7

Category Learning

JOHN K. KRUSCHKE

> What have we here? a man or a fish? dead or alive? A fish: he smells like a fish; a very ancient and fish-like smell; a kind of not of the newest Poor-John. A strange fish!
>
> Shakespeare, *The Tempest*, Act II, Scene II, line 22; spoken by Trinculo

Categorization is Central to Cognition

Beachcombers categorize flotsam as man or fish. Players of 20 questions categorize things as animal, vegetable or mineral. Guards categorize approachers as friend or foe. Bystanders categorize flying objects: 'Look, up in the sky! It's a bird; it's a plane! No, it's Superman!' Categorization permeates cognition in myriad protean variations. Our categorization of an object we encounter determines what we do with it. If it is an old dead fish, we walk away in disgust, but if it is just an unkempt rascally scoundrel, we … Well, at least in *some* cases our categorization of an object affects our reaction to it.

The overriding purpose of categorization is inference of unseen attributes, especially for novel stimuli. Thus, we classify a handwritten squiggle as the letter 'A' because we then infer its sound and meaning in the context of other letters. The sound and meaning are not visible in the squiggle itself. We classify animals as tigers or zebras in order to infer unseen attributes such as being threatening or innocuous. The potential threat is not explicit in the visible features of the animal. Classification takes us from the information given in the stimulus to previously learned, associated information. The classification is itself an inference of an unseen attribute. As anyone knows who has tried to read messy handwriting, the classification of squiggles into letter categories is often not at all obvious. Occasionally also we encounter unusual animals, or flying objects, or flotsam, that are difficult to identify. In principle, any classification is a non-trivial inference of an unseen attribute.

Induction of invisible features is called categorization when it applies to novel stimuli that are not exact replicas of the stimuli experienced during learning. That is, categorization depends on generalizing from particular learned instances to novel situations. Categorization is sometimes defined merely as dividing a set of items into subsets. Typically, however, such a division is only of interest to the extent that novel items are inferred to be in one subset or another. If learned knowledge consisted merely of isolated facts with no generalization, then the knowledge would be useless except for the unlikely exact recurrence of the learned situation. For example, learning that a 4 cm tall, round-capped, beige-colored mushroom is edible would not generalize to 3 cm tall mushrooms. This failure to generalize could result in a starved categorizer. On the other hand, if generalization is too liberal, then complementary problems can arise. For example, inferring that a flat-capped, 4 cm tall, beige mushroom is edible might result in a poisoned categorizer. Thus, generalization from learned cases must be appropriately tuned, not too narrow and not too broad.

Generalization is not the only goal when learning categories. Just as important is retaining previously learned knowledge while quickly acquiring new knowledge. For example, after having learned about edible mushrooms, it could prove catastrophic if learning about poisonous mushrooms required

TABLE 7.1 *Examples of categorization models that instantiate various options for representation and matching processes*

	Matching process	
Representation	Graded Similarity	Strict Match/Mismatch
Content, piecemeal	(A) Context model (Medin & Schaffer, 1978); Generalized context model (Nosofsky, 1986); ALCOVE/RASHNL (Kruschke, 1992; Kruschke & Johansen, 1999); rational model (Anderson, 1991).	(B) Exemplar subsystem of Smith and Minda (2000); Sparse Distributed Memory (Kanerva, 1988); discrete dim. RULEX (Nosofsky, Palmeri, & McKinley, 1994); disjunctive featural rules (Bourne, 1970; Levine, 1975).
Content, global	(C) Modal features (Reed, 1972); central tendency (Smith & Minda, 2000); component-cue network (Gluck & Bower, 1988); ADIT/EXIT (Kruschke, 1996a, 2001a).	(D) Conjunctive featural rules (Bourne, 1970; Levine, 1975).
Boundary, piecemeal	(E) ATRIUM (Erickson & Kruschke, 1998, 2002).	(F) Continuous dim. RULEX (Nosofsky & Palmeri, 1998); COVIS (Ashby et al., 1998).
Boundary, global	(G) PRAS (Vandierendonck, 1995).	(H) Quadratic bound (Ashby, Waldron, Lee, & Berkman, 2001).

dozens of exposures. It could also be disastrous if the learning about poisonous mushrooms erased still-valid knowledge about edible mushrooms (French, 1999; Kruschke, 1993; McCloskey & Cohen, 1989; Mirman & Spivey, 2001; Ratcliff, 1990). Yet edible and poisonous mushrooms share many prominent features, so learning that they should be treated dramatically differently, while also generalizing appropriately to other mushrooms, could be very challenging.

Because classification, i.e. naming the category label of an object, is a paramount example of inference of an unseen attribute, many laboratory experiments in categorization examine people's ability to learn novel classification labels. In a typical experiment, a participant is shown a stimulus and asked to guess which of several category labels is correct. After his or her guess, the correct label is displayed, and the participant studies the stimulus and correct label for a few seconds before moving on to the next case. After many cases (typically but not always with repetition of individual cases), the participant learns the correct category labels of the stimuli. The experiment might then test the learner's generalization with novel stimuli, or his or her ability to learn new categorizations.

For the cognitive scientist, the key questions then are the following: What has the participant learned and how? What sort of representation best describes the learner's knowledge? What sort of processes best describe the learning and classifying activities? Answers to these questions can be constrained by data from the learned items, from generalization to novel items, from learning of new categories, from inference of features from category labels (e.g. Anderson, Ross, & Chin-Parker, 2002; Anderson & Fincham, 1996; Thomas, 1998; Yamauchi & Markman, 2000a, 2000b) and from other types of information.

Category learning is critically important because it underlies essentially all cognitive activities, yet it is very difficult because learned categories must generalize appropriately, learning must occur quickly, and new learning must not overwrite previous knowledge. Moreover, categorization occurs on different dimensions and at different levels of abstraction simultaneously. For example, a cardinal (i.e. the bird) can evoke the color category red or the part category feather or the object category animal, and so on. Within these dimensions there are levels of abstraction, such as scarlet, red or warm within the 'color' dimension, or cardinal, bird or animal within the 'object' dimension.

Varieties of Theories of Categorization

Theories of categorization vary on three aspects. Table 7.1 shows theories that instantiate each of the combinations of the three dimensions. First, the theory specifies what is explicitly represented about the category. Some theories assert that the *contents* within each category are explicitly specified whereas other theories assert that the *boundaries* between categories are explicitly specified. For example, the category 'skyscraper' might be described by a boundary that separates it from 'low-rise', as follows: if the ratio of height to width is greater than 1.62 (the golden ratio), then the building is a skyscraper. The specific ratio of 1.62 determines the boundary in height/width space that separates skyscrapers from low-rises. Alternatively, the skyscraper and low-rise categories might be specified by their contents, e.g. if a building is more similar to Philip Johnson's AT&T Headquarters (now the Sony Building) than to Frank Lloyd Wright's Taliesin, then it is a skyscraper.[1]

Second, for either type of representation, the contents or boundary can be specified by a *global summary* or by *piecemeal components*. For example, the two descriptions of skyscrapers given above were global summaries, insofar as a single condition defined the boundary or the content of each category. By contrast, a piecemeal definition of a boundary might be the following: a building is a skyscraper if it is greater than 20 stories tall or if it is taller than 7 stories and less than 100 feet wide. A piecemeal definition of content might be the following: a building is a skyscraper if it is more like the Sears Tower *or* the John Hancock Building than the O'Hare Airport Terminal *or* Navy Pier.

Third, whatever type of representation is used for the categories, it must be compared with the incoming item that is to be categorized. This comparison process can yield a *strict match versus mismatch*, or a degree of *graded similarity*. For example, the height/width ratio to define skyscraper is strict, but the content definition ('it's like the Sears Tower') uses graded similarity.

In the remainder of this chapter, a few combinations of representation and matching process will be described in detail. Notice that the matching processes are distinct from *learning* processes. Learning is the process that actually generates the representations. Learning will also be discussed en route.

The three dimensions of categorization shown in Table 7.1 were described intuitively in the previous paragraphs. Upon further reflection, however, the dimensions are subtle and demand more careful clarification. Perhaps the clearest distinction is between content and boundary. A representation of content specifies what is in the category, but the extent of the category might be vague. A representation of boundary specifies exactly what the limit of the category is, and any item within that limit is a full member of the category. The only situation in which the distinction between content and boundary breaks down is when the stimulus attributes are nominal features. In this situation a specification of the category's features is tantamount to a specification of the boundary around the category, because the features themselves are assumed to be sharply bounded categories. For example, if we define a bachelor as a human who is male, unmarried and eligible, then we have specified a description of the content of the category. But because the attribute 'male' lies on a dichotomous dimension that has a sharp, genetically specified boundary between it and 'female', a specification of content is informationally equivalent to a specification of boundary.[2] Only when the stimulus dimensions vary continuously, rather than discretely, is there a difference between content and boundary representations.

The distinction between global and piecemeal specification is intuitively acceptable but needs rigorous definition for maximal usefulness. For example, a piecemeal specification of the category 'dog' could be this: a dog is something like Lassie or Rin-tin-tin or Benjie or Pongo. A global specification of dog might be this: a dog is an animal with four legs and a tail and fur and toenails (not claws or hooves) and a height between 0.5 and 1.0 meters and the ability to bark. (This is not an accurate specification of 'dog', but it illustrates the idea.) From these examples it appears that global specifications use a conjunction of properties, whereas piecemeal specifications use a disjunction of properties. The problem with this distinction is that conjunctions and disjunctions can be exchanged via negation: 'A or B' is equivalent to 'Not (Not A and Not B)'. For example, to say that 'a dog is like Lassie or like Benjie' is logically equivalent to saying that 'a dog is not both unlike Lassie and unlike Benjie'. Despite this subtlety in the distinction between global and piecemeal specifications, for the purposes of this chapter I shall define a piecemeal representation as one that involves disjunctions in stimulus characteristics that are natural or primitive in the theory. The disjunctions could be formally realized as logical or as other mathematical expressions.

Finally, there is the distinction between strict and graded matching of stimulus to category specification. All theories of categorization must account for the fact that human (and other species') categorization is probabilistic or graded. For example, a letter drawn as '/-\' could be interpreted as an 'H' or as an 'A', as in 'T/-\E C/-\T'. As another example of gradedness in category membership, a Labrador retriever is usually rated as a more typical dog than a Pekinese. Theories that use a graded matching process naturally account for gradedness in categorization. But theories that use strict matching cannot explain gradedness unless other mechanisms are included. Some strict-matching theories assume that the encoding of the stimulus is probabilistic (i.e. imperfect or distorted) or the specification of the category conditions is probabilistic or the generation of the categorical response is probabilistic. Some graded-matching theories also incorporate probabilistic mechanisms.

In summary, any theory of category learning must specify what information from the world is actually retained in the mind and the format in which that information is structured, how that information is used, and how that information is learned. Hopefully the theory also motivates why that particular learning procedure is useful. In this chapter these issues are addressed in turn, for a few different theories from Table 7.1. Each type of theory is initially described informally, to convey the basic motivating principles of the theory. Each theory is also described with formal, mathematical terms. By expressing a theory mathematically, the theory gains quantitative precision rather than merely

vague verbal description. Mathematical formulation allows publicly derivable predictions rather than theorist-dependent intuitively derived predictions. Mathematical derivations permit stronger support when predictions are confirmed in quantitative detail. Formal models also engender greater explanatory power when the formal mechanisms in the model have clear psychological interpretation. Specification in formal terms sometimes permits clearer applicability because of precise specification of relevant factors.

EXEMPLAR THEORIES

One sure way to learn a category is merely to memorize its instances. For example, a learner's knowledge of the category bird might consist of knowing that the particular cases named Tweety, Woody and Polly are exemplars of birds. There is no derived representation of a prototypical bird, nor is there any abstracted set of necessary and sufficient features that define what a bird is. As new cases of birds are experienced, these cases are also stored in memory. Notice, however, that just because these exemplars of birds are in memory, the learner need not be able to distinctly recall every bird ever encountered. Retrieving a specific memory might be quite different than using it for categorization.

So how are these stored exemplars used to categorize novel items? A new stimulus is classified according to how similar it is to all the known instances of the various candidate categories. (Some exemplar models use only a subset of the stored exemplars, e.g. Sieck & Yates, 2001.) For example, a newly encountered animal is classified as a bird if it is more similar to known exemplars of birds than it is to known exemplars of squirrels or bats, etc. The notion of similarity, therefore, plays a critical role in exemplar theories.

This kind of exemplar theory falls in the top left cell (A) of Table 7.1. The category is specified by its contents, in this case by its exemplars. The specification of contents is not a global summary but is instead a collection of piecemeal information. The process for matching a stimulus to the category representation relies on graded similarity to exemplars, not on strict match or mismatch with exemplars.

Selective attention

Not all features are equally relevant for all category distinctions. For example, in deciding whether an animal is a duck or a rabbit (a potential confusion highlighted by Wittgenstein, 1953), it might be important to pay attention to whether it can fly, but to determine whether the animal is a crow or a bat, it might be important to pay attention to whether it has feathers (cf. Gelman & Markman, 1986). Selective attention plays an important role not only in exemplar theory but in many theories of categorization. Selective attention will be further discussed in a subsequent section.

Learning for error reduction

In principle, exemplar encoding can accurately learn any possible category structure, no matter how complicated, because the exemplars in memory directly correspond with the instances in the world. This potential computational power of exemplar models is one rationale for their use. But realizing this potential can be difficult, because exemplar encoding can retard learning if highly similar instances belong to different categories. One way around this problem is to associate exemplars with categories only to the extent that doing so will reduce errors in categorization. Analogously, various features or stimulus dimensions can be attended to only to the extent that doing so will reduce error. Thus, error reduction is one important rationale for theories of learning.

The formal model described next expresses the notion of error reduction in precise mathematical notation. Variants of this exemplar model have been shown to fit a wide range of phenomena in category learning and generalization (e.g. Choi, McDaniel, & Busemeyer, 1993; Estes, 1994; Kruschke & Johansen, 1999; Lamberts, 1998; Nosofsky & Kruschke, 1992; Nosofsky, Gluck, Palmeri, McKinley, & Glauthier, 1994; Nosofsky & Palmeri, 1997; Palmeri, 1999; Pearce, 1994), but it is not without challenges (e.g. Cohen, Nosofsky, & Zaki, 2001, and others cited later). Exemplar models have also been used to characterize individual differences in learning and attention (e.g. Treat, McFall, Viken, & Kruschke, 2001). Some exemplar models have been extended to involve dynamic processes that yield predictions of response latencies (e.g. Cohen & Nosofsky, 2000; Lamberts, 2000b).

Formal models of exemplar theory

A perceived stimulus can be formally represented by its values on various psychological dimensions. For example, among birds, an eagle might be represented by a large numerical value on the dimension of perceived size, and by another large numerical value on the dimension of ferocity, along with other values on other dimensions. The psychological value on the dth dimension is denoted Ψ_d^{stim}. These psychological scale values can be determined by methods of multidimensional scaling (e.g. Kruskal & Wish, 1978; Shepard, 1962).

In a prominent exemplar-based model (Kruschke, 1992; Kruschke & Johansen, 1999; Medin & Schaffer, 1978; Nosofsky, 1986), the *m*th exemplar in memory is formally represented by its psychological values on the various dimensions; the value on the *d*th dimension is denoted Ψ^{ex}_{md}. The similarity of the stimulus to a memory exemplar corresponds to their proximity in psychological space. The usual measure of distance between the stimulus, *s*, and the *m*th memory exemplar is given by $\text{dist}(s, m) = \sum_i \alpha_i \left| \Psi^{stim}_i - \Psi^{ex}_{mi} \right|$, where the sum is taken over the dimensions indexed by *i*, and $\alpha_i \geq 0$ is the attention allocated to the *i*th dimension. (This assumes, of course, that the psychological dimensions can be selectively attended to.) When attention on a dimension is large, then differences on that dimension have a large impact on the distance, but when attention on a dimension is zero, then differences on that dimension have no impact on distance. The distance is converted to similarity by an exponentially decaying function: $\text{sim}(s, m) = \exp(-\text{dist}(s, m))$. Therefore, when the stimulus exactly matches the memory exemplar so that the distance between them is zero, then the similarity is 1.0, and as the distance between the stimulus and the memory exemplar increases, the similarity drops off toward zero. Shepard (1987) provided a review of the exponential function as a model of human (and other animal) similarity gradients. Lee and Navarro (2002) describe a variation of the model in which the stimulus dimensions are nominally scaled instead of interval scaled.

The exemplars then influence the categorization of the stimulus by 'voting'. The strength of an exemplar's vote is its similarity to the stimulus, such that exemplars that are highly similar to the stimulus cast a strong vote, while exemplars that are remote from the stimulus cast only a weak vote. Each exemplar votes for candidate categories according to associative strengths from the exemplars to the categories. The associative strength from exemplar *m* to category *k* is denoted w_{km}, and the total 'voting' for category *k* is $v_k = \sum_m w_{kn} sim(s, m)$. The overall probability of classifying the stimulus into category *k* is the total vote for category *k* *relative* to the total votes cast overall. Formally, the probability of classifying stimulus *s* into category *k* is given by $p_k = v_k / \sum_c v_c$. Often this response rule is generalized by including a parameter to adjust the decisiveness of the choice probability; two such forms are $p_k = v_k^{\gamma} / \sum_c v_c^{\gamma}$ and $p_k = \exp(\gamma v_k) / \sum_c \exp(\gamma v_c)$, whereby high values of γ map small differences in voting to large differences in choice probability. Wills, Reimers, Stewart, Suret, and McLaren (2000) discuss problems with these response rules and propose an alternative.

Participants in laboratory learning experiments get corrective feedback on each learning trial. The same procedure applies to learning in the model: the model 'sees' the stimulus, votes for what it deems to be the best classification, and then is presented with the correct categorization. The model adjusts its associative weights and attention strengths to reduce the error between its vote and the correct answer. Error is defined as $E = \sum_k (t_k - v_k)^2$, where t_k is the 'teacher' value: $t_k = 1$ if k is the correct category, and $t_k = 0$ otherwise. There are many possible methods by which the associative weights and attention strengths could be adjusted to reduce this error, but one sensible method is *gradient descent* on error. According to this procedure, the changes that make the error decrease most rapidly are computed according to the derivative of the error with respect to the associative weights and attention strengths. The resulting formula for weight changes is $\Delta w_{km} = \lambda_w (t_k - v_k) \text{sim}(s, m)$, where λ_w is a constant of proportionality called the learning rate. This formula states that the associative weight between exemplar *m* and category *k* increases to the extent that the exemplar is similar to the current input and the category teacher is underpredicted. Notice that after the weight changes according to this formula, the predicted category will be closer to the correct category, i.e. the error will have been reduced. The analogous formula for attentional changes is a little more complicated, but it essentially combines information from all the exemplars to decide whether attention on a dimension should be increased or decreased (see Kruschke, 1992; Kruschke & Johansen, 1999).

Prototype Theories

It might seem inefficient or wasteful to remember every instance of a category. Perhaps some sort of summary could be abstracted during learning, and then the individual cases could be safely jettisoned. The summary, also called a prototype, should be representative of the various instances of the category. There are several possible forms of this prototype. One option is for the prototype to be the *central tendency*, or *average*, of all the known cases of the category. For example, the mental representation of dog might be a medium-sized mutt that blends the features of all the experienced instances of dog. The dog prototype need not necessarily correspond to any actually experienced individual dog.

Another option for the prototype could be an *idealized caricature* or extreme case that is maximally distinctive from other categories. For example, Lynch, Coley, and Medin (2000) reported that expert foresters thought of trees in terms of ideal extreme height, rather than in terms of typical medium height. Palmeri and Nosofsky (2001) report a similar finding with random dot patterns.

Alternatively, a prototype could be the most frequent, i.e. modal, *instance*; or a prototype might consist of a combination of the most frequent or modal *features* of the instances.

Whatever the specific nature of the prototypes, a new stimulus is classified according to how similar it is to the prototypes of the various candidate categories. A newly encountered animal is classified as a dog to the extent that it is more similar to the dog prototype than to other category prototypes.

This type of theory appears in the second row and left column of Table 7.1, cell C. The theory specifies a single global summary for the content of each category. The process of matching a stimulus to the category representation uses graded similarity, not strict match or mismatch.

One rationale for this approach to categorization is that it is efficient: an entire set of members in a category is represented by just the small amount of information in the prototype. The economy of representation does not eliminate the need for learning: as new instances appear, a prototype must be updated to reflect current information.

In several studies that compare prototype and exemplar models, it has been found that prototype models do not fit data better than exemplar models (e.g. Ashby & Maddox, 1993; Busemeyer, Dewey, & Medin, 1984; Busemeyer & Myung, 1988; Nosofsky, 1992; Reed, 1972). Other studies have found evidence that is difficult for basic exemplar models to address, but which can (or might) be better addressed by prototype models (e.g. Blair & Homa, 2001; Homa, Sterling, & Trepel, 1981; Minda & Smith, 2001; Smith & Minda, 1998). One conclusion is that human behavior is best described as using a combination of exemplar and prototype representations (e.g. Storms, DeBoeck, & Ruts, 2001), and also rule representations, which will be described later. The challenge for cognitive scientists is carefully discerning the conditions under which each type of representation is used and how they interact.

Formal models of prototype theory

Because a prototype has a value on every dimension of the stimulus, it can be formally represented much like an exemplar, although a prototype need not correspond with any actually experienced instance. The prototype for category k has psychological value on dimension i denoted by Ψ_{ki}^{proto}. For the particular prototype model defined here, this value represents the central tendency of the category instances on that dimension. Other prototype models might instead use the ideal value or modal value on each dimension. The model classifies a stimulus as category k in a manner directly analogous to the exemplar model, such that the probability of classifying stimulus s into category k is given by: $p_k = \text{sim}(s, k)/\sum_m \text{sim}(s, m)$. The sum in the denominator is over all category prototypes, instead of over all exemplars. In principle, the probability choice formula for prototype models could include a decisiveness parameter like the exemplar model, but in the prototype model such a decisiveness parameter trades off with the specificity parameter in the similarity computation so that it has no independent influence.

As new instances are experienced during learning, the prototypes are adjusted to reflect the instances. For the first experienced instance of a category, the prototype is created and set to match that instance. For subsequently experienced instances of the category, the prototype changes from its current values slightly toward the new case. By gradually moving toward the instances of the category as they are experienced, the prototype gradually progresses toward the central tendency of those instances.

The learning of central tendencies can be formalized in the following algorithm, closely related to so-called 'competitive learning' or 'clustering' methods. The idea is that a prototype should be adjusted so that it is as similar as possible to as many instances as possible; in this way the prototype is maximally representative of the stimuli in its category. Define the total similarity of the prototypes to the instances as $S = \sum_{k,s} \text{sim}(s, k)$, where $\text{sim}(s, k) = (-\exp(\sum_i \alpha_i [\Psi_i^{\text{stim}} - \Psi_{ki}^{\text{proto}}]^2)$. The question then is how best to adjust Ψ_{ki}^{proto} so that the total similarity increases. One way to do this is gradient ascent: the prototype values are adjusted to increase the total similarity as quickly as possible. The resulting formula, determined as the derivative of the total similarity with respect to the coordinates, yields $\Delta\Psi_{ki}^{\text{proto}} = \lambda \, \text{sim}(s, k)\alpha_i(\Psi_i^{\text{stim}} - \Psi_{ki}^{\text{proto}})$. This formula causes each prototype's values to move toward the currently experienced stimulus, but only to the extent that the prototype is already similar to the stimulus, and only to the extent that the dimension is being attended to. In this way, prototypes that do not represent the stimulus very well are not much influenced by the stimulus.

An intermediate scheme between prototype representation and exemplar representation is the use of multiple prototypes per category, which can be useful to capture multimodal distributions (e.g. Rosseel, 2002). Alternative learning schemes have been used, such as methods derived from Bayesian statistics (Anderson, 1991). As the number of prototypes per category increases, there can eventually be one prototype per exemplar, and such models become equivalent to exemplar models (Nosofsky, 1991). In exemplar models, however, the coordinates of the exemplars typically do not get adjusted from one trial to the next.

RULE THEORIES

Other ways of representing categories are with *rules*. There are a variety of so-called rule-based models in the literature, so that it can be difficult to define exactly what a rule-based model is (e.g. Hahn & Chater, 1998; Smith & Sloman, 1994). One candidate rule that defines membership in the category of rule-based models is this: a rule-based model uses either a strict match/mismatch process or a boundary representation, i.e. a rule-based model does not use graded matching to content. In Table 7.1 this corresponds to cells in the right column or the bottom two rows (i.e. all cells except A and C). The usage of the term 'rule' is merely conventional, however. What really defines the nature of a categorization model is the type of representation and matching process (and learning process) it uses, as analyzed in Table 7.1.

An example of a rule-based model is one that uses *featural rules* that specify strict necessary and sufficient conditions that define category membership. For example, something is a member of the category 'bachelor' if it is human, male, unmarried and eligible. This type of featural rule theory falls in the second row and right column of Table 7.1, cell D. In this case, the category is specified by a global summary of its content, and the summary must be perfectly matched or else the stimulus is not in the category.

One subtlety of features in a rule is that each feature is itself a category. For example, for something to be a bachelor it must have the features of being human, male, unmarried and eligible. Each of these features is itself a category that must be defined in terms of content or boundary, globally or piecemeal, etc. This echoes a theme mentioned above, that categories occur at many levels of analysis simultaneously.

It is possible for strict rules of content to be specified piecemeal. This type of theory falls in the first row and right column of Table 7.1, cell B. For example, a pitch in baseball is a strike if it is swung at by the batter and missed, *or* it is in the strike zone but not hit.

Another example of such a piecemeal rule theory is a strictly matching exemplar model. In this type of theory, a category is represented by its instances, but a stimulus is classified as a member of the category if and only if it exactly matches one of the instances. The exemplar subsystem of Smith and Minda (2000) is one such example of this kind of theory. The Sparse Distributed Memory (SDM) model developed by Kanerva (1988) uses an internal representation that consists essentially of randomly distributed exemplars that have a strict threshold for matching: a stimulus 'matches' the exemplar if and only if it lies within a specific distance from the exemplar.

Instead of specifying content, rule theories can instead specify boundaries that separate categories. Rules for category definition are typically a single threshold on a single dimension, e.g. a building is a skyscraper if it is more than ten stories tall. In some rule-based theories, rules can be more complicated boundaries, e.g. a building is a skyscraper if its height divided by its width is greater than 1.62.

Rule models that use strict match/mismatch processes predict that human performance should be perfect, i.e. classification should be strictly all or nothing depending on whether the conditions (content or boundary) are perfectly matched or not. Yet human classification is imperfect. Rule models that use a strict match/mismatch process must accommodate the 'fuzziness' of human classification performance through mechanisms such as perceptual randomness or decisional randomness. Thus, the rule is strict, but the stimuli are imperfectly perceived or the classification dictated by the rule is imperfectly produced.

Some boundary-based (rule) models do not assume strict match/mismatch processes, and instead used graded similarity to the boundary. One example of this is the PRAS model of Vandierendonck (1995), which combines rectangular decision boundaries with exponentially decaying similarity gradients.

Rules are computationally attractive as category representations because they can be uniformly applied to all stimuli, regardless of the instances actually experienced. For example, the rule for 'bachelor' can be applied with equal facility to all stimuli, regardless of the specific bachelors we have previously encountered. This uniformity of applicability might not mimic human categorization (Allen & Brooks, 1991). Rules are also computationally attractive because they can describe feature combinations that are not tied to the specific featural realizations. For example, a rule could be that an item is a member of a category if the item has either one of two features but not both. This 'exclusive-or' rule can be applied to *any* two features, and is not tied to particular features such as color or shape. Shanks and Darby (1998) showed that people can indeed learn such abstract rules that are not tied to specific features. The hypothesis-testing model of Levine (1975) is one model that uses structural similarities of abstract rules to predict ease of shifting from one rule to another during learning.

Many natural categories are very difficult to specify in terms of content rules, however (e.g. Rosch & Mervis, 1975). For example, the category 'game' appears to have no necessary and sufficient features (Wittgenstein, 1953). Nevertheless, people are prone to look for features that define category distinctions, and people tend to believe that such defining features exist even if in fact they do not (Ashby, Queller, & Berretty, 1999; Brooks, 1978). The propensity to focus on single dimensions might

depend on the context and content domain; for example, in social conditions people might be more prone to sum evidence across dimensions (Wattenmaker, 1995).

Formal models of rule theory

In the 1960s and 1970s, popular rule models included 'hypothesis testing' or 'concept learning' models (for a review, see Levine, 1975). The emphasis at the time was on how people learn logical combinations of features that define a category. The models therefore emphasized strict matching to content (features), and these models fall into the top two rows of the right column of Table 7.1, cells B and D. In these sorts of models, individual features are tested, one at a time, for the ability to account for the correct classifications of the stimuli. For example, the model might test the rule, 'If it's red it's in category K'. As long as the rule works, it is retained, but when an error is encountered, another rule is tested. As simple rules are excluded, more complicated rules are tried. A recent incarnation of this type of model also is able to learn exceptions to rules, by testing additional features of instances that violate an otherwise successful rule (Nosofsky, Palmeri, & McKinley, 1994). This model is also able to account for differences in behavior across people, because there can be different sets of rules and exceptions that equally well account for the classifications of the stimuli. A different rule-based model is presented by Miller and Laird (1996), for which typicality and similarity effects are addressed by probing discrete prediction rules in a series of steps.

When stimuli vary on continuous dimensions instead of discrete features, rule models can specify the boundary that separates the categories (cell H of Table 7.1). In one class of models, the decision boundary is assumed to be have a shape that can be described by a quadratic function, because a quadratic describes the optimal boundary between two multivariate normal distributions, and natural categories are sometimes assumed to be distributed normally (e.g. Ashby, 1992). In this approach, there are three basic postulates. First, the stimulus is represented as a point in multidimensional space, but the exact location of this point is variable because of perceptual noise. Second, a stimulus is classified according to which side of a quadratic decision boundary it falls on. Third, the decision boundary is also subject to variability because of noise in the decision process. Thus, although the classification rule is strict and there is no explicit role in the model for similarity gradients, the model as a whole produces a gradation of classification performance across the boundary because of noise in perception and decision. There are many variations on this scheme of models, involving different shapes of boundaries, deterministic versus probabilistic decision rules, etc. (Ashby & Alfonso-Reese, 1995; Ashby & Maddox, 1993).

It is possible to combine multiple boundaries to specify a category distinction (cell F of Table 7.1). For example, a model by Ashby, Alfonso-Reese, Turken, and Waldron (1998) combines linear decision boundaries that involve single dimensions, corresponding to verbalizable rules, with linear decision boundaries that combine two or more dimensions, corresponding to implicitly learned rules. These decision boundaries are then weighted to generate an overall linear decision boundary. Thus, while several component boundaries are involved, the overall system behaves as if it had a single globally defined boundary.

Hybrid Representation Theories

Is human category learning completely described by any one of the types of representation in Table 7.1? It is not likely. Recent theories combine different representations to account for complex patterns in human behavior. A variety of work has shown that neither rule-based nor prototype models can fully account for human categorization (e.g. Ashby & Waldron, 1999; Kalish & Kruschke, 1997). In particular, exemplar representation must be supplemented with rules to account for human learning and generalization (e.g., Erickson & Kruschke, 2002; Shanks & St John, 1994; Smith, Patalano, & Jonides, 1998). Some evidence apparently for multiple systems can, it turns out, be explained by single-representation systems (e.g. Lamberts, 2001; Nosofsky & Johansen, 2000; Nosofsky & Kruschke, 2002), but follow-up work continues to challenge the single-representation approach (e.g. Ashby & Ell, 2002; Erickson & Kruschke, 2002).

When multiple representations are combined into a single model, a significant challenge to the theorist is determining how the two representations interact or compete during learning and categorization. One approach to his problem is presented in a model by Erickson and Kruschke (1998, 2002; Kruschke & Erickson, 1994), which combines exemplars with single-dimension rules. The exemplar representation falls in the top left cell (A) of Table 7.1, and the rule representation falls within the third row and left column (cell E). What is important about this model is how it decides when to apply which representation. The model does this with a gating mechanism that learns which representation to pay attention to, depending on the exemplar. Thus, the attentional distribution (attend more to exemplars versus attend more to rules) is itself a categorization before the final categorization of the stimulus. When a stimulus appears, the model first classifies it as a rule-governed stimulus or an exemplar-governed stimulus,

and then the model accordingly classifies the stimulus using the rule-based or exemplar-based subsystem.[3]

Hybrid representation models are proliferating (e.g. Anderson & Betz, 2001), and are sure to appear in creative new approaches in the future. There is also evidence for multiple learning processes, which might also interact with the types of representations being learned (e.g. Raijmakers, Dolan, & Molenaar, 2001; VanOsselaer & Janiszewski, 2001).

Role of Similarity

Exemplar models depend on the computation of similarity between stimuli and items in memory. Prototype models also rely on the determination of similarity between stimuli and memory representations (namely, the prototypes). Even some rule models compute similarity between stimuli and boundaries (e.g. Vandierendonck, 1995), or can be re-conceptualized as doing so.

Some researchers have criticized the notion of similarity as being internally incoherent, and some critics have argued that similarity does not always correlate with categorization. If similarity as a theoretical construct is decrepit and dilapidated, then category learning models founded on similarity are also in peril of collapse.

There have been a number of demonstrations that suggest that similarity is incoherent. In these situations, similarity seems to change depending on how it is measured. Psychological similarity can be empirically assessed by a number of methods. One method simply asks people to rate the similarity of two items on a scale from 1 to 10; another method measures discriminability or confusability of items. Usually these different assessments agree: items measured to be more similar than other items by one method are also measured to be more similar by a different method. But sometimes different assays do not agree (e.g. Tversky, 1977). Similarity can also be context-specific: in the context of hair, gray is more similar to white than to black, but in the context of clouds, gray is more similar to black than to white (Medin & Shoben, 1988). In general, models of psychological similarity presume which features or dimensions are used for comparing the objects, without any explanation of why those features or dimensions are selected. Models of similarity do have parameters for specifying the attention allocated to different features, but the models do not describe how these attentional values come about (Goodman, 1972; Murphy & Medin, 1985).

Other research suggests that similarity is not always an accurate predictor of categorization. Consider the category, *things to remove from a burning house*. The items *heirloom jewelry* and *children* are both central members of this category, yet they have little similarity in terms of visual appearance (Barsalou, 1983). On the other hand, if attention is directed only to the features *valuable*, *irreplaceable* and *portable*, then children and heirloom jewelry bear a strong similarity. As another example, this one taken from personal experience, consider an actual label on a product: 'Great for sleeping, gun shooting, studying, aircraft.' What is the product? Earplugs. The applications of the product, i.e. the members of this category, are highly similar only when attention is directed to the critical feature of undesired noise. Once again the issue of what to attend to is a theoretical crux, not addressed by current theories of similarity.

Thus, similarity is itself a complex psychological phenomenon in need of theoretical explication. Despite the complexities, there are strong regularities in similarity and categorization data that should yield to formal treatment. Excellent reviews of these topics have been written by Goldstone (1994) and by Medin, Goldstone, and Gentner (1993). In particular, a comprehensive theory of similarity will need a theory of attention. The role of attention will be discussed again, below.

More Complex Representations

The previous discussion has assumed that items are represented by collections of features. These representations are 'flat' in the sense that no feature is a combination of other features, and every item within a category has the same universe of candidate features. But psychological representations of complex categories might involve structured representations, involving hierarchical combinations of features and dimensions that are present in some instances but not others. For example, the category *vehicle* has instances such as *car* that have windshields, but also has instances such as *bicycle* that have no windshields. Future models of category learning will need to address these non-flat representations (Lassaline & Murphy, 1998).

Palmeri (1999) showed that an exemplar-based model can account for some learning of hierarchically organized categories, structured by superordinate, basic and subordinate levels analogous to vehicle, car and Ford. But it is an open question as to whether any single-representation model can comprehensively account for learning of categories at multiple levels. Lagnado and Shanks (2002) discuss multiple levels of representation and suggest that a dual-component model is needed. Theories of category learning will eventually have to address other varieties of complex representation, e.g. Markman and Stilwell (2001) discuss the notion of a 'role'-governed category which specifies the relational role played by its members.

Whereas complex representations will eventually need to be addressed by category-learning models, at this time even the simplest types of cue combination are not fully understood. For example, there is active research in determining the conditions under which two features are processed as independently additive components in a representation or as a conjunctive compound that is distinct from the components (e.g. Shanks, Charles, Darby, & Azmi, 1998; Shanks, Darby, & Charles, 1998; Williams & Braker, 1999; Young, Wasserman, Johnson, & Jones, 2000). Lachnit, Reinhard, and Kimmel (2000) discuss the need for a representation of the abstract relational dimension of 'separate' versus 'together', distinct from the numerical quantification of one versus two.

Other Forms of Induction

Until this point in the chapter, it has been emphasized that the function of categorization is to infer an unseen feature from given information. The inferred feature has been a nominal category or characteristic, and the process of inference has consisted of some kind of matching of a 'flat' stimulus vector to a stored vector or set of vectors. But the inferred information could be richer than a simple nominal category, and the inference process could use representations and processes more complicated than matching of flat vectors.

If the inferred information is a value on a continuous scale, then the mapping from stimuli to inferred values constitutes what is called a *function*. For example, a physician who must prescribe an appropriate amount of medication might know the functional relationship between (a) the observed values of body weight and blood pressure and (b) the inferred value of amount of medication. A baseball outfielder might know the functional relationship between (a) the observed values of distance to infielder and urgency and (b) the inferred values of amount of force and angle of throw.

The study of function learning is relatively nascent compared to category learning. A very useful review is provided by Busemeyer, Byun, Delosh, and McDaniel (1997). There are many analogous findings in the two areas, in terms of relative ease of learning different types of dimensional combinations. Theories of function learning can also be differentiated according to the dimensions in Table 7.1, but the most prominent difference among function-learning theories is whether the function is specified globally or piecemeal. Koh and Meyer (1991), for example, have proposed a function-learning model in which globally defined functions are gradually regressed onto the observed instances during learning, much like gradual tuning of hidden nodes in backpropagation (Rumelhart, Hinton, & Williams, 1986; but cf. Kruschke, 1993). Delosh, Busemeyer, and McDaniel (1997) have described a model that learns exemplars, i.e. piecemeal input–output value combinations, analogous to the exemplar-based category-learning model ALCOVE (Kruschke, 1992), and then uses a linear extra-/interpolation response strategy between learned exemplars. Kalish, Lewandowsky, and Kruschke (in press) have described a sophisticated model that applies different globally defined functions to different specific regions of the input space, i.e. piecemeal application of global functions, analogous to the piecemeal application of global rules in the category-learning model ATRIUM (Kruschke & Erickson, 1994; Erickson & Kruschke, 1998, 2002).

The induction process might be based on representations more complicated than flat vectors. People have rich theoretical knowledge about many domains. For example, people know that birds are more than a collection of feathers, beak, wings, chirps, etc. People also know that wings enable flight, that flight has to do with air flow and air pressure; that some birds compete for territory with other birds of the same species, that bird songs can mark territory, etc. Rich knowledge such as this is packaged into *theories* of how behavior works and how the features are causally interrelated. This kind of knowledge might not be felicitously represented as a flat vector of features, and instead might need to be represented as a hierarchy of features in nested relations. Inference based on these complex representations must then also be more complex than feature–vector matching. Many interesting effects in inductive reasoning can be captured by simple feature–vector matching (Sloman, 1998), but more complex representations and processes are needed to account for other situations, such as those involving causal relations (e.g. Ahn, Kim, Lassaline, & Dennis, 2000). Nevertheless, just as categorization theorists debate exemplar versus rule-based models, inductive-reasoning theorists also debate exemplar versus rule-based models (Hahn & Chater, 1998; Heit, 2000; Sloman, 1996).

Toward a Comprehensive Theory? The Role of Attention

The many complexities of categorization and inference will not easily yield to a comprehensive theory, but it does seem that a crucial issue in the path of progress is *attention*. In its broadest definition, attention simply refers to the selectivity of information usage in inference. People learn that out of the plethora of available information, only some aspects should be attended to in certain situations. Attention refers to both enhanced or amplified

processing of some information and diminished or suppressed processing of other information. What often poses the biggest challenges to theories of categorization is the fluidity and context specificity of information selection. Moreover, this selection can occur at multiple levels of information simultaneously, so that it is not only the input features that are selected, but the various higher-level re-representations as well. A variety of perplexing phenomena in category learning might be solved by appropriate theories of attention.

Why should there be attentional selection? One reason is that selection of relevant information can accelerate learning. This speed is engendered by the attentionally enhanced discriminability of information that requires different responses. For example, suppose that color is critical to discriminate edible from poisonous mushrooms, such that darker shades of brown are to be avoided. By attending to color, the discrimination between darker and lighter brown is improved, therefore accuracy and learning are improved.

Not only does the amplifying of relevant information benefit learning, so does the quashing of irrelevant information. This benefit occurs because learned associations will generalize across instances that differ on the irrelevant dimension if that dimension is ignored. For example, suppose that size is irrelevant for discriminating edible from poisonous mushrooms. If you learn that a certain 4 cm tall mushroom is edible, then when you see a 8 cm tall mushroom of the same color you will not starve. That is, by learning that size is irrelevant, learning about a 4 cm mushroom immediately benefits inference about an 8 cm mushroom.

A classic example of enhanced learning due to attentional quashing of irrelevant dimensions comes from the work of Shepard, Hovland, and Jenkins (1961). They studied the relative ease with which people learned different classifications of a set of objects. The objects varied on three binary-valued dimensions. For example, the objects could be simple geometric forms that vary on dimensions of shape (circle or triangle), size (large or small) and color (red or green). The so-called 'Type II' structure has *two* relevant dimensions; an example of such a structure can be described by this rule: a form is in category 1 if it is circular or large but not both, otherwise it is in category 2. A condition of that form, i.e. one feature or another but not both, is called an exclusive-or, also abbreviated as 'xor'. Notice in this example that shape and size are relevant, but not color. The so-called 'Type IV' structure, on the other hand, has *three* relevant dimensions; an example is as follows: a form is in category A if it has any two or more of circular, large and red. Shepard et al. (1961) found that the Type II structure, involving just two relevant dimensions, was easier to learn than the Type IV structure, involving three dimensions, despite the fact that instances in the Type IV structure are more compactly distributed in the three-dimensional stimulus space. Nosofsky and Palmeri (1996) found that this advantage for Type II occurs only for stimuli that have dimensions that can be selectively attended to. For the 'integral' dimensions of hue, saturation and brightness of colors, which are extremely difficult to selectively attend to, Type II is more difficult than Type IV. Nosofsky and colleagues (Nosofsky et al., 1994; Nosofsky & Palmeri, 1996) found that a model that incorporates learned selective attention (ALCOVE: Kruschke, 1992) accommodated the human learning data very well, but models without selective attention failed.

Attentional selection applies not only to the input end of information processing, but also to the output end. Of the many actions a person might take, typically only one action can be carried out at a time. Thus, cognition must ultimately be selective at its output, and this required selectivity might be enhanced by selectivity earlier in the processing stream. Moreover, action influences the input to cognition. Our eyes can be directed to only a limited part of the world, and our hands can feel only a limited extent of a surface. If the world imposes a cost for lingering too long on irrelevant information (e.g. by not detecting predators or competitors for limited resources before they strike), then it is adaptive to learn what to attend to. The selectivity imposed upon perceptions and actions by our effectors suggests that sources of information throughout the processing stream should *compete* for attention. Maddox (2002), for example, has tried to assess attention at different perceptual and decisional levels in the context of category learning. Thus, a third rationale for attentional learning is that perception and action have limited scope and the organism should therefore learn what to attend to when in a competitive environment.

Numerous studies have reported evidence that attention during learning is indeed of limited capacity, such that attending to one source of information detracts from utilizing another source (examples of recent relevant articles include Ashby & Ell, 2002; Gottselig, Wasserman, & Young, 2001; Kruschke & Johansen, 1999; Nosofsky & Kruschke, 2002; Waldron & Ashby, 2001). Many theories of learning include some attentional capacity constraint (e.g. Pearce, 1994) or cue competition (e.g. Rescorla & Wagner, 1972), but relatively few theories include mechanisms for attentional learning (Kruschke, 2001a). Evidence abounds for cue competition in associative or category learning, and some research has also found cue competition in function learning (e.g. Birnbaum, 1976; Busemeyer, Myung, & McDaniel, 1993; Mellers, 1986). If category learning and function learning share similar mechanisms, as discussed earlier, then attention learning should be robustly evident in function learning as well. In summary so far, attentional learning explains a

number of phenomena in human learning because attention improves discrimination on relevant dimensions, improves generalization over irrelevant dimensions, and addresses the problem of limited capacity in perception and action.

Attentional shifting is assessed not only by its influence on learning in novel domains, but also by its influence on subsequent learning. Much research has examined transfer of learning when there are shifts in the relevance of information. The motivation for this type of assessment is straightforward: if a person has learned that some dimensions are relevant but others are irrelevant, then a subsequent categorization that retains the same relevant dimensions should be relatively easy to learn but a subsequent categorization that changes the relevance of dimensions should be relatively difficult. For example, if people initially learn that red indicates category 1 and green indicates category 2 whereas shape is irrelevant, then it should be easy subsequently to learn that blue indicates category 1 and yellow indicates category 2, but it should be difficult to learn that circle indicates category 1 and triangle indicates category 2. The former type of shift is called an 'intradimensional shift' (IDS) because the relevant values for the categorization remain within the same dimension. The latter type of shift is called an 'extradimensional shift' (EDS) because the relevant values change to a different dimension.

A useful notation for the structure of an IDS is A(B) → An(Bn), where the relevant dimension is denoted by a letter without parentheses, the irrelevant dimension is denoted by a letter with parentheses, the shift is denoted by the arrow, and the 'n' indicates novel values of the dimensions. An EDS is denoted by A(B) → Bn(An). It is interesting to ask whether the advantage of IDS over EDS is due to learned perseveration on the relevant dimension or to learned inhibition of the irrelevant dimension, or both. Owen, Roberts, Hodges, Summers, Polkey, and Robbins (1993) attempted to isolate those two influences with two shifts that eliminated one of the initial dimensions. One shift eliminated the initially irrelevant dimension so that it could no longer influence the learning of the shift stage. In abstract notation, the design was A(B) → Cn(An). This shift would be difficult only if people had learned to perseverate on dimension A. Another design was A(B) → Bn(Cn), which would be difficult only if people had learned to inhibit dimension B. Owen et al. (1993) found that both types of shift were difficult, indicating that people learn both perseveration on the relevant dimension and inhibition of the irrelevant dimension.

Kruschke (1996b) studied more complex designs so that relevance shifts could be conducted without introducing novel values of the dimensions. This allowed comparing relevance shifts with *reversal* shifts, wherein the mapping to categories is simply reversed without changing the stimuli. The initially learned structure was an xor on dimensions A and B, with dimension C irrelevant, denoted AxorB(C). The reversal shift, denoted AxorB(C) → AxorB(C)rev, was extremely easy for people to learn. A shift that retained one of the relevant dimensions and was therefore a type of IDS, denoted AxorB(C) → A(B)(C), was next easiest, and a shift to the initially irrelevant dimension, a type of EDS denoted AxorB(C) → C(A)(B), was more difficult. Kruschke (1996b) found that the difficulty of shifting to a previous irrelevant dimension could be nicely captured by a model with learned attention, but the great facility of reversal shifting demanded an additional remapping mechanism. Neural evidence for the distinct processing of relevance and reversal shifts comes from Rogers, Andrews, Grasby, Brooks, and Robbins (2000). Further variations of IDS and EDS structures were studied by Oswald, Yee, Rawlins, Bannerman, Good, and Honey (2001), using rats as subjects. The researchers found that AxorB(C) → AxorBn(Cn), i.e. a type of IDS, was easier than AxorB(C) → AxorCn(Bn), a type of EDS. Not only was the structure more complex than the traditional IDS/EDS, but the dimensions were in three separate modalities (auditory, visual and tactile) and the A dimension was phasic (a brief tone) whereas the B and C dimensions were tonic (static patterns on the walls and floors). The robustness and generality of the advantage of IDS over EDS is generally interpreted as strong evidence for attentional learning in the initial phase of training, with perseveration of that learning into the subsequent phase.

Attentional learning does more than accelerate learning in completely novel domains. Attentional shifting also preserves previous learning for similar stimuli, whenever possible. Thus, when new stimuli appear that share some aspects with previously learned stimuli but also have some novel aspects, if the previously relevant aspects continue to correctly predict appropriate behavior, then the novel aspects will quickly be learned to be irrelevant. That is, the previously learned knowledge about the relevant aspects will be respected, to the extent that it continues to be successful. On the other hand, if the new stimuli demand different behavior, then attention will quickly shift to the novel aspects, thereby protecting the previously learned stimuli from being 'overwritten' by the new items. Thus, attention shifting protects previously learned categorizations by reducing interference when new items demand different categorizations. This is achieved by attentionally highlighting novel aspects of the stimuli, and associating these aspects with the new category.

An example of this attentional highlighting comes from the otherwise perplexing phenomenon known as the *inverse base-rate effect*, wherein a cue that indicates a rare category is apparently overweighted (Medin & Edelson, 1988). In this procedure,

people initially learn that cues A and B indicate category 1, and subsequently learn that cues A and C indicate category 2. When tested with the new cue combination B and C, people tend to classify it as category 2, despite the facts that the cues are really equally predictive of the two categories and the first category was more frequent overall. The apparently irrational behavior is naturally explained by attentional shifting: when learning the second category, people shift attention away from cue A because they have already learned that it predicts category 1, and they shift attention to cue C because it does not conflict with previous learning. Therefore they learn a strong association from cue C to category 2, and, moreover, they learn that when cue C appears they should attend to it, especially in the context of cue A. The quantitative details of data from many such experiments are accurately accounted for by models that have rapid shifts of attention during learning (Dennis & Kruschke, 1998; Fagot, Kruschke, Depy, & Vauclair, 1998; Kalish, 2001; Kalish & Kruschke, 2000; Kruschke, 1996a, 2001b).

Another phenomenon, called apparent base-rate neglect (Gluck & Bower, 1988), can also be explained by attentional shifting of this nature (Kruschke, 1996a). The essential idea is that categories that occur more frequently (i.e. have higher base rates) are learned first. Subsequently the less frequent categories are learned, and attention highlights the distinctive features of the rare categories in order to protect the previous learning about the frequent categories. The highlighting of distinctive features of the rare categories is difficult for some popular exemplar-based models to account for (e.g. Lewandowsky, 1995), but it can be well accommodated by an exemplar-based model that has rapidly shifting, learned selective attention (Kruschke & Johansen, 1999). Another related phenomenon is the 'contrast effect' reported by Kersten, Goldstone, and Schaffert (1998). Essentially, the contrast effect occurs with an EDS using *novel* categories in the shift phase. Kersten et al. (1998) suggested that the ease of an EDS for novel categories was best explained by a distinct type of attention. Alternatively, it can be interpreted as a case of attentional highlighting. When the categories are novel, attention does not perseverate on previously relevant dimensions because doing so would contradict previously learned categories. Instead, attention highlights distinctive dimensions to rapidly accommodate the newly demanded outcomes. Thus, an EDS is relatively difficult when the categories are the same as the initial learning, but it is relatively easy when the categories are novel.

The attentional shift learned during the inverse base-rate effect is context-specific, i.e. attention shifts away from cue A to cue C especially in the context of those two cues and the corresponding responses. The theory of attention shifting propounded here asserts that attention shifts are context- and exemplar-specific, i.e. attentional redistributions are a learned response from particular cue combinations, with some degree of graded generalization from those learned cases. The context specificity of learned attention can address results of various other category-learning experiments. Macho (1997), for example, had people learn a prototype structure divided in two phases. In this structure, all dimensions had two values, denoted 1 and 2. If the stimuli had three dimensions (there were actually more dimensions in Macho's experiments), then the prototype of one category had values of 1 on all three dimensions, denoted 111. The prototype of the other category had the opposite values on all three dimensions, i.e. 222. Other instances were symmetrically distributed around the prototypes, e.g. the first category also included exemplars 112, 121 and 211, whereas the second category also included exemplars 221, 212 and 122. Training on the category instances was split across phases such that each phase made different dimensions more relevant than others despite the fact that when collapsed over the course of both phases the dimensions were equally relevant. Using our three-dimensional example, the first phase could have consisted of instances 111, 112, 221 and 212, for which only the first dimension is perfectly predictive of the correct categorization, and the second phase could have consisted of instances 121, 211, 122 and 222, for which only the third dimension is perfectly predictive of the correct categorization. Macho (1997) found that at least one exemplar-based model with attentional shifts (ALCOVE: Kruschke, 1992) could not accommodate the results. That model's problems were that it could not shift attention quickly enough, nor could it learn exemplar-specific attentional redistributions. It is likely that more recent models with rapid shifts of attention, and with exemplar-specific learned attention (e.g. Kruschke & Johansen, 1999; Kruschke, 2001a), could more accurately account for the results.

In another experiment demonstrating the context specificity of learned attention, Aha and Goldstone (1992) showed that people can learn that one dimension is relevant for categorization in one region of a stimulus space, but a different dimension is relevant in a different region of the space. Erickson and Kruschke (2001) extended their results by showing there are individual differences in which regions of the space are learned first, and that a model that shifts attention among rules, depending on the exemplar, fits the data well.

In general, the sequence of learning plays an important role in what is learned. The variety of experiments summarized above indicated that subsequent learning can be strongly influenced by previous learning. Base rates influence what is learned by influencing the sequence with which items and

categories are learned. The order of learning influences what is learned in large part because attention will shift to those attributes that facilitate learning. Goldstone (1996) showed in several insightful experiments that people will learn more non-diagnostic features of a category if many instances of the category are presented consecutively than if the instances are interleaved with cases from a contrasting category. Spalding and Ross (1994) showed that the particular instances that people analyze early in learning have a strong influence on what is learned about the categories. In particular, attributes common to early-learned cases of a category will tend to dominate knowledge of those categories. Billman and Davila (2001) reported that it is easier for people to discriminate categories that contrast consistently on a few features.

These influences of training order are not merely laboratory curiosities, but apply also to real-life learning in clinical settings (Welk, 2002) and language acquisition (Smith, Jones, Landau, Gershkoff-Stowe, & Samuelson, 2002). Individual differences might also be usefully explained as differences in training history, or as individual differences in generalization gradients, attentional shifting rates or dimensional saliences (e.g. Dixon, Koehler, Schweizer, & Guylee, 2000; Niedenthal, Halberstadt, & Innes-Ker, 1999; Treat et al., 2001). The phenomena reviewed above are consonant with the general notions of attentional learning propounded here: attention focuses on features that consistently indicate a category because doing so facilitates learning of that category. When a contrasting category is presented, attention highlights the consistently distinctive (diagnostic) features of the new category, because doing so facilitates new learning and protects previous knowledge.

For attentional theories to have broad applicability to category learning, they must eventually address more complex representational structures. As an example of this need, Lassaline and Murphy (1998) showed that people learn categories differently depending on whether the features that differ across instances are 'alignable'. Alignable features are values from a common dimension, e.g. red vs. green, whereas non-alignable features are from different dimensions, e.g. wing vs. branch. (Alignability is itself a subtle psychological construct.) Current models of attentional learning do not address the alignability of dimensions. Even if this issue is sidestepped for now, there are still debated questions about how complex knowledge structures influence the distribution of attention over dimensions. Some researchers have argued that prior knowledge does more than simply reallocate attention across stimulus dimensions (e.g. Hayes & Taplin, 1995), whereas other investigators have concluded that background concepts and theories do influence new learning through redistribution of attentional weights (e.g. Vandierendonck & Rosseel, 2000). Whether or not attentional redistribution, combined with a variety of representational constructs, can comprehensively accommodate the wealth of phenomena in category learning, there is little doubt that sophisticated theories of attention shifting will play an important role in understanding category learning.

Summary

Categorization in its broadest definition is simply the inference of unseen attributes from observable features. The unseen attribute could be a category label or some other characteristic of the item. Because inference of appropriate action is perhaps the fundamental goal of cognition, categorization and category learning can be viewed as a core research domain in cognitive science.

Different theories of categorization hypothesize different representations. Various theories also assume different processes for matching stimulus representations and memory representations. Some theories posit representations of content, such as exemplars or prototypes. Other theories posit representations of boundaries between categories. For either type of representation, the specification can be global or piecemeal. Once the representation is established, then the theories can assume that the matching of stimulus and memory representations is done with a graded degree of match, or with a strict match versus no match.

Research is moving toward the conclusion that no one type of representation can accommodate the full complexity of human category learning. The challenge facing researchers now is determining which representations are used in what situations and how the representations trade off in learning and inference. Theories are also moving away from simple 'flat' vector representations to more complex structured representations. Regardless of the representation, there is ample research showing the robust influence of attention in category learning. Theories will have to address how prior knowledge influences attention which influences subsequent learning which influences attention for future learning.

Further Reading

While writing this chapter, the author was frequently tempted to give up and simply point the reader to the many excellent previous summaries of research in categorization. Perhaps that would have been the better course of action! Here, then, are some pointers to previous summaries of the topic.

The book by Smith and Medin (1981) provides a highly readable introduction to the issues in categorization research. The collection of articles by Rosch and Lloyd (1978) are classic statements of fundamental results and theoretical perspectives. The book by Shanks (1995) is a lucid review of issues in associative learning in humans, and a summary of associative learning in animals has been written by Pearce and Bouton (2001). Estes (1994) presents a more mathematically oriented survey of theories of classification. An accessible collection of tutorials is presented by Lamberts and Shanks (1997), see also the review of concepts by Lamberts (2000a), and Goldstone and Kersten (2003) provide an excellent review of the field, all of which you should probably read instead of this chapter.

Acknowledgment

Supported in part by Grant BCS-9910720 from the National Science Foundation.

Notes

1 'Wherever human life is concerned, the unnatural stricture of excessive verticality cannot stand against more natural horizontality' (Frank Lloyd Wright on skyscrapers).

2 'Male' and 'female' are clearly defined genetically in virtually all individuals. There are extremely rare exceptions for whom their chromosomes are neither XX (female) nor XY (male).

3 Mirman and Spivey (2001) described a mixture-of-experts model similar to Erickson and Kruschke's (1998, 2002; Kruschke & Erickson, 1994), in which the rule module is instead a standard backprop network (Rumelhart et al., 1986). The likely problem with Mirman and Spivey's approach is that the model will suffer the same problems as standard backprop, as pointed out by Kruschke (1993).

References

Aha, D. W., & Goldstone, R. (1992). Concept learning and flexible weighting. In *Proceedings of the Fourteenth Annual Conference of the Cognitive Science Society* (pp. 534–539). Hillsdale, NJ: Erlbaum.

Ahn, W. K., Kim, N. S., Lassaline, M. E., & Dennis, M. J. (2000). Causal status as a determinant of feature centrality. *Cognitive Psychology*, *41*, 361–416.

Allen, S. W., & Brooks, L. R. (1991). Specializing the operation of an explicit rule. *Journal of Experimental Psychology. General*, *120*, 3–19.

Anderson, A. L., Ross, B. H., & Chin-Parker, S. (2002). A further investigation of category learning by inference. *Memory and Cognition*, *30*, 119–128.

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

Anderson, J. R., & Betz, J. (2001). A hybrid model of categorization. *Psychonomic Bulletin and Review*, *8*, 629–647.

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

Ashby, F. G. (1992). Multidimensional models of categorization. In F. G. Ashby (ed.), *Multidimensional models of perception and cognition* (pp. 449–483). Hillsdale, NJ: Erlbaum.

Ashby, F. G., & Alfonso-Reese, L. (1995). Categorization as probability density estimation. *Journal of Mathematical Psychology*, *39*, 216–233.

Ashby, F. G., Alfonso-Reese, L. A., Turken, A. U., & Waldron, E. M. (1998). A neuropsychological theory of multiple systems in category learning. *Psychological Review*, *105*, 442–481.

Ashby, F. G., & Ell, S. W. (2002). Single versus multiple systems of category learning. *Psychonomic Bulletin and Review*, *9*, 175–180.

Ashby, F. G., & Maddox, W. T. (1993). Relations between prototype, exemplar and decision bound models of categorization. *Journal of Mathematical Psychology*, *37*, 372–400.

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

Ashby, F. G., & Waldron, E. M. (1999). On the nature of implicit categorization. *Psychonomic Bulletin and Review*, *6*, 363–378.

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

Barsalou, L. (1983). Ad hoc categories. *Memory and Cognition*, *11*, 211–227.

Billman, D., & Davila, D. (2001). Consistent contrast aids concept learning. *Memory and Cognition*, *29*, 1022–1035.

Birnbaum, M. H. (1976). Intuitive numerical prediction. *American Journal of Psychology*, *89*, 417–429.

Blair, M., & Homa, D. (2001). Expanding the search for a linear separability constraint on category learning. *Memory and Cognition*, *29*, 1153–1164.

Bourne, L. E. (1970). Knowing and using concepts. *Psychological Review*, *77*, 546–556.

Brooks, L. R. (1978). Nonanalytic concept formation and memory for instances. In E. Rosch & B. B. Lloyd (Eds.), *Cognition and categorization* (pp. 169–211). Hillsdale, NJ: Erlbaum.

Busemeyer, J. R., Byun, E., Delosh, E. L., & McDaniel, M. A. (1997). Learning functional relations based on experience with input-output pairs by humans and artificial neural networks. In K. Lamberts & D. Shanks (Eds.), *Knowledge, concepts and categories* (pp. 405–437). Cambridge, MA: MIT Press.

Busemeyer, J. R., Dewey, G. I., & Medin, D. L. (1984). Evaluation of exemplar-based generalization and the abstraction of categorical information. *Journal of Experimental Psychology: Learning, Memory, and Cognition*, *10*, 638–648.

Busemeyer, J. R., & Myung, I. J. (1988). A new method for investigating prototype learning. *Journal of Experimental Psychology: Learning, Memory, and Cognition*, *14*, 3–11.

Busemeyer, J. R., Myung, I. J., & McDaniel, M. A. (1993). Cue competition effects: empirical tests of adaptive network learning models. *Psychological Science*, *4*, 190–195.

Choi, S., McDaniel, M. A., & Busemeyer, J. R. (1993). Incorporating prior biases in network models of conceptual rule learning. *Memory and Cognition*, *21*, 413–423.

Cohen, A. L., & Nosofsky, R. M. (2000). An exemplar-retrieval model of speeded same-different judgments. *Journal of Experimental Psychology: Human Perception and Performance*, *26*, 1549–1569.

Cohen, A. L., Nosofsky, R. M., & Zaki, S. R. (2001). Category variability, exemplar similarity, and perceptual classification. *Memory and Cognition*, *29*, 1165–1175.

Delosh, E. L., Busemeyer, J. R., & McDaniel, M. A. (1997). Extrapolation: the sine qua non for abstraction in function learning. *Journal of Experimental Psychology: Learning*, *Memory*, *and Cognition*, *23*, 968–986.

Dennis, S., & Kruschke, J. K. (1998). Shifting attention in cued recall. *Australian Journal of Psychology*, *50*, 131–138.

Dixon, M. J., Koehler, D., Schweizer, T. A., & Guylee, M. J. (2000). Superior single dimension relative to 'exclusive or' categorization performance by a patient with category-specific visual agnosia: empirical data and an ALCOVE simulation. *Brain and Cognition*, *43*, 152–158.

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

Erickson, M. A., & Kruschke, J. K. (2001). Multiple representations in inductive category learning: evidence of stimulus- and time-dependent representation. Available from authors.

Erickson, M. A., & Kruschke, J. K. (2002). Rule-based extrapolation in perceptual categorization. *Psychonomic Bulletin and Review*, *9*, 160–168.

Estes, W. K. (1994). *Classification and cognition*. New York: Oxford University Press.

Fagot, J., Kruschke, J. K., Depy, D., & Vauclair, J. (1998). Associative learning in baboons (*Papio papio*) and humans (*Homo sapiens*): species differences in learned attention to visual features. *Animal Cognition*, *1*, 123–133.

French, R. M. (1999). Catastrophic forgetting in connectionist networks. *Trends in Cognitive Sciences*, *3*, 128–135.

Gelman, S. A., & Markman, E. M. (1986). Categories and induction in young children. *Cognition*, *23*, 183–209.

Gluck, M. A., & Bower, G. H. (1988). From conditioning to category learning: an adaptive network model. *Journal of Experimental Psychology: General*, *117*, 227–247.

Goldstone, R. L. (1994). The role of similarity in categorization: providing a groundwork. *Cognition*, *52*, 125–157.

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

Goldstone, R. L., & Kersten, A. (2003). Concepts and categorization. In A. F. Healy & R. W. Proctor (Eds.), *Handbook of psychology: Vol. 4*. *experimental psychology* (pp. 599–621). New York: Wiley.

Goodman, N. (1972). Seven strictures on similarity. In N. Goodman (ed.), *Problems and projects* (pp. 437–447). New York: Bobbs-Merrill.

Gottselig, J. M., Wasserman, E. A., & Young, M. E. (2001). Attentional trade-offs in pigeons learning to discriminate newly relevant visual stimulus dimensions. *Learning and Motivation*, *32*, 240–253.

Hahn, U., & Chater, N. (1998). Similarity and rules: distinct? exhaustive? empirically distinguishable? *Cognition*, *65*, 197–230.

Hayes, B. K., & Taplin, J. E. (1995). Similarity-based and knowledge-based processes in category learning. *European Journal of Cognitive Psychology*, *7*, 383–410.

Heit, E. (2000). Properties of inductive reasoning. *Psychonomic Bulletin and Review*, *7*, 569–592.

Homa, D., Sterling, S., & Trepel, L. (1981). Limitations of exemplar-based generalization and the abstraction of categorical information. *Journal of Experimental Psychology: Human Learning and Memory*, *7*, 418–439.

Kalish, M. L. (2001). An inverse base rate effect with continuously valued stimuli. *Memory and Cognition*, *29*, 587–597.

Kalish, M. L., & Kruschke, J. K. (1997). Decision boundaries in one-dimensional categorization. *Journal of Experimental Psychology: Learning, Memory, and Cognition*, *23*, 1362–1377.

Kalish, M. L., & Kruschke, J. K. (2000). The role of attention shifts in the categorization of continuous dimensioned stimuli. *Psychological Research (Psychologische Forschung)*, *64*, 105–116.

Kalish, M. L., Lewandowsky, S., & Kruschke, J. K. (in press). Population of linear experts: knowledge partitioning and function learning. *Psychological Review*.

Kanerva, P. (1988). *Sparse distributed memory*. Cambridge, MA: MIT Press.

Kersten, A. W., Goldstone, R. L., & Schaffert, A. (1998). Two competing attentional mechanisms in category learning. *Journal of Experimental Psychology: Learning, Memory, and Cognition*, *24*, 1437–1458.

Koh, K., & Meyer, D. E. (1991). Function learning: induction of continuous stimulus-response relations. *Journal of Experimental Psychology: Learning, Memory, and Cognition*, *17*, 811–816.

Kruschke, J. K. (1992). ALCOVE: an exemplar-based connectionist model of category learning. *Psychological Review*, *99*, 22–44.

Kruschke, J. K. (1993). Human category learning: implications for backpropagation models. *Connection Science*, *5*, 3–36.

Kruschke, J. K. (1996a). Base rates in category learning. *Journal of Experimental Psychology: Learning, Memory, and Cognition*, *22*, 3–26.

Kruschke, J. K. (1996b). Dimensional relevance shifts in category learning. *Connection Science*, *8*, 201–223.

Kruschke, J. K. (2001a). Toward a unified model of attention in associative learning. *Journal of Mathematical Psychology*, *45*, 812–863.

Kruschke, J. K. (2001b). The inverse base-rate effect is not explained by eliminative inference. *Journal of Experimental Psychology: Learning, Memory, and Cognition*, *27*, 1385–1400.

Kruschke, J. K., & Erickson, M. A. (1994). Learning of rules that have high-frequency exceptions: new empirical data and a hybrid connectionist model. In *Proceedings of the Sixteenth Annual Conference of the Cognitive Science Society* (pp. 514–519). Hillsdale, NJ: Erlbaum.

Kruschke, J. K., & Johansen, M. K. (1999). A model of probabilistic category learning. *Journal of Experimental Psychology: Learning, Memory, and Cognition*, *25*, 1083–1119.

Kruskal, J. B., & Wish, M. (1978). *Multidimensional scaling*. Beverly Hills, CA: Sage. (Sage University Paper series on Quantitative Applications in the Social Sciences, 07–011)

Lachnit, H., Reinhard, G., & Kimmel, H. D. (2000). Further investigations of stimulus coding in nonlinear discrimination problems. *Biological Psychology*, *55*, 57–73.

Lagnado, D. A., & Shanks, D. R. (2002). Probability judgment in hierarchical learning: a conflict between predictiveness and coherence. *Cognition*, *83*, 81–112.

Lamberts, K. (1998). The time course of categorization. *Journal of Experimental Psychology: Learning*, *Memory*, *and Cognition*, *24*, 695–711.

Lamberts, K. (2000a). Concepts: core readings. *American Journal of Psychology*, *113*, 663–667.

Lamberts, K. (2000b). Information-accumulation theory of speeded categorization. *Psychological Review*, *107*, 227–260.

Lamberts, K. (2001). Category-specific deficits and exemplar models. *Behavioral and Brain Sciences*, *24*, 484–485.

Lamberts, K., & Shanks, D. (Eds.) (1997). *Knowledge*, *concepts and categories*. Cambridge, MA: MIT Press.

Lassaline, M. E., & Murphy, G. L. (1998). Alignment and category learning. *Journal of Experimental Psychology: Learning*, *Memory*, *and Cognition*, *24*, 144–160.

Lee, M. D., & Navarro, D. J. (2002). Extending the ALCOVE model of category learning to featural stimulus domains. *Psychonomic Bulletin and Review*, *9*, 43–58.

Levine, M. (1975). *A cognitive theory of learning: research on hypothesis testing*. Hillsdale, NJ: Erlbaum.

Lewandowsky, S. (1995). Base-rate neglect in ALCOVE – a critical reevaluation. *Psychological Review*, *102*, 185–191.

Lynch, E. B., Coley, J. D., & Medin, D. L. (2000). Tall is typical: central tendency, ideal dimensions, and graded category structure among tree experts and novices. *Memory and Cognition*, *28*, 41–50.

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.

Maddox, W. T. (2002). Learning and attention in multidimensional identification and categorization: separating low-level perceptual processes and high-level decisional processes. *Journal of Experimental Psychology: Learning*, *Memory*, *and Cognition*, *28*, 99–115.

Markman, A. B., & Stilwell, C. H. (2001). Role-governed categories. *Journal of Experimental and Theoretical Artificial Intelligence*, *13*, 329–358.

McCloskey, M., & Cohen, N. J. (1989). Catastrophic interference in connectionist networks: the sequential learning problem. In G. Bower (ed.), *The psychology of learning and motivation, vol. 24* (pp. 109–165). New York: Academic Press.

Medin, D. L., & Edelson, S. M. (1988). Problem structure and the use of base-rate information from experience. *Journal of Experimental Psychology: General*, *117*, 68–85.

Medin, D. L., Goldstone, R. L., & Gentner, D. (1993). Respects for similarity. *Psychological Review*, *100*, 254–278.

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

Medin, D. L., & Shoben, E. J. (1988). Context and structure in conceptual combination. *Cognitive Psychology*, *20*, 158–190.

Mellers, B. A. (1986). Test of a distributional theory of intuitive numerical prediction. *Organizational Behavior and Human Decision Processes*, *38*, 279–294.

Miller, C. S., & Laird, J. E. (1996). Accounting for graded performance within a discrete search framework. *Cognitive Science*, *20*, 499–537.

Minda, J. P., & Smith, J. D. (2001). Prototypes in category learning: the effects of category size, category structure, and stimulus complexity. *Journal of Experimental Psychology: Learning, Memory, and Cognition*, *27*, 775–799.

Mirman, D., & Spivey, M. (2001). Retroactive interference in neural networks and in humans: the effect of pattern-based learning. *Connection Science*, *13*, 257–275.

Murphy, G. L., & Medin, D. L. (1985). The role of theories in conceptual coherence. *Psychological Review*, *92*, 289–316.

Niedenthal, P. M., Halberstadt, J. B., & Innes-Ker, A. H. (1999). Emotional response categorization. *Psychological Review*, *106*, 337–361.

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

Nosofsky, R. M. (1991). Relation between the rational model and the context model of categorization. *Psychological Science*, *2*, 416–421.

Nosofsky, R. M. (1992). Exemplars, prototypes, and similarity rules. In A. F. Healy, S. M. Kosslyn, & R. M. Shiffrin (Eds.), *Essays in honor of William K. Estes, vol. 2. From learning processes to cognitive processes* (pp. 149–167). Hillsdale, NJ: Erlbaum.

Nosofsky, R. M., Gluck, M. A., Palmeri, T. J., McKinley, S. C., & Glauthier, P. (1994). Comparing models of rule-based classification learning: a replication of

Shepard, Hovland, and Jenkins (1961). *Memory and Cognition*, *22*, 352–369.

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

Nosofsky, R. M., & Kruschke, J. K. (1992). Investigations of an exemplar-based connectionist model of category learning. In D. L. Medin (ed.), *The psychology of learning and motivation* (Vol. 28, pp. 207–250). San Diego: Academic Press.

Nosofsky, R. M., & Kruschke, J. K. (2002). Single-system models and interference in category learning. *Psychonomic Bulletin and Review*, *9*, 169–174.

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

Nosofsky, R. M., & Palmeri, T. J. (1997). An exemplar-based random walk model of speeded classification. *Psychological Review*, *104*, 266–300.

Nosofsky, R. M., & Palmeri, T. J. (1998). A rule-plus-exception model for classifying objects in continuous-dimension spaces. *Psychonomic Bulletin and Review*, *5*, 345–369.

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

Oswald, C. J., Yee, B. K., Rawlins, J. N., Bannerman, D. B., Good, M., & Honey, R. C. (2001). Involvement of the entorhinal cortex in a process of attentional modulation: evidence from a novel variant of an ids/eds procedure. *Behavioral Neuroscience*, *115*, 841–849.

Owen, A. M., Roberts, A. C., Hodges, J. R., Summers, B. A., Polkey, C. E., & Robbins, T. W. (1993). Contrasting mechanisms of impaired attentional set-shifting in patients with frontal lobe damage or Parkinson's disease. *Brain*, *116*, 1159–1175.

Palmeri, T. J. (1999). Learning categories at different hierarchical levels: a comparison of category learning models. *Psychonomic Bulletin and Review*, *6*, 495–503.

Palmeri, T. J., & Nosofsky, R. M. (2001). Central tendencies, extreme points, and prototype enhancement effects in ill-defined perceptual categorization. *Quarterly Journal of Experimental Psychology*, *54A*, 197–235.

Pearce, J. M. (1994). Similarity and discrimination – a selective review and a connectionist model. *Psychological Review*, *101*, 587–607.

Pearce, J. M., & Bouton, M. E. (2001). Theories of associative learning in animals. *Annual Review of Psychology*, 111–139.

Raijmakers, M. E., Dolan, C. V., & Molenaar, P. C. (2001). Finite mixture distribution models of simple discrimination learning. *Memory and Cognition*, *29*, 659–677.

Ratcliff, R. (1990). Connectionist models of recognition memory: constraints imposed by learning and forgetting functions. *Psychological Review*, *97*, 285–308.

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

Rescorla, R. A., & Wagner, A. R. (1972). A theory of Pavlovian conditioning: variations in the effectiveness of reinforcement and non-reinforcement. In A. H. Black & W. F. Prokasy (Eds.), *Classical conditioning: Vol II . Current research and theory* (pp. 64–99). New York: Appleton-Century-Crofts.

Rogers, R. D., Andrews, T. C., Grasby, P. M., Brooks, D. J., & Robbins, T. W. (2000). Contrasting cortical and subcortical activations produced by attentional-set shifting and reversal learning in humans. *Journal of Cognitive Neuroscience*, *12*, 142–162.

Rosch, E., & Lloyd, B. B. (Eds.) (1978). *Cognition and categorization*. Hillsdale, NJ: Erlbaum.

Rosch, E. H., & Mervis, C. B. (1975). Family resemblances: studies in the internal structure of categories. *Cognitive Psychology*, *7*, 573–605.

Rosseel, Y. (2002). Mixture models of categorization. *Journal of Mathematical Psychology*, *46*, 178–210.

Rumelhart, D. E., Hinton, G. E., & Williams, R. J. (1986). Learning internal representations by error propagation. In J. L. McClelland & D. E. Rumelhart (Eds.), *Parallel distributed processing* (Vol. 1, pp. 318–362). Cambridge, MA: MIT Press.

Shanks, D. R. (1995). *The psychology of associative learning*. Cambridge: Cambridge University Press.

Shanks, D. R., Charles, D., Darby, R. J., & Azmi, A. (1998). Configural processes in human associative learning. *Journal of Experimental Psychology: Learning, Memory, and Cognition*, *24*, 1353–1378.

Shanks, D. R., & Darby, R. J. (1998). Feature- and rule-based generalization in human associative learning. *Journal of Experimental Psychology: Animal Behavior Processes*, *24*, 405–415.

Shanks, D. R., Darby, R. J., & Charles, D. (1998). Resistance to interference in human associative learning: evidence of configural processing. *Journal of Experimental Psychology: Animal Behavior Processes*, *24*, 136–150.

Shanks, D. R., & St John, M. F. (1994). Characteristics of dissociable human learning-systems. *Behavioral and Brain Sciences*, *17*, 367–395.

Shepard, R. N. (1962). The analysis of proximities: multidimensional scaling with an unknown distance function, I and II. *Psychometrika*, *27*, 125–140, 219–246.

Shepard, R. N. (1987). Toward a universal law of generalization for psychological science. *Science*, *237*, 1317–1323.

Shepard, R. N., Hovland, C. L., & Jenkins, H. M. (1961). Learning and memorization of classifications. *Psychological Monographs*, *75* (13) (Whole No. 517).

Sieck, W. R., & Yates, J. F. (2001). Overconfidence effects in category learning: a comparison of connectionist and exemplar memory models. *Journal of Experimental Psychology: Learning, Memory, and Cognition*, *27*, 1003–1021.

Sloman, S. A. (1996). The empirical case for two systems of reasoning. *Psychological Bulletin*, *119*, 3–22.

Sloman, S. A. (1998). Categorical inference is not a tree: the myth of inheritance hierarchies. *Cognitive Psychology*, *35*, 1–33.

Smith, E. E., & Medin, D. L. (1981). *Categories and concepts*. Cambridge, MA: Harvard University Press.

Smith, E. E., Patalano, A. L., & Jonides, J. (1998). Alternative strategies of categorization. *Cognition*, *65*, 167–196.

Smith, E. E., & Sloman, S. A. (1994). Similarity- versus rule-based categorization. *Memory and Cognition*, *22*, 377–386.

Smith, J. D., & Minda, J. P. (1998). Prototypes in the mist: The early epochs of category learning. *Journal of Experimental Psychology: Learning, Memory, and Cognition*, *24*, 1411–1436.

Smith, J. D., & Minda, J. P. (2000). Thirty categorization results in search of a model. *Journal of Experimental Psychology: Learning, Memory, and Cognition*, *26*, 3–27.

Smith, L. B., Jones, S. S., Landau, B., Gershkoff-Stowe, L., & Samuelson, L. (2002). Object name learning provides on-the-job training for attention. *Psychological Science*, *13*, 13–19.

Spalding, T. L., & Ross, B. H. (1994). Comparison-based learning – effects of comparing instances during category learning. *Journal of Experimental Psychology: Learning, Memory, and Cognition*, *20*, 1251–1263.

Storms, G., DeBoeck, P., & Ruts, W. (2001). Categorization of novel stimuli in well-known natural concepts: a case study. *Psychonomic Bulletin and Review*, *8*, 377–384.

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

Treat, T. A., McFall, R. M., Viken, R. J., & Kruschke, J. K. (2001). Using cognitive science methods to assess the role of social information processing in sexually coercive behavior. *Psychological Assessment*, *13*, 549–565.

Tversky, A. (1977). Features of similarity. *Psychological Review*, *84*, 327–352.

Vandierendonck, A. (1995). A parallel rule activation and rule synthesis model for generalization in category learning. *Psychonomic Bulletin and Review*, *2*, 442–459.

Vandierendonck, A., & Rosseel, Y. (2000). Interaction of knowledge-driven and data-driven processing in category learning. *European Journal of Cognitive Psychology*, *12*, 37–63.

VanOsselaer, S. M. J., & Janiszewski, C. (2001). Two ways of learning brand associations. *Journal of Consumer Research*, *28*, 202–223.

Waldron, E. M., & Ashby, F. G. (2001). The effects of concurrent task interference on category learning: evidence for multiple category learning systems. *Psychonomic Bulletin and Review*, *8*, 168–176.

Wattenmaker, W. D. (1995). Knowledge structures and linear separability – integrating information in object and social categorization. *Cognitive Psychology*, *28*, 274–328.

Welk, D. S. (2002). Designing clinical examples to promote pattern recognition: nursing education-based research and practical applications. *Journal of Nursing Education*, *41*, 53–60.

Williams, D. A., & Braker, D. S. (1999). Influence of past experience on the coding of compound stimuli. *Journal of Experimental Psychology*: *Animal Behavior Processes*, *25*, 461–474.

Wills, A. J., Reimers, S., Stewart, N., Suret, M., & McLaren, I. P. (2000). Tests of the ratio rule in categorization. *Quarterly Journal of Experimental Psychology*, *53A*, 983–1011.

Wittgenstein, L. (1953). *Philisophical investigations*. New York: Macmillan.

Yamauchi, T., & Markman, A. B. (2000a). Inference using categories. *Journal of Experimental Psychology: Learning*, *Memory*, *and Cognition*, *26*, 776–795.

Yamauchi, T., & Markman, A. B. (2000b). Learning categories composed of varying instances: the effect of classification, inference, and structural alignment. *Memory and Cognition*, *28*, 64–78.

Young, M. E., Wasserman, E. A., Johnson, J. L., & Jones, F. L. (2000). Positive and negative patterning in human causal learning. *Quarterly Journal of Experimental Psychology*, *53B*, 121–138.