ACADEMIC PRESS

Available online at www.sciencedirect.com

SCIENCE DIRECT®

Cognitive Psychology 45 (2002) 482–553

Cognitive Psychology

www.elsevier.com/locate/cogpsych

Are there representational shifts during category learning?

Mark K. Johansen[a,1] and Thomas J. Palmeri[b,*,1]

[a] *Department of Psychology, Indiana University, 1101 E. 10th Street, Bloomington, IN 47405, USA*

[b] *Department of Psychology, Vanderbilt University, 301 Wilson Hall, 111 21st Avenue South, Nashville, TN 37203, USA*

Accepted 14 March 2002

Abstract

Early theories of categorization assumed that either rules, or prototypes, or exemplars were exclusively used to mentally represent categories of objects. More recently, hybrid theories of categorization have been proposed that variously combine these different forms of category representation. Our research addressed the question of whether there are representational shifts during category learning. We report a series of experiments that tracked how individual subjects generalized their acquired category knowledge to classifying new critical transfer items as a function of learning. Individual differences were observed in the generalization patterns exhibited by subjects, and those generalizations changed systematically with experience. Early in learning, subjects generalized on the basis of single diagnostic dimensions, consistent with the use of simple categorization rules. Later in learning, subjects generalized in a manner consistent with the use of similarity-based exemplar retrieval, attending to multiple stimulus dimensions. Theoretical modeling was used to formally corroborate these empirical observations by comparing fits of rule, prototype, and exemplar models to the observed categorization data. Although we provide strong evidence for shifts in the kind of information used to classify objects as a function of categorization experience, interpreting these results in terms of shifts in representational systems underlying perceptual categorization is a far thornier issue. We provide a discussion of the

* Corresponding author. Fax: 1-615-343-8449.

E-mail address: thomas.j.palmeri@vanderbilt.edu (T.J. Palmeri).

[1] Both authors contributed equally to the work presented in this paper, and the writing duties were divided between the authors, so the order of authorship was decided arbitrarily.

PII: S0010-0285(02)00505-4

challenges of making claims about category representation, making reference to a wide body of literature suggesting different kinds of representational systems in perceptual categorization and related domains of human cognition.

1. Introduction

An enduring debate in cognitive science is whether key aspects of human cognition are rule-based or similarity-based. Intuitively, some decisions seem to require deliberate, analytic reasoning by applying abstract rules, whereas other decisions seem to spring to mind automatically based on similarity to prior experiences (Sloman, 1996). These qualitatively different modes of cognition have been studied in such varied domains as language processing (e.g., Pinker, 1999), reasoning (e.g., Sloman, 1996; Smith, Langston, & Nisbett, 1992), skill acquisition (e.g., Anderson, Fincham, & Douglass, 1997; Logan, 1988), problem solving (e.g., Medin & Ross, 1989; Ross, 1987), categorization (e.g., Brooks, 1978; Medin & Smith, 1981; Shanks & St. Johns, 1994), and other aspects of human cognition. The present paper examines the use of rules and similarity to examples in perceptual categorization with a focus on how experience might modulate the use of these different types of category knowledge.

Early theories assumed that people represent categories by forming abstract logical rules, and research focused on what kinds of rules people found more or less difficult to learn (e.g., Bourne, 1970; Bruner, Goodnow, & Austin, 1956; Hunt, Marin, & Stone, 1966). Subsequent research instead assumed that people formed abstract category representations based on prototypes, statistical central tendencies of experienced category exemplars (e.g., Homa, 1978; Posner & Keele, 1968; Reed, 1972). Later developments showed that theories that assume similarity to stored category exemplars could account for many phenomena allegedly demonstrating formation of abstract rules or prototypes (e.g., Busemeyer, Dewey, & Medin, 1984; Choi, McDaniel, & Busemeyer, 1993; Hintzman, 1986; Nosofsky, 1986; Shin & Nosofsky, 1992). A large body of subsequent research demonstrated the theoretical success of exemplar-based models in accounting for a wide range of categorization phenomena (e.g., Estes, 1994; Kruschke, 1992; Lamberts, 1995; Nosofsky, 1988; Nosofsky & Palmeri, 1997).

More recently, however, investigators have begun to reexamine the potential role of more abstract forms of representation, such as rules or prototypes, in category learning. Various hybrid theories have been proposed recently that involve mixtures of rules and exemplars (e.g., Anderson & Betz, 2001; Erickson & Kruschke, 1998; Johnstone & Shanks, 2001; Nosofsky, Palmeri, & McKinley, 1994; Palmeri, 1997; Smith, Patalano, & Jonides, 1998; Thomas, 1998), prototypes and exemplars (e.g., Anderson, 1990; Love, Medin, & Gureckis, in press; Smith & Minda, 1998), and various kinds of linear and nonlinear decision boundaries (e.g., Ashby, Alfonso-Reese, Turken, & Waldron, 1998). The research presented in this paper examines the questions of whether multiple kinds of category representation are used and whether categorization experience modulates the use of these different representations.

Let us intuitively motivate the difference between rule-based and similarity-based categorization with this example. We can imagine that a novice searching the woods for prized Chanterelle mushrooms must refer to a set of fairly complex rules for telling them apart from many similar, yet quite poisonous species, such as the Jack O' Lantern mushroom (Phillips, 1991). Although these rules may become internalized, without requiring reference to a field guide, categorizing mushrooms as edible versus poisonous may still appear to involve deliberate use of explicit rule-based knowledge. With experience, however, a mushroom gatherer eventually seems to shift from this potentially slow, deliberate, attention-demanding mode of categorizing to a far more rapid and automatic mode of processing that seems to characterize more expert-like performance. What kinds of changes occur that allow someone to become a skilled mushroom gatherer who can recognize the prized Chanterelle so quickly and effortlessly, with little thought or conscious awareness, and without needing to make recourse to explicit rules?

A way of characterizing these shifts is to view categorization as another domain in which people develop skills with experience. According to Logan's (1988) instance theory, automaticity in a range of cognitive skills is attributed to a shift from strategic and algorithmic processes, such as the use of explicit rules, to the retrieval of exemplars from memory. Could such shifts characterize the development of automaticity in categorization whereby people initially use simple rules but eventually come to rely on similarity-based retrieval of exemplars? Palmeri (1997) found evidence for shifts from rules to exemplars in a paradigm in which subjects were supplied an explicit counting rule for initially classifying objects into different categories. In a different paradigm, Brooks and colleagues (Allen & Brooks, 1991; Regehr & Brooks, 1993) found evidence for intrusions of similarity-based retrieval even when subjects were supplied an explicit categorization rule. But in many experimental paradigms and in many real world situations, people are not supplied categorization rules prior to learning about categories of objects. Do people adopt an analytic strategy of developing simple rules at the outset of category learning? If so, do these rules eventually give way to exemplar retrieval, or other forms of category representation, with increased experience? Although a number of current theories have posited combinations of rules and exemplars, little research has investigated how the use of these different kinds of category knowledge might change with experience. Indeed, an alternative view has recently been proposed that people initially abstract prototypes when learning about categories but may eventually rely on memory for specific examples with no generalization to those stored exemplars (Smith & Minda, 1998).

In this paper, we have followed other investigators (e.g., Erickson & Kruschke, 1998; Kalish & Kruschke, 1997; Nosofsky & Palmeri, 1998) by limiting our use of the term *rule* to refer to single-dimension categorization rules. In addition to being the same operational definition used by others, such unidimensional rules can be verbalized, another operationalization of rule-based categorization (Ashby et al., 1998). Obviously, we would not want to preclude learning more complex conjunctive, disjunctive, or other logical categorization rules, for people can clearly learn such rules (e.g., Salatas & Bourne, 1974) and can use complex rules when explicitly provided (e.g., Allen & Brooks, 1991; Nosofsky, Clark, & Shin, 1989; Smith et al., 1998). But it is clear that subjects do not find such rules as easy to learn as unidimensional rules (e.g., Bourne, 1970; Feldman, 2000; Nosofsky et al., 1994). Also, during unsupervised categorization, where subjects are free to form their own categories without any corrective feedback, unidimensional rules clearly dominate categorization performance (e.g., Ahn & Medin, 1992; Ashby, Queller, & Berretty, 1999; Regehr & Brooks, 1995). Importantly, all of the category structures that were used in the experiments reported in this paper permitted the formation of imperfect single-dimension rules. We will return to a more general discussion of rule-based categorization at the end of this paper.

This paper begins by reviewing previous empirical and theoretical evidence suggesting that people may learn perceptual categories by forming simple rules. A closer examination of this prior research will reveal evidence for exemplar-based categorization as well, with the relative proportion of presumed rule-based and exemplar-based categorization modulated by experience. To directly test for possible shifts from rules to exemplars, we then report three experiments tracking how people generalized their acquired category knowledge as a function of learning. The results show that early in learning, subjects generalized on the basis of single diagnostic dimensions, consistent with the use of simple categorization rules. Later in learning, subjects generalized in a manner consistent with exemplar-based categorization, utilizing multiple dimensions to assess similarity to stored exemplars. Theoretical modeling is then used to formally corroborate these empirical observations by comparing the fits of rule, prototype, and exemplar models to the observed categorization data. The combination of empirical and theoretical results provides strong evidence for shifts in the kind of information that is used to classify stimuli as a function of categorization experience. However, interpreting these results in terms of shifts in the kind of representational system underlying perceptual categorization is far more difficult. In the final section of this paper, we discuss the challenges of making claims about representation on the basis of empirical data and theoretical modeling and review various aspects of the categorization literature and related domains suggesting multiple representational systems.

2. Background and overview

Perceptual categorization has been studied using a variety of experimental paradigms. This paper focuses on a widely used paradigm in which subjects learn to classify stimuli composed of binary-valued dimensions with corrective feedback but without being provided any explicit rules beforehand. Other categorization studies have provided explicit rules at the outset of learning (e.g., Allen & Brooks, 1991; Nosofsky et al., 1989). Some have used stimuli that vary along continuous rather than discrete dimensions (e.g., Ashby & Gott, 1988; Homa, Sterling, & Trepel, 1981). Others have examined unsupervised category learning without corrective feedback (e.g., Ahn & Medin, 1992; Ashby et al., 1999), implicit learning without knowledge that categories are even being acquired (e.g., Berry & Dienes, 1993; Johnstone & Shanks, 2001), and other learning modes such as category learning by feature inference (e.g., Yamauchi & Markman, 1998). Although there have been many paradigms used to investigate categorization, we have chosen a paradigm that has one of the longest histories in categorization research and that has been a dominant paradigm for developing and testing formal models of categorization.

The category structure we used in our first experiment is shown in Table 1 and will be used here to introduce some key points we will be addressing in this paper. This is the structure first used in the classic studies by Medin and Schaffer (1978) and has been used in many studies since then (e.g., Lamberts, 1995; Medin & Smith, 1981; Nosofsky, 2000; Nosofsky et al., 1994; Palmeri & Nosofsky, 1995; Smith & Minda, 2000). Each stimulus is composed of four dimensions and each dimension can take on one of two possible values.

Table 1
Category structure used by Nosofsky et al. (1994), Palmeri and Nosofsky (1995), and in our Experiment 1 (from Experiments 2 and 3 of Medin & Schaffer, 1978). Each stimulus was composed of four dimensions. Each dimension could take on one of two possible values

Category A		Category B		Transfer	
A1	1 1 1 2	B1	1 1 2 2	*T1	1 2 2 1
A2	1 2 1 2	B2	2 1 1 2	*T2	1 2 2 2
A3	1 2 1 1	B3	2 2 2 1	T3	1 1 1 1
A4	1 1 2 1	B4	2 2 2 2	*T4	2 2 1 2
A5	2 1 1 1			*T5	2 1 2 1
				*T6	2 2 1 1
				T7	2 1 2 2

Note. * Highlights the five critical transfer items.

In the original Medin and Schaffer experiment, the dimensions were size, color, form, and position. So, for example, the stimulus 1 1 1 2 might correspond to a large red square to the right, whereas the stimulus 1 2 1 1 might correspond to a large blue square to the left. Five items are assigned to category A, four are assigned to category B, and the remaining seven are designated transfer items. The categories are ill-defined in that no single feature along a dimension can be used to perfectly classify the items. Rather, the categories have a family resemblance structure in that category A items tend to have value 1 along each dimension, and category B items tend to have value 2 along each dimension.

In this paradigm, subjects are presented training items one at a time, randomly drawn from category A or category B, they categorize the presented item as a member of one of the two categories, and they receive corrective feedback. At the start of training, subjects have no idea how to correctly categorize the training items and must resort to guessing, but they are told that by paying attention to the corrective feedback provided by the computer they should be able to eventually learn the categories. Our key empirical measure was how subjects applied their acquired category knowledge to classifying the transfer items. In this paradigm, subjects are typically presented transfer items along with training items to be categorized without corrective feedback after some fixed number of training blocks. Because we were interested in how category knowledge develops with training, we instead presented test blocks of transfer and training items at various stages throughout category learning.

In the original Medin and Schaffer (1978) studies, prototype (additive independent cue) and exemplar (multiplicative interactive cue) models were compared on their ability to account for the average probability that each stimulus was classified as a member of category A or B. According to a prototype model, subjects form prototypes 1 1 1 1 for category A and 2 2 2 2 for category B with items classified according to their relative similarity to the prototypes. According to an exemplar model, subjects store the specific training instances of the categories with items classified according to their relative summed similarity to the exemplars of the two categories. Medin and Schaffer reported better qualitative and quantitative accounts of the observed data by an exemplar model (the context model) than a prototype model, a finding that was instrumental in launching the (arguable) theoretical dominance of exemplar models of categorization over the last two decades (see Nosofsky, 2000; Smith & Minda, 2000, for a recent debate regarding the theoretical implications of these classic results).

Most investigations of the Medin and Schaffer paradigm have focused on average probabilities of classifying items as members of the two categories. Although averaging performance across numerous subjects may have statistical appeal, there have been arguments that such averaging obscures important individual differences in the kinds of strategies subjects employ to categorize stimuli. Indeed, questions of averaging may have important theoretical consequences in that some models of categorization may be able to provide superior accounts of the average subject data but may show systematic weaknesses in accounting for individual subject data (e.g., Ashby, Maddox, & Lee, 1994; Maddox, 1999; Martin & Caramazza, 1980; Nosofsky et al., 1994; Palmeri & Nosofsky, 1995; Smith & Minda, 1998).

As a particularly relevant example, Nosofsky et al. (1994) reported that both an exemplar model and a rule-plus-exception model provided comparable accounts of average data in a replication and extension of the classic Medin and Schaffer (1978) study. However, when they systematically examined categorization at the individual subject level, they obtained results suggesting the use of simple rules, not the storage and retrieval of exemplars, as the basis for learning these categories. Indeed, an exemplar model provided an exceedingly poor account of the individual subject data, even though the model had provided an excellent account of the average data. By contrast, the rule-plus-exception model they formulated provided a very good account of both the individual subject data and the average data.

There are two primary approaches to examining individual differences in categorization. One approach is to test a small number of subjects over many sessions, collecting thousands of observations per individual (e.g., Ashby & Gott, 1988; Nosofsky, 1986; Nosofsky & Palmeri, 1997). Models are tested on how well they can account for detailed aspects of the responses made by each individual subject. A quite different approach is to collect a small amount of data from a large number of individuals tested within a single session and then to try to characterize whether clusters of different response profiles may be present across those many individuals (e.g., Nosofsky et al., 1989; Nosofsky & Palmeri, 1998; Nosofsky et al., 1994). Models are tested on how well they account for the variability in responses across those numerous individuals. (The final approach of testing many individuals for many sessions may be both logistically and financially prohibitive.) Testing many subjects a limited number of times each is especially beneficial when the focus is on how subjects generalize their acquired category knowledge to classifying new transfer items. When subjects are tested on new stimuli multiple times, those stimuli may influence category judgments in unexpected ways because the transfer stimuli may become part of the learned category representations (e.g., Nosofsky, 1986). In other words, new transfer stimuli may no longer be "new" if subjects are required to categorize those stimuli too many times.

In the present experiments, we systematically examined performance by large numbers of individual subjects using an empirical measure called a *distribution of generalization patterns* (Nosofsky et al., 1989; Nosofsky et al., 1994; Pavel, Gluck, & Henkle, 1988). A *generalization pattern* for an individual subject is defined by how that subject classified each new transfer item. For example, for the category structure shown in Table 1, a subject who classified T1-T3 as members of category A and T4-T7 as members of category B would be said to have exhibited a generalization pattern AAABBBB, whereas a subject who classified T3, T4, and T6 as members of category A and the remaining transfer items as members of category B would have exhibited pattern BBAABAB. The distribution of generalization patterns is a tally of the number of subjects who displayed each of the possible generalization patterns. With seven transfer items, each of which could be classified into one of two categories, there are $2^7 = 128$ possible generalization patterns that subjects could exhibit.

As we will see, examining such distributions of generalization patterns can be particularly informative as to what information subjects are using to categorize items. Before discussing how these distributions can be used to understand categorization, we must briefly digress to discuss how the distributions will be visually displayed. Displaying the full distribution of generalization patterns for the Medin and Schaffer category structure would require generating a bar graph with observed proportions for all of the 128 possible generalization patterns. Clearly, a graph with that many entries would be quite difficult to read and interpret. Therefore, for illustrative purposes, throughout this paper we decreased the number of generalization patterns that needed to be displayed in a figure to just those generalizations involving the most critical transfer items. To do this, we collapsed across responses to what reasonably can be considered noncritical transfer items. These noncritical items were those which most subjects should and did consistently classify into the same category. For example, T3 (1 1 1 1) in Table 1 is the prototype of category A; whether subjects have acquired rules, prototypes, or exemplars, this item should be classified into category A with high probability. Similarly, T7 (2 1 2 2) is very similar to the category B prototype, differing along the least diagnostic dimension; regardless of what kind of information subjects have acquired, this item should be classified as a member of category B with high probability. Categorization judgments for the remaining five items (T1, T2, T4, T5, and T6) are critical transfer items in that individual subjects do categorize these items differently and in that different theoretical models make different predictions as to how these items should be classified. Thus, for example, the full generalization patterns AAABBBB, AABBBBB, AABBBBA, and AAABBBA were combined into a single generalization pattern AABBB, collapsing across responses to the noncritical items T3 and T7. For this category structure, focusing on just the five critical transfer items decreases the number of generalization patterns that need to be displayed in a figure down to $2^5 = 32$ patterns. That said, in all of the quantitative tests of formal models reported later in this paper, the models were fitted to the full distributions of generalization patterns, not the abbreviated distributions displayed in the figures.

Now, how can these distributions of generalization patterns be used to distinguish between models of categorization? Nosofsky et al. (1994) suggested that subjects learn categories, such as those of Medin and Schaffer (1978), by forming simple rules and perhaps remembering exceptions to those rules, rather than storing and retrieving detailed exemplar information. To test this idea, they formalized a rule-plus-exception (RULEX) model of categorization. One of the key assumptions of RULEX is that different subjects may learn different rules and probabilistically store exceptions to those rules to varying degrees. Averaging across the idiosyncratic behaviors of different subjects resulted in predictions at the level of average data that were indistinguishable from those of an exemplar model. However, RULEX does predict systematic differences in categorization behavior at the individual-subject level. To examine individual differences in categorization, Nosofsky et al. (1994) replicated the Medin and Schaffer experiment but tested over 200 subjects and examined the distribution of generalization patterns.

Turning to the structure shown in Table 1, although no perfect single-dimension rule can be used to classify the training items, several imperfect single-dimension rules can be formed that work quite well. For example, a subject who formed a rule that value 1 along dimension one signals category A and value 2 along dimension one signals category B would need to learn the exceptions A5 and B1. A subject who instead formed a rule that value 1 along dimension three signals category A and value 2 along dimension three signals category B would need to learn the exceptions A4 and B2. Although numerous other single-dimension or conjunctive rules are also possible, these other rules require memorizing more exceptions. Rules along dimensions one and three were predicted to be formed more frequently because they are far more efficient rules. The use of such simple single-dimension categorization rules can be discovered by examining distribution of generalization patterns. Specifically, application of rules along dimension one or dimension three would result in those subjects exhibiting generalization patterns AABBB and BBABA, respectively (using the truncated notation described above).

The left panel of Fig. 1 displays the distribution of generalization patterns observed by Nosofsky et al. (1994). The two most prominent generalization patterns, AABBB and BBABA, are those consistent with rules along dimensions one and three, respectively. RULEX accounted for 99.0% of the variance in the average transfer data and 85.7% of the variance in the distributions of generalization patterns. By contrast, the context model only accounted for 35.9% of the variance in the distribution of generalization patterns. Palmeri and Nosofsky (1995) conducted another replication and extension of the Medin and Schaffer (1978) experiment in order to additionally test predictions made by RULEX and the context model regarding recognition memory and speeded categorization. The right panel of Fig. 1 displays the distribution of generalization patterns from that later study, again showing the prominent rule-based generalization patterns,

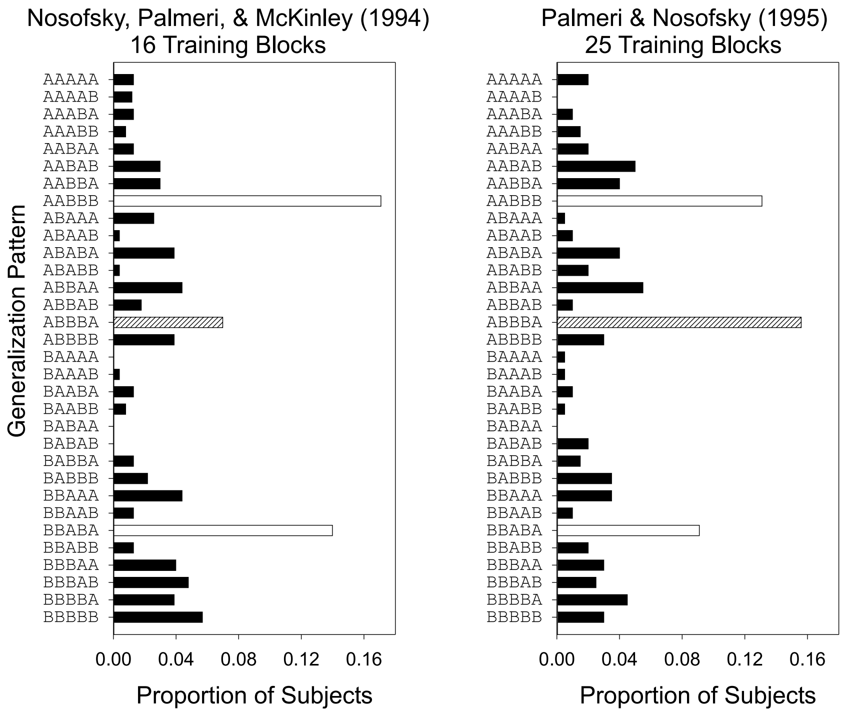

Fig. 1. The left panel displays the distributions of generalization patterns from Nosofsky et al. (1994), in which subjects were given 16 training blocks. The right panel displays the distributions of generalization patterns from Palmeri and Nosofsky (1995), in which subjects were given 25 training blocks. Each of the 32 generalization patterns is one of the possible ways that a subject could classify the five critical transfer items (T1, T2, T4, T5, and T6, respectively) from the category structure shown in Table 1. For example, pattern AABBB denotes classifying T1–T2 as members of category A and T4–T6 as members of category B. White bars highlight the two prominent rule-based generalization patterns. The hatched bar highlights the prominent exemplar-based generalization pattern.

AABBB and BBABA. The predominance of these rule-based patterns appears inconsistent with the exemplar-based categorization process that had previously been thought to underlie the learning of these categories (but see Nosofsky & Johansen, 2000).

However, as can be seen in both panels of Fig. 1, in both of these previous studies, another prominent generalization was also observed, ABBBA. This generalization is consistent with exemplar-based categorization in that models that assume generalization by similarity to stored exemplars predict ABBBA to be the most prominent generalization across a wide set of parameters. By contrast, this pattern is not predicted by RULEX to show any particular prominence in the predicted distribution of generalizations across any set of parameter values (in fact, other patterns, in addition to the two rule-based patterns, are predicted to be more prominent than this exemplar pattern, as shown by Nosofsky et al., 1994). Examining the left panel of Fig. 1, we see that this exemplar-based generalization was quite prominent in the Nosofsky et al. (1994) experiment, in which subjects were provided 16 training blocks. Examining the right panel of Fig. 1, we see that ABBBA was the most prominent pattern in the Palmeri and Nosofsky (1995) experiment, where 25 training blocks were provided.

Relatively little attention was paid to this exemplar-based generalization in either of these previous studies, apart from noting its presence and suggesting that it could reflect use of exemplar retrieval. Nor was it suggested that the relative prominence of this generalization might be influenced by experience. By contrast, our initial study was motivated by what appears to be an increased prominence in apparent exemplar-based generalizations as a function of categorization experience across these two independent studies. Indeed, might this increased prominence in the ABBBA pattern with increased training reflect a shift from rules to exemplars as a function of experience, as suggested by some automaticity theories? The first experiment attempted to address that issue.

3. Experiment 1

In this first experiment, subjects were trained on the Medin and Schaffer (1978) category structure for a total of 32 training blocks. Unlike previous studies, we provided single transfer blocks at various points throughout category learning, testing subjects on all stimuli without feedback after 2, 4, 8, 16, 24, and 32 blocks of training.[2] The key empirical measure was how the distribution of generalization patterns evolved as a function of learning. Would there be evidence of rule use early in training, indexed by the relative prominence of the rule-based generalizations AABBB and BBABA? Would there be evidence for a gradual emergence of exemplar retrieval with more training, indexed by the relative prominence of the presumed exemplar-based generalization, ABBBA?

[2] We acknowledge that there are potential problems raised by testing subjects multiple times during training. As noted by one of the reviewers, exposure to transfer items could encourage rule formation because it makes subjects aware of items outside the training set, emphasizing the requirements to generalize beyond the training set. Alternatively, interleaving categorization tests within category training could hinder rule formation because it disrupts any hypothesis-testing strategies subjects may be using to learn the categories. Unfortunately, given the large number of subjects required to measure distributions of generalization patterns, it was simply unfeasible to conduct these experiments using between-subjects designs.

3.1. Method

3.1.1. Subjects

Subjects were 198 undergraduate students who received credit in an introductory psychology course. All subjects were tested individually.

3.1.2. Stimuli

Stimuli were computer-generated drawings of rockets that varied along four binary-valued dimensions: The shape of the wing (triangular or rectangular), tail (jagged or boxed), nose (staircase or half-circle), and porthole (circular or star). The rockets were adapted from those used by Hoffman and Ziessler (1983) and were the same as those used by Nosofsky et al. (1994) and Palmeri and Nosofsky (1995). As shown in Table 1, five stimuli belonged to category A, four belonged to category B, and seven were new transfer items. The assignment of physical dimensions to abstract dimensions and physical values to abstract values along dimensions was randomized for every subject. In all experiments, stimulus presentation and response recording were controlled by customized computer programs.

3.1.3. Procedure

In all experiments, subjects were trained to categorize stimuli as members of one of two categories with feedback. At varying points during training, they were tested with a single transfer block in which they categorized the sixteen training and transfer items without corrective feedback.

In this experiment, subjects received a total of 32 training blocks. Each of the nine training items was presented once per block. Stimuli were presented in random order within each block, subject to the constraint that the same stimulus was not shown on consecutive trials. Subjects classified each item as a member of category A or B and then received corrective feedback for 2 s. There was a 500 ms interval between trials. Subjects responded by pressing one of two response keys on a computer keyboard.

One transfer block was presented after 2, 4, 8, 16, 24, and 32 blocks of training (labeled TB2, TB4, TB8, TB16, TB24, and TB32 in tables and figures). In each block, all 16 stimuli from Table 1 were presented just once, in randomized order, without corrective feedback. After the subject classified an item as a member of category A or category B, the computer responded "OK" for one second and proceeded to the next item. There was a 500 ms blank interval between trials.

3.2. Results

Sixty-eight subjects with an error rate of more than 25% on the last four training blocks were removed from further analyses. One justification for this strict criterion is that we were interested in examining performance that approached asymptotic levels of accuracy by the end of training. We could not train our subjects for more than 32 blocks within a single experimental session. Providing more training would have required bringing subjects back for multiple sessions, which was logistically impossible given such large numbers of subjects. Excluding poor performers seemed a reasonable solution. That said, we did examine the data excluding no subjects and observed qualitatively similar results to those we report here. Moreover, 75.4% of excluded subjects did not perform significantly greater than chance when categorizing during the last four training blocks, providing further justification for our exclusion of these subjects. This exclusion criterion was used in all three experiments.

An α level of .05 was established for all statistical tests used in this paper.

3.2.1. Average transfer data

The average probabilities of categorizing each of the 16 training and transfer stimuli as a member of category A as a function of the six transfer blocks is shown in the bottom portion of Fig. 2. An important goal of our later theoretical modeling will be to attempt to account for the full spectrum of these average probabilities as well as the distributions of generalization patterns using various formal models of categorization. Because our empirical focus will be on the generalizations, we will just highlight a couple of important aspects of the transfer data. Throughout training, more errors were made on A5 (2 1 1 1) and B1 (1 1 2 2), and A4 (1 1 2 1) and B2 (2 1 1 2), which can be characterized as the exceptions to rules along dimensions 1 and 3, respectively; significantly more errors were made on the exceptions than the nonexceptions in each transfer block [$t(129) > 5.214$]. Not surprisingly, we also found clear evidence for learning as a function of training in the decrease in the number of errors made categorizing old training items; a within-subjects ANOVA on the proportion of categorization errors within each block revealed a significant effect of block [$F(5, 645) = 85.94$, $MSe = .019$] and a significant linear contrast on blocks [$F(1, 129) = 247.43$, $MSe = .031$].

3.2.2. Distributions of generalization patterns

We defined a generalization pattern for each subject as the pattern of responses that subject gave to each of the seven new transfer items. For purposes of fitting computational models to the distributions of generalization patterns, discussed later in this paper, we used the full empirical distributions of generalization patterns from all experiments (these can be obtained from the authors). As outlined earlier, to simplify the illustrations, we used a truncated generalization that included only the five critical transfer items T1, T2, T4, T5, and T6. The top portion of Fig. 2 displays the distributions of generalization patterns observed within each transfer block.

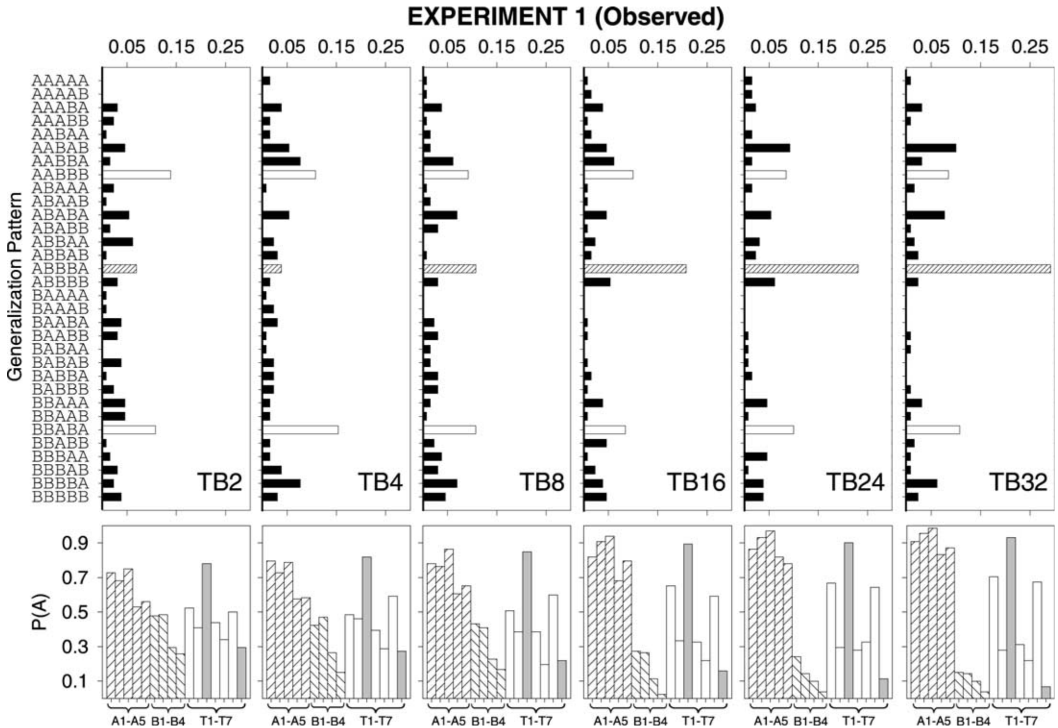

Fig. 2. The top row shows distributions of generalization patterns and the bottom row shows average categorization probabilities $P(\mathrm{A})$ for training and transfer items after 2 (TB2), 4 (TB4), 8 (TB8), 16 (TB16), 24 (TB24), and 32 (TB32) blocks of training in Experiment 1. For the distributions, white bars highlight the two prominent rule-based generalization patterns and the hatched bar highlights the prominent exemplar-based generalization pattern. For the average categorization probabilities, hatched bars are training items, white bars are critical transfer items, and gray bars are noncritical transfer items.

In this empirical analysis, we will focus on the presence of prominent generalizations in the distributions as a function of learning. In the theoretical section, we will test the ability of models to account for the entire set of observed distributions. First, as discussed earlier, we can define simple rule-based generalizations for the two most likely single-dimension rules. Application of a dimension-1 rule yields generalization AABBB, whereas application of a dimension-3 rule yields generalization BBABA. Examining Fig. 2, we see that these two highlighted rule-based generalizations were most prominent during the early stages of learning but became proportionately less prominent in later stages of learning (although their absolute proportions did not change considerably).

Indeed, with more training, pattern ABBBA emerged as the most prominent generalization. One way of characterizing this pattern is that it may serve as a marker for exemplar-based categorization in that this generalization is predicted by exemplar models to be the most prominent one in the distribution. To show this, we conducted a large set of simulations of two exemplar models, the context model (Medin & Schaffer, 1978) and ALCOVE (Kruschke, 1992). As with all formal models, these models have free parameters. In the context model, parameters are attention weights along the four dimensions used to compute similarities (Nosofsky, 1984). ALCOVE is a connectionist version of the context model that learns dimensional attention weights and learns association weights between exemplars and categories; parameters govern the rate of learning and map output activations onto response probabilities. Ideally, one would like to report parameter-free predictions. Because both the context model and ALCOVE have free parameters, we instead generated predicted distributions of generalization patterns across over 10,000 different sets of parameter values, as described in Appendix A. Summarizing the results of those simulations, both models were highly constrained in terms of which generalizations they could predict to be maximally prominent in the distributions. In fact, the models were relatively inflexible in that they never predicted certain generalizations to be maximally prominent. Indeed, for both models, pattern ABBBA emerged as the most prominent pattern across a wide range of possible parameter values under optimal conditions that produced high accuracy at categorizing training items (Nosofsky, 1998a,b). Both models do make a priori predictions about the distributions of generalization patterns; they are not just sophisticated curve fitting techniques that can fit any pattern of observed data. Given these simulations, we felt justified in characterizing ABBBA as a potential marker of exemplar-based generalization for this structure. Turning again to the data reported in Fig. 2, we can confirm one of our key predictions, that exemplar-based generalizations would gradually gain prominence with training. In the later theoretical analyses, we will corroborate this observation by demonstrating excellent fits of an exemplar model to the generalizations observed during the later stages of category learning.

Our next analysis provided a further summary of the proportion of rule- and exemplar-based generalizations as a function of experience. Because of the discrete nature of category responding, subjects may be using a simple rule or exemplar strategy but may classify one of the transfer items in the "wrong" category because responses are not entirely deterministic. Therefore, as a conservative measure, we tallied the proportion of dimension-1 generalizations as the proportion of AABBB generalizations and those that differed by one response (e.g., AABBA, AABAB, etc.), the proportion of dimension-3 generalizations as BBABA and those that differed by one response, and the proportion of exemplar generalizations as ABBBA and those that differed by one response. Note that some of the generalizations (e.g., BBBBA) may be counted in more than one group (analyses in which we ignored generalizations that counted in multiple groups were qualitatively similar). Because each subject categorized each transfer item just once during each block, we believed it was sensible to permit generalizations that differed by one response into our tally from our target generalization; also, we believed that our approach was a simpler and more theoretically neutral approach than tallying specific patterns as rule- or exemplar-based.[3] Fig. 3 displays the proportion of dimension-1, dimension-3, and exemplar generalizations as a function of learning. Again, the most prominent finding was that exemplar generalizations became more prominent as a function of training. However, in this experiment, rule generalizations did not diminish with training.

One final question emerging from an analyses of the distribution of generalization patterns is whether they really do reveal any information that could not be discovered by simply examining the average data. Specifically, is there information presented in the top row of Fig. 2 that is not already present in the bottom row of Fig. 2? Clearly, not every subject will display the same generalization pattern. But that is a necessary consequence of how the transfer items are classified on average. Suppose a particular item has a .5 probability of being classified into category A. If all subjects respond on the basis of this probability and if each subject provides only a single categorization response, then half of the subjects will classify that item as an A and half will classify it as a B. Extending this to two items, each classified with .5 probability, all of the possible generalizations from two items (AA, AB, BA, and BB) should be observed with equal likelihood. Examin-

[3] We used this scoring system in all three experiments. Although it may be theoretically neutral, it could possibly misrepresent the categorization strategies used by subjects. Our intent in presenting these analyses was to provide a simple alternative summary of the empirical results that are present within the evolving distributions of generalization patterns. The conclusions gleaned from the figures showing these simple analyses are corroborated by direct examination of the distributions of generalization patterns and by the theoretical modeling results presented later in this paper.

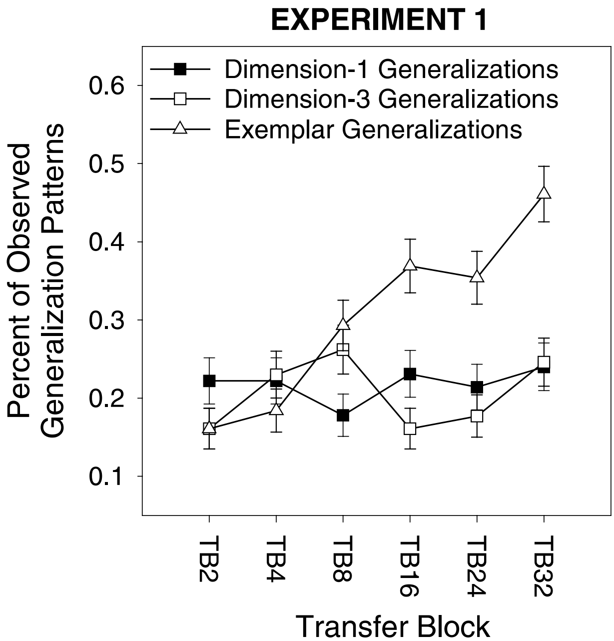

Fig. 3. The percentage of observed generalization patterns consistent with dimension-1 rule generalizations (filled squares), dimension-3 rule generalizations (open squares), and exemplar generalizations (open triangles) as a function of transfer block in Experiment 1.

ing distributions of generalization patterns allows us to detect deviations from this possibility, which can be theoretically important. For example, two items may be classified with probability of .50 when averaged across subjects, but in the extreme, all subjects may show generalization AA or BB and no subjects may show generalization AB or BA.

For the following analysis, consider the null hypothesis that the average transfer probabilities shown in the bottom of Fig. 2 govern the behavior of every individual subject. Given that each subject produced a single response for each item in each transfer block, we can directly calculate the expected probability of observing any given generalization from those average probabilities. For example, the expected probability of observing generalization ABBBA is given by $P(\mathrm{A}|\mathrm{T1}) \times P(\mathrm{B}|\mathrm{T2}) \times P(\mathrm{B}|\mathrm{T4}) \times P(\mathrm{B}|\mathrm{T5}) \times P(\mathrm{A}|\mathrm{T6})$. The statistical question is whether any observed generalization probability in the distribution significantly exceeds the expected generalization probability calculated from the average transfer probabilities. Using a bootstrapping procedure, we generated 50,000 simulated distributions of generalization patterns using the observed transfer probabilities. For each generalization pattern, we then statistically determined whether the observed probability of that generalization fell within the tail of the distribution of probabilities across those 50,000 simulations. Finding probabilities in the observed distri-

Table 2
Generalization patterns that were observed significantly more frequently than expected from the observed average transfer probabilities in Experiment 1 (see text for details)

Pattern	TB2 (2)	TB4 (4)	TB8 (8)	TB16 (16)	TB24 (24)	TB32 (32)
BBABA	***	***	*	*	***	***
BBAAB	*					
BBAAA	*			**	**	*
ABBBA				**	**	**
AABBB	***	***	*	**	**	
AABAB	*	*		**	***	***

Note. *** Indicates significance at the $p < .001$ level; ** indicates significance at the $p < .01$ level; * indicates significance at the $p < .05$ level.

bution of generalization patterns that exceed this statistical criterion provides evidence that distributions provide information that is not present in the average transfer probabilities.

As shown in Table 2, six of thirty-two generalizations were observed significantly more frequently than expected from the average probabilities. Most importantly, the proportions of the two rule generalizations (BBABA and AABBB) were significant early in training, and the proportion of the exemplar generalization (ABBBA) was significant later in training. One of the rule-based generalizations (BBABA) was also significant later in training, but as observed by Nosofsky et al. (1994), exemplar models do predict this high proportion. Other patterns were also significant. Pattern BBAAB was only significant during the first block. Pattern BBAAA was significant at various stages of learning, but the observed probabilities were small. Only pattern AABAB was significant and observed with relatively high probability. Indeed, at the end of learning, generalization AABAB was observed with comparable probability to the rule generalization AABBB. As can be seen in the theoretical section, our representational shift model accounted for the observed probability of this particular generalization.

3.3. Discussion

The goal of this first experiment was to find evidence for a shift from rule-based to exemplar-based categorization as a function of category learning. The experiment was partially motivated by a reexamination of earlier published data extending the classic Medin and Schaffer (1978) study. Nosofsky et al. (1994) observed that with 16 training blocks, rule-based generalizations were dominant, a result consistent with a rule model but inconsistent with an exemplar model, but an exemplar-based generalization was also relatively prominent. Palmeri and Nosofsky (1995) observed that with 25 training blocks, the single-dimension rule-based generalizations were also relatively prominent, but an exemplar-based generalization was dominant.

The present experiment examined the evolution of the distribution of generalization patterns as a function of training, testing participants at various stages of category learning.

We observed that early in learning there was evidence for a relative prominence of generalizations based on single dimensions (BBABA and AABBB) with little evidence for exemplar-based generalizations. With additional training, however, an exemplar-based generalization (ABBBA) gradually gained prominence. These empirical results are consistent with the hypothesis of a shift from rule-based to exemplar-based category representations as a function of training, as suggested by some theories of automaticity. In the theoretical modeling section at the end of this paper, we will attempt to corroborate this descriptive finding by comparing the fits of various models to the observed data.

We should emphasize that the Medin and Schaffer category structure was designed to contrast predictions of independent-cue (prototype) models and interactive-cue (exemplar) models of categorization. Perhaps this fact led to our finding little evidence for a decrease in rule generalizations over learning, even though we did find evidence for an increase in exemplar generalizations. Indeed, some of the purported rule generalizations are actually predicted to show moderately high levels of prominence by exemplar-based models (Nosofsky et al., 1994). The structures developed for the next two experiments specifically were designed to contrast predictions of rule-based and exemplar-based models. Because of the reemerging interest in prototype models (e.g., Smith & Minda, 1998), we also aimed to contrast predictions of a prototype model with those of rule-based and exemplar-based models as well.

4. Experiments 2 and 3

These experiments aimed to contrast predictions by rule and exemplar models regarding distributions of generalization patterns. Developing theoretically diagnostic category structures can be best characterized as a trial and error process. What is especially challenging is that predictions of exemplar models can change dramatically with subtle changes to the category structures (e.g., Nosofsky, 2000), something not nearly so true for rule or prototype models. So, the approach that we used was to develop a number of category structures meeting certain "design principles" and then to test predictions regarding distributions of generalizations for rule-based and exemplar-based models. For these two experiments, we retained category structures (see Tables 3 and 4) that proved especially diagnostic with respect to the generalizations predicted by rule-based and exemplar-based models. We also discovered that the category structure used in Experiment 3 proved to be additionally diagnostic for contrasting prototype- and exemplar-based categorization as well.

Table 3
Category structure used in Experiment 2

Category A		Category B		Transfer	
A1	2 1 1 1	B1	1 2 2 2	T1	2 2 2 2
A2	1 2 1 1	B2	2 1 2 2	T2	1 1 1 1
A3	1 1 2 1	B3	2 2 1 2	*T3	2 2 1 1
A4	1 1 1 2	B4	2 2 2 1	*T4	2 1 1 2
A5	1 2 1 2	B5	2 1 2 1	*T5	1 2 2 1
				*T6	1 1 2 2

Note. * Highlights the four critical transfer items.

Table 4
Category structure used in Experiment 3

Category A		Category B		Transfer	
A1	1 1 1 2	B1	1 1 2 2	T1	2 2 2 2
A2	1 2 1 2	B2	2 2 1 2	*T2	2 2 1 1
A3	1 1 1 1	B3	2 2 2 1	T3	2 1 2 2
A4	1 2 2 1	B4	2 1 2 1	*T4	2 1 1 2
A5	2 1 1 1			*T5	1 2 2 2
				T6	1 2 1 1
				*T7	1 1 2 1

Note. * Highlights the four critical transfer items.

Various design principles were established for constructing category structures used in these experiments. First, because our goal was to examine distributions of generalization patterns, and because displaying distributions with fewer generalizations is preferred, we constructed structures with only four critical transfer items, highlighted by asterisks in Tables 3 and 4. This yields only sixteen possible generalizations which needed to be displayed. As in Experiment 1, the ‘‘noncritical’’ transfer items were those deemed theoretically less diagnostic. In Experiment 2, the noncritical items were the category prototypes; in Experiment 3, the noncritical items were a prototype and items that differed from a prototype along a relatively non-diagnostic dimension. A subsidiary benefit of limiting the total number of generalizations is that it allowed us to test fewer subjects than we tested in the first experiment.

Another design principle was to have critical transfer items that were classified into the two categories with roughly equal probability throughout training. Providing transfer items that produce average categorization probabilities of .5 maximizes the amount of individual subject variability in how those items could be classified. Maximizing individual subject variability maximizes the information that can be revealed in the distributions of generalization patterns.

Another design principle was to vary the likelihood that subjects would form rules along particular dimensions. In both experiments, rules were far more likely along dimensions one and three than dimensions two and four. As can be seen in Tables 3 and 4, rules along dimensions one and three produce fewer exceptions, making those dimensions more likely to be selected. Use of rules along these dimensions yields two rule-based generalizations for the critical transfer items. For both experiments, rule-based generalizations are BBAA and AABB for rules along dimensions one and three, respectively.

We also required category structures for which a rule model and an exemplar model would predict contrasting distributions of generalization patterns. Generating predictions by an exemplar model for the category structures used in Experiments 2 and 3 is complicated because predictions, even at a qualitative level, are sensitive to the parameter values used to generate those predictions. As in Experiment 1, in order to generate a priori predictions of an exemplar model, we simulated the context model and ALCOVE across a wide spectrum of the parameter space, as reported in Appendix A. For each of the two category structures, both the context model and ALCOVE converged on the same two generalization patterns emerging as most prominent ones in the distributions. These exemplar-based generalizations were ABAB and BABA in Experiment 2, and ABBA and BAAB in Experiment 3. We discovered that a prototype-based model could account for the "exemplar-based" generalizations from Experiment 2 under certain parameter settings. However, the prototype model could not account for the "exemplar-based" generalizations from Experiment 3, regardless of parameter settings.

Finally, although a rule-based model, such as RULEX, predicts that a high proportion of subjects will display single-dimension rule-based generalizations listed above, other generalizations are predicted as well. In addition to rules, RULEX assumes that subjects will attempt to memorize exceptions to those rules, which can also be used to generalize when classifying new transfer items. We systematically examined the predictions of the RULEX model formalized by Nosofsky et al. (1994) and found that it could not predict the prominence of the so-called exemplar-based generalizations in the distributions of generalization patterns for either category structure. As will be discussed at the end of this paper, although it may be possible to formalize a rule-based model that does predict these prominent "exemplar-based" generalizations, to our knowledge there does not currently exist a rule-based model that does so.

Following procedures established in Experiment 1, subjects were trained to categorize stimuli with corrective feedback. At various stages during learning, subjects were tested on training and transfer items without corrective feedback. The critical data for consideration were the distributions of generalization patterns observed at each stage of training, with a specific eye to how those distributions evolved with experience. The aim of our initial analysis will be to examine the relative prominence of the rule-based and exemplar-based generalizations in the distributions as a function of learning.

Our first step will be a descriptive analysis of the observed distributions of generalization patterns. We will later test the ability of various formal models to account for the observed distributions of generalization patterns in all three experiments. Our goal in these theoretical analyses was not to test specifically the predictive power of these models per se, but to use the models to test specific hypotheses regarding which of the representations may be used at various stages of category learning.

4.1. Methods

4.1.1. Subjects

Subjects in Experiment 2 were 63 undergraduates, and subjects in Experiment 3 were 121 undergraduates. Subjects received credit in a psychology course, and all were tested individually.

4.1.2. Stimuli

The stimulus set was the same in Experiments 2 and 3. Stimuli were computer-generated cartoon faces that varied along four binary-valued dimensions: Hair color (red or yellow), nose size (small or large), ear size (small or large), and mouth shape (line or circle). These faces were similar in some respects to stimuli used by Lamberts (1995).

The category structures used in Experiments 2 and 3 are given in Tables 3 and 4, respectively. For both experiments, the categories were linearly separable. In addition, for both experiments, each dimension was partially diagnostic (with value 1 along each dimension associated with category A and value 2 along each dimension associated with category B), but dimensions 1 and 3 were more diagnostic than dimensions 2 and 4. The assignment of physical dimensions to abstract dimensions and physical values to abstract values along dimensions was randomized for every subject.

4.1.3. Procedure

The experimental procedures for Experiments 2 and 3 were identical to the experimental procedures for Experiment 1, except for the schedule of transfer blocks. In Experiment 2, a transfer block was presented after 4, 6, 8, 12, 16, and 32 blocks of training (these are labeled TB4, TB6, TB8, TB12, TB16, and TB32 in the tables and figures in the subsequent analyses). In Experiment 3, a transfer block was presented after 4, 8, 12, 16, 24, and 32 blocks of training (these are labeled TB4, TB8, TB12, TB16, TB24, and TB32 in the tables and figures in the subsequent analyses).

4.2. Experiment 2 results

Following the exclusion criteria established in Experiment 1, eighteen subjects with an error rate of more than 25% on the last four training blocks were removed from further analyses; 27.8% of these excluded subjects did not categorize greater than chance during the last four training blocks.

4.2.1. Average transfer data

Average probabilities of categorizing each stimulus as a member of category A as a function of the six transfer blocks are shown in the bottom row of Fig. 4. Early in training, more errors were made classifying A1 and B1, and A3 and B3, which can be characterized as exceptions to rules along dimensions 1 or 3, respectively; significantly more errors were made on the exceptions than nonexceptions in the first four blocks [$t(44) > 3.014$]. Categorization probabilities for the four critical transfer items, T3–T6, showed little change over the course of learning, and never differed significantly from a .5 probability by a z-test for proportions. If distributions of generalization patterns simply reflected average transfer probabilities, we would expect to observe all of the sixteen possible generalizations to be equally likely throughout training. As shown below, this possibility was not observed. Finally, we also saw evidence for learning in the decrease in categorization errors made on the old training items in each block; a within-subjects ANOVA revealed a significant effect of block [$F(5, 220) = 43.49$, $MSe = .014$] and a significant linear contrast on blocks [$F(1, 44) = 212.851$, $MSe = .013$].

4.2.2. Distributions of generalization patterns

Fig. 4 displays the distributions of generalization patterns observed within each transfer block. Early in training we saw evidence for rule use by the relative prominence of the rule-based generalizations AABB and BBAA. Later in training we saw evidence for exemplar use by the relative prominence of the exemplar-based generalizations ABAB and BABA. As a summary, we tallied the proportion of rule-based generalizations as the proportion of AABB and BBAA generalizations and those differing by one response, and we tallied the proportion of exemplar-based generalizations as the proportion of BABA and ABAB generalizations and those differing by one response. The left panel of Fig. 5 displays these proportions and shows additional corroborating evidence of early rule use which gave way to later exemplar use.

Recall that average probabilities of around .5 were observed for classifying all of the critical transfer items throughout training. Distributions of generalization patterns generated directly from these average probabilities would yield distributions in which all generalizations were equally likely. As shown in Fig. 4, this possibility was not observed. For statistical corroboration, we found that in each transfer block except the first one, the observed distributions were significantly different from "null distributions" in which all generalizations were equally likely using a χ^2 test [$\chi^2(15) > 25.033$]; the observed distribution for the first block was margin-

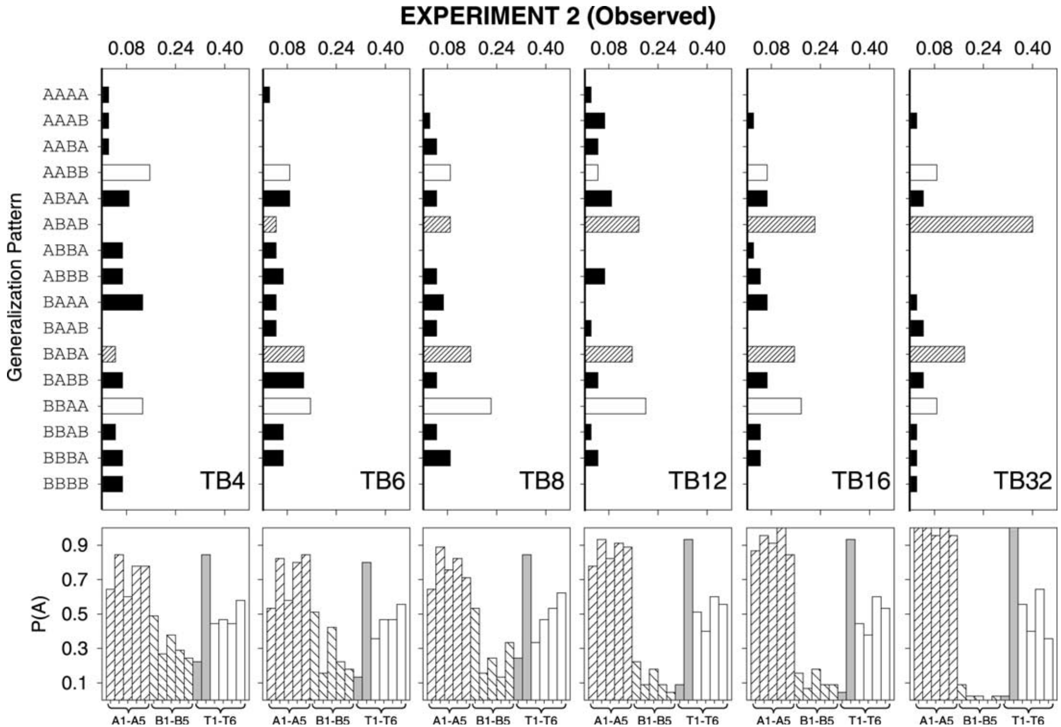

Fig. 4. The top row shows distributions of generalization patterns and the bottom row shows average categorization probabilities $P(\mathrm{A})$ for training and transfer items after 4 (TB4), 6 (TB6), 8 (TB8), 12 (TB12), 16 (TB16), and 32 (TB32) blocks of training in Experiment 2. For the distributions, white bars highlight the two prominent rule-based generalization patterns and hatched bars highlight the two prominent exemplar-based generalization patterns. For the average categorization probabilities, hatched bars are training items, white bars are critical transfer items, and gray bars are noncritical transfer items.

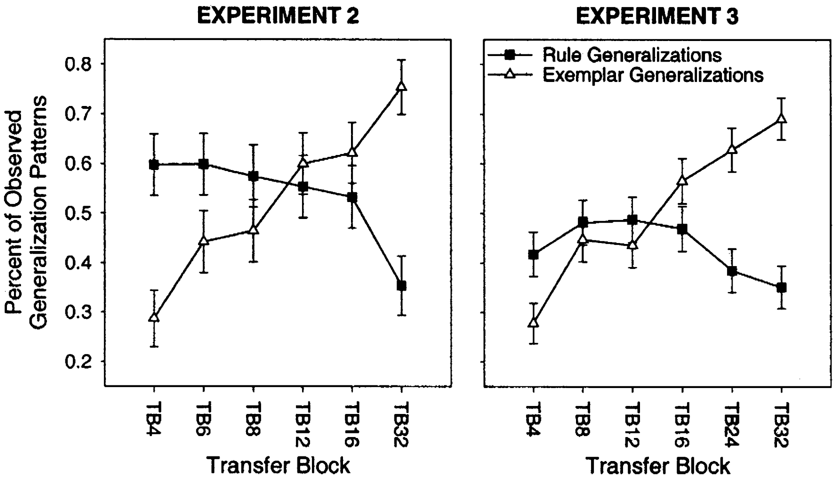

Fig. 5. The percentage of observed generalization patterns consistent with rule-based generalizations (filled squares) and exemplar-based generalizations (open triangles) as a function of transfer block in Experiment 2 (left panel) and Experiment 3 (right panel).

ally significantly different from a null distribution ($\chi^2(15) = 23.629$, $p = .072$).

Finally, following procedures from Experiment 1, we tested whether generalization probabilities in the distributions were significantly different from those predicted from the average transfer data. With the bootstrapping procedure, we found that four of the 16 generalizations were observed significantly more frequently than expected from average probabilities at some point in training, as shown in Table 5. The dimension-3 rule generalization, AABB, was significant relatively early in training; for some reason, the dimension-1 rule generalization, BBAA, only emerged as significant in block TB16. The two exemplar generalizations, BABA and ABAB, were signifi-

Table 5
Generalization patterns that were observed significantly more frequently than expected from the observed average transfer probabilities in Experiment 2 (see text for details)

Pattern	TB4 (4)	TB6 (6)	TB8 (8)	TB12 (12)	TB16 (16)	TB32 (32)
BBAA				*		
BABA				***	**	***
ABAB			*	*	**	***
AABB	**		*			

Note. *** Indicates significance at the $p < .001$ level; ** Indicates significance at the $p < .01$ level; * Indicates significance at the $p < .05$ level.

cant relatively late in training. These results are consistent with the hypothesis of early rule use and later exemplar use.

4.3. Experiment 3 results

Twenty-seven subjects with error rates greater than 25% on the last four blocks were removed from analyses; 57.5% of those subjects did not perform greater than chance during those blocks.

4.3.1. Average transfer data

Average probabilities of categorizing each stimulus as a member of category A as a function of the six transfer blocks are shown in the bottom row of Fig. 6. In each block, significantly more errors were made classifying A5 and B1, and A4 and B2, which can be characterized as exceptions to rules along dimensions 1 and 3, respectively [$t(93) > 4.158$]. The four critical transfer items, T2, T4, T5, and T7, showed little change over learning and did not significantly differ from a .5 probability by a z test at any point in training. Evidence for learning was established by a decrease in errors over blocks; a within-subjects ANOVA revealed a significant effect of block [$F(5,465) = 54.59$, $MSe = .017$] and a significant linear contrast on blocks [$F(1,93) = 182.87$, $MSe = .024$].

4.3.2. Distributions of generalization patterns

The top row of Fig. 6 displays the distributions of generalization patterns observed in the six transfer blocks using generalizations composed of critical transfer items T2, T4, T5, and T7. Early in learning, we saw evidence for rule use by the relative prominence of generalizations AABB and BBAA. Later in learning, we saw evidence for exemplar use by the relative prominence of generalizations ABBA and BAAB. As an additional summary, we tallied the proportion of rule generalizations as the proportion of AABB and BBAA generalizations and those differing by one response, and tallied the proportion of exemplar generalizations as the proportion of ABBA and BAAB generalizations and those differing by one response. The right panel of Fig. 5 displays the proportion of rule-based and exemplar-based generalizations as a function of training. The figure shows clear evidence of early rule use and later exemplar use.

Because we observed .5 categorization probabilities for all four critical transfer items, we next conducted a χ^2 test of the hypothesis that that all sixteen possible generalization patterns were observed with equal likelihood. Confirming the impressions obtained from visually examining Fig. 6, we found that within each transfer block, the observed distribution of generalization patterns was significantly different from a null distribution in which all patterns were equally likely [$\chi^2(15) > 42.845$].

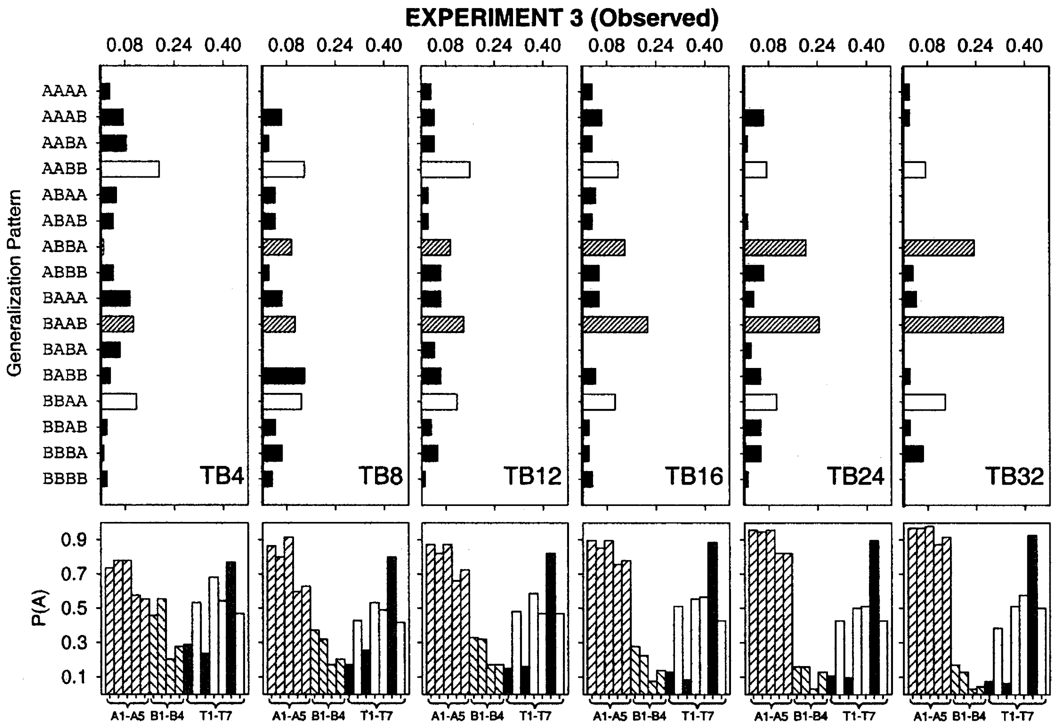

Fig. 6. The top row shows distributions of generalization patterns and the bottom row shows average categorization probabilities $P(\mathrm{A})$ for training and transfer items after 4 (TB4), 8 (TB8), 12 (TB12), 16 (TB16), 24 (TB24), and 32 (TB32) blocks of training in Experiment 3. For the distributions, white bars highlight the two prominent rule-based generalization patterns and hatched bars highlight the two prominent exemplar-based generalization patterns. For the average categorization probabilities, hatched bars are training items, white bars are critical transfer items, and gray bars are noncritical transfer items.

Table 6
Generalization patterns that were observed significantly more frequently than expected from the observed average transfer probabilities in Experiment 3 (see text for details)

Pattern	TB4 (4)	TB8 (8)	TB12 (12)	TB16 (16)	TB24 (24)	TB32 (32)
BBAA	**	**	**	*		
BAAB			*	***	***	***
ABBA		*		***	***	***
AABB	***	**	**			

Note. *** Indicates significance at the $p < .001$ level; ** indicates significance at the $p < .01$ level; * indicates significance at the $p < .05$ level.

We next tested whether any generalizations were observed with probabilities significantly greater than expected from the average transfer probabilities. With the bootstrapping procedure, we found that four of the sixteen generalizations were observed significantly more frequently than expected from the transfer data at some point in training, as shown in Table 6. The proportion of rule generalizations was significant relatively early in training. The proportion of exemplar generalizations was significant relatively late in training, consistent with the hypothesis of early rule use and later exemplar use.

4.4. Discussion

Experiment 1 found evidence for a gradual emergence of exemplar-based representations through the growth of the predicted exemplar-based generalization pattern, but that experiment did not find evidence for a de-emphasis of rule-based generalizations with training. By contrast, the category structures used in Experiments 2 and 3 were specifically designed to contrast the predictions of models assuming rule-based representations and models assuming exemplar-based representations. Systematic analyses of the evolution of the distributions of generalization patterns obtained in both experiments revealed a prominence of rule-based generalizations early in training and a shift to a prominence of exemplar-based generalizations later in training, consistent with our a priori hypotheses.

5. Theoretical modeling

The results from these experiments suggest a shift from rules to exemplars as a function of experience. That said, it may be imprudent to base these broad conclusions on just a cursory examination of patterns of observed data. It is well known that complex nonlinear models, such as the framework used in the forthcoming theoretical modeling, can produce nonintuitive predictions. As stated by Hintzman (1990, pp. 110–111), formal models "force the theorist to be explicit, so that assumptions are publicly accessible and the reliability of derivations can be confirmed... Surprises are likely when the model has properties that are inherently difficult to understand, such as variability, parallelism, and nonlinearity—all, undoubtedly, properties of the brain." So, although our results seem to suggest a shift from rules to exemplars, it could be the case that a model that assumes a single category representation throughout the entire course of category learning could account for the full spectrum of our observed results. So, the following theoretical analyses formulate rule-based, prototype-based, and exemplar-based representational assumptions within the same general theoretical framework.

There are a variety of approaches to using formal modeling to understand categorization behavior. Common to all approaches is instantiating hypothetical psychological processes involved in representing category information, retrieving category information, and making decisions within a computational or mathematical formalism. One approach to employing formal modeling is aimed at developing a new model to account for some novel set of psychological phenomena, testing that model at accounting for the observed findings, and perhaps contrasting that model with existing models. Another related approach is to select a number of existing formal models and to compare them on their ability (or inability) to account for a set of empirical data. One potential shortcoming of both of these approaches is that various models may differ on a number of critical assumptions and it may be difficult to decide just on the basis of fits to data which of these fundamental assumptions underlies the adequacy (or inadequacy) of a particular model. For example, models may differ in their representational assumptions (say exemplar storage versus decision boundaries) but may also make different assumption regarding the form of the decisions process (say probabilistic versus deterministic); success (or failure) of particular models may be attributed to representations but they could just as well be attributed to the form of the decision rule (e.g., see Maddox & Ashby, 1998; Nosofsky, 1998a,b).

The modeling described here aims to provide converging tests of hypotheses regarding representational shifts during category learning and to compare the results of those tests with our empirical analyses. Toward this end, we required a modeling framework with formal mechanisms that allowed for category learning and that allowed for different types of category representation to be instantiated within the same general architecture. Numerous models of category learning have recently been proposed. A non-exhaustive, but representative list includes Nosofsky et al.'s (1994) rule-plus-exception model (RULEX), Anderson's (1990) rational model, Erickson and Kruschke's (1998) attention to rules and instances model (ATRIUM), Love et al.'s (in press) SUSTAIN model of category learning, Vandierendonck's

(1995) parallel rule activation and synthesis model (PRAS), and Ashby et al.'s (1998) competition between verbal and implicit systems model (COVIS). Although each of these models has been successful in accounting for particular empirical results, they each tend to be quite committed to a specific, perhaps mixed, form of category representation. Because our purpose was not to develop yet another model of category learning whose performance could be compared and contrasted with other models, we required a theoretical framework in which we could compare the predictions of prototype, rule, and exemplar representations while keeping all other aspects of the models as identical as possible.

We believe that Kruschke's (1992) ALCOVE model may provide an ideal framework for our purposes. Arguably, ALCOVE (Kruschke, 1992) has been subjected to some of the most rigorous tests and has proved capable of accounting for a variety of categorization phenomena (e.g., Nosofsky & Kruschke, 1992). ALCOVE is a connectionist category learning model that formally instantiates key assumptions of a successful exemplar model of categorization, the generalized context model (Nosofsky, 1984, 1986). However, as will be outlined below, the ALCOVE architecture is flexible in that prototypes or rules can be substituted for the exemplars without changing most other aspects of the model. The ALCOVE architecture thus provides a framework for testing representational assumptions (prototypes, rules, or exemplars) while keeping all other aspects of the learning process and the decision process identical across different models. Our aim, then, is not to test ALCOVE per se, but to use the ALCOVE architecture as a means of providing formal tests of our hypotheses regarding category representations.

5.1. The ALCOVE architecture

ALCOVE is a three-layered feedforward connectionist network in which activation is passed from a stimulus input layer with a node for each psychological dimension, which is gated by a selective attention weight, to a representational hidden layer, and on to a category output layer via association weights. Category learning in the model involves adjusting both the selective attention weights on each dimension and the association weights between hidden layer representational nodes and category output nodes through the process of gradient descent on error.

In its traditional form, ALCOVE represents categories in terms of stored exemplars in the representational hidden layer. The exemplar nodes are activated based on their similarity to the input stimulus presented along the attentionally weighted psychological dimensions. A version of ALCOVE using prototype representations can be created by simply replacing the exemplar nodes with a single prototype node for each category. Likewise, a version of ALCOVE using simple single-dimension rules can be created by forcing the model to attend to only a single dimension, but to allow different simulated subjects to attend to different dimensions. Hence exemplar, prototype, and rule models within the general ALCOVE architecture can be created that use the same stimulus input representations, the same category response mechanisms, and the same learning mechanisms.

In our simulations, the stimulus input layer consisted of four binary-valued dimension nodes, with a given input node i denoted by a_i^{in}; a stimulus is represented as a vector of activations across those inputs. For the standard exemplar version of ALCOVE, each hidden node corresponds to an exemplar. A particular hidden exemplar node (a_j^{hid}) is activated according to its similarity to the input stimulus, where similarity is computed as in the generalized context model (Nosofsky, 1984). Similarity between a presented item and a stored exemplar is inversely related to the psychological distance between them. For separable-dimension stimuli, the activation of exemplar node a_j^{hid} is given by

$$a_j^{\text{hid}} = \exp\left[-c\sum_{\text{in } i}\alpha_i|h_{ji} - a_i^{\text{in}}|\right], \tag{1}$$

where a_i^{in} is the value of the input stimulus along dimension i, h_{ji} is the value of the hidden exemplar along dimension i, α_i is the attention to dimension i, and c is a similarity scaling parameter. When optimally allocated, the selective attention weights tend to emphasize differences along diagnostic dimensions and de-emphasize differences along nondiagnostic dimensions (Nosofsky, 1984, 1998a,b). Because the present experiments used binary-valued dimensions, the above equation can be simplified to

$$a_j^{\text{hid}} = \exp\left[-c\sum_{\text{in } i}\alpha_i D_{ji}\right], \tag{2}$$

where D_{ji} equals 0 if input stimulus and exemplar j match on dimension i and equals 1 if they mismatch.

ALCOVE learns to associate exemplars with category outputs. The activation of category output node k, a_k^{out}, is given by

$$a_k^{\text{out}} = \sum_{\text{hid } j} w_{kj}a_j^{\text{hid}}, \tag{3}$$

where w_{kj} is the learned association weight between exemplar j and output node k. The probability of categorizing the input stimulus as a member of category K is given by

$$P(K) = \frac{\exp(\varphi a_K^{\text{out}})}{\sum_{\text{out } k}\exp(\varphi a_K^{\text{out}})}, \tag{4}$$

where φ is a response mapping parameter.

Both attention weights (α_i) and association weights (w_{kj}) are learned by gradient descent on error (see Kruschke, 1992). A parameter λ_α is the learning rate for attention weights. We employed attention weight normalization in which the sum of the weights was constrained to sum to 1 (see Nosofsky & Johansen, 2000). A parameter λ_w is the learning rate for association weights. Following Kruschke (1992), so-called *humble teachers* were used in that the teaching signal, t_k, was set to the maximum of +1 and a_k^{out} for the correct category and was set equal to the minimum of −1 and a_k^{out} for the incorrect category. Over the course of learning, the association weights tend to represent the degree to which each exemplar is associated with each category. Over the course of learning, the attention weights tend to reflect an optimal allocation of dimensional attention, which has the effect of stretching the psychological space along diagnostic dimensions and shrinking it along nondiagnostic ones. In the simulations, attention learning and associate weight learning only took place during the learning blocks in which corrective feedback was provided on every trial with no learning allowed during the transfer blocks.

To create a prototype model within the ALCOVE architecture, the exemplar nodes in the hidden layer were replaced with just two prototype nodes, one for each of the two learned categories. The positions of each prototype in psychological space corresponded to the central tendencies of the two categories. In each of our three experiments, these two prototypes were 1 1 1 1 and 2 2 2 2 (versions of the prototype model in which we allowed the location of the prototypes to be learned did not result in qualitatively different predictions). Activation of a prototype node in the hidden layer was calculated the same way as the activation of an exemplar node in the hidden layer of the exemplar version of ALCOVE. We should emphasize that the use of such a multiplicative similarity function is not the only way to formulate a prototype model. Indeed, others have suggested an additive similarity function instead (e.g., Medin & Schaffer, 1978; Smith & Minda, 1998). We chose the multiplicative version because it has provided a superior account of some recent categorization data (Minda & Smith, 2000).

There may be several ways of instantiating a rule model in the ALCOVE architecture. We chose a simple approach which was surprisingly effective in accounting for some of the empirical data, a criteria we deemed sufficient for our goal of evaluating potential rule-use early in learning. This simple rule model started with the original version of ALCOVE. A rule along dimension N was achieved by fixing the selective attention weight on dimension N (α_N) to 1 and fixing all of the other attention weights to 0; no conjunctive, disjunctive, or more complex rules were investigated (see Choi et al., 1993). Like RULEX (Nosofsky et al., 1994), different simulated subjects were assumed to form rules along different dimensions. We did not attempt to directly incorporate any specific rule-learning mechanism into the model. Instead, we assumed that the number of simulated subjects who formed rules along a given dimension was proportional to the relative diagnosticity of that dimension; specifically, the number of subjects forming a rule along a dimension was proportional to the correlation between the feature value along that dimension and the correct category label across all stimuli from both categories. Essentially, the number of simulated subjects using a rule along a dimension was roughly proportional to how accurately the categories could be learned by just forming a simple rule along that dimension.

5.2. Details of the modeling procedures

In the various versions of ALCOVE, four parameters were allowed to freely vary: The scaling parameter, c, the response mapping parameter, φ, the attention learning rate, λ_α, and the association weight learning rate, λ_w; because the rule version of ALCOVE had fixed attention weights, the attention learning rate parameter was not used. A standard hill-climbing parameter fitting procedure was used to adjust these four parameters so as to maximize the fits of the models to the entire set of observed data.

Because we were simultaneously fitting both the average transfer data and the distributions of generalization patterns, we needed to use a fit statistic that combined fits to both subsets of the observed data. The average transfer data and the distributions of generalizations have different numbers of data points (e.g., in Experiment 1, there were 16 transfer probabilities but there were 128 generalizations probabilities in each block) and the variances of these two subsets of data were also different, so simply combining the two subsets directly (e.g., in Experiment 1, fitting 864 undifferentiated data points) was not acceptable for our purposes; for example, simply calculating the root mean squared error (RMSE) combined across the two subsets of data may cause the fits to one subset of data to dominate the overall fits at the expense of the fits to the other subset of data. There are numerous ways of combining the fits across both subsets of data. We chose the following procedure: For a given set of parameters at some point in the hill-climbing procedure, we calculated the percentage of variance accounted for by the predicted average transfer data and the percentage of variance accounted for by the predicted distributions of generalization patterns for each of the six transfer blocks individually. These twelve summary fits were then added together to yield an omnibus measure of the overall fit of the model to the entire set of observed data. This method of combining the individual fits puts equal weight on the fits to the average transfer data and the fits to the distributions and equally weights the fits for each transfer block as well. One important reason for simultaneously fitting both the average data and the distributions is that the model must be able to account for the average categorization probabilities for the training stimuli, which are not included in the generalization patterns tallied in the distributions.

The hill-climbing procedure adjusted the four parameters incrementally to find the maximum value of this derived omnibus measure of fit. For all of the model fits, we began the hill-climbing procedure at several different initial values of the parameters so as to decrease the possibility of settling upon a local maxima. In Experiments 1 and 3, the four-parameter prototype and exemplar models, and the three-parameter rule model, needed to account for 6×16 transfer probabilities and 6×128 generalization proportions, or a total of 864 data points. In Experiment 2, the models needed to account for 6×16 transfer probabilities and 6×64 generalization proportions, or a total of 480 data points.

Like many category learning models (e.g., Anderson, 1990; Nosofsky et al., 1994), and presumably like human subjects, ALCOVE is sensitive to the particular sequence of training items. At each point in the hill-climbing search procedure, a set of parameters was used to generate predictions from 800 simulated subjects, each of which was provided a different randomized sequence of training items (like the actual subjects in our experiments). Model predictions were generated by averaging across these 800 simulated subjects. Unlike human subjects, who produce discrete category responses, each simulated ALCOVE subject generates a set of continuous categorization response probabilities. To generate the predicted average transfer data, the predicted transfer probabilities were simply averaged across all 800 simulated subjects. To generate the predicted distribution of generalization patterns, we first needed to generate a distribution of generalization patterns from the predicted transfer probabilities from each individual simulated subject. The predicted distribution of generalization patterns from a single simulated subject was computed from the average probability vector predicted by the model. Let p_1 denote the probability that transfer item T1 is classified into category A, then the proportion p_1 of the generalization patterns had an A in their first position and the proportion $1 - p_1$ of the generalization patterns had a B in their first position; for example, the predicted probability for generalization AAABBBB would be $p_1 \times p_2 \times p_3 \times (1 - p_4) \times (1 - p_5) \times (1 - p_6) \times (1 - p_7)$ (see Nosofsky et al., 1994). We then averaged the individual predicted distributions of generalization patterns across all 800 simulated subjects to generate an overall predicted distribution of generalization patterns that we compared with the observed data.

5.3. Theoretical modeling results

Using the procedure described above, we fitted the exemplar, prototype, and rule versions of ALCOVE to the average transfer data and the distributions of generalization patterns from each of the three experiments. We will begin by describing the fits of the models to the transfer blocks of each experiment. Then we will describe the fits of a *shift model* that instantiates shifts from single-dimension rules to exemplar retrieval within a single framework.

5.3.1. Theoretical accounts of Experiment 1

The best-fitting parameters for the exemplar, prototype, and rule models for fits to data from Experiment 1 are shown in Table 7. The predicted average transfer probabilities and the predicted distributions of generalization patterns for the three models in each of the six transfer blocks are shown in Figs. 7 and 8, respectively. The summary fit statistics of the models in terms of percent of variance accounted for by the predicted transfer data and the predicted distributions as a function of transfer block are shown in the left and right panels of Fig. 9, respectively.

As shown in Fig. 9, the rule model provided a good quantitative account of the data from the first two transfer blocks, accounting for over 90% of the variance in the average transfer data and about 70% of the variance in the distributions of generalizations. In these early blocks, the rule model performed at least as well as the other two models in accounting for the average transfer data and much better than the other models in accounting for the

Table 7
Best-fitting parameters for the exemplar, prototype, rule, and shift model

Parameter	Exemplar	Prototype	Rule	Shift
Experiment 1				
c	17.858	3.367	39.000	7.020
λ_w	.290	.335	.059	.056
λ_α	1.033	.040	—	.201
φ	1.410	1.005	1.510	2.995
κ'				1.859
ρ				.142
Experiment 2				
c	18.734	1.108	40.000	20.734
λ_w	.170	.033	.027	.170
λ_α	.875	2.342	—	.675
φ	1.920	9.028	1.388	2.535
κ'				.800
ρ				.138
Experiment 3				
c	9.989	11.804	35.000	16.033
λ_w	.476	.688	.053	.125
λ_α	.290	.966	—	.125
φ	2.420	3.289	1.488	3.084
κ'				1.075
ρ				.032

Note. c = sensitivity parameter, λ_w = association weight learning rate, λ_α = attention weight learning rate, φ = response mapping parameter, κ' = initial attention weight reallocation parameter, ρ = annealing rate for κ.

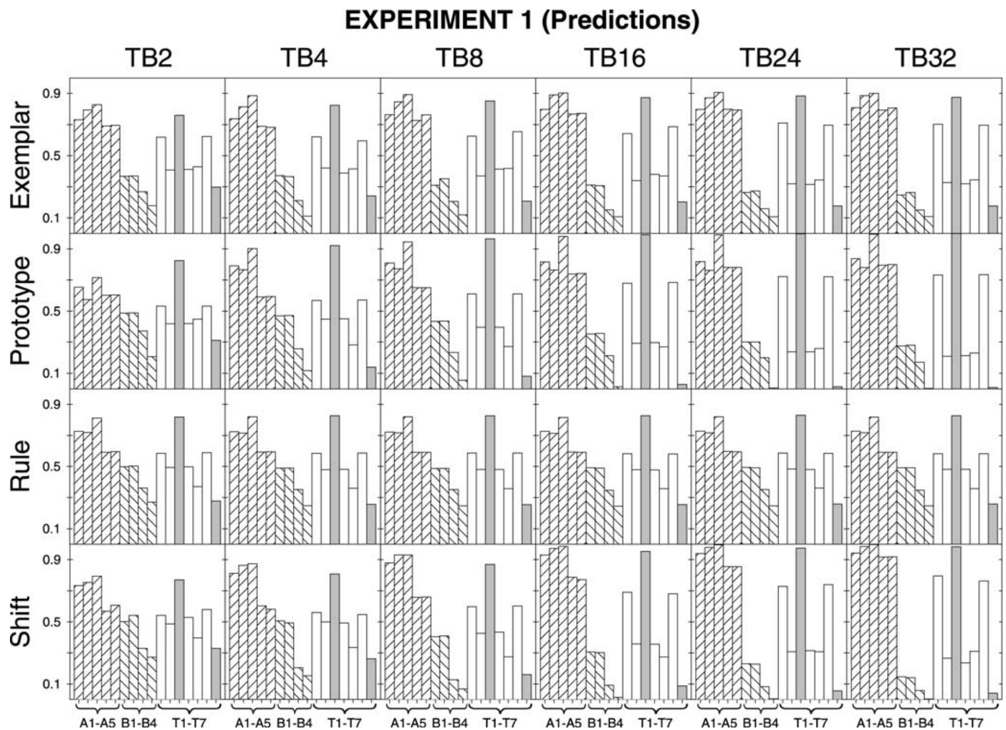

Fig. 7. Predicted average transfer probabilities from Experiment 1. Each column contains predicted transfer probabilities after 2, 4, 8, 16, 24, and 32 blocks of training. The four rows display the predicted probabilities for the exemplar, prototype, rule, and shift models. Hatched bars are training items, white bars are critical transfer items, and gray bars are noncritical transfer items.

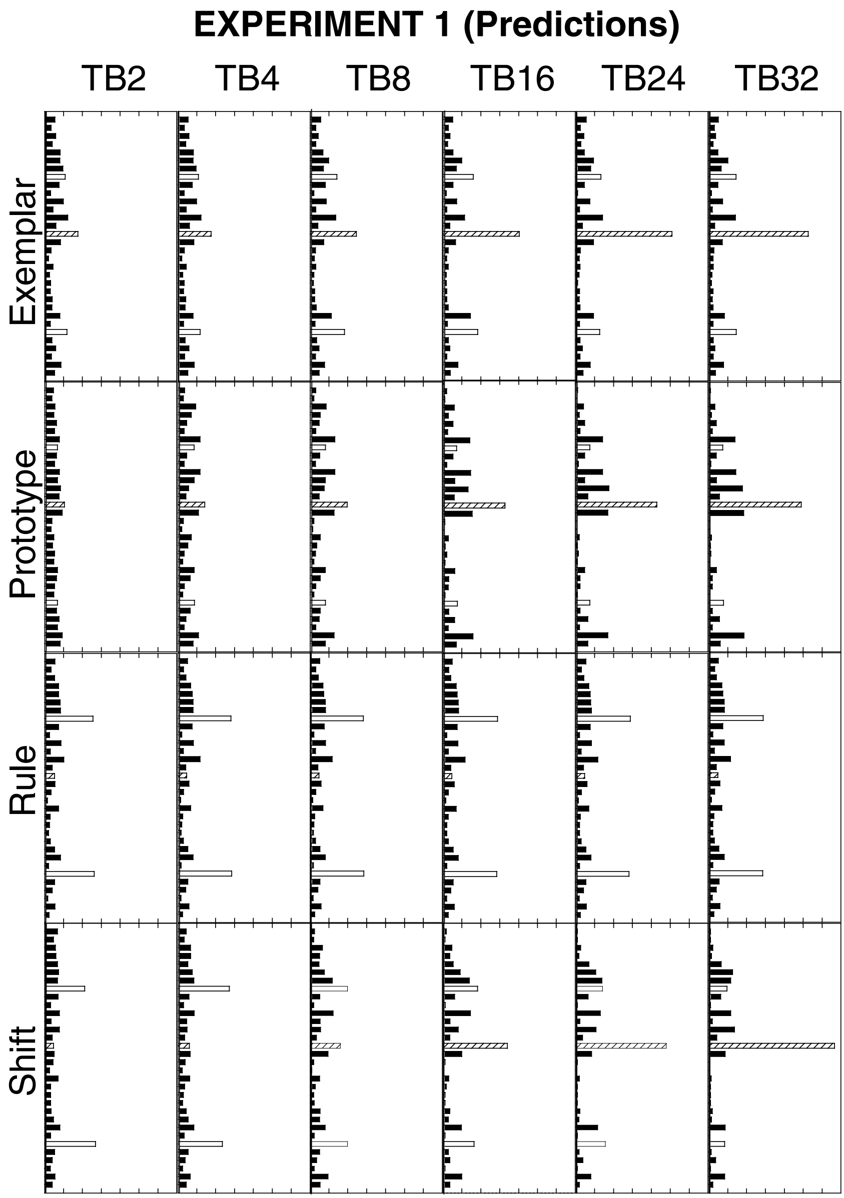

Fig. 8. Predicted distributions of generalization patterns from Experiment 1. The generalization patterns within each distribution are ordered the same as those shown in Fig. 2, so refer to that figure for generalization pattern labels. Each column contains predicted distributions as a function of transfer after 2, 4, 8, 16, 24, and 32 blocks of training. The four rows display the predicted distributions of generalization patterns for the exemplar, prototype, rule, and shift models, respectively.

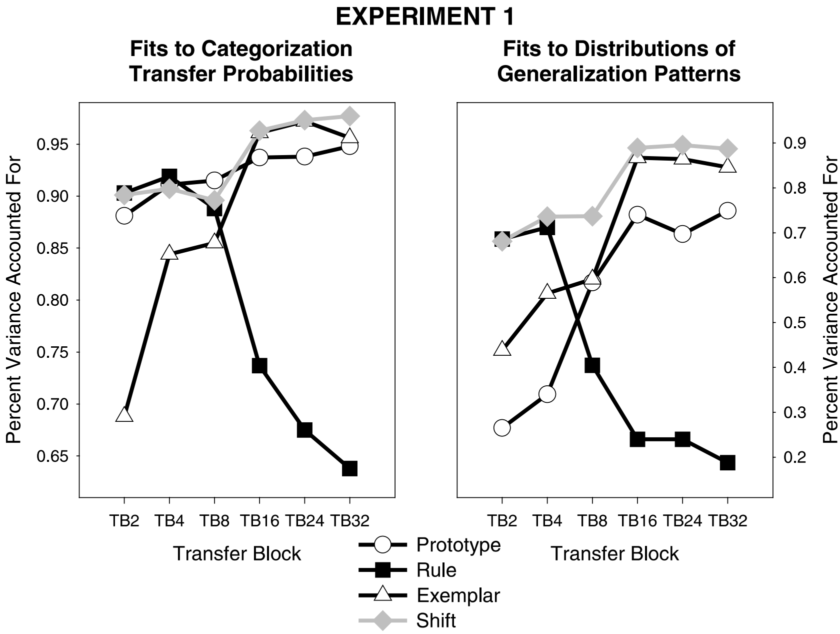

Fig. 9. Percent of variance accounted for in fitting the categorization transfer probabilities (left panel) and distributions of generalization patterns (right panel) as a function of transfer block for the prototype, rule, exemplar, and shift model in Experiment 1.

distribution data. As can be seen in Fig. 8, the rule model accounted for the two rule-based generalization peaks that were observed in this experiment early in training. However, in the later transfer blocks, the performance of the rule model declined rapidly; by the final transfer block, the rule model accounted for 64% of the variance in the average transfer data and only 19% of the variance in the distributions of generalization patterns. Not surprisingly, as can be seen in Fig. 8, the rule model completely failed to account for the prominent exemplar-based generalization peak later in learning. Consistent with our a priori hypotheses, the rule model provided a good account of the data early in training but provided a very poor account of the data later in training.

Mirroring the theoretical results for the rule model, the exemplar model performed poorly in the first three transfer blocks, accounting for only 69% of the transfer data and only 44% of the distribution of generalizations in the first transfer block. As can be seen in Fig. 8, the exemplar model completely failed to account for the prominent rule-based generalizations early in learning. However, in the last three transfer blocks, the exemplar model accounted for over 95% of the variance in the average transfer data and over 85% of the variance in the distributions of generalization patterns. As can be seen in Fig. 8, the exemplar model nicely accounted for the gradually increasing prominence of the exemplar-based generalization pattern as a function of learning. So, the performance of the exemplar model can be characterized as nearly opposite to that of the rule model; the exemplar model did a poor job of accounting for the observed data from the early transfer blocks but did an excellent job of accounting for the observed data from the later transfer blocks, again consistent with our a priori hypotheses.

Finally, for comparison, the performance of the prototype model was intermediate to that of the exemplar and rule models. The prototype model performed almost as well as the rule model in the first two transfer blocks in terms of accounting for the average transfer data, but it performed considerably worse than the simple rule model in accounting for the distributions of generalization patterns. As can be seen in Fig. 8, the prototype model was unable to account for the qualitative trends in the distributions of generalization patterns early in learning. In the later transfer blocks, the performance of the prototype model improved considerably, as shown in Fig. 9. Although the prototype model was able to qualitatively account for the distributions of generalization patterns, predicting the growing peak "exemplar-based" generalization pattern ABBBA, the overall quantitative fits to both the transfer data and the distributions were considerably worse than that of the exemplar model.

5.3.2. Theoretical accounts of Experiments 2 and 3

The results from the modeling of Experiments 2 and 3 converged with those of Experiment 1, so we combined the theoretical modeling of these two experiments into a single section. The best-fitting parameter values for the exemplar, prototype, and rule models for fits to all six transfer blocks in Experiments 2 and 3 are shown in Table 7. For Experiments 2 and 3, the predicted average transfer data for the three models are shown in Figs. 10 and 13, respectively, and the predicted distributions of generalization patterns are shown in Figs. 11 and 14, respectively. The summary fit statistics for the predicted average transfer data and for the distributions of generalization patterns as a function of transfer block are shown in Figs. 12 and 15, respectively.

For both experiments, the simple rule model provided a better account of both the average transfer data and the distributions of generalization patterns early in category learning than either the exemplar model or the prototype model. As shown in Figs. 11 and 14, the rule model successfully accounted for the prominent rule-based generalization patterns in the initial transfer blocks. However, as shown in Figs. 12 and 15, the fits of the rule model to the data from both experiments dropped precipitously in later transfer blocks, accounting for only around 10% of the variance in the distributions of generalization patterns by the last transfer block in both experiments. Not surprisingly, the rule model was completely unable to

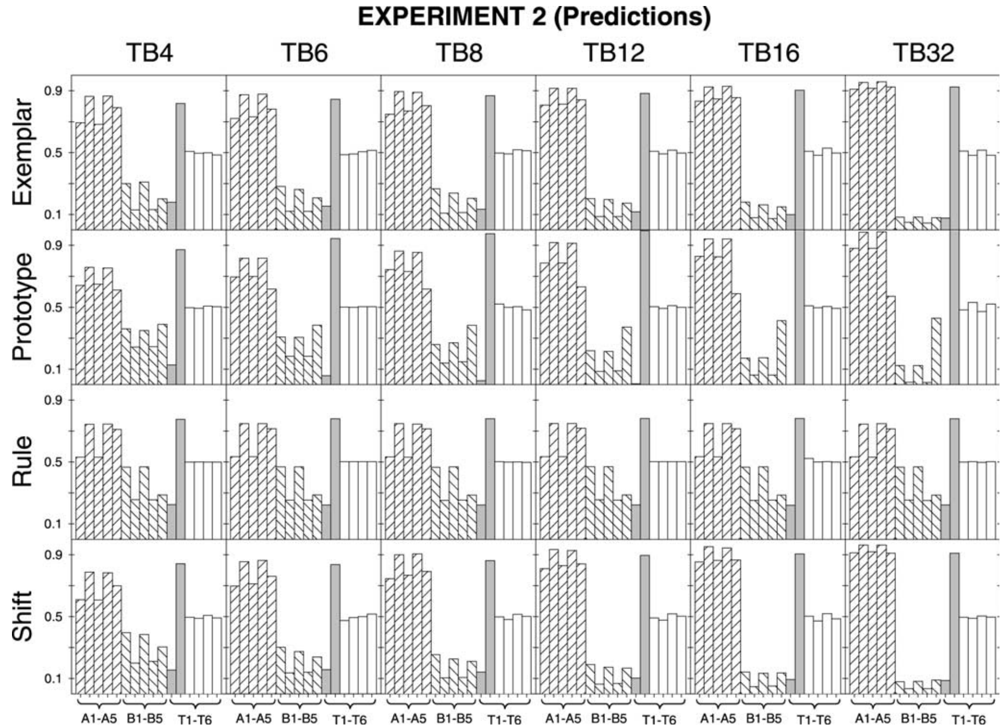

Fig. 10. Predicted average transfer probabilities from Experiment 2. Each column contains predicted transfer probabilities after 4, 6, 8, 12, 16, and 32 blocks of training. The four rows display the predicted probabilities for the exemplar, prototype, rule, and shift models. Hatched bars are training items, white bars are critical transfer items, and gray bars are noncritical transfer items.

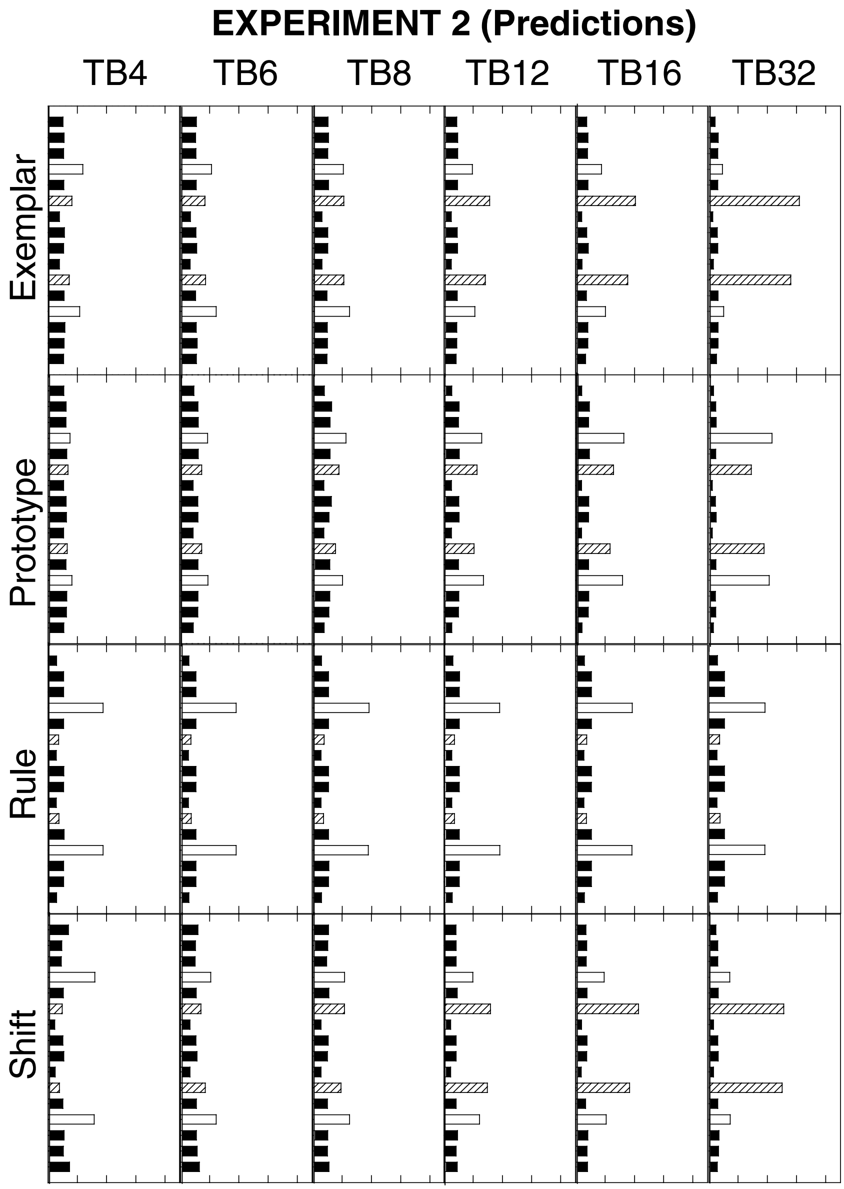

Fig. 11. Predicted distributions of generalization patterns from Experiment 2. The generalization patterns within each distribution are ordered the same as those shown in Fig. 4, so refer to that figure for generalization pattern labels. Each column contains predicted distributions as a function of transfer after 4, 6, 8, 12, 16, and 32 blocks of training. The four rows display the predicted distributions of generalization patterns for the exemplar, prototype, rule, and shift models, respectively.

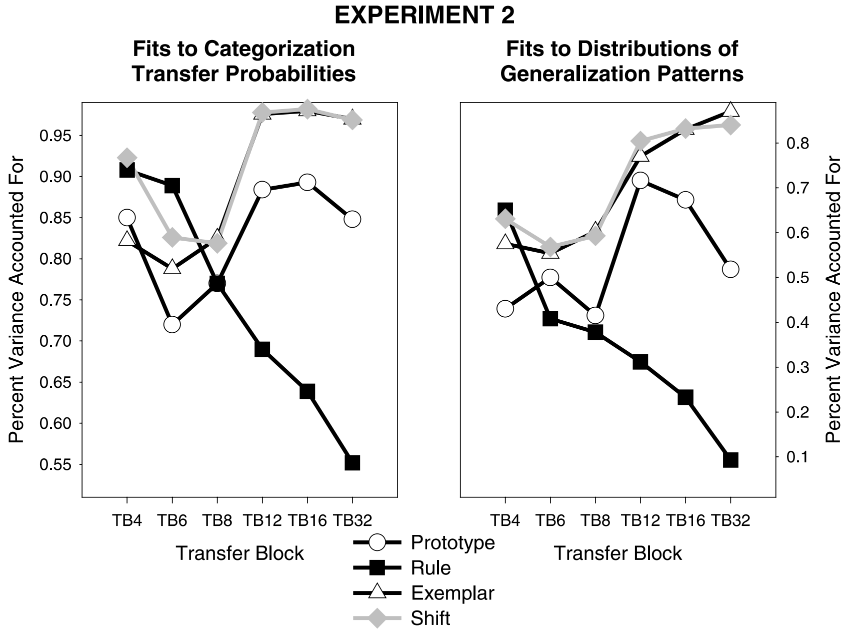

Fig. 12. Percent of variance accounted for in fitting the categorization transfer probabilities (left panel) and distributions of generalization patterns (right panel) as a function of transfer block for the prototype, rule, exemplar, and shift model in Experiment 2.

qualitatively account for the distributions of generalization patterns and also provided a very poor account of the average transfer data as well.

As in Experiment 1, the exemplar model provided a poor fit to data observed in the early transfer blocks compared to the rule model. Although the exemplar model did predict the rule generalizations to be somewhat prominent early in learning, the exemplar model systematically underpredicted the magnitude of these rule generalizations and somewhat overpredicted the magnitude of the exemplar generalizations for both experiments. For Experiment 3, even though the exemplar model provided a reasonable qualitative account of the distributions of generalizations observed in the early transfer blocks, it provided a very poor account of the average transfer data in the first transfer block of that experiment. By contrast, the exemplar model provided a very good account of both the average transfer data and the distributions of generalizations in the final three transfer blocks. As can be seen in Figs. 11 and 14, the exemplar model successfully accounted for the growing exemplar-based generalization peaks observed in both experiments. As we observed in the theoretical fits to Experiment 1, in fits to both Experiments 2 and 3, the exemplar model provided a relatively poor account of the data early in learning but provided a very good account of the data later in learning, consistent with our hypotheses.

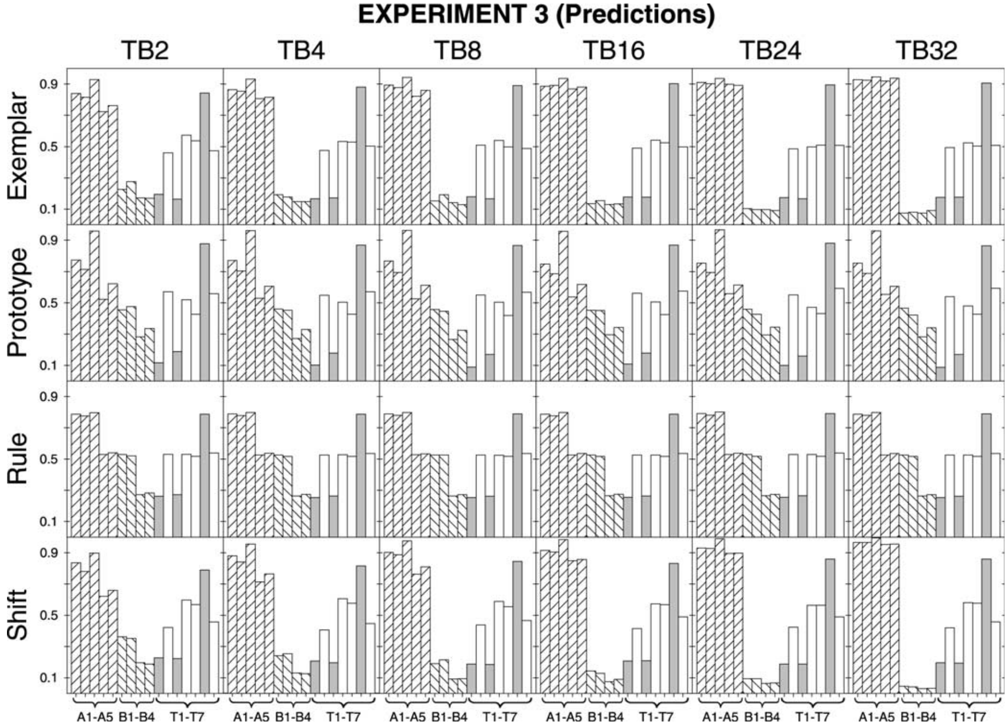

Fig. 13. Predicted average transfer probabilities from Experiment 3. Each column contains predicted transfer probabilities after 4, 8, 12, 16, 24, and 32 blocks of training. The four rows display the predicted probabilities for the exemplar, prototype, rule, and shift models. Hatched bars are training items, white bars are critical transfer items, and gray bars are noncritical transfer items.

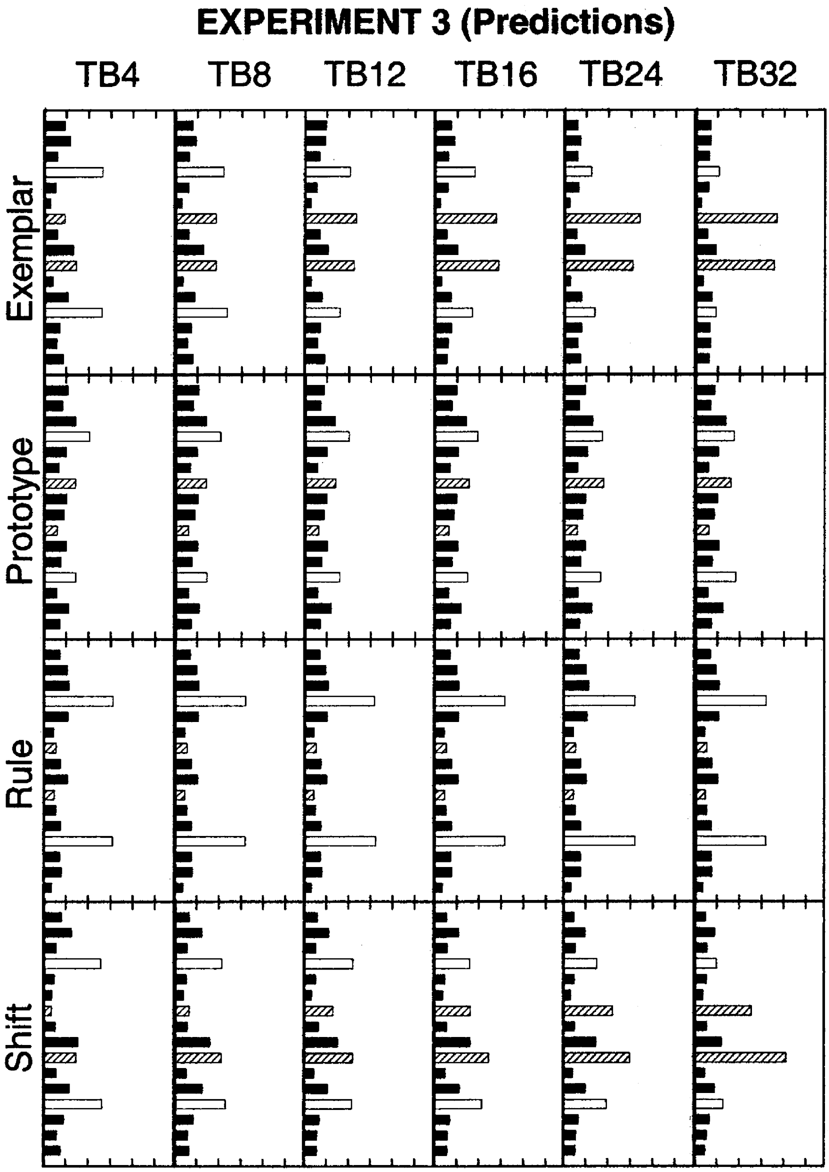

Fig. 14. Predicted distributions of generalization patterns from Experiment 3. The generalization patterns within each distribution are ordered the same as those shown in Fig. 6, so refer to that figure for generalization pattern labels. Each column contains observed and predicted distributions as a function of transfer after 4, 8, 12, 16, 24, and 32 blocks of training. The four rows display the predicted distributions of generalization patterns for the exemplar, prototype, rule, and shift models, respectively.

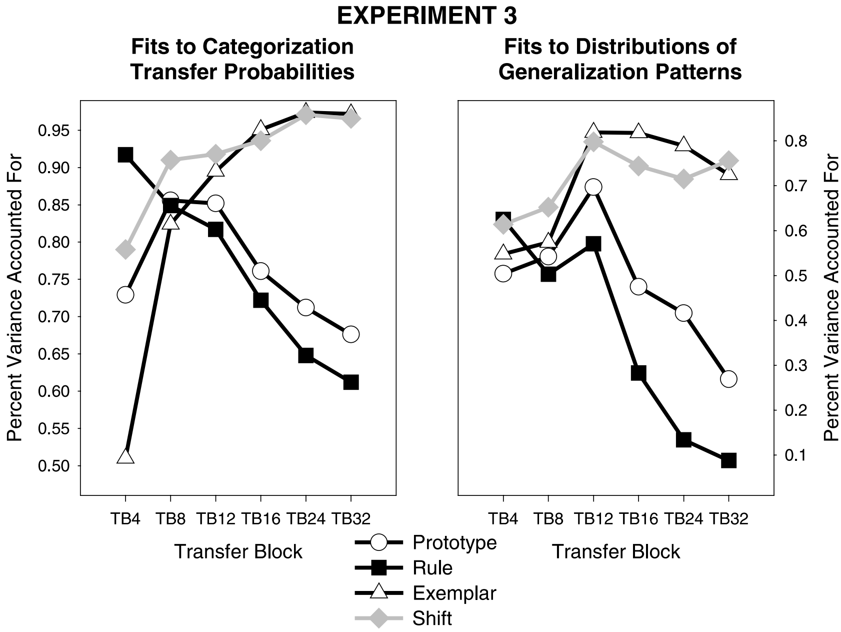

Fig. 15. Percent of variance accounted for in fitting the categorization transfer probabilities (left panel) and distributions of generalization patterns (right panel) as a function of transfer block for the prototype, rule, exemplar, and shift model in Experiment 3.

By contrast, the prototype model provided a relatively poor account compared to the rule model early in training and the exemplar model later in training. For both experiments, the prototype model systematically underpredicted the prominence of the rule-based generalization patterns early in learning. More problematic was that the prototype model could not account for the qualitative trends in the observed distributions of generalization patterns later in learning. For both experiments, the prototype model did not predict the so-called exemplar-based generalization patterns to become most prominent, but instead predicted the rule-based generalization patterns to remain prominent throughout learning. This poor performance of the prototype model for the data from experiment 3 was not improved by just fitting the model to the distributions of generalization patterns. Hence, even a prototype plus exemplar memorization model of the sort put forth by Smith and Minda (1998) would have difficulty accounting for our observed data since generalization to new items in their model is based solely on similarity to the category prototypes.

5.3.3. Theoretical accounts of individual transfer blocks

In the above theoretical modeling, we required each of the three models to attempt to account for the observed data throughout the entire course of category learning; in no case did any of the three models perform well throughout learning. Since we are testing the possibility that category representations may shift during learning, it seemed prudent to test the three models on their ability (or inability) to account for data from individual transfer blocks. Perhaps the poor performance of some of the models, most notably the prototype model, was a result of forcing the models to account for the full course of category learning instead of just that portion of category learning where that particular representational model may indeed dominate performance.

The prototype, rule, and exemplar versions of ALCOVE were individually fitted to just the first, the third, and the sixth transfer blocks of each experiment; the parameters of the models were allowed to vary freely in the fits to each block. The quality of the fits to the average transfer data and distributions of generalization patterns as a function of transfer block in each of the three experiments are shown in Fig. 16. Overall, the trends in these fits to individual transfer blocks were similar to the fits to the entire dataset. By the final block, the exemplar model provided a far better fit than the prototype or rule model; the advantage of the exemplar model over the prototype model was largely of a quantitative nature in Experiments 1 and 2, but the prototype model provided a poor qualitative and quantitative account of Experiment 3. The fit of the rule model was very good for the early transfer blocks and very poor for the later transfer blocks. The most notable difference between the individual-block fits and the overall fits is that the three models provided very similar qualitative and quantitative accounts of the first transfer block. This perhaps surprising result can easily be explained by examining the distribution of selective attention weights for individual simulated subjects within the exemplar and prototype frameworks—essentially, these two models were behaving very much like our simple rule model in that most of the learned selective attention was placed along a single dimension, with different simulated subjects maximally attending to different dimensions. To summarize, the model fits to individual transfer blocks from the three experiments provided further converging evidence for the use of simple single-dimension rules early in category learning and for the retrieval of exemplars later in category learning.

5.3.4. A representational shift model

Our results are consistent with the hypothesis that there is a representational shift in category learning from rules to exemplars. In the theoretical modeling described thus far, the rule model and the exemplar model were both instantiated within a common ALCOVE framework. Building on this previous theoretical modeling, in this section, we attempted to instantiate a simple representational shift model within the ALCOVE framework and test whether such a model could account for the entire set of observed data from the three experiments.

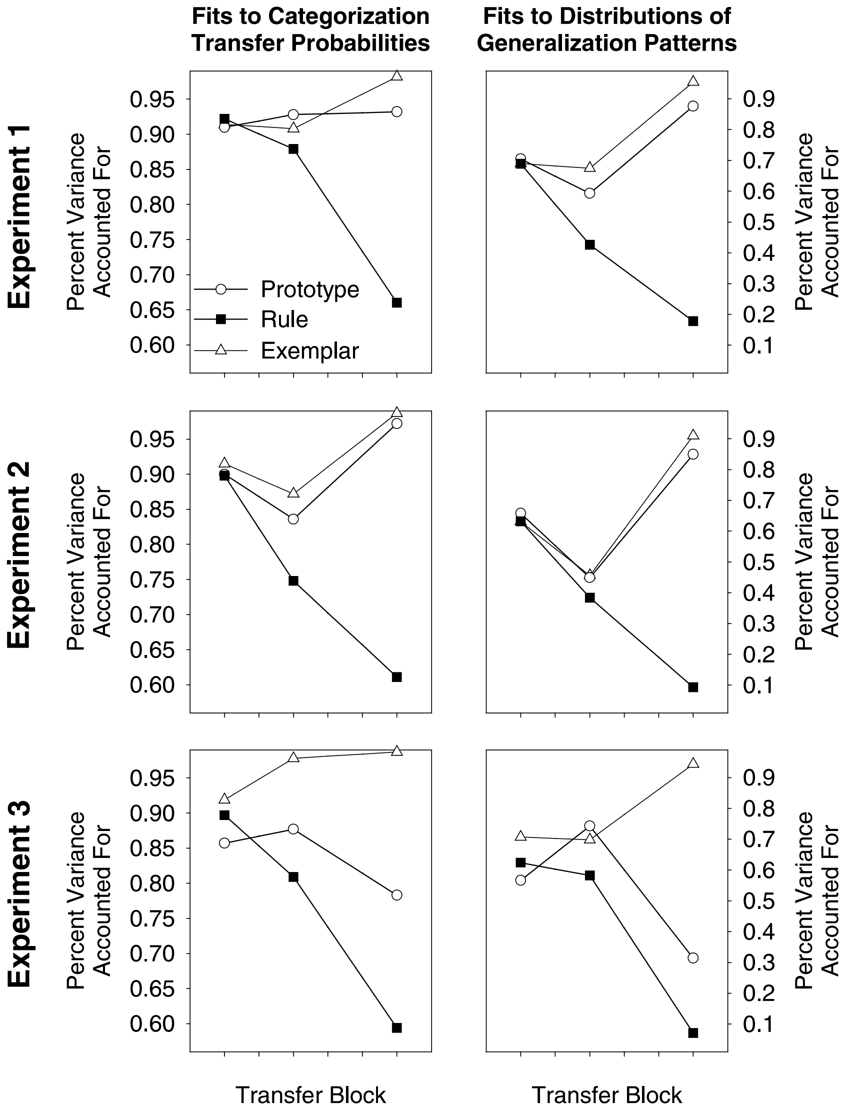

Fig. 16. Percent of variance accounted for in fitting the categorization transfer probabilities (left panel) and distributions of generalization patterns (right panel) from the first, third, and last transfer block for the prototype, rule, exemplar, and shift models in Experiments 1, 2, and 3. In these fits, the models were separately fitted to each individual transfer block.

There may be many ways of instantiating the hypothesized representational shifts within the ALCOVE framework. We chose a relatively simple approach of instantiating a shift model by gradually relaxing a winner-take-all allocation of dimensional attention over the course of training.

Early in learning, most attention is reallocated to the dimension with the greatest selective attention weight (forming something like a simple rule). Later in learning, dimensional attention is allowed to spread across multiple dimensions. As in the original ALCOVE model, dimensional selective attention weights were learned via gradient descent on error. In this representational shift model, attention weights were reallocated dynamically on every trial in such a way that the dimension with the highest attention weight is additionally allocated much of the attention that would have gone to the other dimensions. Formally, the reallocated attention on dimension N, α_N^*, is given by

$$\alpha_N^* = \frac{\alpha_N^\kappa}{\sum_m \alpha_m^\kappa}, \tag{5}$$

where α_m is the original attention weight on dimension m, and κ is an attention weight reallocation parameter; the new reallocated attention weights, α_m^*, were then used in computing similarities in Eq. (1). Large values of κ cause most of the attention to be reallocated to the single dimension with the most attention; with all of the attention allocated to a single dimension, the shift model reduces to the simple rule model we tested. Setting κ equal to 1 reduces the shift model to the original version of ALCOVE in which attention was allowed to spread across multiple dimensions. The initial value of the attention weight reallocation parameter, κ', was an additional free parameter of the model.

On every learning trial, in addition to adjusting the attention weights and the association weights, the attention weight reallocation parameter was incrementally annealed (e.g., Kruschke & Johansen, 1999) from its starting value κ' using the following schedule

$$\kappa_t = \kappa' \frac{1}{1 + \rho t} + 1, \tag{6}$$

where κ_t is the value of the attention weight reallocation parameter on trial t (the value of κ on the first trial is $\kappa' + 1$), and ρ is the annealing rate for that parameter. This has the effect of gradually shifting representation from rule-like when κ is high to exemplar-like when κ approaches one over training.

In fits to Experiments 1–3, the best-fitting parameter values for the shift model are shown in Table 7; the predicted average transfer data are shown in Figs. 7, 10, and 13; the predicted distributions of generalization patterns are shown in Figs. 8, 11, and 14; the summary fit statistics are shown in Figs. 9, 12, and 15. As expected, the shift model combined the successes of the rule model and the exemplar model. For each experiment, the shift model performed as well as the rule model early in learning and performed as well as the exemplar model later in learning. Examining the predicted distributions of generalizations for each experiment, we see that the shift model accurately predicted the early rule generalizations, which were gradually replaced by exemplar generalizations later in learning. It should be emphasized that this simple model used the number of training trials as the sole modulator of the shift from a single dimension to multiple dimensions; a more realistic model might also use some internal criterion to modulate that shift, such as categorization accuracy.

6. General discussion

For the general discussion, we will begin by summarizing the empirical and theoretical results we have reported in this paper. We will then move to a discussion of how to interpret our results with respect to the central question regarding representational shifts in category learning. We provide a discussion of the challenges of making claims about representation, referring to a body of literature suggesting different kinds of representation in categorization and related domains of human cognition.

6.1. Summary

In three experiments, subjects learned to categorize stimuli with feedback. We tracked how our subjects generalized their category knowledge by testing them on critical transfer items without feedback. Specifically, we examined how each individual subject categorized each of the transfer items to create a generalization pattern for that subject. Doing this for all subjects, we created a distribution of generalization patterns and tracked how the distributions evolved over learning. Summarizing across the three experiments, we observed systematic shifts in the distributions of generalization patterns as a function of learning, with different prominent generalizations appearing early in learning versus later in learning. The presence of these shifts was almost entirely masked when average categorization probabilities were examined rather than distributions of generalization patterns.

Our empirical results revealed systematic changes in how subjects generalized their category knowledge as a function of category learning. The first challenge confronting us was how best to characterize the differences between early generalizations and later generalizations. For each of the category structures subjects were trained on, two of the stimulus dimensions were highly diagnostic in that a particular value along that dimension tended to be associated with a single category; but these dimensions were not perfectly diagnostic in that there was an "exception" stimulus in each category. In all three experiments, the prominent generalizations observed early in learning were those that followed one of these highly diagnostic dimensions. Specifically, for some subjects, if a transfer item had the value along dimension one that was associated with category A then that item was classified as a member of category A, otherwise it was classified as a member of category B. For other subjects, if a transfer item had the value along dimension three that was associated with category A then that item was classified as a member of category A, otherwise it was classified as a member of category B.

Our empirical observations were based on examining the prominent generalizations in the distributions early in learning. But other generalizations were also observed and subjects also categorized training items as well. To test whether this single-dimension generalization hypothesis could account for the entire set of observed data, and not just the most prominent generalizations, we fitted a simple model in which all attention was allocated to a single dimension, with different simulated subjects attending to different dimensions in proportion to their diagnosticity. This highly restricted version of ALCOVE provided a surprisingly good account of the early learning data, including both the distributions of generalization patterns as well as the average categorization probabilities for all items. Thus, both the direct examination of the prominent generalizations combined with the model-based analysis provided strong evidence for single-dimension generalizations during the early stages of category learning. One reason these results are important is that an alternative hypothesis has recently been proposed that subjects initially may learn categories by forming prototypes (Smith & Minda, 1998; Smith, Murray, & Minda, 1997). We found little evidence for early prototype abstraction in our three experiments. Another reason why these results are important is that they may bear on the issue of a rules to exemplars shift in the development of automaticity in categorization, a point we will return to later in this discussion.

A more challenging question is how to characterize the generalizations observed later in learning. In all three experiments, the prominent generalizations observed later in learning were clearly inconsistent with generalization along a single diagnostic dimension. RULEX (Nosofsky et al., 1994) assumes that subjects form simple, typically single-dimension, rules. But these rules can be supplemented by the probabilistic storage of exceptions. Coupling rules and memory for exceptions, RULEX can predict a wide variety of generalizations, not just those based on a single dimension. However, RULEX in no way predicts, a priori, any prominence of the particular generalizations that we observed to be most prominent. That is, although the prominent generalizations we observed at the end of learning might be consistent with some particular hand-selected set of rules and probabilistically stored exceptions, the process by which RULEX learns rules and probabilistically stores exceptions does not predict these particular selections of rules and exceptions to be any more likely than many other possible selections of rules and exceptions. Thus, the rule-plus-exception process formalized in RULEX does not provide a viable candidate for explaining the prominent generalizations we observed at the end of learning.

We also considered the possibility that the prominent generalizations might be consistent with prototype abstraction. Although the multiplicative prototype model we investigated could qualitatively account for the prominent generalizations observed in the first two experiments, it was unable to account for the prominent generalizations observed in the third experiment. Thus, prototype abstraction also does not provide a viable candidate for explaining the generalizations observed at the end of learning either.

Instead, we argue that the prominent generalizations observed later in learning are consistent with exemplar-based categorization. First, an exemplar model, ALCOVE, provided excellent accounts of both the distributions of generalization patterns and the average categorization probabilities at the end of learning, providing a far better account than the prototype and rule models. However, one reasonable question emerging from this model-based analysis is whether an exemplar model is simply a sophisticated curve-fitting algorithm that can fit any pattern of data or whether it can actually predict, in a priori manner, the prominence of particular generalizations over other ones. To address this, we systematically examined the predictions of two exemplar models (see Appendix A), the context model and ALCOVE, across a wide range of their possible parameters. First, both exemplar models never predicted many of the possible generalizations to ever be most prominent in their predicted distributions of generalization patterns. Moreover, under conditions that maximized the accuracy with which training items are classified into their respective categories, both exemplar models predicted particular generalization patterns to be most prominent. These were the same generalizations that we observed to be most prominent by the end of learning in all three experiments. Although these model-based analyses cannot unequivocally prove that generalizations in the later stages of learning are based on exemplar retrieval, they are certainly consistent with the predictions of formalized exemplar models of categorization.

To summarize, across three experiments, we observed shifts in how subjects generalized their learned category knowledge. Early in learning, subjects appeared to generalize on the basis of single diagnostic dimensions. Later in learning, subjects generalized on the basis of multiple dimensions in a manner consistent with the use of exemplar-based category knowledge. We believe our results provide strong evidence for shifts in the information that is used to categorize stimuli as a function of learning. We will soon turn to the more controversial issue of whether these results reveal something about shifts in the representations and processes used to categorize stimuli as a function of learning: Is there a rules-to-exemplars shift in category learning as suggested by some theories of the development of automaticity? But we first need to review some of the extant evidence for a distinction between rule-based and exemplar-based categorization more generally.

6.2. Rules and exemplars in categorization

The question of whether human cognition is rule-based or similarity-based predates modern psychology and continues to be a topic of great interest and debate. In a recent review of the empirical case for two systems of human reasoning, Sloman (1996) contrasted one mode of cognition that seems largely symbolic and logical, based on the use of abstract rules, with one that is largely associative, of which exemplar-similarity is one possible manifestation. To motivate the distinction between these two modes of cognition, Sloman suggested that "sometimes conclusions simply appear at some level of awareness, as if the mind goes off, does some work, and then comes back with the result, and sometimes coming to a conclusion requires doing the work oneself, making an effort to construct a chain of reasoning . . ." (1996, p. 3). Intuitively, these different modes of cognition seem to characterize different ways in which people categorize objects as well. Indeed, one motivation for the present work was our own intuitions about how we learned many perceptual categories ourselves. When first learning categories, we seemed to begin by testing hypotheses about various explicit rules for categorizing items. As training continued, our subjective impression changed to one of "knowing" which category an item belonged in without needing to make recourse to those explicit rules. These impressions have been confirmed by informal interviews we have conducted with subjects across a variety of categorization paradigms over the years. Although these intuitions may be useful for generating hypotheses regarding the various bases for human cognition, they constitute exceedingly weak evidence for the existence of two different systems underlying human cognition. Indeed, awareness itself may provide a fallible heuristic for telling apart rule-based from similarity-based cognition (e.g., Shanks & St. Johns, 1994). Fortunately, we can marshal some empirical and theoretical evidence for differences between presumably rule-based and similarity-based categorization. In this section we will review some of this evidence across a variety of experimental domains. We will then return to the specific issue of whether there may be shifts from rule-based to exemplar-based categorization with experience.

6.2.1. Empirical evidence for rules versus exemplar similarity

One classic paradigm that was used to distinguish between apparent rule-based categorization and similarly based categorization involved a simple sorting task. In this task, three stimuli (A–C) vary along two separable dimensions (e.g., color and size). Stimuli A and B match along one dimension but differ considerably along the other dimension, whereas stimuli B and C are similar along both dimensions but match along neither dimension. When asked which two stimuli "go together," children sort B and C together, presumably because they are perceptually more similar to one another along both dimensions. By contrast, adults sort A and B together, possibly because they are using an analytic rule based on dimensional identity, irrespective of overall similarly (e.g, Smith & Kemler, 1977). Adults can be driven to sort based on overall similarity, rather than a dimensional rule, by requiring fast responses or by providing a concurrent task load (e.g., Smith, 1981; Smith & Kemler Nelson, 1984; Ward, 1983). One interpretation of these results has been that sorting by similarity is the default mode of processing that is overridden by a presumably slower and more attention-demanding rule-based mode. Forcing rapid responses and providing a concurrent task have been claimed to disrupt this analytic mode of sorting.

This bias by adults to sort on the basis of a single analytic dimension has been further investigated using richer stimulus sets in other experimental paradigms. Across a number of experiments, Ahn and Medin (1992) provided subjects with a large number of stimuli with discrete features that could be sorted according to their family resemblance, creating clusters of stimuli that differed somewhat from unseen prototype stimuli. Although this similarity-based sorting was available to subjects, most instead sorted stimuli on the basis of a single diagnostic dimension, producing sorts that violated the family resemblance structure embedded within the stimulus set. Regehr and Brooks (1995) followed up on these perhaps surprising results using other stimuli and other sorting methods, showing that this bias to sort on the basis of a single diagnostic dimension was quite resilient. Ashby et al. (1999) further investigated this phenomena by presenting subjects with continuous-dimension stimuli sampled from two bivariate normal distributions. Although the stimuli could be sorted into two clear clusters based on overall similarity, subjects instead tended to sort on the basis of a single diagnostic dimension. As with the Ahn and Medin (1992) results, the resulting sorts fractured the similarity-based family resemblance structure embedded within the stimulus set.

Together, these sorting studies suggest that people approach the task of creating categories through unsupervised sorting by adopting an analytic, perhaps rule-based, strategy of sorting on the basis of a single diagnostic dimension. Perhaps these results may help explain why subjects in our experiments seem to approach the task of learning categories through explicit supervision with corrective feedback through a similar analytic strategy of searching for simple rules that can distinguish members of one category from members of another category.

Further evidence for a distinction between apparent rule-based and exemplar-based categorization comes from a recent study by Waldron and Ashby

(2001; but see Nosofsky & Kruschke, 2001).[4] They had subjects learn categories either defined by a single-dimension rule or by family resemblance along multiple dimensions. Subjects either learned categories under a concurrent task load (performing a concurrent numeric Stroop task) or under no task load. The load manipulation had little effect on learning categories defined by family resemblance but had a large effect on learning categories defined by a simple rule. As suggested by some of the sorting results described earlier, the hypothesis testing (or selective attention) required to learn single-dimension, rule-based categories seems to require resources of some sort that are tapped by this concurrent load manipulation.

Thomas (1998) also provided evidence for a distinction between rule-based and similarity-based categorization using a feature prediction task. She had subjects learn two categories with feedback that were defined by bivariate normal distributions. After learning, she presented subjects with stimuli possessing just one dimension and asked them to predict the value of the missing dimension. Some subjects were able to predict the missing dimension in a manner consistent with the underlying category distributions, suggesting that they had learned the categories by remembering information about the statistical properties of the category distributions (perhaps by storing exemplars). Other subjects, while having also learned the categories, could not predict the missing stimulus dimensions. Presumably, these subjects had learned the categories by forming simple single-dimension rules, without storing information about the individual category members, thereby being unable to predict missing stimulus dimensions. Although these results do not show evidence for use of both rules and exemplars within an individual, they do suggest that two different modes of categorizing objects, by similarity or by rules, may exist, with some subjects showing a preference for one learning mode over another.

Other studies have shown that when subjects are explicitly provided a complex categorization rule, similarity to stored examples may still exert considerable influence. Brooks and colleagues (Allen & Brooks, 1991; Regehr & Brooks, 1995) provided subjects with a complex multidimensional rule for categorizing stimuli into one of two categories. After given experience applying this rule to training items, subjects were asked to classify

[4] Ashby and colleagues (Ashby et al., 1998; Ashby & Waldron, 1999) have distinguished between verbal (rule-based) and implicit (procedural-memory-based) category learning systems, not rule-based and exemplar-based category learning systems. Although Ashby and colleagues have dismissed exemplar models as viable candidates for the "implicit" category learning system, the nonparametric procedural learning model proposed by Ashby and Waldron (1999) is extremely similar to both a "covering map" version of ALCOVE (Kruschke, 1992) and the rational model (Anderson, 1990, 2001). Both of these models are extremely similar to exemplar models and are formally identical to exemplar models under certain conditions (see Nosofsky, 1991).

transfer items without feedback. Brooks and colleagues observed that subjects showed a sensitivity to similarity to observed examples that interacted with how well those examples followed the categorization rule. Although "good" and "bad" transfer items followed the categorization rule equally well, subjects were more likely and more rapid to classify the "good" examples as members of a category because they were more similar to the training examples. Thus, even when subjects are supplied an explicit categorization rule, similarity to examples exerts an influence. In a more realistic experimental setting, Brooks, Norman, and Allen (1991) also showed that expert dermatologists, who presumably could make recourse to complex rules for categorizing various skin disorders, showed systematic influences of similarity to previously viewed cases.

Finally, the distinction between rule-based and exemplar-based categorization has played a central role in understanding the task of artificial grammar learning. Originating in classic studies by Reber (1967, 1969), subjects view letter strings created using a complex finite state grammar. The grammar specifies the rules for initial letters and all subsequent letters of a string. These grammars are complex and their rules cannot be easily verbalized. In a typical experiment, subjects view a series of strings without being told that they were generated by rules of any kind. Often they are simply told to remember the strings for a later memory test. After viewing the strings, subjects are told that the strings were generated by a complex grammar and are asked to discriminate new grammatical strings from ungrammatical strings, which subjects can do better than chance. A central question in this literature is whether subjects learn the strings by forming complex implicit rules or by remembering exemplars (or exemplar fragments). The literature surrounding this task is vast, and the issues surrounding it are controversial, but the emerging picture is that people use both simple explicit rules and similarity to exemplars (or exemplar fragments) to make their grammaticality judgments (e.g., Johnstone & Shanks, 2001).

6.2.2. Neuropsychological evidence for rules versus exemplar similarity

There is also some emerging evidence for a distinction between rule-based and exemplar-based categorization in studies testing neuropsychological patient populations and in studies using functional brain imaging. For example, Smith, Tracy, and Murray (1993), compared depressed and nondepressed individuals on two different category learning tasks. As in the study by Waldron and Ashby (2001), described earlier, subjects either learned categories that had a family resemblance structure or learned categories that were defined by a perfect single-dimension rule. Depressed subjects were significantly impaired at learning categories defined by a simple rule, but were not impaired at learning categories defined by family resemblance. Presumably, hypothesis-testing strategies required to learn rule-defined categories were selectively impaired by the biochemical, structural, or strategic changes underlying depression, which may include changes in the operation of prefrontal working memory areas. Some further evidence for this conjecture is that depressed individuals also show deficits on the Wisconsin Card Sorting task (e.g., Franke et al., 1993), a test that requires verbal rule following and rule switching.

Parkinson's disease appears to lead to the impaired functioning of the basal ganglia, prefrontal cortex, and the anterior cingulate, all areas that may be critical for rule-based categorization (see Ashby et al., 1998). Indeed, Parkinson's patients are impaired at the Wisconsin Card Sorting task (Brown & Marsden, 1988). Furthermore, Parkinson's patients have also been shown to be impaired at learning categories defined by probabilistic cues (Knowlton, Mangels, & Squire, 1996). Although one interpretation of this deficit has been one of impaired probabilistic learning, this task may also require hypothesis testing in that the probabilistic cues that were used in the task varied considerably in their individual diagnosticity (Flanery & Palmeri, 2001). By contrast, Parkinson's patients are not impaired at category learning tasks that appear not to demand hypothesis testing, such as dot pattern categorization and artificial grammar learning (Reber & Squire, 1999). Such neuropsychological studies suggest that there may be different brain areas critical for analytic rule-based categorization and for more "implicit" exemplar-based categorization, though they do not necessarily imply that rule-based and exemplar-based systems are independent, nor that they even need to be separate systems.

Some recent functional brain imaging studies have also suggested differences between rule-based and exemplar-based categorization. Smith et al. (1998) provided subjects with complex categorization rules in an extension of the Brooks paradigm described earlier (Allen & Brooks, 1991). They used subjects' performance on critical transfer items to separate "rule-based" and "exemplar-based" categorizers; the "exemplar-based" categorizers were those that showed more pronounced effects of similarly on categorization of "good" transfer items in the Brooks paradigm. Using positron emission tomography (PET), they localized a variety of brain areas that were significantly activated for the rule-based categorizers but not for the exemplar-based categorizers. One of the prominent areas of activation was a parietal lobe area thought to be involved in critical aspects of selective attention, a process that may be more critical for explicit rule use than exemplar retrieval. In addition, areas of prefrontal cortex thought to be involved in working memory were also active for rule-based categorizers.

Flanery and Palmeri (2001) used functional magnetic resonance imaging (fMRI) to compare brain activity elicited by two classic categorization paradigms that have been used to test brain damaged individuals. Within a single fMRI run, subjects categorized dot patterns with feedback (adapting the paradigm originally used by Knowlton & Squire, 1993), they categorized probabilistic cues with feedback (adapting the paradigm originally used by Knowlton et al., 1996) and were shown a series of appropriate controls and baseline stimuli. In one experiment, all categories were learned within the fMRI run, referred to as early learning trials. In another experiment, all categories were prelearned prior to the fMRI run, referred to as later learning trials. The results most relevant to the present discussion were the presence of significant activation of both the anterior cingulate and the caudate of the basal ganglia during the early learning trials of probabilistic cues, but not dot patterns. Again, recall that these are two of the critical brain areas that have been argued to play an important role in rule abstraction and rule use during category learning (Ashby et al., 1998). Thus, these fMRI results suggest that the deficits shown by Parkinson's patients in the probabilistic task may not necessarily emerge from the probabilistic nature of the task, as argued by Knowlton et al. (1996), but rather because the task may require hypothesis testing. This possibility is under exploration in current brain imaging studies.

To summarize, results emerging from neuropsychological studies, including patient studies and functional brain imaging studies, also suggest that there may be differences between rule-based and exemplar-based categorization. Although a typical interpretation of these neuropsychological results is that these are functionally independent neural systems, we prefer the far more conservative interpretation that they just indicate some kind of anatomical distinction between the demands of apparently rule-based and exemplar-based categorization tasks. Whether these are independent systems, interacting systems, or a single system with dissociable components is an issue that requires further investigation (Palmeri & Flanery, 2002), as we will elaborate later.

6.2.3. Theoretical evidence for rules versus exemplar similarity

This brief review has highlighted some of the empirical differences between apparent rule-based and similarity-based categorization. But to truly say that some aspect of categorization behavior is based on rules or based on similarity to exemplars is a statement about particular kinds of mental representation. However, neither behavior of normal subjects, nor behavior of neuropsychological patients, nor patterns of brain activity can truly reveal the contents of the mental representation of categories. Functional brain imaging reveals where information may be processed in the human brain, but cannot directly reveal how that information is represented nor what processes act on that represented information. And observed behavior is necessarily a combination of representations and the processes that act upon those representations. No behavioral study can uniquely reveal representation without process (see Barsalou, 1990). Indeed, Barsalou has argued that on the basis of behavioral data alone "trying to determine whether people use exemplars or abstracted representations is futile" (p. 62). That said, although patterns of behavior by themselves cannot be used to conclude whether people use rules or exemplars, we can test formal mathematical and computational models that combine specific representations and processes on how well (or how poorly) they can account for observed behavior. "Perhaps, the best empirical research can do is to test particular models of each kind, not 'rules' and 'similarity' generally" (Hahn & Chater, 1998, p. 199). "We can only conclude that particular *models* (i.e., representation-process pairs) are either supported or rejected" (Barsalou, 1990, p. 63).

Representations and processes underlying exemplar-based categorization have been reasonably well specified. This paper has reviewed a series of related models based on relatively low-dimensional spatial representations (e.g., Kruschke, 1992; Medin & Schaffer, 1978; Nosofsky, 1984); there are also classes of exemplar models based on high-dimensional feature representations as well (e.g., Hintzman, 1986). Rule-based models constitute a far broader class of potential theories, so we need to narrow the scope of possible candidates. Without specifying a clear operational definition for what can count as a categorization rule, a "rule" could include just about any method for categorizing items. Indeed, categorizing according to exemplar similarity could be characterized by the following rather unrule-like rule: "put item i in category J if it is more similar to exemplars of category J than it is to exemplars of any other category." Therefore, in order to make the distinction between rule-based and exemplar-based categorization at all meaningful, we need to preclude such a broad, all-encompassing scope for possible categorization rules (see Hahn & Chater, 1998; Nosofsky et al., 1989; Sloman, 1996).

One simple operational definition of rule-based categorization is that the rules are verbalizable, unlike exemplar-based and other more "implicit" modes of categorization (e.g., Ashby et al., 1998). The use of verbal rules seems to characterize the explicit intuitions people have when first learning categories. The functional imaging and neuropsychological studies described above seem to implicate a variety of frontal brain areas involved in verbal rule selection and rule switching (see Ashby et al., 1998; Waldron & Ashby, 2001). Verbalization also emerges as one potential criterion for rule-based reasoning more generally (e.g., Smith et al., 1992). Although verbalization is not directly instantiated within most formal models of rule-based categorization, this verbalization criterion seems to have had an indirect influence on how these models have been formalized. Rules, unlike exemplars, are typically assumed to rely on just a limited number of stimulus dimensions, perhaps because people can only verbalize limited combinations of features/dimensions. For example, the original RULEX model (Nosofsky et al., 1994) assumes that people begin category learning by testing various single-dimension rules, but then attempt to learn conjunctive rules if single-dimension rules fail. The continuous-dimension version of RULEX (Nosofsky & Palmeri, 1998) also emphasizes single-dimension rules. Rule modules in ATRIUM (Erickson & Kruschke, 1998) and COVIS (Ashby et al., 1998) assume single-dimension rules as well. Specifically, a rule in ATRIUM and COVIS is operationally defined as a boundary that is orthogonal to a single psychological dimension (e.g., "large objects are A's, and small objects are B's," or "squares are A's, and circles are B's"). Clearly, such unidimensional rules can be verbalized, unlike the "rules" that may underlie exemplar-based categorization.

Numerous architectures combining such rule-based and exemplar-based representations have been proposed recently. Comparisons between many of these various models have been more fully reported in other papers (e.g., Ashby et al., 1998; Erickson & Kruschke, 1998; Nosofsky & Palmeri, 1998), so we will just briefly outline some of them here. At one extreme are models that posit functionally independent rule-based and exemplar-based systems that race to completion (e.g., Logan, 1988; Palmeri, 1997); exemplar-based representations gain strength with repeated exposure to category exemplars and eventually win the race. Alternatively, rule-based and exemplar-based modules may be functionally independent, but the outputs of these systems may compete based on strength of evidence rather than completion time (e.g., Ashby et al., 1998). Erickson and Kruschke (1998) proposed a related architecture (ATRIUM) in which there are separate rule-based and exemplar-based modules that compete, but the exemplar information serves to gate the relative contribution of those individual modules; in this way, the model learns whether rule-based or exemplar-based information should be used to select the appropriate response for a particular category instance. Finally, there have been architectures proposed that combine rules, exemplars, and perhaps other representations within a single representational system; these include the parallel rule activation and synthesis model (PRAS) of Vandierendonck (1995), the SUSTAIN model of Love et al. (in press), and the ACT-R framework of Anderson and Betz (2001). Anderson and Betz specifically implemented the exemplar-based EBRW model of Nosofsky and Palmeri (1997) and the rule-based RULEX model of Nosofsky et al. (1994) within the ACT-R framework. Following Nosofsky and Palmeri (1997), this implementation assumed that exemplar representations became stronger with repeated experience with category instances and eventually dominated categorization performance. Indeed, Anderson and Betz suggested that this ACT-R implementation could account for the shifts from rules to exemplars that we reported in this paper. Although the architectures of these mixed models vary considerably, the conclusions of the papers summarizing these models each argue for mixed category representations combining rules and exemplars rather than a single representational system (see, however, Nosofsky & Johansen, 2000).

For most of these formal models, the distinction between "rules" and "exemplars" is probably best characterized as describing different contents of category knowledge rather than qualitatively different kinds of representational media. The models differ in their assumptions about where rule-based and exemplar-based information is stored—distinct modules (e.g., ATRIUM and COVIS) or a single module (e.g., SUSTAIN, RULEX, PRAS, and ACT-R)—but the representational medium for rules and for exemplars are typically quite similar within any particular model. Generally, rules are more abstract than exemplars, and rules may require a stricter matching criteria than exemplars (e.g., Hahn & Chater, 1998). For example, RULEX (Nosofsky et al., 1994) assumes a combination of rules and memory for exceptions. The notation $1 * ** \rightarrow A$ would indicate a single-dimension rule, the notation $2 * *1 \rightarrow A$ would indicate a conjunctive rule, and the notation $2 * 12 \rightarrow A$ would indicate a memorized exception (where $*$ matches any stimulus value). Although rules and memory for exceptions play a distinct conceptual role in RULEX, the primary difference between them is their level of abstraction, not a fundamental difference in their underlying representational medium. In the PRAS model (Vandierendonck, 1995), rules are rectangular regions in psychological space whereas exemplars are individual points in psychological space. In the limit, a very small rule region becomes formally indistinguishable from a point exemplar representation, thus rules and exemplars differ quantitatively not qualitatively. SUSTAIN (Love et al., in press) can form clusters based on single dimensions (rules) or based on multiple dimensions (prototypes or exemplars). Again the distinction is based on the number of dimensions utilized in creating a category representation, not a fundamental difference in how information is essentially represented. In the ACT-R framework of Anderson and Betz (2001), rules in the RULEX module and exemplars in the EBRW module are both represented as chunks within the same declarative memory with various production rules deciding which module to execute on a given categorization trial. As in RULEX, rules are one-dimensional mappings from stimulus features to categories. As in EBRW, exemplars are multidimensional mappings from stimulus representations to categories. In COVIS, a verbal rule partitions psychological space with a linear decision boundary orthogonal to a single psychological dimension. Implicit categorizations permit partitions with decision boundaries that can have any orientation in psychological space and can include nonlinear boundaries. Thus "rules" and "implicit" representations differ quantitatively, not qualitatively. Similarly, when ATRIUM is applied to binary-valued stimuli, such as we used in the present experiments, rules are single-dimension abstractions whereas exemplars are multidimensional representations. As described below, our distinction between rules and exemplars similarly follows the approach used by other investigators: Rule-based category representations are single-dimension abstractions and exemplar-based category representations are multidimensional category instances.

6.3. Is there a rules to exemplars shift in categorization?

Rather than commit ourselves to a particular functional architecture of competing, interacting, or mixed representations of rules and exemplars, we instead chose to examine the use of single-dimension rules versus multidimensional exemplars within a single system based on the well-known ALCOVE model. In our theoretical modeling, we implemented rule-based categorization in individual simulated subjects by forcing a version of the ALCOVE model to selectively attend to a single psychological dimension. We also implemented a hybrid model that shifted from more rule-like to more exemplar-like representation by gradually unrestricting selective attention away from a single dimension over the course of learning. In this way, our theoretical modeling seems to imply a single system that is constrained in various ways to behave in a more "rule-like" or more "exemplar-like" manner at different stages of category learning. Yet, irrespective of our particular approach to modeling, our description of rule-based and exemplar-based categorization at several points in this paper may seem to imply quasi-independent processing modules. So, are rules and exemplars largely independent systems (Ashby et al., 1998; Erickson & Kruschke, 1998; Palmeri, 1997) or are they manifestations of the same underlying system (Nosofsky & Johansen, 2000)?

Although our results demonstrate more rule-like and more exemplar-like behavior at different stages in category learning, our results probably cannot be used to unequivocally decide the relative functional independence or dependence of rule-based and exemplar-based representations and processes in a theoretically neutral manner. Indeed, on the basis of behavioral data alone, it may be quite difficult to completely distinguish between multiple- and single-system accounts (e.g., Nosofsky & Johansen, 2000; Nosofsky & Zaki, 1998). For example, Nosofsky and Johansen showed that much of the evidence for multiple systems for categorization (e.g., Ashby et al., 1998; Erickson & Kruschke, 1998; Knowlton & Squire, 1993; Nosofsky et al., 1994; Smith & Minda, 1998) could be given an alternative interpretation in terms of a plausibly elaborated single-system exemplar model. As a particularly relevant example, they reexamined the observed distributions of generalization patterns reported by Nosofsky et al. (see Fig. 1). Nosofsky and Johansen extended the generalized context model (Nosofsky, 1984, 1986) to allow four subgroups of "rule-based" subjects that each primarily attended to just one of the four stimulus dimensions and another "exemplar-based" subgroup that optimally attended to all of the stimulus dimensions (Nosofsky, 1984, 1998a,b); this relatively high-parameter static model of categorization is similar in some important respects to the shift model of category learning that we investigated in the theoretical results section. Their elaborated version of the generalized context model with five subgroups performed as well as RULEX and much better than the simple context model in accounting for the observed distributions of generalization patterns.

Both this elaborated version of the generalized context model and our elaborated version of ALCOVE (the shift model) assume that subgroups of subjects largely attend to single stimulus dimensions. Both models assume the formation of simple single-dimension categorization rules, according to our operational definition of a rule. But what causes individual subjects to decide to attend to just a single dimension when the optimal categorization strategy is to divide attention across multiple dimensions? Arguably, a hallmark theoretical assumption of the generalized context model (Nosofsky, 1986) is that selective attention weights are eventually set by subjects in such a way so as to optimize categorization performance. The optimality assumption has been verified in a variety of experiments (e.g., Nosofsky, 1984, 1986, 1998a,b). And we observed in the present experiments that when subjects are supplied sufficient categorization training, they show generalization patterns that are consistent with a more optimal distribution of attention weights across multiple dimensions. Arguably, this optimality assumption is also theoretically important because it increases the testability (and falsifiability) of the generalized context model by significantly constraining the freely varying attention weight parameters.

So, what mechanism might cause individuals to choose a suboptimal attentional allocation when learning categories such as those that we investigated? Certainly, one possibility is that distributing attention across multiple dimensions is resource demanding in some way, causing subjects to adopt a less resource demanding strategy of attending to just a single dimension. Although plausible, this seems rather theoretically unsatisfying because it requires a new assumption that has no a priori justification and that has not been required in previous applications of exemplar models. Moreover, if attention weight distribution is resource demanding, why do subjects seem to eventually shift attention across multiple dimensions with additional training? If this attentional allocation is resource demanding early in learning, it should be resource demanding later in learning as well, unless yet another set of assumptions about an interaction between resource demands and training is also adopted.

An alternative possibility is that subjects are indeed engaging in hypothesis testing to form simple categorization rules. The rules that are formed may then be used to explicitly set the distribution of attention weights in a top–down manner, thereby initially biasing the attentional allocation to appear rule-based (see Choi et al., 1993). Indeed, in the original formulation of the context model, Medin and Schaffer (1978) suggested that the dimensional attention weights may reflect the operation of explicit hypothesis testing that subjects may engage in when learning categories. It may be that as subjects gain more experience with the categories, this top–down influence may gradually be relaxed so as to allow a more optimal distribution of attention weights to be learned, as suggested by Nosofsky (1984, 1998a,b). Lin and Murphy (1997) have also suggested that background knowledge can act to modify patterns of selective attention to reflect expectations based on prior experience or based on high-level causal knowledge. So, although hypothesis testing may be one mechanism that influences attention to dimensions, expectations based on other kinds of knowledge can be another source for such top–down influences on selective attention to stimulus dimensions.

In a recent model of visual attention, Logan and Gordon (2001) suggested that high-level executive processes may operate by setting parameters of a relatively autonomous attention, categorization, and response selection system. Perhaps in the context of learning categories, high-level processes may engage in hypothesis testing or knowledge retrieval with the aim of setting parameters of a relatively autonomous low-level exemplar-based category learning system; these parameters may include the initial selective attention weights as well as other parameters such as learning rates, response biases, and response criteria (e.g., Nosofsky & Johansen, 2000). In this characterization, rule-based (read high-level executive agent) and exemplar-based systems may be separate systems, but they interact in a complex manner. One theoretical implication of this possibility is that the boundary between single and multiple systems may become rather blurred. These may not be independent modules engaged in rule-based or exemplar-based categorization. They may instead be highly interactive subsystems, perhaps operating at different levels of explicit awareness or with different levels of executive control, that operate in conjunction to learn novel perceptual categories.

A number of the various formal models described earlier could possibly account for our results showing a shift from rules to exemplars, perhaps with some quite reasonable modifications (see Anderson & Betz, 2001). Although our results do not conclusively dictate which of these various architectures is the appropriate one to model human category learning, we would argue that our results do provide support for the notion that people employ multiple representations when learning categories and that the relative dominance of rules and exemplars changes with categorization experience. To our knowledge, such a finding has not previously been documented in category learning paradigms in which subjects are not explicitly provided a categorization rule. We believe our results will serve as an important benchmark for evaluating various proposed models of category learning. Determining which of these various architectures, if any, best describes human category learning will likely require a combination of basic behavioral experiments and theoretical modeling, along with evidence obtained using various cognitive neuroscience techniques. A convergence of evidence seems to be suggesting that there are multiple ways categories can be represented, and there are multiple component processes that can be engaged during category learning and category use. What remains to be determined is how these various pieces fit together so as to develop a complete picture of how humans learn novel perceptual categories and how this basic aspect of human cognition might be implemented in a neural architecture. Are there functionally independent categorization systems, or are there instead highly interacting systems, or is there a single system with mixed representations and multiple processes?

Acknowledgments

Order of authorship was decided arbitrarily because both authors contributed equally to the work presented in this paper and the writing duties were divided between the authors. This work was supported by NIMH Grant R01 MH61370 to Vanderbilt University, NSF Grant BCS-9910756 to Vanderbilt University, Vanderbilt University Research Council Direct Research Support Grants, an Indiana University Cognitive Science Program Summer Research Fellowship, NIMH Training Grant T32 MH19879 to Indiana University, and NIMH Grant R01 MH48494 to Indiana University. We thank Shannon Bader and Karinne Damadio for help preparing some of the figures and tables. We thank Woo-kyoung Ahn, Dawson Creek, Isabel Gauthier, Marci Flanery, Brad Love, Gert Storms, and two anonymous reviewers for their extensive comments and criticisms on earlier versions of this paper. We are also grateful to John Kruschke and Michael Erickson for their helpful suggestions on the formal modeling. Finally, we thank Robert Nosofsky for his support during the early stages of this project.

Appendix A

A.1. Context model and ALCOVE predictions of Experiments 1–3

In this section, we attempted to generate a priori predictions of the distributions of generalization patterns made by the context model (Medin & Schaffer, 1978) and ALCOVE (Kruschke, 1992).

A.1.1. Context model predictions

According to the context model, the probability of classifying item i into category A, $P(\mathrm{A}|i)$, is given by the summed similarity of i to exemplars of category A divided by the summed similarity of i to all exemplars from all learned categories. The similarity between i and j is a multiplicative function of matches and mismatches along each dimension m

$$S_{ij} = \prod_{m}^{M} S_m^{\delta_m(i,j)}, \tag{A.1}$$

where M is the number of dimensions, $0 \leqslant s_m \leqslant 1$ are free parameters representing the similarity of mismatching values along dimension m, and $\delta_m(i,j)$ is an indicator function equal to one when i and j mismatch along dimension m and equal to zero otherwise. In generating predictions for the experiments reported in this paper, four free similarity parameters were required for the four stimulus dimensions. Diagnostic dimensions require small values of s_m, causing mismatches along a diagnostic dimension to have a relatively large influence on similarity, and relatively nondiagnostic dimensions require large values of s_m, causing mismatches along those dimensions to have little influence on similarity.

Our goal was to determine whether the context model predicted prominent "exemplar-based" generalizations under assumptions of optimal parameter selection (Nosofsky, 1998a,b). In other words, for similarity parameters that yielded high levels of predicted accuracy on training items, were there particular generalizations that were predicted to be prominent in the distribution? If so, we will characterize those patterns as "exemplar-based" because they are consistent with exemplar generalization. Of course, an alternative possibility is that, through judicious parameter selection, the context model could predict any generalization to be maximally prominent in the distribution, a scenario not supported by our analyses.

For each experiment, our first step was to generate predictions by the context model across a full spectrum of parameter values. Each of the four similarity parameters was incremented in a geometric progression across twenty different values in a range from 0 to 1, yielding a $20 \times 20 \times 20 \times 20$ grid of possible vectors of similarity parameters. For each vector of similarities, predicted classification probabilities were then generated using the equations provided above. We then ordered the context model predictions according to overall accuracy at classifying the training items. Optimal parameter settings were defined as those that gave high levels of accuracy on training items. Finally, we generated predicted distributions of generalization patterns, focusing our attention on the most prominent generalizations in the distribution.

Let us begin with the context model predictions for Experiment 1. For vectors of similarity parameters yielding predicted accuracies greater than 97.5% on training stimuli (nearly 11,000 parameter sets met this criterion), the exemplar-based generalization pattern, ABBBA, was predicted to be the most prominent generalization for 56.4% of the parameter sets; some other patterns were also predicted to be maximally prominent for some other parameter sets, but 26 generalization patterns never emerged as the most prominent pattern. Moreover, in addition to emerging as the most frequently predicted peak generalization based on this systematic grid search for optimal parameter settings, ABBBA most often emerged as the peak generalization pattern if the similarity parameters were explicitly set as a function of the diagnosticity of each dimension (i.e., dimensions 1 and 3 receive high weight, dimension 4 receives intermediate weight, and dimension 2 receives low weight).

For Experiment 2, for vectors of similarity parameters yielding predicted accuracies greater than 97.5% (over 17,000 parameter sets met this criterion), the two exemplar-based generalizations, ABAB and BABA, were predicted to be most prominent for 82.4% of the parameter sets. For comparison, the two rule-based generalizations, AABB and BBAA, were predicted to be most prominent for just 17.2% of the parameter sets. And for Experiment 3, for vectors of similarity parameters yielding predicted accuracies greater than 97.5% (over 15,000 parameter sets met this criterion), the two exemplar-based generalization patterns, ABBA and BAAB, were predicted to be most prominent for 85.0% of the parameter sets. For comparison, the two rule-based generalization patterns, AABB and BBAA, were predicted to be most prominent for just 3.7% of the parameter sets.

So, under those conditions that yielded optimal classification of the training items, particular generalizations were predicted to be maximally prominent in the distributions of generalizations. We characterized these as the "exemplar-based" generalizations in the three experiments. Although the context model can also predict prominent "rule-based" generalization as well, similarity parameters that yielded these generalizations produced sub-optimal learning of the training items.

A.1.2. ALCOVE predictions

As outlined in Section 5, ALCOVE learns categories by adjusting dimensional selective attention weights and association weights via gradient descent on error (Kruschke, 1992). The rate of learning is governed by the attention learning rate parameter, λ_α, and the association learning rate parameter, λ_w. The outputs of the network are also governed by the similarity scaling parameter, c, in Eq. (1), and the response mapping parameter, φ, in Eq. (4). As with the simulations conducted using the context model, our goal in the following simulations was to determine whether ALCOVE predicts particular prominent generalization patterns across numerous combinations of possible parameters. Unlike the context model simulations, we did not need to focus on model parameters that led to optimal classification performance because the connectionist learning algorithm employed by ALCOVE is explicitly designed to adjust attention weights and association weights in a way that minimizes classification errors on training items.

For each experiment, our initial step was to generate ALCOVE predictions across a large spectrum of parameter values. We selected a range of parameters that contained values previously found to produce best-fitting predictions by ALCOVE. The similarity scaling parameter, c, was varied between 3 and 30 in steps of 3; the response mapping parameter, φ, was varied between .5 and 5.0 in steps of .5; the attention learning rate parameter, λ_α, was varied between .1 and 1.0 in steps of .1; and the association weight learning rate parameter, λ_w, was varied between .05 and .50 in steps of .05. We also trained the network for 32, 48, or 64 training blocks since some combinations of parameters led to overall slower learning rates than other combinations of parameters. In addition, some combinations of parameters led to degenerate predictions in that categorization performance did not exceed chance; for example, if the learning rate parameters were relatively too large, the attention weights or association weights tended to change value dramatically from trial to trial, causing a complete lack of convergence during learning. So, we excluded any simulations where the average classification performance on the training items by the end of training did not significantly exceed chance. We used the same procedures for generating predicted distributions of generalization patterns that we used in the Theoretical Modeling section. Generating predictions from these 30,000 combinations of parameter values, each of which was used in averaging across 800 simulated runs of the ALCOVE model, required several weeks of computer time on a dual-processor workstation.

We focused on the most prominent generalizations predicted for each vector of parameters for each experiment. For Experiment 1, the exemplar-based generalization pattern, ABBBA, was the most prominent pattern for 72.4% of the parameter sets. For comparison, the two rule-based generalization patterns, AABBB and BBABA, were the most prominent patterns for 25.6% of the parameter sets. For Experiment 2, the two exemplar-based generalization patterns, ABAB and BABA, were the most prominent patterns for 72.6% of the parameter sets; the two rule-based generalization patterns, AABB and BBAA, were the most prominent patterns for 27.4% of the parameter sets. And for Experiment 3, the two exemplar-based generalization patterns, ABBA and BAAB, were the most prominent patterns for 66.1% of the parameter sets; the two rule-based generalization patterns, AABB and BBAA, were the most prominent patterns for 33.5% of the parameter sets. As was the case for the context model predictions, parameters that yielded the "rule-based" generalizations produced relatively sub-optimal learning of the training items.

Our simulation results with ALCOVE converge with our simulation results using the context model. In both cases, particular "exemplar-based" generalization patterns were predicted to be the most prominent generalization patterns in the predicted distribution of generalization patterns across a wide spectrum of possible parameter values. Thus, we can characterize these particular patterns as "exemplar-based" generalizations in that are predicted by the exemplar models in an a priori manner.

References

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

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

Anderson, J. R. (1990). *The adaptive character of thought*. Hillsdale, NJ: Erlbaum.

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

Anderson, J. R., Fincham, J. M., & Douglass, S. (1997). The role of examples and rules in the acquisition of a cognitive skill. *Journal of Experimental Psychology: Learning, Memory, and Cognition, 23*, 932–945.

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

Ashby, F. G., & Gott, R. E. (1988). Decision rules in the perception and categorization of multidimensional stimuli. *Journal of Experimental Psychology: Learning, Memory, and Cognition, 14*, 33–53.

Ashby, F. G., Maddox, W. T., & Lee, W. W. (1994). On the dangers of averaging across subjects when using multidimensional scaling or the similarity-choice model. *Psychological Science, 5*, 144–151.

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

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

Barsalou, L. W. (1990). On the indistinguishability of exemplar memory and abstraction in memory representation. In T. K. Srull & R. S. Wyer (Eds.), *Advances in social cognition* (pp. 61–88). Hillsdale, NJ: Erlbaum.

Berry, D. C., & Dienes, Z. (1993). *Implicit learning: Theoretical and empirical issues*. Hillsdale, NJ: Lawrence Erlbaum.

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

Brooks, L. R., Norman, G. R., & Allen, S. W. (1991). Role of specific similarity in a medical diagnostic task. *Journal of Experimental Psychology: General, 120*, 278–287.

Brown, R. G., & Marsden, C. D. (1988). Internal versus external cues and the control of attention in Parkinson's disease. *Brain, 111*, 323–345.

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

Bruner, J. S., Goodnow, J. J., & Austin, G. A (1956). *A study of thinking*. New York: Wiley.

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.

Choi, S., McDaniel, M. A., & Busemeyer, J. R. (1993). Evaluation of exemplar-based and network models of conceptual rule learning. *Memory & Cognition, 21*, 413–423.

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

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

Feldman, J. (2000). Minimization of Boolean complexity in human concept learning. *Nature, 407*, 630–633.

Flanery, M. A., & Palmeri, T. J. (2001). fMRI studies of perceptual and probabilistic categorization. *Journal of Cognitive Neuroscience, 107B*(Suppl. S), 62.

Franke, P., Maier, W., Hardt, J., Frieboes, R., Lichtermann, D., & Hain, C. (1993). Assessment of frontal-lobe functioning in schizophrenia and unipolar major depression. *Psychopathology, 26*, 76–84.

Hahn, U., & Chater, N. (1998). Similarity and rules: Distinct? Exhaustive? Empirically distinguishable? *Cognition, 65*, 197–230.

Hintzman, D. L. (1986). "Schema abstraction" in a multiple-trace model. *Psychological Review, 93*, 411–428.

Hintzman, D. L. (1990). Human learning and memory: Connections and dissociations. *Annual Review of Psychology, 41*, 109–139.

Hoffman, J., & Ziessler, C. (1983). Objectidentifikation in kunstlichen begriffshierarchien. *Zeitscrift fur Psychologie, 16*, 243–275.

Homa, D. (1978). Abstraction of ill-defined form. *Journal of Experimental Psychology: Human Learning and Memory, 4*, 407–416.

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.

Hunt, E. B., Marin, J., & Stone, P. J (1966). *Experiments in induction*. San Diego: Academic Press.

Johnstone, T., & Shanks, D. R. (2001). Abstractionist and processing accounts of implicit learning. *Cognitive Psychology, 42*, 61–112.

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

Knowlton, B. J., Mangels, J. A., & Squire, L. R. (1996). A neostriatal habit learning system in humans. *Science, 273*, 1399–1402.

Knowlton, B. J., & Squire, L. R. (1993). The learning of categories: Parallel brain systems for item memory and category knowledge. *Science, 262*, 1747–1749.

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

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

Lamberts, K. (1995). Categorization under time pressure. *Journal of Experimental Psychology: General, 124*, 161–180.

Lin, E. L., & Murphy, G. L. (1997). Effects of background knowledge on object categorization and part detection. *Journal of Experimental Psychology: Human Perception and Performance, 23*, 1153–1169.

Logan, G. D. (1988). Toward an instance theory of automatization. *Psychological Review, 95*, 492–527.

Logan, G. D., & Gordon, R. D. (2001). Executive control of visual attention in dual task situations. *Psychological Review, 108*, 393–434.

Love, B. C., Medin, D. L., & Gureckis, T. M. (in press). SUSTAIN: A network model of category learning. *Psychological Review*.

Maddox, W. T. (1999). On the dangers of averaging across observers when comparing decision bound models and generalized context models of categorization. *Perception & Psychophysics, 61*, 354–375.

Maddox, W. T., & Ashby, F. G. (1998). Selective attention and the formation of linear decision boundaries: Comment on McKinley and Nosofsky (1996). *Journal of Experimental Psychology: Human Perception and Performance, 24*, 301–321.

Martin, R. C., & Caramazza, A. (1980). Classification in well-defined and ill-defined categories: Evidence for common processing strategies. *Journal of Experimental Psychology: General, 109*, 320–353.

Medin, D. L., & Ross, B. H. (1989). The specific character of abstract thought: Categorization, problem solving, and induction. In R. Sternberg (Ed.), *Advances in the psychology of human intelligence* (Vol. 5, pp. 189–223). San Diego, CA: Academic Press.

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

Medin, D. L., & Smith, E. E. (1981). Strategies and classification learning. *Journal of Experimental Psychology: Human Learning and Memory, 7*, 241–253.

Minda, J. P., & Smith, J. D. (2000). 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.

Nosofsky, R. M. (1984). Choice, similarity, and the context theory of classification. *Journal of Experimental Psychology: Learning, Memory, and Cognition, 10*, 104–114.

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

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

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

Nosofsky, R. M. (1998a). Selective attention and the formation of linear decision boundaries: Reply to Maddox and Ashby (1998). *Journal of Experimental Psychology: Human Perception and Performance, 24*, 322–339.

Nosofsky, R. M. (2000). Exemplar representation without generalization? Comment on Smith and Minda's (2000) "Thirty categorization results in search of a model". *Journal of Experimental Psychology: Learning, Memory, and Cognition, 26*, 1735–1743.

Nosofsky, R. M (1998b). Optimal performance and exemplar models of classification. In M. Oaksford & N. Chater (Eds.), *Rational models of cognition*. Oxford: Oxford University Press.

Nosofsky, R. M., Clark, S. E., & Shin, H. J. (1989). Rules and exemplars in categorization, identification, and recognition. *Journal of Experimental Psychology: Learning, Memory, and Cognition, 15*, 282–304.

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

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

Nosofsky, R. M. & Kruschke, J. K. (2001). Single-system models and interference in category learning: Commentary on Waldron and Ashby (2001). *Psychonomic Bulletin & Review, 9*, 169–174.

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 & 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.

Nosofsky, R. M., & Zaki, S. R. (1998). Dissociations between categorization and recognition in amnesic and normal individuals: An exemplar-based interpretation. *Psychological Science, 9*, 247–255.

Palmeri, T. J. (1997). Exemplar similarity and the development of automaticity. *Journal of Experimental Psychology: Learning, Memory, and Cognition, 23*, 324–354.

Palmeri, T. J., R. H. & Flanery, M. A. (2002). Memory systems and perceptual categorization. In B. Ross (Ed.), *The psychology of learning and motivation* (Vol. 41). Elsevier, San Diego.

Palmeri, T. J., & Nosofsky, R. M. (1995). Recognition memory for exceptions to the category rule. *Journal of Experimental Psychology: Learning, Memory, and Cognition, 21*, 548–568.

Pavel, M., Gluck, M. A., & Henkle, V. (1988). Generalization by humans and multi-layer networks. In *Proceeding of the tenth annual conference of the cognitive science society*. Hillsdale, NJ: Erlbaum.

Phillips, R (1991). *Mushrooms of North America*. Boston, MA: Little, Brown and Company.

Pinker, S. (1999). *Words and rules*. Basic Books.

Posner, M. I., & Keele, S. W. (1968). On the genesis of abstract ideas. *Journal of Experimental Psychology, 77*, 353–363.

Reber, A. S. (1967). Implicit learning of artificial grammars. *Journal of Verbal Learning and Verbal Behavior, 6*, 855–863.

Reber, A. S. (1969). Transfer of syntactic structure in synthetic languages. *Journal of Experimental Psychology, 81*, 115–119.

Reber, P. J., & Squire, L. R. (1999). Intact learning of artificial grammars and intact category learning by patients with Parkinson's disease. *Behavioral Neuroscience, 113*, 235–242.

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

Regehr, G., & Brooks, L. R. (1993). Perceptual manifestations of an analytic structure: The priority of holistic individuation. *Journal of Experimental Psychology: General, 122*, 92–114.

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

Ross, B. H. (1987). This is like that: The use of earlier problems and the separation of similarity effects. *Journal of Experimental Psychology: Learning, Memory, and Cognition, 13*, 629–639.

Salatas, H., & Bourne, L. E., Jr. (1974). Learning conceptual rules: III. Processes contributing to rule difficulty. *Memory & Cognition, 2*, 549–553.

Shanks, D. R., & St. Johns, M. F. (1994). Characteristics of dissociable human learning systems. *Behavioral and Brain Sciences, 17*, 367–447.

Shin, H. J., & Nosofsky, R. M. (1992). Similarity-scaling studies of dot-pattern classification and recognition. *Journal of Experimental Psychology: General, 121*, 278–304.

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

Smith, E. E., Langston, C., & Nisbett, R. E. (1992). The case for rules in reasoning. *Cognitive Science, 16*, 1–40.

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

Smith, L. B. (1981). The importance of the overall similarity of objects for adults' and children's classification. *Journal of Experimental Psychology: Human Perception and Performance, 1*, 811–824.

Smith, L. B., & Kemler, D. G. (1977). Developmental trends in free classification: Evidence for a new conceptualization. *Journal of Experimental Child Psychology, 24*, 705–715.

Smith, J. D., & Kemler Nelson, D. G. (1984). Overall similarity in adults' classification: The child in all of us. *Journal of Experimental Psychology: General, 113*, 137–159.

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, J. D., Murray, M. J., & Minda, J. P. (1997). Straight talk about linear separability. *Journal of Experimental Psychology: Learning, Memory, and Cognition, 23*, 659–680.

Smith, J. D., Tracy, J. I., & Murray, M. J. (1993). Depression and category learning. *Journal of Experimental Psychology: General, 122*, 331–346.

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.

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

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

Ward, T. B. (1983). Response tempo and separable-integral responding: Evidence for an integral-to-separable processing sequence in visual perception. *Journal of Experimental Psychology: Human Perception and Performance, 9*, 103–112.

Yamauchi, T., & Markman, A. B. (1998). Category learning by inference and classification. *Journal of Memory and Language, 39*, 124–148.

Statement of ownership, management, and circulation required by the Act of October 23, 1962, Section 4369, Title 39, United States Code: of

COGNITIVE PSYCHOLOGY

Published monthly (except January, May, July and October) by Academic Press, 6277 Sea Harbor Drive, Orlando, FL 32887-4900. Number of issues published annually: 8. Editor: Dr. Gordon Logan, Vanderbilt University, Department of Psychology, Nashville TN 37240.

Owned by Academic Press, 525 B Street, Suite 1900, San Diego, CA 92101-4495. Known bondholders, mortgagees, and other security holders owning or holding 1 percent or more of total amount of bonds, mortgages, and other securities: None.

Paragraphs 2 and 3 include, in cases where the stockholder or security holder appears upon the books of the company as trustee or in any other fiduciary relation, the name of the person or corporation for whom such trustee is acting, also the statements in the two paragraphs show the affiant's full knowledge and belief as to the circumstances and conditions under which stockholders and security holders who do not appear upon the books of the company as trustees, hold stock and securities in a capacity other than that of a bona fide owner. Names and addresses of individuals who are stockholders of a corporation which itself is a stockholder or holder of bonds, mortgages, or other securitie s of the publishing corporation have been included in paragraphs 2 and 3 when the interests of such individuals are equivalent to 1 percent or more of the total amount of the stock or securities of the publishing corporation.

Total no. copies printed: average no. copies each issue during preceding 12 months: 1342; single issue nearest to filing date: 1500. Paid circulation (a) to term subscribers by mail, carrier delivery, or by other means: average no. copies each issue during preceding 12 months: 602; single issue nearest to filing date: 775. (b) Sales through agents, news dealers, or otherwise: average no. copies each issue during preceding 12 months: 401; single issue nearest to filing date: 576. Free distribution (a) by mail: average no. copies each issue during preceding 12 months: 51; single issue nearest to filing date: 51. (b) Outside the mail: average no. copies each issue during preceding 12 months: 9; single issue nearest to filing date: 9. Total no. of copies distributed: average no. copies each issue during preceding 12 months: 1063; single issue nearest to filing date: 1411. Percent paid and/or requested circulation: average percent each issue during preceding 12 months: 94%; single issue nearest to filing date: 96%.

(Signed) Stephanie Smith, Assoc. Manager, Global Sales Operations