shifting is constrained to be zero, the model cannot reproduce the response preferences shown by humans, even when it is provided with separate learning rates for present and absent categories. Insofar as the attention shifting mechanism in ADIT is critical to account for the human data, we have additional evidence that comparable attention shifting is indeed occurring in human learning.

General discussion

The current study compared category learning in humans and baboons using a task derived from the work of Kruschke (1996 a). Consistent with Kruschke (1996 a), results from humans demonstrated that when two categories that share a feature are learned successively, the shared feature is associated with the earlier-learned category, and the distinctive feature of the later-learned category is strongly associated with the later category. In the terms of Kruschke (1996 a), this result suggests that, when learning new associations, humans shifted their attention away from the stimulus features already associated with a response, and focused their attention on uncommitted distinctive features. This conclusion is confirmed by our modeling, showing that ADIT with attentional shifting fits human data well. ADIT with zero attention shifting failed to account for the human data, but fit the baboon data.

One obvious advantage of attentional shifting in the context of our experiment is to protect previously learned associations and to facilitate new learning. Humans are remarkably proficient among animals in their ability to rapidly learn arbitrary new associations without catastrophically interfering with previously learned associations. It is proposed, therefore, that attentional shifting mechanisms are an important component of this ability.

Regarding the baboons, two main conclusions can be drawn from their results. First, baboons changed their associations during the learning process. Second, baboons selectively processed stimulus features, rather than exclusively the stimuli as wholes. The first conclusion, regarding changing associations, derives from the inspection of transfer trials involving the imperfect predictor "I". Results showed that this elementary feature was not differentially associated with either E or L during the test phase, while it was more strongly associated with E than with L early in training phase 2. The second conclusion, regarding the selective processing of stimulus features, derives from the test phase in which the elementary stimulus features were presented individually or in novel combinations. If these novel combinations were processed as compound wholes, then responses should have been at chance level for the five transfer stimuli. Contrasting with this expectation, responses to the test items PE and PL differed from random choices, suggesting that these test stimuli were associated to either E or L, although they were never presented as such during the training phases. Note also that the test items PE and PL were as strongly associated with their respective categories as the typical stimuli of E and L (see Table 1), ruling out the possibility that performance in the test phase reflected an effect of stimulus generalization alone. Altogether, the results imply that baboons processed the compound stimuli used in this experiment as sets of elementary features, rather than as compounds.

Baboons' focusing on stimulus features might depend on the experimental procedure and nature of stimulus configuration, and it is likely that the use of other types of stimuli, such as continuous bi-dimensional forms instead of discontinuous pairs of forms, might have favored the processing of the two stimulus dimensions. It should be noted, however, that the processing of stimulus feature, as opposed to compounds, has previously been observed in baboons for both stimulus discrimination (e.g., Fagot and Deruelle 1997; Deruelle and Fagot, in press) and categorization tasks (e.g., Dépy et al. 1997). For instance, Fagot and Deruelle (1997) presented the same baboons as in the current study with large (i.e., global) forms made of smaller forms (e.g., a large square made of small squares). After presentation of these compound stimuli as samples, the subject had to match the samples with comparison forms by considering either the local or global stimulus level. Results demonstrated a strong advantage (i.e., shorter response times) for processing the local aspects of the stimulus compared to global processing.

In other animal species, such as birds, rodents and fish, Mackintosh (1969) argued that results from probability learning and from serial reversal shift learning could be accounted for if subjects learned to selectively attend to relevant cues, and if this selective attention was learned to different degrees by different species. Mackintosh (1969) argued that "The simplest explanation ... of the behavioural differences between rat, bird and fish, is to suggest that the three classes of animal differ in the extent to which they can learn to attend to a given cue ..." (pp. 148–149). Our experiment and modeling demonstrate that baboons and humans also differ in their ability to rapidly shift attention among cues, suggesting that differences in selective attention may account for a broad range of species differences in learning and behavior.

Acknowledgements John Kruschke's research was supported in part by NIMH FIRST Award 1-R29-MH51572–01. We wish to thank D. Washburn who initiated the collaboration between the French and American groups. We also thank B. Arnaud, G. Argenton, M. Chiambretto, R. Fayolle and F. Lavergne for technical assistance. P. Lemaire is acknowledged for useful comments on an earlier version of the manuscript. The data presented in this paper were reported at the 88th Annual Meeting of the Southern Society for Philosophy and Psychology, 4 April 1996, Nashville, Tennessee. This study complied with the current French laws on animal treatment and experimentation.

References

Anderson JR (1991) The adaptive nature of human categorization. Psychol Rev 98:409–429

Cerella J (1979) Visual classes and natural categories in the pigeon. J Exp Psychol Hum Percept Perform 5:68–77

D'Amato MR, Sant P van (1988) The person concept in monkeys (*Cebus apella*). J Exp Psychol Anim Behav Proc 14:43–55

Dépy D, Fagot J, Vauclair J(1997) Categorisation of three-dimensional stimuli by humans and baboons: search for a prototype effect. Behav Proc 39:299–306

Deruelle C, Fagot J (in press) Visual search for global/local stimulus features in humans and baboons. Psychonom Bull Rev

Fagot J, Deruelle C (1997) Processing of global and local visual information and hemispheric specialization in humans (*Homo sapiens*) and baboons (*Papio papio*). J Exp Psychol Hum Percept Perform 23:429–442

Gluck MA, Bower GH (1988) From conditioning to category learning: an adaptive network model. J Exp Psychol Gen 117: 227–247

Goffe WL, Ferrier GD, Rogers J (1994) Global optimization of statistical functions with simulated annealing. J Econometrics 60:65–99

Kruschke JK (1992) ALCOVE: an exemplar-based connectionist model of category learning. Psychol Rev 99:22–44

Kruschke JK (1996a) Base rates in category learning. J Exp Psychol Learning Mem Cogn 22:3–26

Kruschke JK (1996b) Dimensional relevance shifts in category learning. Connection Sci 8:201–223

Kruschke JK (1997) Attention in learning: relating Mackintosh's (1975) theory to connectionist models and human categorization. Talk presented at the Eighth Australasian Mathematical Psychology Conference, Perth, Australia, 29 November 1997

Mackintosh NJ (1969) Comparative studies of reversal and probability learning: rats, birds and fish. In: Gilbert RM, Sutherland NS (eds) Animal discrimination learning. Academic Press, New York, pp 137–162

Mackintosh NJ (1975) A theory of attention: variations in the associability of stimuli with reinforcement. Psychol Rev 82: 276–298

Medin DL, Schaffer MM (1978) Context theory of classification learning. Psychol Rev 85:207–238

Miller RR, Barnet RC, Grahame NJ (1995) Assessment of the Rescorla-Wagner model. Psychonom Bull 117:363–389

Nosofsky RM (1986) Attention, similarity and the identification-categorization relationship. J Exp Psychol General 115:39–57

Pearce JM (1994) Similarity and discrimination: a selective review and a connectionist model. Psychol Rev 101:587–607

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

Shanks DR (1992) Connectionist accounts of the inverse base-rate effect in categorization. Connection Sci 4:3–8

Siegel S, Allan LG (1996) The widespread influence of the Rescorla-Wagner model. Psychonom Bull Rev 3:314–321

Sutherland NS, Mackintosh NJ (1971) Mechanisms of animal discrimination learning. Academic Press, New York

Vauclair J, Fagot J, Hopkins WD (1993) Rotation of mental images in baboons when the visual input is directed to the left cerebral hemisphere. Psychol Sci 4:99–103

Wickens TD (1989) Multiway contingency tables analysis for the social sciences. Erlbaum, Hillsdale

Witte EA, Villareal M, Marroco RT (1996) Visual orienting and alerting in rhesus monkeys: comparison with humans. Behav Brain Res 82:103–112