

Knowledge Partitioning in Categorization: Constraints on Exemplar Models

Lee-Xieng Yang and Stephan Lewandowsky
University of Western Australia

The authors present 2 experiments that establish the presence of knowledge partitioning in perceptual categorization. Many participants learned to rely on a context cue, which did not predict category membership but identified partial boundaries, to gate independent partial categorization strategies. When participants partitioned their knowledge, a strategy used in 1 context was unaffected by knowledge demonstrably present in other contexts. An exemplar model, attentional learning covering map, was shown to be unable to accommodate knowledge partitioning. Instead, a mixture-of-experts model, attention to rules and instances in a unified model (ATRIUM), could handle the results. The success of ATRIUM resulted from its assumption that people memorize not only exemplars but also the way in which they are to be classified.

In this article, we address the representation of complex perceptual categories. Contrary to the conventional and widespread assumption that people's representations are homogeneous and integrated, we show in two experiments that people often master a complex categorization task by forming independent components, or *parcels*, of knowledge. We also show that once a knowledge parcel is chosen, it provides the sole basis of performance to the exclusion of knowledge demonstrably present in other parcels. These results support and extend related findings of knowledge partitioning in the areas of expertise (Lewandowsky & Kirsner, 2000), function learning (Kalish, Lewandowsky, & Kruschke, in press; Lewandowsky, Kalish, & Ngang, 2002), and categorization involving numeric predictors (Yang & Lewandowsky, 2003). In addition, we show that a pure exemplar model, attentional learning covering map (ALCOVE, Kruschke, 1992), cannot readily accommodate knowledge partitioning, whereas a hybrid rule-plus-exemplar approach, attention to rules and instances in a unified model (ATRIUM, Erickson & Kruschke, 1998, 2001), provides a quantitative account of our results. In addition, ATRIUM accounts for the large individual differences in our studies by assuming variability in the distribution of attention at the outset.

The structure of this article is as follows: We first briefly review existing approaches to categorization, with a particular focus on hybrid rule-plus-exemplar models. Hybrid models can accommo-

date several existing results that point toward considerable heterogeneity in people's category representations. We next survey those empirical demonstrations of heterogeneity, with a particular focus on knowledge partitioning. The survey reveals that although some people always learn an integrated category representation, a significant number of individuals partition complex tasks into seemingly independent parcels of knowledge that are characterized by little cross-linking or integration. The existence of knowledge partitioning presents a strong challenge to models of categorization. We then present two experiments that reveal knowledge partitioning with a perceptual category-learning task in a significant number of participants and conclude with the application of ALCOVE and ATRIUM to the results.

Approaches to Categorization

Contemporary approaches to categorization can be separated into three broad classes. In one class, categories are considered to be represented by a single summary of the learning experiences, such as a prototype or a rule. In the second class of theories, it is assumed that categories are represented by the entire collection of previous experiences through memorization of all exemplars. In the third class of theories, the coexistence of both types of representation are acknowledged, and the circumstances under which each is used are specified.

Examples of the first type of theory include the conventional prototype view (Homa, 1984; Posner & Keele, 1968; Reed, 1972) as well as rule-based models (e.g., general recognition theory, Ashby, 1988). General recognition theory, which instantiates rules as boundaries through the perceived psychological category space, has been shown to account for a large number of phenomena, for example, the fact that people tend to respond deterministically and often create optimal decision bounds (e.g., Ashby, 1988; Ashby & Gott, 1988; Ashby & Maddox, 1990, 1992; Ashby & Perrin, 1988; Ashby & Townsend, 1986).

The second class of theories involves exemplar-based models (Kruschke, 1992; Medin & Schaffer, 1978; Nosofsky, 1986). Exemplar models can account for a large number of categorization phenomena, including prototype effects (Nosofsky & Johansen, 2000; Nosofsky & Kruschke, 1992), effects of specific exemplars

Lee-Xieng Yang and Stephan Lewandowsky, School of Psychology, University of Western Australia, Crawley, Western Australia, Australia.

Lee-Xieng Yang is now at the Department of Psychology, National Chung Cheng University, Chia-Yi, Taiwan.

Preparation of this article was facilitated by a Large Research Grant from the Australian Research Council to Stephan Lewandowsky and Mike Kalish. We thank Leo Roberts for assistance during the preparation of this article and Matt Duncan for his comments on a draft of this article. We are also indebted to Mike Kalish for his comments and contributions throughout the project. We thank Davina French for suggesting we model individual differences by randomly varying the initial attention to context.

Correspondence concerning this article should be addressed to Stephan Lewandowsky, School of Psychology, University of Western Australia, Crawley, Western Australia 6009, Australia. E-mail: lewan@psy.uwa.edu.au

(Nosofsky & Kruschke, 1992), effects of frequency and typicality (Nosofsky, 1988, 1991), and the different learning difficulty of various category structures (Kruschke, 1992; Nosofsky, Gluck, Palmeri, McKinley, & Glauthier, 1994). Exemplar models can account for performance even when the category structures conform to simple rules or boundaries (Nosofsky, 1984, 1986, 1987; Nosofsky, Clark, & Shin, 1989). The success of exemplar theory and its perceived theoretical dominance were succinctly summarized by Nosofsky and Johansen (2000): "A single-system, exemplar-similarity approach appears adequate to account for the major phenomena of interest" (p. 395).

Notwithstanding the visible success of exemplar models, there is also considerable evidence for rule-based representations in at least some circumstances. For example, Allen and Brooks (1991) asked participants to classify drawings of hypothetical creatures that were built from simple graphical features. Stimuli were constructed such that they could be classified by an additive rule (i.e., if a creature had two of three features associated with a category, it always belonged to that category). One group of participants was explicitly given this rule before training commenced, whereas another group learned to categorize the stimuli on the basis of corrective feedback alone. Allen and Brooks found that participants who learned the task without awareness of the rule overwhelmingly misclassified critical items that resembled members of one category but that, by the rule, belonged to the other one. Of particular interest here is the additional finding that even rule-informed participants misclassified those critical items nearly half the time. This result suggests that people remain sensitive to exemplar similarity even when they apply a rule during categorization (see also Nosofsky et al., 1989), a phenomenon that can be readily accommodated by the third type of theory, which postulates the parallel existence of different modes of categorization.

The view that categorization may rely on different representations has been forcefully articulated by Ashby and colleagues (e.g., Ashby, Alfonso-Reese, Turken, & Waldron, 1998; Ashby, Maddox, & Bohil, 2002; Maddox, Ashby, & Bohil, 2003; Waldron & Ashby, 2001). Ashby and colleagues suggest that people approach categorization in two fundamentally different ways, depending on the nature of the categorization task: Rule-based categorization tasks are readily described by verbal rules (that are often, but not always, one-dimensional) and can be learned by way of an explicit reasoning and hypothesis-testing system. Information-integration tasks, by contrast, require the perceptual integration of information across dimensions (and hence are necessarily multidimensional), typically cannot be described verbally, and are mediated by a procedural-learning-based system. The further assumption that these two categorization systems rely on different neural circuitry (Ashby et al., 1998) is supported by several dissociations between the two modes of categorization that were predicted from their hypothesized neural architecture. For example, the presence of feedback has been shown to be more crucial for the learning of information-integration tasks than it has for rule-based tasks (Ashby et al., 2002). Delay of feedback has been shown to disrupt information-integration learning but not rule-based tasks (Maddox et al., 2003); and, conversely, a secondary task has been found to interfere more with rule-based learning than it does with information-integration tasks (Waldron & Ashby, 2001). All of these dissociations involved comparisons between two different category structures, each of which was thought to preferentially

elicit one mode of processing. Comparatively little emphasis has been placed on the potential heterogeneity of representation within a single category structure. This differentiates the approach taken by Ashby and colleagues from other theories that similarly acknowledge the coexistence of rules and exemplars but explore their parallel roles within a single category structure. We focus primarily on these *hybrid* approaches.

Hybrid Rule-and-Exemplar Representations

Several computational models have been put forward recently that postulate the parallel use of rules and exemplars during categorization (e.g., ATRIUM, Erickson & Kruschke, 1998, 2001; rule-plus-exception [RULEX], Nosofsky, Palmeri, & McKinley, 1994; Palmeri & Nosofsky, 1995; parallel rule activation and synthesis [PRAS], Vandierendonck, 1995). Erickson and Kruschke (1998) reported a particularly diagnostic experiment in which most stimuli in a two-dimensional (x - y) space could be classified by a single linear boundary perpendicular to the y -axis. Within each rule-defined region, one exceptional training instance appeared that formed the single member of another category. The results showed that although participants classified most new transfer items on the basis of the rule, stimuli in immediate proximity of the exceptional training items tended to be assigned to the corresponding exceptional category.

Erickson and Kruschke (1998) accounted for their data with a computational model, which they called ATRIUM, that combined an exemplar module with one (or more) rule module(s). In ATRIUM, stimuli are classified either on the basis of their similarity to previously encountered exemplars or on the basis of a rule instantiated by a dimensional threshold. Far from just providing a specific explanation of a circumscribed finding involving rules and exceptions, the mixture-of-experts approach instantiated by ATRIUM has several powerful properties that are widely applicable. Perhaps the most crucial property is that the extent to which each module contributes to categorization is uniquely weighted for each stimulus. Those weights are adjusted during learning and, in consequence, different subsets of stimuli may be classified according to different types of information. Erickson and Kruschke (2001) illustrate this concept with an intuitive analogy: ". . . when classifying different members of the violin family, one might use rules based on size, whereas when classifying different types of guitars (e.g., electric vs. acoustic), shape-based rules might prove more fruitful" (p. 2).

The fact that modules are uniquely weighted for each stimulus potentially enables ATRIUM to account for two manifestations of representational variability, both of which are explored in this article. First, by assuming variability in initial conditions, the theory may be able to account for the large individual differences that sometimes characterize category-learning experiments. In support, Erickson and Kruschke (2001) showed that ATRIUM could accommodate the performance of five qualitatively different subgroups of participants in their experiment if they varied the relative weighting of the various rule modules.

Second, because ATRIUM formalizes the idea that stimulus features can be differentially emphasized in different contexts, it may account for heterogeneous categories in general, including those that do not follow a rule-plus-exception structure. For example, Aha and Goldstone (1992) sampled training stimuli from

two distinct clusters in a two-dimensional category space, each of which was bisected by its own uniquely oriented boundary. When extended further from their cluster, the boundaries dictated opposite classifications for the same test items. Aha and Goldstone found that participants classified transfer stimuli using the closest partial boundary, thus revealing their sensitivity to differences between subsets of stimuli. Recent research on knowledge partitioning has extended this phenomenon further, to the point of showing that people's knowledge can be divided into mutually contradictory independent components.

Knowledge Partitioning in Categorization

The theoretical framework of knowledge partitioning, first proposed by Lewandowsky and Kirsner (2000), states that knowledge need not be integrated—as is commonly assumed; see Lewandowsky et al. (2002) for a review—but may be separated into different parcels according to the context in which a problem is presented. To give an illustrative anecdotal example, consider a problem from statistics involving the general linear model (GLM). Most statistical practitioners would (correctly) reject a regression analysis for a set of x - y pairs when the level of measurement of the independent variable (x) is nominal or ordinal (e.g., small, medium, and large). However, judging by the number of published occurrences, statistical practitioners are less likely to reject the idea of a trend analysis in the context of an analysis of variance (ANOVA) under similar circumstances. That is, although the GLM constrains the required level of measurement, and even though most practitioners are conversant with the model, they seem to apply it differently in the regression and the ANOVA contexts.

The knowledge-partitioning framework has been supported by research with experts solving domain-relevant problems (Lewandowsky & Kirsner, 2000) as well as with novice participants during function learning (Kalish et al., in press; Lewandowsky et al., 2002) and category learning (Yang & Lewandowsky, 2003). Yang and Lewandowsky's (2003) study forms the departure point for the present experiments and thus deserves to be presented in more detail.

Yang and Lewandowsky (2003) presented participants (in their Experiments 3 and 4) with the category structure shown in Figure 1. The true category boundaries are represented by the parallel ascending lines. The area between these boundaries, labeled 2 in the figure, represents Category A, and the regions outside the boundaries (Areas 1 and 3) represent Category B. All training items fell within the area enclosed by dotted lines. Like they were in the studies by Aha and Goldstone (1992) and Erickson and Kruschke (2001), training items can be considered in two clusters, each of which contains its own local boundary. In contrast to those earlier studies, Yang and Lewandowsky introduced a third dimension, called *context*, that by itself never predicted category membership but that, in the condition of interest, consistently identified each cluster of training items. Thus, training stimuli within the upper cluster always appeared with one type of context, whereas stimuli in the lower cluster appeared with the other type of context. At test, all transfer items (represented by solid diamonds in the figure) appeared twice, once with each context cue. The results revealed that about one third of the participants consistently applied the upper boundary in the upper context and the lower

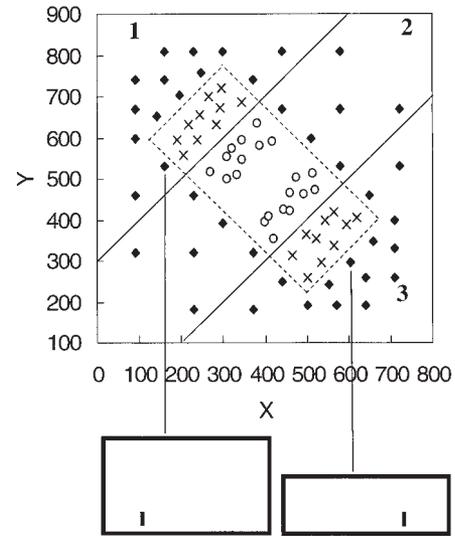


Figure 1. Stimulus space used in Experiments 1 and 2. The abscissa denotes the vertical segment position, and the ordinate denotes the height of the rectangle. Two sample stimuli are shown below the stimulus space to illustrate the nature of the task. The training items are shown as open circles for Category A and crosses for Category B. The transfer items are shown as solid diamonds.

boundary in the lower context. For example, a stimulus presented in Area 3 (the bottom right of Figure 1) would be classified as belonging to Category A in the upper context despite its proximity in x - y space to training items in Category B, whereas the same stimulus presented in the lower context would be classified as belonging to Category B.

The fact that those participants exclusively applied one rule to the entire category space in one context and another rule to the same category space in another context differentiates the results of Yang and Lewandowsky (2003) from those of related experiments (Aha & Goldstone, 1992; Erickson & Kruschke, 2001) and extends the applicability of the knowledge-partitioning framework to categorization. However, the study by Yang and Lewandowsky also raised three empirical and theoretical challenges whose resolution is the focus of this article.

First, Yang and Lewandowsky (2003) did not quantitatively model their results. It is therefore unclear whether any models of categorization, including ATRIUM and other hybrid approaches, can handle the seemingly complete partitioning of category knowledge. Below, we resolve this issue by applying two computational models to knowledge-partitioning data.

Second, Yang and Lewandowsky (2003) found that one third of the participants learned the true parallel boundaries even when context uniquely identified each cluster (with the strategy of the remaining one third of the participants not being clearly identifiable). The presence of two clearly identifiable, but very different, subgroups of participants after training under identical conditions poses an additional challenge to candidate models that we also take up in Experiment 1.

Finally, Yang and Lewandowsky (2003) used alpha-numeric stimuli that were presented as lists of numbers and/or verbal labels. Although this practice is not without precedent in category learn-

ing (e.g., Erickson & Kruschke, 1998; Lewandowsky, Kalish, & Griffiths, 2000), two limitations can be identified: First, Nosofsky and Johansen (2000) argued that “. . . numeric labels . . . may alter . . . the presumed similarity relations among the objects” (p. 386). In support, Nosofsky and Johansen obtained strikingly different results when they repeated the study by Erickson and Kruschke (1998) without any numeric labels. Second, within the earlier framework of multiple categorization systems (e.g., Ashby et al., 1998), the study by Yang and Lewandowsky clearly involved a rule-based task and did not require information-integration.

We first take up the final issue by presenting two experiments that extend the work of Yang and Lewandowsky (2003) to a perceptual category-learning task. Although the category structure continues to support rule-based learning (according to the criteria suggested by Ashby et al., 1998), the use of a perceptual task addresses the concern raised by Nosofsky and Johansen (2000). Hence, we knew that if we obtained partitioning in our experiments, the results would extend the generality of the knowledge-partitioning framework and also provide benchmark data for model comparison.

Experiment 1

In this experiment, we used the category structure shown in Figure 1. We instantiated stimuli using the same graphical elements that were used in the studies by Erickson and Kruschke (1998, 2001, 2002); stimuli were composed of a rectangle and a vertical line segment within it. Stimuli varied along two dimensions, the position of the vertical line segment (x) and the height of the rectangle (y). The context cue was instantiated by the color (i.e., red or green) of the rectangle and line segment.

The experiment included two conditions: In the randomized-context condition, context cues were randomly assigned to an equal number of training instances from each category and each cluster, and in the systematic-context condition, context was consistently mapped to clusters. Thus, context on its own did not predict category membership (because each cluster contained an equal number of items from each category) but did identify whether a stimulus was close to the upper or lower partial boundary.

Method

Participants

Forty-eight undergraduate volunteers from the University of Western Australia, Crawley, Western Australia, participated in exchange for partial course credit. An equal number of participants were randomly assigned to each condition.

Stimuli

All stimuli were sampled from the two-dimensional pseudocontinuous category space shown in Figure 1. The partial category boundaries are described by $|y - x - 100| = 200$. Stimuli within and outside the two boundaries belong, respectively, to Categories A and B. We use *boundary* to refer to a design feature of the stimulus space. In contrast, we use *rule* to refer to participants' imputed categorization strategy or to a process within a computational model.

The two categories were arbitrarily instantiated as different kinds of space shuttle, and the dimensions x and y were, respectively, the height of

the rectangle and the position (offset from the bottom-left corner) of the vertical segment within it. The dimensional values shown in Figure 1 were rescaled to map onto the physical extent of the stimulus display. In addition, stimuli were differentiated by color (thus instantiating context) and were presented in either red or green.

Training items. Each participant received a unique set of 40 training items that were randomly sampled from within the dotted rectangle in Figure 1. The figure includes a representative sample of items, with open circles and crosses denoting membership in Categories A and B, respectively. The set of 40 training items was divided into two clusters of 20 stimuli located above (upper cluster) and below (lower cluster) the positive diagonal. If participants relied on either Dimension x or y by itself, they would be able to classify the training items with 58% accuracy.

In the randomized-context condition, the context dimension (i.e., color) was randomly assigned to training stimuli with the constraint that, in each cluster, there were equal numbers of training stimuli in each category presented in the two contexts.

In the systematic-context condition, context predicted the cluster to which a training item belonged, without, however, predicting category membership. Thus, stimuli within the upper cluster were all presented in one color, and those within the lower cluster were all presented in the other color, with an equal number of items in each color belonging to each category. Assignment of color to cluster was counterbalanced across participants, and we label the contexts generically as *upper* and *lower*. In both conditions, there were eight blocks of training trials, each of which involved a different random sequence of the 40 items.

Transfer items. There were 60 transfer items in total. Of those, 40 novel items were the same for all participants and are represented by filled diamonds in Figure 1. The remaining 20 items were randomly chosen training instances (5 items from each category in each cluster). The sequence of 60 items was presented twice in a different random order within each sequence and with context alternating between sequences.

Diagnosticity of transfer items. Novel transfer items came from the three areas identified by number in Figure 1. The effect of context on responding across those three areas can be used to infer the presence of knowledge partitioning. If partitioning is absent, as would be expected in the randomized-context condition, context should have no effect in any of the areas. All items in Area 2 should be preferentially classified as belonging to Category A, whereas those in Area 1 and Area 3 should be classified as belonging to Category B.

In contrast, if knowledge partitioning is present, context should affect classification in Areas 1 and 3. Specifically, participants should classify items in Area 1 as belonging to Category B in the upper context and to Category A in the lower context. Conversely, items in Area 3 would be classified as A in the upper context and B in the lower context. Finally, items in Area 2 should be unaffected by context, because both partial boundaries dictate a Category A response.

Procedure

Participants were tested individually in a quiet booth. Each trial commenced with the display of a fixation signal (a plus sign) in the center of the screen for 500 ms followed by presentation of the stimulus, which remained visible until participants responded by keypress. The width of rectangle was kept constant at 12.5 cm for all stimuli, and the height of rectangle varied from 2.8 cm to 13.2 cm. The distance of the segment position to the left margin of rectangle varied from 1 cm to 11 cm. All lines were 0.2 cm wide and were colored red or green as dictated by context.

In the training phase, participants were given written feedback (the word *correct* or *wrong* shown in the center of the screen) for 1,000 ms after each response and before the next trial commenced. No feedback was presented during transfer.

Self-paced breaks were inserted after every 40 training trials and after every 60 transfer trials. In addition, there was a break between the training and transfer phases. The experiment lasted just under 1 hr.

Results and Discussion

Training Performance

Performance improved considerably with training for both conditions, from .53 (Block 1) to .77 (Block 8). Participants performed slightly better in the systematic-context condition ($M = .70$) than they did in the randomized-context condition ($M = .66$). Accordingly, a 2 (condition) \times 8 (block) between- and within-subjects ANOVA revealed a main effect of block, $F(7, 322) = 42.21$, $MSE = 0.01$, $p < .01$, no main effect of condition, $F(1, 46) = 2.18$, $MSE = 0.08$, $p < .15$; and no interaction between those two variables, $F(7, 322) < 1$.

Although participants in the systematic-context condition performed slightly better than did participants in the randomized-context condition, the difference (unlike that in the studies by Yang & Lewandowsky, 2003) was not statistically significant. In terms of absolute level of performance, the present training results were comparable to those observed by Yang and Lewandowsky (2003).

The relatively high accuracy at the conclusion of training suggests that people can learn the parallel category boundaries. We confirm this in the Computational Modeling section, in which we show that a rule based on the parallel boundaries describes performance better than alternative ways in which people may have learned to classify the training items.

Transfer Performance

The dependent variable for all transfer analyses in this article (including the computational modeling and all data reported in tables and figures) was the probability of classifying a stimulus as belonging to Category A. Responses were aggregated across items within each of the three diagnostic areas. To facilitate presentation, for all ANOVAs in this article, we report only the highest order significant effects and do not enumerate significant component effects unless there is compelling reason to do so. Performance on the 20 training items shown at transfer was analyzed separately from performance on the 40 novel items.

Transfer responses to training items. Classification probabilities for training items are shown by context and area in Table 1, with the randomized-context and systematic-context conditions in the top and bottom panel, respectively. In the randomized-context condition, there is evidence of consistent application of the parallel boundaries, with items in Area 2 predominantly classified as A and items in Area 1 and Area 3 predominantly classified as B irrespective of whether test and training contexts matched. In confirmation, none of the simple comparisons between contexts were significant in this condition.

In the systematic-context condition, in contrast, the parallel boundaries were applied only when the test context was congruent with the training context. When the test context was incongruent (i.e., a trained combination of x - y values was presented in a new context), participants responded to the training items quite differently, and there was little evidence that the parallel boundaries were applied. In confirmation, there was an effect of context (congruent vs. incongruent) that narrowly missed conventional significance in Area 1, $F(1, 23) = 3.89$, $MSE = 0.09$, $p = .06$; $\omega^2 = .104$, and a significant effect of context in Area 3, $F(1, 23) = 5.11$, $MSE = 0.09$, $p < .05$. These effects suggest that participants

Table 1
Mean Probability of Category A Responses in Experiment 1 for Training Items Presented at Transfer in Each Area and Test Context

Context	Area		
	1	2	3
Randomized-context condition			
Congruent	.29	.82	.46
Incongruent	.28	.81	.54
Systematic-context condition			
Congruent	.29	.80	.37
Incongruent	.47	.75	.57

treated training items presented in a different context as novel stimuli.

Novel transfer items. Transfer responses to novel items are shown in Table 2. As expected, test context had no effect on classification probabilities in any area in the randomized-context condition. In contrast, context had a strong effect in Areas 1 and 3 in the systematic-context condition.

A 2 (condition) \times 2 (test context) \times 3 (area) between- and within-subjects ANOVA revealed a main effect of area, $F(2, 92) = 48.08$, $MSE = 0.07$, $p < .01$, but not condition, $F(1, 46) = 1.32$, or test context, $F(1, 46) = 1.16$. The overarching three-way interaction was significant with $F(2, 92) = 10.19$, $MSE = 0.05$, $p < .01$. The interaction was explored by separate two-way within-subjects ANOVAs for each of the two conditions. In the randomized-context condition, there was no trace of a Test Context \times Area interaction, $F(2, 46) < 1$. That interaction was, however, highly significant in the systematic-context condition, $F(2, 46) = 15.48$, $MSE = 0.06$, $p < .01$. The corresponding simple comparisons between test contexts were significant in Area 1, $F(1, 23) = 12.44$, $MSE = 0.06$, $p < .01$; Area 3, $F(1, 23) = 16.47$, $MSE = 0.07$, $p < .01$; and in Area 2, $F(1, 23) = 6.89$, $MSE = 0.01$, $p < .05$.

As expected under knowledge partitioning, classification probabilities differed substantially—and in opposing directions—between contexts in Areas 1 and 3. This suggests that participants divided the categorization space into two spaces, each of which had a partial boundary that was applied on the basis of context. The significant context effect in Area 2 was unexpected; further analysis not reported here revealed that the effect resulted from the (theoretically uninteresting) asymmetrical learning of the two partial boundaries. Therefore, we do not consider it any further.

However, although response probabilities in Area 1 and Area 3 differed between contexts, some of their absolute magnitudes deviated only moderately from chance (i.e., .50). One possible reason for this is that different subgroups of participants used different strategies during transfer. Some participants may have learned the parallel boundaries, whereas others may have learned to partition their knowledge.

Individual differences. We explored individual differences by entering participants' response profiles (i.e., the vector of responses, coded as 0s and 1s, to the 40 novel transfer items) into a

Table 2
Mean Probability of Category A Responses in Experiment 1 for Novel Transfer Items in Each Area and Test Context

Participant group	Condition					
	Randomized-context			Systematic-context		
	Area 1	Area 2	Area 3	Area 1	Area 2	Area 3
Overall						
Upper	.27	.67	.28	.25	.69	.50
Lower	.31	.65	.30	.50	.60	.19
True-boundary						
Upper				.18	.64	.20
Lower				.19	.64	.19
Knowledge partitioning						
Upper				.17	.83	.87
Lower				.84	.66	.17

k-means cluster analysis with 3 predefined clusters (cf. Lewandowsky et al., 2000). Cluster centroids were predefined to represent strategies that consisted, respectively, of application of (a) the parallel boundaries, (b) the upper partial boundary, and (c) the lower partial boundary. Response profiles were entered into the cluster analysis for each condition and test context separately. Using a Euclidean distance measure, the analysis assigned each participant's profile to the closest cluster. Participants who were equidistant from all predefined patterns were considered to be unrecognizable.

Table 3 shows the assignment of participants to clusters for each condition and test context. In confirmation of the aggregate analysis, most participants in the randomized-context condition applied the parallel boundaries in both contexts. In the systematic-context condition, in contrast, the distribution of categorization strategies differed greatly between contexts. Although there was still a considerable number of participants who applied the parallel boundaries in both contexts, 46% (11) of participants used the upper partial boundary in the upper context (but none used it in the lower context), and, conversely, 50% (12) of participants used the lower boundary in the lower context (but only 2 participants used it in the upper context). Statistical support for the association between test context and classification was provided by Cramer's coefficient (ϕ). Cramer's coefficient is a transformation of chi-square and is readily interpretable as a measure of association that ranges from 0 to unity (Wickens, 1989). For the systematic-context condition, Cramer's coefficient was highly significant ($\phi = .616, p < .01$).¹

For the remaining analyses and the forthcoming model fitting, two subgroups of participants were formed in the systematic-context condition and the data were aggregated within each group. Participants were assigned to a knowledge-partitioning (KP) group ($n = 9$) if the cluster analysis classified them into the upper rule cluster in the upper context and to the lower rule cluster in the lower context. If the cluster analysis assigned a participant to the true-boundary cluster in both contexts, then he or she was assigned to the true-boundary (TB) subgroup ($n = 10$). This left 5 participants who could not be unambiguously assigned to either group and who were not considered in these analyses.

To illustrate the implications of knowledge partitioning, Figure 2 shows the classification probabilities for each novel transfer item

for participants in the KP group. The top panel shows the upper test context, and the bottom panel shows the lower context. Every plotted square represents a novel transfer item, and the degree of shading indicates the likelihood of the item's being classified into Category A. The figure reveals a distance effect in both contexts, with items further away from each partial boundary classified with more extreme probabilities.

Comparison of novel and training items. Using responses from all participants together, we compared the two classes of items by computing the differences between upper and lower contexts for each item class and each area. In the randomized-context condition, those differences were entered into a 2 (item type; training vs. novel) \times 3 (area) within-subjects ANOVA, which did not yield any significant effects. The parallel analysis of the systematic-context condition, in contrast, yielded a significant main effect of area, $F(2, 46) = 10.52, MSE = 0.26, p < .01$, with mean differences of $-.22, .07$, and $.26$ for Areas 1, 2, and 3, respectively, but no other effects. The absence of any item-type effects confirms that participants did not differentiate between completely novel combinations of x and y and training items presented in a novel context but with an old x - y pairing. This replicates a similar effect reported by Yang and Lewandowsky (2003).

Experiment 2

Experiment 1 provided further evidence for the existence of knowledge partitioning in categorization. This confirms that the earlier results of Yang and Lewandowsky (2003) were not tied to the use of numeric categorization stimuli. Experiment 1 also underscored the presence of large individual differences in the systematic-context condition, with two subgroups of participants using two very different representations or strategies to classify

¹ Although Wickens (1989) suggests that repeated observations on the same participants may well be considered independent, a more conservative approach is to divide the underlying value of chi-square by 2, thus accounting for repetition of individuals across contexts (Wickens, 1989). Even with that correction, the association between context and cluster remained significant ($\phi = .346, p < .02$).

Table 3
Number and Percentage of Participants Identified as Using a Particular Strategy by *k*-Means Cluster Analysis

Cluster	Test context			
	Upper		Lower	
	<i>n</i>	%	<i>n</i>	%
Experiment 1				
Randomized-context condition				
True-boundary	19	79	15	63
Upper-boundary	2	8	2	8
Lower-boundary	2	8	3	12
Unrecognized	1	5	4	17
Systematic-context condition				
True-boundary	11	46	12	50
Upper-boundary	11	46	0	0
Lower-boundary	2	8	12	50
Unrecognized	0	0	0	0
Experiment 2				
Systematic-context condition				
True-boundary	7	44	7	44
Upper-boundary	8	50	1	6
Lower-boundary	0	0	7	44
Unrecognized	1	6	1	6

test items. In Experiment 2, we examined two additional attributes of knowledge partitioning using the same stimulus space we used in Experiment 1.

The first concerns the extent to which the hypothesized knowledge parcels (i.e., representations of the partial boundaries) are independent. In the extreme case, if partitioning is complete, each might be used as if the other had never been learned. To examine this possibility, we included two single-context conditions in which participants only learned one subset of training instances (either above or below the positive diagonal in Figure 1) in its corresponding context (upper or lower, respectively) before being tested on all transfer items in both contexts. Comparison of transfer performance between those two single-context conditions and the matching test context of the systematic-context condition reveals the extent of partitioning. Specifically, should transfer performance be identical between conditions within a given context, then the two knowledge parcels acquired in the systematic-context condition can be considered to be largely independent. Conversely, should those comparisons reveal an effect of condition within each context, this identifies at least some linkage between parcels in the systematic-context condition. The randomized-context condition was not included in Experiment 2.

Our second purpose in Experiment 2 concerned participants' knowledge of the predictive value of each dimension. In Experiment 1, context by itself did not directly predict category membership, although a considerable number of participants nonetheless used it to gate access to partial knowledge. In Experiment 2, we examined whether participants were aware of the fact that context did not directly predict classification by adding a contingency rating task (e.g., Wasserman & Berglan, 1998; Williams, Sagness, & McPhee, 1994). In this task, we asked participants to

provide a numeric estimate of the contingency between dimensional values and the outcome.

Method

Participants

Forty-eight undergraduate volunteers from the University of Western Australia participated in exchange for partial course credit. An equal number of participants (16) were randomly assigned to each condition. None had participated in Experiment 1.

Stimuli and Procedure

Systematic-context condition. The category structure and stimuli were the same as they were in Experiment 1. The usual transfer phase was followed by the new contingency rating task involving 10 trials. On each contingency trial, a single dimensional value was shown, and ratings were obtained for four possible values of *x* (segment position = 200, 400, 600, and 800), four possible values of *y* (height = 100, 300, 500, and 700), and the two possible levels of context (red and green). For example, when the value of *x* (segment position) was to be rated, a white vertical segment was presented whose lower end touched a horizontal line equal in length to the rectangle in the complete stimulus. When the value of *y* (height) was to be

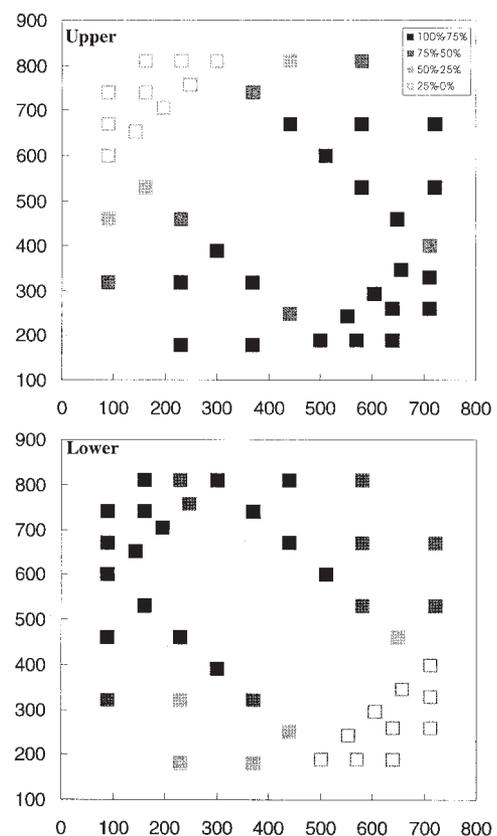


Figure 2. The probability of Category A responses made by participants in the knowledge-partitioning subgroup for all transfer items in Experiment 1. The darker the fill of a point, the more likely the item was classified as belonging to Category A. The upper and lower panels show the results in the upper and lower contexts, respectively.

rated, a white rectangle was presented without a vertical segment. For context, the word *Red* or *Green* was presented in its corresponding color.

To make their contingency judgment, participants used the right- and left-arrow keys to move a vertical bar on the screen along a horizontal scale. The scale was marked by a category label at each end, and the assignment of labels was counterbalanced across participants. Each end of the scale represented 100% certainty that an item with the given dimensional value would belong to that category.

Single-context conditions. In the upper-only condition, the 20 learning instances were all located in the upper half of the stimulus space and were presented with a single-context cue (red or green, counterbalanced across participants). Conversely, training in the lower-only condition involved the other 20 learning instances. There were eight blocks of training trials in both conditions. Each block contained a different random sequence of the training items. In the transfer phase, all 20 learning items were presented together with the 40 novel transfer items from the systematic-context condition. Thus, participants in the single-context conditions experienced half of the novel transfer items in a new color. Participants were not given any information about the role of color.

In all other respects, the procedure was identical to that of Experiment 1.

Results and Discussion

Training

Performance improved with training from .63 (Block 1) to .83 (Block 8). A 3 (condition) \times 8 (block) between- and within-subjects ANOVA confirmed the presence of a significant effect of block, $F(7, 315) = 24.83$, $MSE = 0.01$, $p < .01$, but found no effect of condition and no interaction between the two variables ($F_s \leq 1$). Given that the single-context conditions involved only half the number of training items, the absence of a condition effect is noteworthy.

Transfer

Transfer responses to training items. Table 4 shows transfer responses for all conditions, areas, and test contexts; again, *congruent* and *incongruent* refer to context being the same or differ-

ent, respectively, between training and test. In the systematic-context condition, a 2 (test context: congruent vs. incongruent) \times 3 (area) within-subjects ANOVA revealed a significant interaction between both variables, $F(2, 30) = 9.28$, $MSE = 0.04$, $p < .01$. Further exploration by simple comparisons revealed that the effect of context was significant in Areas 1 and 3, $F(1, 15) = 8.35$, $MSE = 0.13$, $p < .05$, and $F(1, 15) = 8.59$, $MSE = 0.02$, $p < .05$, respectively, but not in Area 2 ($F < 1$). Thus, as in the previous experiment, the effect of context in Areas 1 and 3 suggests that training items were regarded as novel when presented in an incongruent context.

The same analyses were applied to the two single-context conditions, except that the area variable had two levels only (because training items were drawn only from Areas 1 and 2 in the upper-only condition and Areas 2 and 3 in the lower-only condition). In both conditions, only the effect of area was significant, with $F(1, 15) = 36.62$, $MSE = 0.09$, $p < .01$ and $F(1, 15) = 52.42$, $MSE = 0.06$, $p < .01$, for upper- and lower-only, respectively. With the remaining $F_s \leq 1.79$ and $F_s \leq 2.08$ in the upper- and lower-only conditions, respectively, it is clear that participants ignored context. Given that context was invariant during training, thus having the limited role of an unexplained change in color of the stimuli at transfer, the result does not come as a surprise.

Transfer responses to novel items. Table 5 shows the classification responses to novel transfer items for all areas and conditions. As in Experiment 1, the pattern for the systematic-context condition was suggestive of knowledge partitioning. This was confirmed by the 2 (test context) \times 3 (area) within-subjects ANOVA for that condition, which revealed a significant Context \times Area interaction, $F(2, 30) = 8.17$, $MSE = 0.08$, $p < .01$. Exploration of the interaction by simple comparisons between contexts revealed significant effects in Areas 1 and 3, $F(1, 15) = 8.63$, $MSE = 0.09$, $p < .05$, and $F(1, 15) = 7.13$, $MSE = 0.08$, $p < .05$, respectively, but not in Area 2 ($F < 1$).

A *k*-means cluster analysis paralleling that of Experiment 1 was conducted; the results are shown in Table 3. The table shows that performance was quite dependent on context, with half (8 of 16) of the participants applying the upper boundary in the upper context and roughly half (7 of 16) applying the lower boundary in the lower context. A significant Cramer's coefficient ($\phi = .624$, $p < .01$; conservative correction $\phi = .441$, $p \approx .01$) confirmed the strong association between cluster and context. This replicates the knowledge-partitioning effect of Experiment 1. For the subsequent analyses and modeling, participants were assigned to the KP ($n = 5$) and the TB ($n = 5$) groups in the same manner as in Experiment 1.

Turning to the single-context conditions, the table suggests that participants learned the partial boundary appropriate to each condition, although it was applied less rigorously in the incongruent test context (i.e., the upper context for the lower-only condition and vice versa). Specifically, the differences between the extreme areas (1 and 3) were greater when the test context matched the one encountered during training. In confirmation, combined analysis of the single-context conditions in a 2 (condition) \times 2 (test context) \times 3 (area) between- and within-subjects ANOVA revealed a significant interaction involving all three variables, $F(2, 60) = 6.24$, $MSE = 0.04$, $p < .01$.

In summary, the fact that performance on novel transfer items differed between single-context conditions in the diagnostic re-

Table 4
Mean Probability of Category A Responses in Experiment 2 for Training Items Present at Transfer in Each Area and Test Context

Context	Area		
	1	2	3
Systematic-context condition			
Congruent	.16	.83	.36
Incongruent	.53	.77	.53
Upper-only condition			
Congruent	.24	.76	
Incongruent	.34	.74	
Lower-only condition			
Congruent		.70	.18
Incongruent		.62	.22

Table 5
Mean Probability of Category A Responses in Experiment 2 for Novel Transfer Items in Each Area and Test Context

Context	Condition								
	Systematic-context			Upper-only			Lower-only		
	Area 1	Area 2	Area 3	Area 1	Area 2	Area 3	Area 1	Area 2	Area 3
Upper	.20	.57	.46	.16	.68	.90	.75	.50	.25
Lower	.51	.60	.19	.31	.64	.78	.84	.60	.14

gions (Areas 1 and 3) but not in the central area (Area 2) for which both boundaries predict the same response shows that participants learned different linear partial boundaries in the two conditions. Moreover, participants applied the partial boundary less stridently to novel transfer items in the incongruent context; this replicates a similar observation by Yang and Lewandowsky (2003).

Modularization of Classification Rules

Our aim in this experiment was to examine the extent to which knowledge parcels became independently modularized and encapsulated during training. To address this issue, we computed item-wise correlations both within and across conditions (i.e., for each transfer stimulus, we computed the proportion of Category A responses across participants within each context). First, a between-context correlation was computed for all participants in the systematic-context condition. This provides a measure of the extent to which knowledge is integrated across the two contexts. The relatively low value of the correlation ($r = .326$) suggests that participants responded to items quite differently in the two contexts.

Second, a within-context correlation was computed between each single-context condition and responses made to the same items (by different participants) in the corresponding context in the systematic-context condition. This within-context correlation ($r = .677$), despite being computed between individuals, was higher than the correlation between contexts computed within the same participants. When the within-context correlation was computed including only the KP group from the systematic-context condition, the correlation increased considerably to $r = .896$. The high within-context correlation implies that participants who partitioned their knowledge applied each partial boundary with little or no consideration of the other one—their behavior in each context was highly similar to that of participants who only learned one boundary. This conclusion is further supported by a negative between-

context correlation for the KP group ($r = -.738$), which reflects the fact that the two partial boundaries dictate opposite classifications for the majority of transfer items.

Contingency Ratings

We analyzed contingency ratings by expressing responses as the judged probability of an item belonging to Category A. Table 6 shows the average ratings for all tested levels of the three dimensions (x , y , and context) together with their normative values (computed for the overall set of training items). To determine whether participants' judged contingencies departed from chance, we compared the mean response at each value of each dimension with .5 using a single-sample t test. Using a significance level of .05 for each test, six dimensional values were found to depart from chance.

Specifically, there was some suggestion that participants were sensitive to the role of dimensions x and y , but their judged contingencies clustered relatively close to chance and were not nearly as extreme as the normative values. For the context dimension, there was no evidence that participants erroneously judged it to be predictive, as their responses were close to the normatively correct chance value (largest $t \approx 1.00$) and also did not differ between the two contexts, $t(15) < 1$. Thus, participants used context to gate use of partial knowledge despite being aware of its nonpredictive nature.

Summary

The main contribution of Experiment 2 was to show that when people partition their knowledge, there is little or no evidence for integration between knowledge parcels. Instead, people seemingly gate their knowledge on the basis of context, even though they are aware that context by itself does not predict category membership.

Table 6
Contingency Ratings in Experiment 2 and Significance Tests for Deviation From Chance (.50)

Value	x				y				Context	
	100	300	500	700	200	400	600	800	Upper	Lower
Normative	.00	.52	.53	.00	.00	.52	.52	.00	.50	.50
Observed	.28*	.76**	.55	.21**	.33*	.68*	.59	.42*	.45	.57
SE	.08	.07	.08	.08	.07	.06	.06	.07	.06	.07

* $p < .05$. ** $p < .005$.

The findings from Experiments 1 and 2 provide a benchmark for evaluation of computational models. We now examine and compare two candidate models by applying them to the data just reported. The models were an exemplar model (ALCOVE; Kruschke, 1992) and a hybrid rule-plus-exemplar model (ATRIUM, Erickson & Kruschke, 1998). We chose them because they embody two very distinct approaches to categorization.

Computational Modeling

Basic Approach and Overview

In all modeling in this article, we assumed that stimulus dimensions were psychologically separable, which entailed use of the city block metric for calculation of stimulus similarity. This follows relevant precedent for the stimuli used here (Erickson & Kruschke, 2001). Also following Erickson and Kruschke (2001), we fit models by minimizing Akaike's information criterion (AIC; Akaike, 1974). The AIC is defined as $-2\log L + 2N$, where $\log L$ is the logarithm of the likelihood of the data given the model, and N is the number of free parameters in the model. The AIC permits comparisons between models with different numbers of free parameters because it combines a lack-of-fit index with a penalty for model flexibility as measured by the number of free parameters.

We fit models to the data using the level of aggregation that was mandated by the individual-differences analysis. Thus, fits to the randomized-context condition involved aggregate results from all above-chance participants, whereas fits to the systematic-context condition involved the results from the TB and KP subgroups separately. The models were fit either to the transfer or the training data, depending on the purpose of the modeling. When the transfer data were used, AIC was minimized at the level of responses to each test item across a 2 (transfer trial block) \times 40 (novel transfer items) \times 2 (category responses) grid. When the models were fit to the training data, AIC was minimized across a three-way grid comprised of 8 Blocks \times 40 Training Items \times 2 Category Labels.

Stimulus representations for the models were obtained by linearly transforming the x and y values to the range 0–1. The two values of context were represented as 0 and 1.

ALCOVE: An Exemplar Theory

ALCOVE is an exemplar-based connectionist model of category learning (Kruschke, 1992). During training, ALCOVE learns to adjust the connection strengths between nodes that represent the exemplars and the possible responses. ALCOVE also learns how much attention to pay to each dimension (cf. Nosofsky, 1986, 1987), which results in the stretching of highly diagnostic dimensions and the shrinking of others that turn out to be less relevant. ALCOVE adjusts connection strengths and dimensional attention by gradient descent using conventional error-driven network learning. When presented with a novel item for categorization, ALCOVE responds on the basis of the similarity between that item and every previously encountered exemplar, taking into account the amount of attention devoted to each dimension.

ALCOVE has accommodated a variety of results with great quantitative precision (e.g., base-rate neglect, Kruschke, 1992, 1996a, 1996b; though see Lewandowsky, 1995; categorization of correlated dimensions, Kruschke, 1990, 1996b; the different learn-

ing difficulty of filtration and condensation tasks, Kruschke, 1991; prototype effects, Nosofsky & Kruschke, 1992). In all cases, the concept of dimensional attention has been a key ingredient of ALCOVE's success. Unlike early conceptions of dimensional attention (e.g., Nosofsky, 1986, 1987), ALCOVE incorporates a learning algorithm that permits the model to use prediction errors during training to arrive at an optimal distribution of its attention.

In the present context, it follows that ALCOVE might account for knowledge partitioning by devoting far more attention to context than to the other two dimensions: This would be equivalent to a three-dimensional representation of the category space in Figure 1, with the two partial boundaries in x - y space separated widely along a third (context) dimension. Given sufficient attention on context, this would render all generalizations context-specific, much as was observed in the present experiments. However, it is not altogether clear why ALCOVE would pay that much attention to context, given that context did not directly predict category membership in any of the conditions and given that perfect categorization was possible on the basis of the remaining dimensions. In previous applications, ALCOVE was shown to learn to ignore irrelevant dimensions under these circumstances (Kruschke, 1996b).

Applying ALCOVE to Experiment 1

Description of Parameters

A full description of ALCOVE can be found in Kruschke (1992). Here, we briefly summarize its four main parameters:

1. The specificity, c , is a positive constant that determines the steepness of the similarity gradient surrounding each exemplar. The larger the specificity, the more rapidly the activation of a given node falls off for stimuli in its neighborhood.
2. The decision certainty, ϕ , enters into computation of the predicted response probability. If the value of ϕ is close to unity, the model uses probability matching and chooses a category on the basis of the relative summed similarities between a test item and the stored exemplars of each category. As ϕ increases, responding becomes more deterministic until, in the extreme case, the category with the largest summed similarity is chosen exclusively.
3. The learning rate for association weights between exemplars and output nodes is captured by λ_w .
4. The learning rate for dimensional attention strengths is represented by λ_a . Attention strengths were constrained to sum to unity. Unless otherwise noted, attention was set to .33 for each dimension at the outset of training in all simulations.

Randomized-Context Condition

Only data from participants whose performances were reliably above chance during the final training block were included in the modeling. The chance cutoff was .65 (\approx 26 of 40 stimuli), as

derived from a binomial distribution with $p = .50$ and $n = 40$ ($\alpha = .05$). The data from 18 participants were included by this criterion.

ALCOVE provided a good quantitative account of performance when fit to the training data (see Table 7). The best fitting parameter estimates and the fit statistic are provided in Table 8, and the associated predicted transfer responses are shown in the left panel of Figure 3. The data are represented by the bars (with shading indicating context), and ALCOVE's predictions are shown by the symbols (i.e., a diamond represents the upper context, and a cross represents the lower context). Note that we obtained ALCOVE's transfer predictions after fitting the model to the training data and that they thus represent the model's generalization performance. It comes as little surprise that ALCOVE, like participants, responds to transfer items in a manner that suggests knowledge of the true boundaries.

We next fit ALCOVE directly to the transfer responses of the same set of participants. The parameter estimates are also shown in Table 8, and the accompanying predictions are shown in the right panel of Figure 3. It is not surprising that the quantitative account of transfer responses improves if the model is fit directly to those data.

Systematic-Context Condition

For this condition, the two identifiable subgroups of participants were considered separately. When fitting ALCOVE to the training data, all four parameters could be kept identical between the two groups without loss of fit (see Table 8; shared parameters are reported whenever separate estimates for each group did not appreciably improve fit). The associated transfer predictions are shown in the two left panels in Figure 4, with the TB and KP groups shown at the top and bottom, respectively. It is clear that ALCOVE cannot predict knowledge partitioning when fit to the training data.

Table 7
Observed and Predicted Proportion Correct at Learning in Experiment 1

Participant group	n	Training block							
		1	2	3	4	5	6	7	8
Randomized-context condition									
Overall									
Observed	18	.52	.59	.63	.67	.69	.68	.75	.76
ALCOVE		.52	.54	.60	.62	.66	.69	.72	.73
ATRIUM		.50	.54	.62	.66	.67	.70	.71	.71
Systematic-context condition									
KP									
Observed	9	.54	.66	.67	.71	.74	.79	.80	.84
ALCOVE		.49	.56	.60	.69	.74	.78	.81	.83
ATRIUM		.53	.64	.69	.73	.74	.75	.77	.77
TB									
Observed	10	.55	.64	.67	.71	.74	.80	.80	.82
ALCOVE		.49	.54	.59	.66	.72	.76	.79	.82
ATRIUM		.53	.66	.68	.72	.74	.75	.76	.77

Note. KP = knowledge-partitioning group; TB = true-boundary group.

Table 8
Best Fitting Parameter Values for ALCOVE in Experiment 1 and Fit Statistics

Parameter	Condition		
	Randomized-context	Systematic-context	
	Overall	TB	KP
Training			
c	9.90		9.90
ϕ	2.10		1.92
λ_w	0.01		0.01
λ_α	0.08		0.03
$-2\log L$	1,553.28	1,139.96	1,204.09
Transfer			
c	9.11		9.89
ϕ	2.15		6.09
λ_w	0.01	0.06	0.01
λ_α	0.07	0.01	0.15
$-2\log L$	128.64	76.67	213.11

Note. TB = true-boundary group; KP = knowledge-partitioning group.

We next fit ALCOVE to the transfer responses of the two subgroups using a shared estimate of c and ϕ but separate estimates for the learning rates (λ_w and λ_α ; see Table 8). The predictions are shown in the two right panels of Figure 4. The results confirmed our expectation that ALCOVE, even when fit to the transfer data of partitioning participants, would not be able to show knowledge partitioning. Presumably, this failure arose because there was no reason for its learning algorithm to recognize the (indirect) relevance of context. Accordingly, the final dimensional attention weights were .02 for context and .60 and .38 for x and y , respectively.

Individual Differences and ALCOVE

ALCOVE's apparent inability to learn to pay attention to context does not necessarily preclude an exemplar-based account of

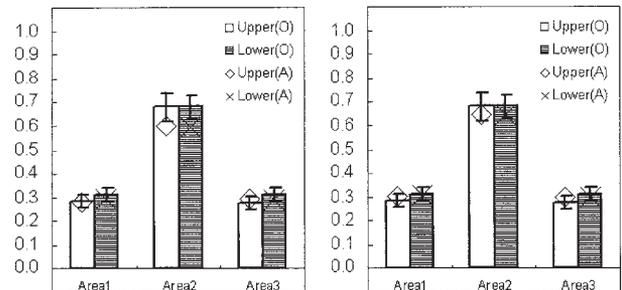


Figure 3. Observed data (O; vertical bars, with standard errors shown) and predictions of the attentional learning covering map model (ALCOVE; A, symbols) for the randomized-context condition in Experiment 1. The left panel shows the predicted transfer performance when the model was fit to the training data, whereas the right panel shows predictions when ALCOVE was fit to the transfer data.

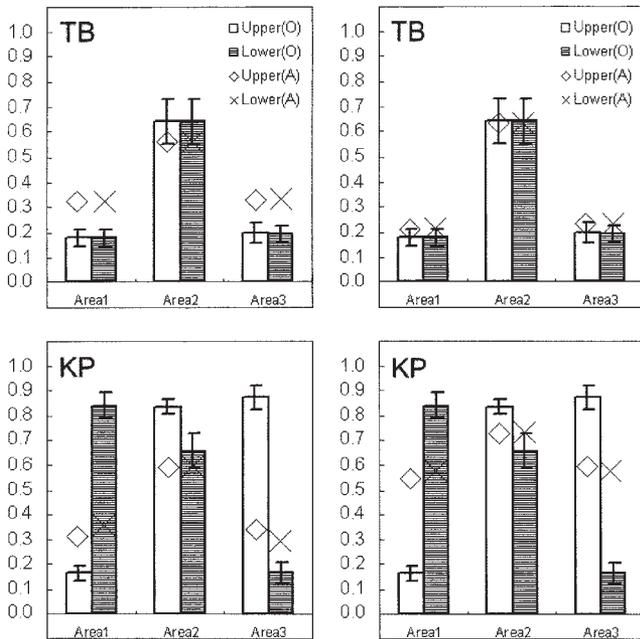


Figure 4. Observed data (O; vertical bars, with standard errors shown) for two groups of participants and predictions of the attentional learning covering map model (ALCOVE; A, symbols) for the systematic-context condition in Experiment 1. The left panels show predicted transfer performance when ALCOVE was fit to the training data, whereas the right panels show predictions when fit to the transfer data. Initial attention strengths were .33 for each dimension for all fits. TB = the true-boundary group; KP = the knowledge-partitioning group.

knowledge partitioning. Specifically, it is conceivable that if the model commenced learning with a strong emphasis on context, it might still be able to predict knowledge partitioning. This, in turn, would suggest that the individual differences observed in both experiments may have arisen from variability among participants' preferences for one dimension over the others at the outset of training.

We therefore examined whether ALCOVE might capture those individual differences and might, after all, predict knowledge partitioning if it approached learning with an uneven distribution of attention across the three dimensions. Using the parameter estimates from the fit to the training data of the KP and TB groups (see top panel of Table 8), we conducted 100 simulation runs in which the initial attention to context was randomly sampled from a normal distribution for each replication. The top panels of Figure 5 show ALCOVE's predictions when initial attention to context was sampled from a distribution with a mean of .15 and a standard deviation of .07 (initial attention weights were constrained to sum to unity and were equal for x and y), and the bottom panels show predictions for a distribution with a mean of .42 and a standard deviation of .15. Within each row, the left and right panels show, respectively, predictions for the top and bottom 30% of the distribution of initial attention to context (and data for the KP and TB group, respectively). From left to right, the obtained average initial attention to context was, respectively, .23 and .05 (top row) and .76 and .25 (bottom row) across panels. Notwithstanding the

wide range of initial attention to context (viz. from .05 to .76), ALCOVE consistently failed to predict knowledge partitioning. Instead, in all cases, the model either produced transfer responses that resembled those of the TB group, or it produced chance-level performance (bottom-left panel). It appears that ALCOVE cannot capture the individual differences in our experiments by manipulating initial attention to context.

In a final simulation, we set the attention given to context at the outset of training to .8, and reestimated parameters using the transfer data from each group separately. ALCOVE captured the transfer responses of the KP group well, with $-2\log L = 84.72$ ($c = 9.90$, $\phi = 9.54$, $\lambda_w = .009$, $\lambda_\alpha = .005$, and final attention strengths for dimensions x , y , and context of .25, .21, and .54, respectively). However, the extreme attentional starting value for context then prevented ALCOVE from accommodating the TB group, with $-2\log L = 247.10$, notwithstanding the renewed estimate of the four free parameters ($c = 9.90$, $\phi = 6.95$, $\lambda_w = .005$, $\lambda_\alpha = .017$, and final attention strengths for x , y , and context .31, .21, and .48, respectively).

This final simulation showed that when initial attention is highly unevenly distributed, parameter values can be estimated that per-

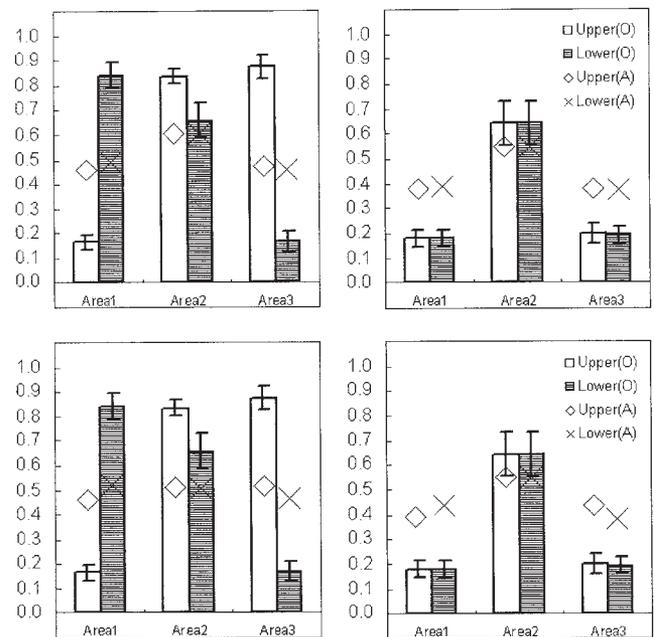


Figure 5. The predicted transfer performance of the attentional learning covering map model (A; represented by diamonds and crosses for upper and lower context, respectively) when initial attention strength to context randomly varied across 100 replications with all parameters kept constant. The left and right columns of panels show predictions for the top and bottom 30%, respectively, of the distribution of initial attention to context. For the top row of panels, initial attention to context was sampled from a normal distribution with mean of .15 and standard deviation of .07, whereas for the bottom panels, those values were mean of .42 and standard deviation of .15. Observed responses and standard errors (O; all from Experiment 1) are represented, respectively, by white and black bars for the upper and lower contexts. Observed results in the left panel represent the knowledge-partitioning group, and those in the right panel represent the true-boundary group.

mit ALCOVE to produce knowledge partitioning. Although this is useful as an existence proof of a suitable exemplar representation, it is qualified by the fact that, under those circumstances, ALCOVE cannot accommodate the TB responses even if parameters are reestimated. We therefore turn to a theoretical alternative that, because it comprises several modules that make independent contributions to a categorization decision, may be a better candidate for accommodating knowledge partitioning.

ATRIUM: Mixtures of Experts

ATRIUM (Erickson & Kruschke, 1998, 2001, 2002; Kruschke & Erickson, 1994), with its hybrid rule- and exemplar-based representations, implements a mixture-of-experts approach (Jacobs, Jordan, Nowlan, & Hinton, 1991). Accordingly, different modules (or experts) learn independently to predict outcomes, with a gating mechanism deciding the extent to which each module contributes to the final outcome for each exemplar. Erickson and Kruschke (1998, 2001) provided a full description of ATRIUM; here, we summarize it briefly.

ATRIUM comprises a minimum of two modules: an exemplar module that consists of ALCOVE plus at least one rule module. A rule module in ATRIUM implements a one-dimensional linear boundary that bisects a dimension into two response regions. If the category space involves multiple dimensions, one rule module exists for each dimension. All modules receive input in parallel, and all produce a candidate categorization response. The candidate responses from all modules—including the exemplar module—are combined by a gating mechanism to determine the final output of the model. The gating mechanism is a crucial feature of ATRIUM that differentiates it from other models of categorization. The mechanism translates the activation of the output units (also called gate nodes; Erickson & Kruschke, 2001) of each module into a gain, or choice probability, for the corresponding module. The module's candidate response probabilities are weighted by that gain, and the final response of the system is the sum of those weighted probabilities across all modules. For example, assuming the presence of three modules, a stimulus might evoke the candidate response probabilities (probability of Category A response) of .8, .6, and .2 from the two rule modules and the exemplar module, respectively. Supposing furthermore that for this particular stimulus the first rule module has a gain of .98, and the remaining two modules have a gain of .01 each, then the final output of the model would be given by the weighted sum $.8 \times .98 + .6 \times .01 + .2 \times .01$.

During training, ATRIUM learns a variety of weights by gradient descent; namely, (a) the association weights between the input and output nodes in the rule modules; (b) dimensional attention in the exemplar module; (c) the weights between output nodes and hidden nodes in the exemplar module; and (d) the weights between gate nodes and hidden nodes in the exemplar module. The latter are of particular importance because they represent knowledge of the extent to which each module contributes to classification of the current item, thus giving rise to what is termed *representational attention* (Erickson & Kruschke, 2002). The activation of the gate nodes is translated into the gain of the associated module by a normalization process that ensures that all gains sum to unity.

Representational attention, the ability to classify different stimuli in different ways, is the most crucial attribute of ATRIUM. Thus, when applied to a rule-plus-exception structure, as we discussed at the outset of this article (Erickson & Kruschke, 1998), ATRIUM classifies transfer stimuli on the basis of two distinct types of representation. First, the majority of stimuli in a given response region are categorized on the basis of the one-dimensional rule. Second, stimuli that are located in proximity to the exceptions encountered during training are classified not on the basis of the rule but by the exemplar module. The information about how stimuli should be classified is associated with each learned exemplar and is retrieved on the basis of the proximity between a transfer stimulus and the closest trained exemplar(s).

Adapting ATRIUM to Model Knowledge Partitioning

One limitation of ATRIUM is that, as currently formulated, the rule modules can only learn one-dimensional boundaries. This creates a difficulty in the present case, in which two dimensions jointly determine a (partial) category boundary. We resolved this issue by representing rules along a compound psychological dimension that was formed by the difference between x and y values. Hence, instead of using x and y values as input for the rule modules, we created a compound dimension z , defined as $z = y - x$, that was bisected into two response regions by each rule module (following a suggestion by M. A. Erickson, personal communication, December 4, 2001). The z dimension thus provided a computational instantiation of the way in which people are known to verbalize certain two-dimensional rules. For example, Ashby et al. (1998) suggested that if rectangles have to be classified on the basis of their height and width, people can readily formulate the simple verbal rule, "Respond A if the stimulus rectangle is taller than it is wide; respond B if the rectangle is wider than it is tall" (p. 444). Note that this verbal rule is isomorphic to forming a single compound dimension that represents the difference between height and width and bisecting it into two response regions. Similarly, in our experiments, participants could have verbalized the lower partial boundary with the rule "Respond A if the rectangle is taller than the offset of the vertical bar from the left; respond B otherwise," with the category assignment reversed for the upper partial boundary (accompanied by an adjustment of the threshold for how much rectangle height had to exceed offset).

The present version of ATRIUM was thus composed of two one-dimensional rule modules, each of which—via the translation $z = y - x$ —learned one of the diagonal partial boundaries. The model additionally contained an exemplar module consisting of ALCOVE, and a gating mechanism that adjudicated between the competing modules.

The parameters in ATRIUM were as follows: The specificity, c , and choice probability, ϕ , were the same as they were in ALCOVE. The module probability scaling constant, ϕ_g , determines how deterministic the choice process between modules is. Larger values of ϕ_g map into a more extreme distribution of contributions from the various modules, such that the module with the largest gain will overwhelmingly contribute to the final output, even if the gains for the other modules are only slightly smaller. The remaining parameters include two learning rates for the two rule modules, λ_{r1} and λ_{r2} : a learning rate for the association weights and dimensional attention, respectively, in the exemplar

module (λ_e and λ_α), and the learning rate λ_g for the associations between the exemplar nodes and the gating node. Additionally, the rule sharpness, γ_r , was set equal for the two rule modules. The rule sharpness determines the steepness of the boundary implemented by the rule modules, with large values corresponding to a steep threshold-like response profile and small values corresponding to a more gradual sigmoid translation from one category to another.

The modeling reported below therefore required nine parameters in total. Unless otherwise noted, the initial attention strengths were set to .33 for all dimensions. Gains for all three modules were equal (set to .33) at the outset of learning.

Applying ATRIUM to Experiment 1

Randomized-context condition. ATRIUM was first fit to the training data of the above-chance participants. As shown in Table 7, ATRIUM provided a good quantitative description of learning behavior. Best fitting parameter estimates and the fit statistic are shown in Table 9. With nine free parameters and attention spread evenly across dimensions at the outset, the performance of ATRIUM (AIC = 1558.29 for training data) was slightly better than that of ALCOVE, which used four free parameters (AIC = 1561.28). The associated transfer predictions are shown in the left panel of Figure 6. Similar to ALCOVE, ATRIUM also predicted a true-boundary response pattern when it was fit to the training data.

Table 9
Best Fitting Parameter Values for ATRIUM in Experiment 1

Parameter	Condition		
	Randomized-context	Systematic-context	
		Overall	TB
Training			
c	8.33		8.47
ϕ	0.67		1.04
ϕ_g	1.08		1.00
γ_r	8.85		8.38
λ_e	0.02		1.76
λ_{r1}	1.21		0.99
λ_{r2}	1.96		0.86
λ_g	0.35		0.67
λ_α	0.95		0.01
$-2\log L$	1,540.29	1,125.27	1,122.71
Transfer			
c	5.10		9.10
ϕ	2.52	4.29	3.68
ϕ_g	0.27	0.31	0.59
γ_r	6.69		3.79
λ_e	0.55		0.47
λ_{r1}	0.06		0.28
λ_{r2}	1.94		0.20
λ_g	0.99	1.89	0.22
λ_α	0.58	0.23	0.83
$-2\log L$	65.51	77.96	65.68

Note. TB = true-boundary group; KP = knowledge-partitioning group.

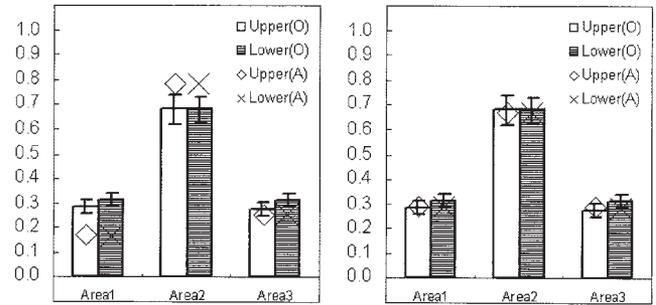


Figure 6. Observed data (O; vertical bars, with standard errors shown) and predictions of the attention to rules and instances unified model (A; symbols) for the randomized-context condition in Experiment 1. The left panel shows the predicted transfer performance when the model was fit to the training data, whereas the right panel shows predictions when the attentional learning covering map model was fit to the transfer data.

We next fit ATRIUM to the transfer responses directly, with its predictions shown in the right panel of Figure 6 (see Table 9 for parameters). Compared with ALCOVE (AIC = 136.64 for transfer data), ATRIUM performed better (AIC = 83.51) despite being penalized for its additional five parameters. We conclude that ATRIUM, much like ALCOVE, can provide a good quantitative account of performance in the randomized-context condition.

We next explored whether some rule other than that of the parallel boundaries might provide a better account of performance in the randomized-context condition. Although our examination of knowledge partitioning does not critically depend on whether participants in the randomized-context condition relied on the parallel boundaries, we could facilitate interpretation of the remaining modeling by showing that alternative categorization rules handle the data less well. One possible alternative rule that participants might have applied to the category space in the randomized-context condition is the conjunctive rule, “Respond A if $x > 300$ and $y > 400$; otherwise respond B.”² We implemented this alternative rule in ATRIUM again by forming a compound dimension z that was set to 1 or 0 as determined by the conjunctive rule. In addition, we removed the exemplar module from ATRIUM in order to focus on the utility of the conjunctive rule by itself, which yielded a total of three parameters that were freely estimated from the aggregate training data: the choice probability ϕ , the learning rate for the rule module, λ_r , and the rule sharpness γ_r . The conjunctive rule yielded AIC = 2121.8 (parameters $\phi = 1.97$, $\lambda_r = 0.01$, and $\gamma_r = 9.23$) when fit to the training data and AIC = 230.33 ($\phi = 2.24$, $\lambda_r = 0.01$, and $\gamma_r = 2.51$) when fit to the transfer data. To permit an exact comparison to the parallel boundaries, we reapplied the standard parallel-boundary version with the role of the exemplar module limited to providing representational attention (i.e., adjudication between the partial rule modules). Exemplars did not contribute to the categorization responses (i.e., there was no learning of exemplar-to-output weights and no attentional learning). This rule-only parallel-boundary version yielded AIC = 1621.84 when fit to the training data ($c = 9.88$, $\phi = 1.21$, $\phi_g = 1.08$, $\lambda_{r1} = 0.54$, $\lambda_{r2} = 1.22$, $\lambda_g = 0.54$, and $\gamma_r = 5.91$) and

² We thank F. Gregory Ashby for pointing out this alternative.

AIC = 85.94 when fit to the transfer data ($c = 6.04$, $\phi = 8.53$, $\phi_g = 1.31$, $\lambda_{r1} = 0.13$, $\lambda_{r2} = 1.49$, $\lambda_g = 0.50$, and $\gamma_r = 2.31$). Clearly, the parallel boundaries provided a better account of the data than did the conjunctive rule. Because of the possibility that this aggregate analysis might hide some noteworthy individual differences, we next applied the conjunctive and parallel-boundaries versions to each of the 18 above-chance participants separately. When fit to the training data, the parallel-boundaries version provided a better account of the data than did the conjunctive rule in 17 of 18 cases and in 12 of 18 cases when fit to the transfer data. We therefore suggest that participants in the randomized-context condition learned the true parallel boundaries rather than some other approach to the task.³

Systematic-context condition. ATRIUM was fit to each group of participants (TB and KP) separately, again with attention spread evenly across dimensions at the outset. When ATRIUM was fit to the training data, best fitting parameter estimates could be identical for both groups without loss of fit (see Table 9). Table 7 shows that ATRIUM provided a very good account of training performance in this condition. Turning to the associated transfer predictions, ATRIUM predicted knowledge partitioning for both groups of participants as shown in the left panels of Figure 7. This is in striking contrast to the true-boundary transfer pattern that was predicted by ALCOVE under identical circumstances. At the end of training, ATRIUM's learned attention strengths for dimension x , y , and context were tightly clustered around .33 (ranging from .32 to .34) for both subgroups.

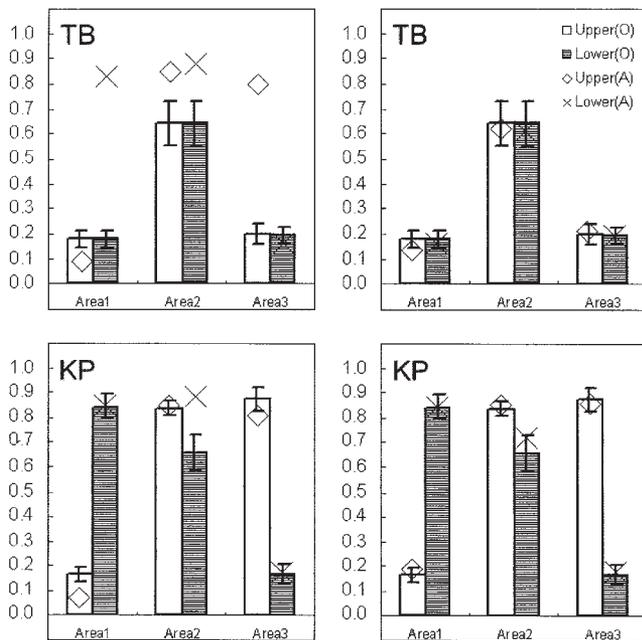


Figure 7. Observed data (O; vertical bars, with standard errors shown) for two groups of participants and predictions of the attention to rules and instances unified model (ATRIUM; A, symbols) for the systematic-context condition in Experiment 1. The left panels show predicted transfer performance when ATRIUM was fit to the training data, whereas the right column shows predictions when fit to the transfer data. TB = the true-boundary group; KP = the knowledge-partitioning group.

We next fitted ATRIUM to the transfer responses of the two groups. It was possible to obtain a good quantitative fit with five parameters kept invariant for both groups and four parameters (decision certainty, ϕ ; module probability scaling constant, ϕ_g ; gate node learning rate, λ_g ; attention learning rate, λ_a) varying between groups (see Table 9). As we show in the right panels of Figure 7, ATRIUM captured the performance of both groups. For the TB group, final attention strengths were .50, .48, and .02 for dimensions x , y , and context. For the KP group, the attention strengths were .04, .25, and .71 for x , y , and context, respectively.

Comparison of ATRIUM to ALCOVE by computing the AIC across both groups revealed that ATRIUM (AIC = 169.64) performed considerably better than ALCOVE (AIC = 301.78) despite having 13 instead of 6 free parameters. If model comparison is limited to the TB group, then ALCOVE performed better (AIC = 84.67) than ATRIUM (AIC = 95.96), although this slight superiority was entirely due to the smaller number of free parameters in ALCOVE rather than to a difference in the quality of fit.

In summary, although both models can accommodate transfer responses in the randomized-context condition, only ATRIUM was able to handle the knowledge partitioning that was observed in some participants in the systematic-context condition. Indeed, the fit of ATRIUM to the training data revealed that knowledge partitioning was its preferred way of approaching this categorization task.

ATRIUM and knowledge partitioning. We next examined the reasons and mechanisms behind ATRIUM's ability to explain knowledge partitioning. We first analyzed the contribution of each module at the end of training by examining its gain. The average gains associated with the novel transfer items in different areas and contexts are shown in Table 10 and represent the probabilities with which each module contributed to the corresponding transfer responses.

For the fit to the KP group, it is clear that the exemplar module was unimportant, irrespective of the area and context in which a transfer stimulus was presented. Instead, classification was primarily based on the rule modules, with the upper rule module being most important in the upper context and the lower rule module crucial in the lower context. Although this result is not unexpected, given the nature of ATRIUM's predictions, the extreme magnitudes of those gains are noteworthy because they suggest that each rule module is used at the virtual exclusion of any other information. That is, although ATRIUM's gating mechanism is designed to blend output from all modules, in this case, there appears to be little blending. Further analysis of the gains during training revealed that the rule modules very rapidly gained prominence, whereas the exemplar module equally quickly lost its importance.

³ A final attempt was made to enable the conjunctive rule to accommodate the data by replacing the fixed cutoffs on dimensions x and y (300 and 400, respectively) with two additional free parameters (C_x and C_y). On fit to the training data, the AIC was 2,018.98 (with parameter values $\phi = 2.02$, $\lambda_r = 0.01$, $\gamma_r = 9.83$, $C_x = 307.26$, $C_y = 399.06$). The two additional parameters provided little improvement over the original fit of the conjunctive rule (AIC = 2,121.8). On fit to the transfer data, the parameter estimates were $\phi = 4.36$, $\lambda_r = 0.04$, $\gamma_r = 7.70$, $C_x = 192.84$, $C_y = 294.19$. The fit improved (AIC = 171.81) compared with that of the original conjunctive rule (AIC = 230.33). However, in both cases, the conjunctive rule still provided a worse fit than the parallel-boundaries version.

Table 10
Average Gains for All Modules in ATRIUM in Both Contexts in Experiment 1

Area	Upper context			Lower context		
	Exemplar	Rule 1	Rule 2	Exemplar	Rule 1	Rule 2
Systematic-context condition (KP participants)						
1	0.03	0.91	0.06	0.05	0.12	0.83
2	0.04	0.89	0.07	0.03	0.08	0.89
3	0.06	0.83	0.11	0.02	0.07	0.91
Systematic-context condition (TB participants)						
1	0.15	0.74	0.11	0.17	0.70	0.13
2	0.22	0.46	0.32	0.23	0.40	0.37
3	0.23	0.16	0.61	0.22	0.14	0.65
Randomized-context condition						
1	0.01	0.93	0.06	0.01	0.93	0.06
2	0.02	0.48	0.50	0.02	0.48	0.50
3	0.01	0.05	0.94	0.01	0.05	0.94

Note. KP = knowledge-partitioning group; TB = true-boundary group.

For the TB group, in contrast, the distribution of gains was less extreme, although there was a tendency for the upper rule module to be favored for items in Area 1 (irrespective of context), whereas the lower rule module was favored for items in Area 3 (again irrespective of context). Thus, it appears that ATRIUM learned the true boundary by partitioning the category space—albeit not on the basis of context but on the basis of the location of transfer items within x - y space. (The table also shows that ATRIUM behaved similarly in the randomized-context condition.) This result is particularly intriguing because it suggests that knowledge partitioning may be central to complex categorization tasks, even if it cannot be detected by behavioral measures. Kalish et al. (in press) presented similar evidence for the centrality of knowledge partitioning in function learning.

Applying ATRIUM to Experiment 2: Independence of Knowledge Parcels

A further characteristic of knowledge partitioning is the apparent independence of knowledge parcels, as is shown by the correlational analysis of Experiment 2. Accordingly, we examined whether ATRIUM can capture this aspect of knowledge partitioning in two ways.

Correlational analysis. We used the best fitting parameter values obtained for the KP group in Experiment 1 and presented ATRIUM with the training stimuli used in the two single-context conditions in Experiment 2. We then used ATRIUM's predictions for individual transfer items in the systematic-context condition and the single-context conditions, obtained under identical parameter settings but after two different training regimes, to compute correlations in the same manner as for the behavioral data. The predicted within-context correlation between conditions was highly positive ($r = .993$), whereas the between-context correlation within the systematic-context condition was negative ($r = -.898$). This mirrors the pattern in the data (the empirical corre-

lations were .896 and $-.738$) and further supports the notion that knowledge parcels, as instantiated by the rule modules in ATRIUM, can be largely independent.

Componential learning. Another way in which we can demonstrate the independence of knowledge parcels is by using ATRIUM to build combined knowledge from partial training. If knowledge parcels are independent of one another, then knowledge gathered separately in different single-context conditions should, when combined, be able to be used to predict the KP responses in the systematic-context condition. On the basis of this hypothesis, we fit ATRIUM to the training data from the single-context conditions in Experiment 2 and then, with the same parameter estimates, asked it to predict the transfer data of KP participants in the systematic-context condition.

When fitting the training data of a single-context condition, the exemplar module and the other rule module were switched off (i.e., only the upper-rule module was active in the upper-only condition, and vice versa for the lower-only condition). We estimated a total of four parameters using a goodness-of-fit index that combined performance across both modules (though each was fit to a different data set). Two of the parameters were shared by both rule modules (the specificity parameter c and the decision constant ϕ), whereas the remaining parameters (the learning rates) varied between modules. The best fit occurred when $c = 9.86$, $\phi = 1.14$, $\lambda_{r(\text{upper})} = 0.99$, and $\lambda_{r(\text{lower})} = 1.16$, and produced a very good quantitative account of learning in both single-context conditions (figures not shown here).

Using these parameter values and learned associative weights within each rule module, we obtained predictions for the KP group in the systematic-context condition by combining the output from both modules under a variety of combinations of gain values (gains could not be learned in this case because, owing to the componential training, the model had no way of associating stimuli to gate nodes for the two competing modules). Gains were varied from 0 to 1 in steps of .10 for each module and were orthogonally combined, yielding a total of 121 gain combinations. The best fit to the KP results occurred when the gain for the context-appropriate module was .9 and the gain for the context-inappropriate module was .1 ($-2 \log L = 93.94$). The fit remained good ($-2 \log L = 124.36$) even when gains were set to 1.00 and .00 for the context-appropriate and context-inappropriate module, respectively.

In summary, this simulation showed that (a) partitioning is complete and (b) knowledge in each parcel corresponds to what is learned in the single-context conditions. Moreover, (c) this analysis also underscored ATRIUM's ability to handle knowledge partitioning even when the exemplar module is absent, thus further clarifying the crucial role of the mixture-of-experts approach.

Individual Differences and ATRIUM

When fit to the training data of the systematic-context condition of Experiment 1, ATRIUM always predicted knowledge partitioning, even when modeling performance of the TB participants. Thus, ATRIUM predicts that all participants would partition their knowledge in the systematic-context condition. This is at odds with the data, which show that only about 30% of participants partition their knowledge, whereas another 30% learned the true boundary. Could ATRIUM capture those individual differences

without being given explicit information about participants' transfer responses?

We once again examined this issue by initiating training with randomly chosen attention strengths for context. Using the common parameter estimates from the fit to the training data of the KP and TB groups in the systematic-context condition (see top panel of Table 9), we conducted 100 simulation runs with the initial attention to context randomly sampled for each replication from a normal distribution with a mean of .15 and a standard deviation of .07 (initial attention weights were constrained to sum to unity and were equal for x and y). Figure 8 shows ATRIUM's transfer responses for the top and bottom 30% (left and right panel, respectively) of the distribution of initial attention to context. The obtained average initial attention to context was .23 and .05, respectively, for the left and right panels. The values of the module gains for the top and bottom 30% were found to parallel the earlier results for the KP and TB groups (see Table 10). That is, for replications within the top 30%, context uniformly determined which rule module was applied, whereas for the bottom 30%, it was the area from which test items were sampled and context was largely irrelevant.

Figure 8 shows very clearly that, unlike ALCOVE, ATRIUM can reproduce the approximate proportion of participants who partition their knowledge (the 30% of the replications underlying the predictions in the left panel) or learn the true boundary (right panel) by making some very simple assumptions about the distribution of initial attention. When the model randomly emphasizes context at the outset, it partitions knowledge. When the model tends to ignore context at the outset, it learns the true boundary. This suggests that the individual differences in our experiments may likewise have reflected random variation of the initial importance that participants attached to the various dimensions.

Another particularly noteworthy aspect of the results in Figure 8 is that the variation in initial attention strength gave rise to qualitatively different generalization patterns. That is, the divergence

between predictions in the left and right panels occurred despite exposure to an identical training regime and to identical parameter settings.

General Discussion

Summary of Results and Limitations

In Experiment 1, we extended the knowledge partitioning observed by Yang and Lewandowsky (2003) to a perceptual category-learning task. Roughly one third of participants in the systematic-context condition were found to rely on context to choose the appropriate rule for categorization, whereas another one third of participants learned the true parallel boundaries. Participants in the randomized-context condition, in contrast, uniformly learned the true boundary. This outcome was consistent with previous demonstrations of knowledge partitioning (e.g., Lewandowsky et al., 2002; Yang & Lewandowsky, 2003), as was the additional fact that training performance was better in the systematic-context condition than it was in the randomized-context condition. In addition, in Experiment 2, we showed that when people partition their knowledge, the components of knowledge are largely independent, and there is little evidence of cross-referencing.

One limitation of the present studies and those reported by Yang and Lewandowsky (2003) is that the stimuli arguably constituted a rule-based categorization task. In light of recent evidence, reviewed at the outset of this article, that rule-based categorization may differ from other tasks in which information must be integrated at a predecisional stage (e.g., Maddox et al., 2003; Waldron & Ashby, 2001), it follows that knowledge partitioning need not necessarily also occur with information-integration tasks. Work is currently underway in our laboratory that examines partitioning in information-integration tasks.

The relationship between knowledge partitioning and several related issues (e.g., the role of correlated features, configural vs. elemental processing) was discussed by Yang and Lewandowsky (2003). Here we focus on (a) the theoretical implications of the observed knowledge partitioning and (b) the large individual differences observed in both experiments.

Theoretical Implications

Exemplar-Based Models of Categorization

The fact that participants used context even though they demonstrably knew that it did not directly predict category membership (contingency ratings in Experiment 2) challenges exemplar models that assume that people learn to pay attention to relevant dimensions only. In accordance, the modeling showed that ALCOVE, an exemplar theory, cannot account for knowledge partitioning except under some very arbitrary circumstances. In examining the model, we identified a shift of attention away from context during learning as responsible for this failure. At first glance, we may seem to attribute the failure to the specific learning mechanism in ALCOVE rather than to exemplar representations in general. However, this possibility was compromised by the fact that when knowledge partitioning was produced in ALCOVE (by its arbitrarily focusing attention on context at the outset and estimating parameters from the KP transfer data), the theory simultaneously

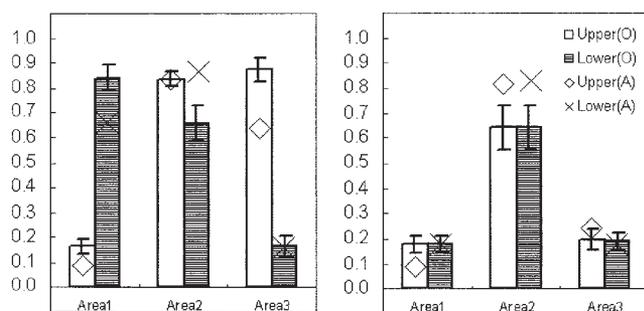


Figure 8. The predicted transfer performance of the attention to rules and instances unified model (A; represented by diamonds and crosses for upper and lower context, respectively) when initial attention strength to context randomly varied across 100 replications with other parameters kept constant. The left and right panels show predictions for the top and bottom 30%, respectively, of the distribution of initial attention to context. Observed responses and standard errors (O; all from Experiment 1) are represented, respectively, by white and black bars for the upper and lower context. Observed results in the left panel represent the knowledge-partitioning group, and those in the right panel represent the true-boundary group.

lost its ability to handle true-boundary behavior. Thus, exemplar representations can account for the behavior of one or the other of our principal groups of participants, but they cannot account for both.

Mixture-Of-Experts Models of Categorization

The modeling identified ATRIUM, a mixture-of-experts model of categorization, as a powerful explanation of knowledge partitioning. Not only did ATRIUM capture partitioning, but it also produced true-boundary behavior when fit to the transfer responses of TB participants. Moreover, the model was able to capture the observed distribution of individual differences by randomly varying the initial attention given to context. Those slight variations in initial attention engendered the qualitatively different ways in which the task was mastered, with the model resembling either the TB or KP group at the end of training. The crucial property of ATRIUM that underpinned its success was the concept of representational attention, or the ability to use different components of knowledge to classify different stimuli.

However, as we show next, not all hybrid models are likely to handle our data and, conversely, there may be several potential versions of ATRIUM that are compatible with the present results. For example, RULEX (Nosofsky et al., 1994) may have difficulty accommodating our data because it first searches for suitable rules on the basis of the predictiveness of single dimensions. It follows that because context was not directly predictive, RULEX is unlikely to include it in one of its rules.

Anderson and Betz (2001) proposed another hybrid model of categorization that implements Nosofsky and Palmeri's (1997, 1998) exemplar-based random walk and continuous RULEX models within the adaptive control of thought-revised (ACT-R) framework (Anderson & Lebiere, 1998). We suggest that the Anderson and Betz model also cannot handle knowledge partitioning for two related reasons. First, the rule module in ACT-R was identical to that used by RULEX, and so our preceding comments about the likely limitations of RULEX also apply here. Second, Anderson and Betz explicitly stated that in their model, the "... decision to use a rule-based or exemplar-based approach is determined by the overall success of these approaches, rather than by the success with respect to a particular stimulus ..." (p. 642). This appears immediately incompatible with our finding that choice of rules is context-dependent and requires a gating mechanism that determines which module to apply to a particular stimulus. Neither of these considerations applies to ATRIUM.

Indeed, it turns out that there are several ways in which ATRIUM could have produced similar results. Erickson and Kruschke (2002) stated that there is no in-principle reason why ATRIUM, instead of combining rules with exemplars, could not instead use several gated exemplar modules. There is little doubt that this alternative conception could also handle our results. In the componential-learning simulation, we had already explored the converse possibility, namely that ATRIUM can accommodate partitioning with multiple rule modules alone and without any contribution from exemplars. We found that the model was able to reproduce the behavior of the KP group.

Because the identity of the potential modules in ATRIUM is not prescribed, we conclude that what is critical for knowledge partitioning is not the combination of rules and exemplars but the

general mixture-of-experts approach (Jacobs et al., 1991) that is embodied in ATRIUM. The defining elements of the mixture-of-experts approach are (a) that each of multiple modules independently contributes partial information and (b) that there is a gating mechanism that adjudicates among different modules in a stimulus-specific manner.

The idea of multiple modules (or experts) also underlies a recent connectionist model of function learning, known as population of linear experts (POLE; Kalish et al., in press). Like ATRIUM, POLE has been shown to handle knowledge partitioning in its explanatory domain, for example, the context-dependent partitioning of function learning that was reported by Lewandowsky et al. (2002) plus several other context-independent partitioning phenomena reported by Kalish et al. (in press).

One particularly noteworthy aspect of POLE is that it learns complex functions by assigning different parts of the task to different partial (linear) functions. For example, when people learn to predict a magnitude (y) that is a quadratic function of a continuous stimulus variable (x), the model assumes that people associate subranges of x with different linear functions (i.e., $y = bx + c$) that are chosen, on the basis of error-correction during learning, from a whole population of such linear experts with different slopes (i.e., different values of b) and intercepts (varying c). POLE therefore implements the idea that knowledge partitioning, rather than being a niche phenomenon, is fundamental to people's ability to learn function concepts. This theme was also present in our application of ATRIUM: The module gain analysis (Table 10) revealed that ATRIUM partitioned its knowledge even in the randomized-context condition—except that the partitioning was performed on the basis of the area from which stimuli were sampled rather than on the basis of context. We therefore suggest that the partitioning of a complex task into several simpler ones that are learned independently of each other is as fundamental to categorization as it is, arguably, to function learning.

In support, the division of knowledge into components is also central to a very recent and innovative model of categorization known as supervised and unsupervised stratified adaptive incremental network (SUSTAIN, Love, Medin, & Gureckis, in press). In SUSTAIN, the internal representation of categories consists of one or more clusters, with additional clusters created during learning on the basis of the surprise created by a novel stimulus. For example, SUSTAIN might start out learning with a single cluster that embodies the knowledge that all things that fly belong to the category *birds*, with another cluster of things that do not fly associated with the category *mammals*. If a surprising stimulus, such as *bat*, is encountered and—withstanding its featural similarity to birds—is assigned to the *mammals* category by corrective feedback, a new cluster is recruited that represents the bat stimulus (and future items that may be similar, such as flying foxes). The category *mammal* is then represented by two clusters, each of which represents knowledge about a subset of exemplars. The similarities between the clustering approach embodied by SUSTAIN and knowledge partitioning are obvious. SUSTAIN's success—it accounts for a large number of classic data in categorization—therefore lends further support to our claim that partitioning is fundamental to categorization.

Explaining Individual Differences

Recent research in categorization has emphasized individual differences (e.g., Lewandowsky et al., 2000; Nosofsky et al., 1989; Nosofsky & Palmeri, 1998; Nosofsky et al., 1994; Yang & Lewandowsky, 2003). Johansen and Palmeri (2002) differentiated between two primary approaches to examining individual differences. One approach relies on a small sample of participants, each of whom provides a large number of observations (perhaps in the thousands), with competing models fitted to each individual separately. The second approach, followed here, involves the testing of a larger number of participants for a single session only, with models evaluated by how well they account for the variability in responses across participants.

We showed that ATRIUM, unlike ALCOVE, was able to explain most of the individual differences in the systematic-context condition with simple assumptions about the initial distribution of attention (see also Erickson & Kruschke, 2001). A particularly noteworthy aspect of this result was the emergence of qualitatively different categorization strategies from small initial differences, which resembles a similar observation made by Kalish et al. (in press) when their POLE model was applied to the function-learning responses of individual participants.

There is much precedent in the literature for the finding that people develop different category structures if they approach learning with different initial states or goals. For example, Barsalou (1983, 1985) investigated the formation of ad hoc categories by presenting four exemplars from different common categories and asking participants to generate an integrative category label. Participants could spontaneously create and use an ad hoc category even if that category involved highly diverse items. Barsalou (1991) showed, in addition, that construction of ad hoc categories can be tied to achievement of a particular goal, a point forcefully extended to categorization and concept use in general by Solomon, Medin, and Lynch (1999). By implication, participants in our experiments, on noticing the grouping of stimuli by color, may have chosen to explore the possibility of classifying the stimuli in some way involving color, thus facilitating the creation of two ad hoc modules that are differentiated by color.

Conclusions

In this article, we made several empirical and theoretical contributions. At an empirical level, we extended the phenomenon of knowledge partitioning from category learning involving numeric predictors to perceptual stimuli. Our stimuli had previously been used to model category-learning data, thus facilitating the application of a pure exemplar-based model (ALCOVE) and a mixture-of-experts model of categorization (ATRIUM) to our results. Both models were able to handle the results of the randomized-context condition, but only ATRIUM could accommodate knowledge partitioning without making extreme assumptions about participants' initial distribution of attention. Moreover, ATRIUM was able to account for the distribution of individual differences observed in our experiments by randomly varying the initial attention paid to context. The success of ATRIUM parallels the explanatory power of a related mixture-of-experts approach in function learning. We suggest that, far from being a niche phenomenon, knowledge partitioning is a fundamental attribute of the learning of many complex tasks.

References

- Aha, D. W., & Goldstone, R. L. (1992). Concept learning and flexible weighting. In J. K. Kruschke (Ed.), *Proceedings of the fourteenth annual conference of the Cognitive Science Society* (pp. 534–539). Hillsdale, NJ: Erlbaum.
- Akaike, H. (1974). A new look at the statistical model identification. *IEEE Transactions in Automatic Control*, *19*, 716–723.
- 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., & Betz, J. (2001). A hybrid model of categorization. *Psychonomic Bulletin & Review*, *8*, 629–647.
- Anderson, J. R., & Lebiere, C. (1998). *The atomic components of thought*. Hillsdale, NJ: Erlbaum.
- Ashby, F. G. (1988). Estimating the parameters of multidimensional signal detection theory from simultaneous ratings on separate stimulus components. *Perception & Psychophysics*, *44*, 195–204.
- Ashby, F. G., Alfonso-Reese, L. A., Turken, A. U., & Waldron, E. M. (1998). A neuropsychological theory of multiple systems in category learning. *Psychological Review*, *105*, 442–481.
- Ashby, F. G., & 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. (1990). Integrating information from separable psychological dimensions. *Journal of Experimental Psychology: Human Perception and Performance*, *16*, 598–612.
- Ashby, F. G., & Maddox, W. T. (1992). Complex decision rules in categorization: Contrasting novice and experienced performance. *Journal of Experimental Psychology: Human Perception and Performance*, *18*, 50–71.
- Ashby, F. G., Maddox, W. T., & Bohil, C. J. (2002). Observational versus feedback training in rule-based and information-integration category learning. *Memory & Cognition*, *30*, 666–677.
- Ashby, F. G., & Perrin, N. A. (1988). Toward a unified theory of similarity and recognition. *Psychological Review*, *95*, 124–150.
- Ashby, F. G., & Townsend, J. T. (1986). Varieties of perceptual independence. *Psychological Review*, *93*, 154–179.
- Barsalou, L. W. (1983). Ad hoc categories. *Memory & Cognition*, *11*, 211–227.
- Barsalou, L. W. (1985). Ideals, central tendency, and frequency of instantiation as determinants of graded structure in categories. *Journal of Experimental Psychology: Learning, Memory, and Cognition*, *11*, 629–649.
- Barsalou, L. W. (1991). Deriving categories to achieve goals. *The Psychology of Learning and Motivation*, *27*, 1–64.
- Erickson, M. A., & Kruschke, J. K. (1998). Rules and exemplars in category learning. *Journal of Experimental Psychology: General*, *127*, 107–140.
- Erickson, M. A., & Kruschke, J. K. (2001). *Multiple representations in inductive category learning: Evidence of stimulus- and time-dependent representation*. Manuscript submitted for publication.
- Erickson, M. A., & Kruschke, J. K. (2002). Rule-based extrapolation in perceptual categorization. *Psychonomic Bulletin & Review*, *9*, 160–168.
- Homa, D. (1984). On the nature of categories. *Psychology of Learning and Motivation*, *18*, 49–94.
- Jacobs, R. A., Jordan, M. I., Nowlan, S. J., & Hinton, G. E. (1991). Adaptive mixtures of local experts. *Neural Computation*, *3*, 79–87.
- Johansen, M. K., & Palmeri, T. J. (2002). Are there representational shifts during category learning? *Cognitive Psychology*, *45*, 482–553.
- Kalish, M. L., Lewandowsky, S., & Kruschke, J. K. (in press). Population of linear experts: Knowledge partitioning and function learning. *Psychological Review*.
- Kruschke, J. K. (1990). *ALCOVE: A connectionist model of category learning* (Cognitive Science Research Rep. No. 19). Bloomington: Indiana University.

- Kruschke, J. K. (1991). *Dimensional attention learning in connectionist models of human categorization* (Indiana University Cognitive Science Research Report, No 50). Bloomington: Indiana University.
- Kruschke, J. K. (1992). ALCOVE: An exemplar-based connectionist model of category learning. *Psychological Review*, *99*, 22–44.
- Kruschke, J. K. (1996a). Base rates in category learning. *Journal of Experimental Psychology: Learning, Memory, and Cognition*, *22*, 3–26.
- Kruschke, J. K. (1996b). Dimensional relevance shifts in category learning. *Connection Science*, *8*, 225–247.
- Kruschke, J. K., & Erickson, M. A. (1994). Learning of rules that have high-frequency exceptions: New empirical data and a hybrid connectionist model. In A. Ram & K. Eiselt (Eds.), *Proceedings of the sixteenth annual conference of the Cognitive Science Society* (pp. 514–519). Hillsdale, NJ: Erlbaum.
- Lewandowsky, S. (1995). Base-rate neglect in ALCOVE: A critical re-evaluation. *Psychological Review*, *102*, 185–191.
- Lewandowsky, S., Kalish, M., & Griffiths, T. L. (2000). Competing strategies in categorization: Expediency and resistance to knowledge restructuring. *Journal of Experimental Psychology: Learning, Memory, and Cognition*, *26*, 1666–1684.
- Lewandowsky, S., Kalish, M., & Ngang, S. K. (2002). Simplified learning in complex situations: Knowledge partitioning in function learning. *Journal of Experimental Psychology: General*, *131*, 163–193.
- Lewandowsky, S., & Kirsner, K. (2000). Expert knowledge is not always integrated: A case of cognitive partition. *Memory & Cognition*, *28*, 295–305.
- Love, B. C., Medin, D. L., & Gureckis, T. M. (in press). SUSTAIN: A network model of category learning. *Psychological Review*.
- Maddox, W. T., Ashby, F. G., & Bohil, C. J. (2003). Delayed feedback effects on rule-based and information-integration category learning. *Journal of Experimental Psychology: Learning, Memory, and Cognition*, *29*, 650–662.
- Medin, D. L., & Schaffer, M. M. (1978). Context theory of classification learning. *Psychological Review*, *85*, 207–238.
- 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. (1987). Attention and learning processes in the identification and categorization of integral stimuli. *Journal of Experimental Psychology: Learning, Memory, and Cognition*, *13*, 87–108.
- 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). Typicality in logically defined categories: Exemplar similarity versus rule instantiation. *Memory & Cognition*, *19*, 131–150.
- 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., Gluck, M. A., Palmeri, T. J., McKinley, S. C., & Glauthier, P. (1994). Comparing models of rule-based classification learning: A replication and extension of Shepard, Hovland, and Jenkins (1961). *Memory & Cognition*, *22*, 352–369.
- 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 D. L. Medin (Ed.), *The psychology of learning and motivation* (Vol. 28, pp. 207–250). San Diego, CA: Academic Press.
- 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.
- 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.
- Posner, M. I., & Keele, S. W. (1968). On the genesis of abstract ideas. *Journal of Experimental Psychology*, *77*, 353–363.
- Reed, S. K. (1972). Pattern recognition and categorization. *Cognitive Psychology*, *3*, 382–407.
- Solomon, K. O., Medin, D. L., & Lynch, E. (1999). Concepts do more than categorize. *Trends in Cognitive Science*, *3*, 99–105.
- 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*, *8*, 168–176.
- Wasserman, E. A., & Berglan, L. R. (1998). Backward blocking and recovery from overshadowing in human causal judgment: The role of within-compound associations. *Quarterly Journal of Experimental Psychology*, *51B*, 121–138.
- Wickens, T. D. (1989). *Multiway contingency tables analysis for the social sciences*. Hillsdale, NJ: Erlbaum.
- Williams, D. A., Sagness, K. E., & McPhee, J. E. (1994). Configural and elemental strategies in predictive learning. *Journal of Experimental Psychology: Learning, Memory, and Cognition*, *20*, 694–709.
- Yang, L.-X., & Lewandowsky, S. (2003). Context-gated knowledge partitioning in categorization. *Journal of Experimental Psychology: Learning, Memory, and Cognition*, *29*, 838–849.

Received October 28, 2003

Revision received January 16, 2004

Accepted January 21, 2004 ■

E-Mail Notification of Your Latest Issue Online!

Would you like to know when the next issue of your favorite APA journal will be available online? This service is now available to you. Sign up at <http://watson.apa.org/notify/> and you will be notified by e-mail when issues of interest to you become available!