Journal of Experimental Psychology:
Learning, Memory, and Cognition
1998, Vol. 24, No. 6, 1411–1436

0278-7393/98/$3.00

Prototypes in the Mist: The Early Epochs of Category Learning

J. David Smith and John Paul Minda
State University of New York at Buffalo

Recent ideas about category learning have favored exemplar processes over prototype processes. However, research has focused on small, poorly differentiated categories and on task-final performances—both may highlight exemplar strategies. Thus, we evaluated participants' categorization strategies and standard categorization models at successive stages in the learning of smaller, less differentiated categories and larger, more differentiated categories. In the former case, the exemplar model dominated even early in learning. In the latter case, the prototype model had a strong early advantage that gave way slowly. Alternative models, and even the behavior of individual parameters within models, suggest a psychological transition from prototype-based to exemplar-based processing during category learning and show that different category structures produce different trajectories of learning through the larger space of strategies.

Categorizing objects into psychological equivalence classes is a basic cognitive task. Descriptions of categorization long favored a generalized prototype principle (Homa, Rhoads, & Chambliss, 1979; Homa, Sterling, & Trepel, 1981; Mervis & Rosch, 1981; Posner & Keele, 1968, 1970; Rosch, 1973, 1975; Rosch & Mervis, 1975). Humans were supposed to average their exemplar experience to derive the category's prototype, compare new items to it, and accept the items as category members if similar enough.

More recently, though, some have argued that prototypes are an insufficient organizing principle for categories (Murphy & Medin, 1985). Formal treatments have shown that prototype models sometimes poorly describe humans' performance (Medin & Schaffer, 1978; Nosofsky, 1987, 1992). Empirical studies have challenged the prediction of prototype theories that humans should find linearly separable categories especially learnable (Medin & Schwanenflugel, 1981). As a result, prototype-based descriptions of categorization performance have been treated critically or marginalized (McKinley & Nosofsky, 1995, 1996; Nosofsky, 1991, 1992; Shin & Nosofsky, 1992), and the literature has come to favor instead a generalized exemplar principle in categorization.

This possibility, that humans store specific exemplars and use these encapsulated episodes as comparative standards by which to categorize new instances, is an important claim about cognition. The more general the claim, the more important. Yet existing explorations of this exemplar principle have not mapped completely the domain of categorization, with its limitless space of different category structures, exemplar pool sizes, observer populations, performance conditions, and performance levels. Research must still specify the appropriate extension of the exemplar principle by evaluating, for example, whether different categorization processes apply during different epochs of learning or for different category structures. This article joins others in exploring these two questions (Homa & Chambliss, 1975; Homa, Dunbar, & Nohre, 1991; Homa et al., 1981; Nosofsky, Gluck, Palmeri, McKinley, & Glauthier, 1994; Nosofsky, Palmeri, & McKinley, 1994). This exploration is important because in both regards the literature may have unintentionally exaggerated the generality of exemplar processes in categorization.

Regarding different epochs of learning, participants often receive extensive training before modeling occurs, with training ending when they achieve a criterion of consecutive correct responses or achieve above-chance performance on all the training stimuli (Medin & Schwanenflugel, 1981; Medin & Smith, 1981; Nosofsky, 1986). Then, the prototype-based and exemplar-based descriptions of performance are compared. This research strategy provides a static snapshot of task-final, mature performance. However, at this stage of learning, strong exemplar traces may have arisen, making the exemplar model especially appropriate. Therefore, this approach cannot show humans' first principles in categorization or the early stages of their category learning. These will have been paved over by long training with a restricted set of exemplars, and they will be invisible in a static, end-of-performance snapshot. To see them, you need something more like a video, a video that explores successive stages of learning (see Estes, 1986a, p. 501). We provide such a video here.

In this regard, our research is allied to other studies that have evaluated performance in different stages of category learning (Ahn & Medin, 1992; Estes, 1986b; Homa et al., 1991; Medin, Wattenmaker, & Hampson, 1987; Nosofsky, Kruschke, & McKinley, 1992; Nosofsky, Palmeri, & McKinley, 1994; Regehr & Brooks, 1995). However, whereas some of this research (Ahn & Medin, 1992; Medin et al., 1987; Regehr & Brooks, 1995) has adopted a distinctive sorting paradigm, here we adopt a typical categorization paradigm.

J. David Smith, Department of Psychology and Center for Cognitive Science, State University of New York at Buffalo; John Paul Minda, Department of Psychology, State University of New York at Buffalo.

Correspondence concerning this article should be addressed to J. David Smith, Department of Psychology, Park Hall, State University of New York at Buffalo, Buffalo, New York 14260-4110. Electronic mail may be sent to psysmith@acsu.buffalo.edu.

Whereas previous research (Ahn & Medin, 1992; Medin et al., 1987; Nosofsky, Palmeri, & McKinley, 1994; Regehr & Brooks, 1995) has focused on rule-based descriptions, here we focus on the value of prototype-based descriptions for describing category learning. Moreover, previous research has suggested that participants' progression of strategies might be general across different category structures (i.e., with poorer or better category differentiation, with smaller or larger exemplar pools—see, e.g., Nosofsky, Gluck, et al., 1994; Nosofsky, Palmeri, & McKinley, 1994). Here we analyze the domain of category structures more finely and examine the idea (Homa et al., 1981; Reed, 1978; Smith, Murray, & Minda, 1997) that different category structures foster different trajectories of learning. Finally, previous research has tried to show the general superiority of one class of model over another (Estes, Campbell, Hatsopoulos, & Hurwitz, 1989; Nosofsky, 1992; Nosofsky, Gluck, et al., 1994; Nosofsky et al., 1992). This emphasis on global fit misses the possibilities that we focus on here—that at different points in learning, participants occupy different positions in the larger space of categorization strategies and that some of these differences have theoretical implications.

Regarding different category structures, those typically used have favored exemplar-based processes because they have featured small exemplar pools (about four items per category) and poorly differentiated categories (Medin, Dewey, & Murphy, 1983; Medin & Schaffer, 1978; Medin & Schwanenflugel, 1981; Medin & Smith, 1981; Nosofsky, 1986, 1989, 1991; Nosofsky, Gluck, et al., 1994; Nosofsky, Palmeri, & McKinley, 1994). Medin and Schwanenflugel (1981, p. 365) understood that this kind of category structure can engender a unique task psychology and specialized strategies. It can even turn a seeming categorization task into an identification task in which participants associatively pair whole exemplars and their category labels but have no sense of coherent categories in doing so. This may happen because exemplar memorization is easier and more obvious when few exemplars repeat frequently (Homa & Chambliss, 1975; Homa, Cross, Cornell, Goldman, & Shwartz, 1973; Homa et al., 1979, 1981). This may happen because less differentiated categories weaken the urge to form prototype-based clusters of exemplars. By either account, exemplar-based strategies should dominate for sparse and difficult category structures, and the literature confirms that they do. However, this approach cannot show the categorization strategies humans favor when they face larger, better differentiated categories. We evaluate these strategies here.

In this regard, our research is allied to that of Smith et al. (1997). They analyzed performance when participants learned both smaller, less differentiated categories and larger, more differentiated categories. The former categories produced performance profiles that were fit profoundly better by an exemplar-based model than by a prototype-based model. The latter categories produced many performance profiles that the standard exemplar model failed in systematic ways to capture and that the prototype model fit better. Accordingly, we consider participants' categorization strategies, and the potential of different categorization models for capturing these strategies, for both kinds of category structures.

In this article, we describe four category-learning experiments that examine the path that participants trace through the larger space of categorization strategies as they learn, and we consider how different formal models describe this trajectory. Early in learning, participants show strategies that standard exemplar models accommodate poorly but that simple prototype-based descriptions of categorization accommodate well. Late in learning, exemplar-based models and exemplar-based descriptions of categorization come into their own. These results illuminate humans' early approaches to category learning and the changing character of performance during learning. They also suggest that the progression of strategies during category learning is strongly affected by different category structures.

From a formal perspective, this trajectory through strategy space during category learning is sometimes so pronounced that both the dominant prototype and exemplar models fail to capture it completely. From a psychological perspective, this trajectory can be simply and productively hypothesized to reflect a progression from a strong reliance on prototypes to a strong reliance on exemplar memorization. Both perspectives encourage models and theories of categorization that incorporate prototype-based and exemplar-based representations and processes. Both perspectives encourage the idea that prototypes and exemplars have changing roles and influences during the different stages of category learning and during the learning of different category structures.

Experiment 1

Experiment 1 explored the possibility that participants' early information-processing strategies may be fit better by assuming prototype-based categorization than by assuming exemplar-based categorization. To this end, participants' performance at different stages of category learning was modeled using the basic and most prominent prototype and exemplar models. We predicted that the basic exemplar model would fit performance better late in learning, as it has in many previous studies. We wondered whether the basic prototype model would fit performance better early in learning. Experiment 1 included both linearly separable (LS) and not linearly separable (NLS) category structures because there is continuing interest in both. (LS categories are those that can be partitioned by a linear discriminant function, and for which one can simply sum up the evidence offered separately by each feature of an item and use that sum to correctly decide category membership.)

Two methodological aspects of Experiment 1 arranged a balanced comparison between models. First, Smith et al. (1997) found that prototype models and exemplar models fit best equal numbers of performance profiles when better stocked, better structured categories were used. Therefore, Experiment 1 adopted these category structures and asked whether different models have a selective advantage during different epochs of learning. Second, Smith et al. found that aggregating data over participants before modeling placed the prototype model at an inherent disadvantage, camouflaging its good fit to the performances of many individual

participants. Therefore, Experiment 1 featured the modeling of individual performance profiles.

Method

Participants. Thirty-two students participated to fulfill a course requirement.

Stimuli and category structures. The present category structures were made possible by using six-dimensional stimuli. The stimuli were pronounceable six-letter nonsense words (CVCVCV). Many articles have been devoted to stimulus materials like these, making this a well-understood class of ill-defined category in the literature, and one that has allowed the replication of categorization phenomena found with other stimuli (Jacoby & Brooks, 1984; Smith & Shapiro, 1989; Smith, Tracy, & Murray, 1993; Smith et al., 1997; Whittlesea, 1987; Whittlesea, Brooks, & Westcott, 1994). By focusing on these stimulus materials in Experiments 1–3, we made the critical variations of category structure and exemplar-set size within one stimulus domain so that our comparisons were as interpretable as possible. By using more pictorial materials in Experiment 4, we generalized the principal results.

Stimulus generation began with the creation of four prototype pairs (*hafudo–nivety, gafuzi–wysero, banuly–kepiro, lotina–gerupy*). The first and second members of each pair were designated as Stimulus 0 0 0 0 0 0 and Stimulus 1 1 1 1 1 1, respectively. These prototypes were created randomly but with several constraints to ensure the pronounceability of all stimuli, the orthographic appropriateness of all stimuli, the identical syllabification of all stimuli, and the roughly equal use of all vowels. For example, *q* (which needs two vowels), *c* (whose pronunciation depends on the following vowel), and a final *e* (that can change syllabification) were disallowed. Each prototype pair contained six different vowels and six different consonants. Appendix A shows the LS and NLS category structures and sample stimulus sets.

Each LS category contained one prototype, two stimuli with five features in common with the prototype, and four stimuli with four typical features. There were no exception items. The LS similarity relations were thus fairly homogeneous, with all items sharing a majority of features with their prototype and clustered around it in six-dimensional stimulus space. A prototype strategy, using an additive rule that summed across independent attributes, could allow perfect categorization if used perfectly.

Each NLS category contained one prototype, five stimuli with five features in common with the prototype, and one stimulus with five features in common with the opposing prototype. The similarity relations in the NLS categories were heterogeneous because they contained a group of similar items tightly clustered around the prototype and one outlier stimulus. The clusters of similar instances, combined with the exception items, balanced overall within- and between-category similarity at the level of the LS categories. Even so, the NLS category structure defeats any categorization strategy that depends on an additive combination of featural evidence. For one thing, such a strategy produces categorization errors for Stimuli A7 and B7, which have more features in common with the opposing prototype. For another thing, the NLS stimulus set contains complementary stimulus pairs within each category (i.e., Stimuli A5 and A7; Stimuli B5 and B7) that have no features in common at all. These stimulus pairs rule out successful categorization using any linear discriminant function. That is, any weighting of the independent cues that allows the successful classification of A7 (e.g., a very heavy weighting on the fifth feature) ensures the misclassification of A5.

These general category structures were predetermined to allow the generation of well-matched LS and NLS stimulus sets. Following these assignments, hundreds of stimulus sets were computer generated and screened for LS and NLS structures that matched in several ways. The two category structures finally chosen had identical exemplar–exemplar similarity, both within category (3.88 features) and between categories (2.12 features), and had identical exemplar–prototype similarity, both within category (4.57 features) and between categories (1.43 features). Thus, the two category structures had identical structural ratios (1.83, using the exemplar–exemplar similarities). These similarity calculations were done additively and assumed equal salience for all features.

In addition, LS and NLS categories were matched in the overall informativeness of all attributes. For the LS and NLS categories, Category A and Category B stimuli took the typical value 5, 5, 5, 6, 6, and 5 times, respectively, for Attributes 1 through 6. There were no criterial attributes available in either stimulus set, and any single-letter strategy should have been identically salient and viable in both of them.

LS and NLS stimulus sets were constructed using each of the four prototype pairs.

Procedure

Participants were tested individually, having been randomly assigned to a category structure and to a prototype pair. Words were presented on a computer terminal in blocks of 14 trials; each block was a random permutation of all 14 stimuli. Each participant received his or her own unique stimulus order. Participants responded using the *1* and the *2* keys on the keypad. Correct responses were rewarded by a brief whooping sound generated by the computer; errors earned a 1-s low buzzing sound. A running total of participants' correct responses was displayed at the top of the screen. Trials continued in an unbroken fashion until 392 trials (28 blocks) had been presented. Participants had unlimited time to view each stimulus before responding. The stimulus remained visible after wrong choices during the 1-s error signal.

Entering the experiment, participants were told that they would see nonsense-word stimuli that could be classified as Group 1 (Category A) words or Group 2 (Category B) words. They were further told to

> look carefully at each word and decide if it belongs to Group 1 or Group 2. Type a "1" on the keypad if you think it is a Group 1 word and a "2" if you think it is a Group 2 word. If you choose correctly, you will hear a "whoop" sound. If you choose incorrectly, you will hear a low buzzing sound. At first, the task will seem quite difficult, but with time and practice, you should be able to answer correctly.

Formal Modeling Procedures

The basic exemplar model. In evaluating the exemplar model, we focus on the context model originated by Medin (1975—see also Medin et al., 1983; Medin & Schaffer, 1978; Medin & Smith, 1981) and generalized by Nosofsky (1984, 1986; McKinley & Nosofsky, 1995). Palmeri and Nosofsky (1995) referred to this model as the standard context model. In the exemplar model, the to-be-classified item in the present tasks would be compared with the seven Category A exemplars (including itself if it is a Category A item) and with the seven Category B exemplars (including itself if it is a Category B item), yielding the overall similarity of the item to Category A members and Category B members. Dividing overall Category A similarity by the sum of overall A and B similarity would essentially yield the probability of a Category A response.

The similarity between the to-be-categorized item and any exemplar was calculated in three steps as follows. First, the values

(1 or 0) of the item and the exemplar were compared along all six dimensions. (This simplifying step let us proceed without psychologically scaling the entire six-dimensional stimulus space.) Matching features made a contribution of 0.0 to the overall psychological distance between the stimuli; mismatching features contributed to overall psychological distance in accordance with the attentional weight that dimension carried. In the present model, each dimensional weight was between 0.0 and 1.0, and the six weights were constrained to sum to 1.0.

Second, this raw psychological distance between item and exemplar was scaled with a sensitivity parameter that could vary from 0.0 to 20.0. Larger sensitivity values essentially magnify psychological space, increasing the differentiation among stimuli, increasing overall performance, and increasing the value the exemplar model places on exact identity between the item and an exemplar. Formally, then, the scaled psychological distance (d) between the to-be-classified item (i) and exemplar (j) is given by

$$d_{ij} = c\left(\sum_{k=1}^{N} w_k |x_{ik} - x_{jk}|^1\right)^1,$$

where x_{ik} and x_{jk} are the values of the item and exemplar on dimension k, w_k is the attentional weight granted dimension k, and c is the sensitivity parameter. The use of the exponent 1 in this equation incorporated the plausible idea that the psychological space underlying the present stimuli was organized according to a city-block similarity metric (Nosofsky, 1987, p. 89).

Third, the similarity (η_{ij}) between the item and exemplar was calculated by taking $\eta_{ij} = e^{-d_{ij}^1}$, with d_{ij} being the scaled psychological distance between the stimuli. The use of the exponent 1 in this equation incorporated the plausible idea that similarity in the present case was an exponential-decay function of psychological distance (Nosofsky, 1987, p. 89; Shepard, 1987).

These three steps were repeated for calculating the psychological similarity between a to-be-categorized item and each A and B exemplar. Then, summing across the Category A exemplars ($j \epsilon C_A$) and Category B exemplars ($j \epsilon C_B$), we calculated the total similarity the item had to Category A members and Category B members. In the standard exemplar model, these quantities would yield directly the probability (P) of a Category A response (R_A) for stimulus i (S_i) by taking

$$P(R_A/S_i) = \frac{\sum_{j \epsilon C_A} \eta_{ij}}{\sum_{j \epsilon C_A} \eta_{ij} + \sum_{j \epsilon C_B} \eta_{ij}}.$$

Repeating this process for each of the 14 items, one would derive the performance profile predicted by the model. However, for reasons to be described, an additional guessing-rate parameter was added to the exemplar model as follows. It was assumed that some proportion of the time (G) participants simply guessed Category A or B haphazardly. It was assumed that the rest of the time participants used exemplar-based categorization in the way already described. With the guessing parameter added, the context model had eight parameters—six dimensional weights constrained to sum to 1.0, a sensitivity parameter, and a guessing-rate parameter. We count parameters conceptually in this article (i.e., eight here). However, the constraint on the sum of the dimensional weights means that the number of free parameters in the various models is one less (i.e., seven here).

To find the best-fitting parameter settings of the exemplar model, we seeded the space with a single parameter configuration and calculated predicted categorization probabilities for the 14 stimuli according to that configuration. The measure of fit was the sum of the squared deviations between the 14 predicted probabilities and the 14 observed categorization probabilities of some participant's performance. This measure was minimized during an analysis by using a fine-grained hill-climbing algorithm that constantly altered slightly the provisional best-fitting parameter settings and chose the new settings if they produced a better fit (i.e., a smaller sum of squared deviations between predicted and observed performance). In this way, the algorithm moved toward the best-fitting configuration. To ensure that local minima were not a serious problem in the present parameter spaces, this analysis was repeated by seeding the space with four more quite different configurations of the exemplar model and hill climbing from there. The variance among the five fits tended to be very small, indicating that the minima found were close to global ones.

The basic prototype model. To evaluate the prototype model, we supposed that each to-be-categorized item would be compared with the category prototype along the six independent dimensions, using additive similarity calculations. We assumed additive similarity in order to follow most closely the influential research of Medin and his colleagues (Medin & Schaffer, 1978; Medin & Smith, 1981) and to evaluate a simple and intuitive prototype model. Mismatching features on a dimension contributed 0.0 similarity; matching features contributed the amount of their dimension's weight. The six attentional weights were again constrained to sum to 1.0. So, for example, if an observer allocated attention homogeneously (.166 to each dimension), and a stimulus shared five or four features in common with the prototype, the judged similarity would be .83 or .67, respectively. If a stimulus differed from the prototype in only one sharply attended feature (e.g., a .300 attentional weight), its judged similarity would be .70. In the simplest case, the item's similarity to the prototype could be taken to be the probability of a correct categorization and its complement to be the probability of an error.

An additional guessing-rate parameter was also added to the prototype model. This parameter is especially useful for modeling participants' halting performance during the early epochs of category learning. Without it, for example, the prototype (with perfect self-similarity) would be predicted to be categorized perfectly. Thus, it was assumed that some proportion of the time (G) participants simply guessed Category A or B haphazardly, while using prototype-based similarity as already described the rest of the time (see also Medin & Smith, 1981). So, for example, if a stimulus shared five or four features with the prototype, and a homogeneously attending observer had a guessing proportion of .20, the performances predicted by the prototype model were, respectively, $[(.20/2) + [(1 - .20) \times (5 \times .166)]] = .76$ and $[(.20/2) + [(1 - .20) \times (4 \times .166)]] = .63$. The equivalent guessing-rate parameter for the exemplar model has already been described.

To analyze the behavior of the prototype model with seven parameters and find its best-fitting configuration, we seeded the space with a single parameter configuration and calculated its predicted categorization probabilities for the 14 stimuli. Again we hill climbed toward the best-fitting configuration by minimizing the sum of the squared deviations between the predicted probabilities and some observed performance. Again the space was seeded with four additional parameter settings. The small variance among the five best fits indicated that the minima found were close to global ones.

The basic prototype and exemplar models just described have been extremely influential in the categorization literature. In fact, the basic exemplar model has dominated the literature (Lamberts, 1994, 1995; Medin et al., 1983; Medin & Schaffer, 1978; Medin & Smith, 1981; Nosofsky, 1984, 1985, 1986, 1987, 1988, 1989, 1992; Nosofsky, Clark, & Shin, 1989; Nosofsky et al., 1992; Palmeri &

Nosofsky, 1995; Shin & Nosofsky, 1992; Smith et al., 1997). The two models typify the large family of models in which exemplars are categorized more accurately in accordance with their perceptual similarity to the underlying category representation or representations. They equivalently incorporate this perceptual similarity into the Shepard–Luce choice rule that has long served cognitive psychology. They instantiate with transparency the different commitments of processing using prototypes or stored exemplars, allowing these commitments to be evaluated fairly against humans' performances. Their dominance, their typicality, their equivalent choice rules and their transparent representational assumptions explain their favored status and their use here. We consider variants on both models in the section on Additional Modeling Perspectives.

Results

To analyze performance at different stages of learning, we divided the 392 trials (28 blocks of 14 stimuli) into seven 56-trial segments. This analytic strategy balanced two competing goals. Aggregating more data creates more stable estimates of performance. Aggregating less data isolates purer performance strategies instead of averaging successive strategies. The 56-trial segment is our compromise between more and less.

Smith et al. (1997) demonstrated that fitting models to data aggregated over participants averaged away individual differences in performance. Accordingly, each participant's performance for each trial segment was modeled individually, using both the seven-parameter prototype model and the eight-parameter exemplar model. The resulting best-fitting configurations specified guessing and sensitivity parameters and six attentional weights for the six stimulus features. They also allowed us to assess the degree of fit between predicted and observed performance. Where desired, the 16 outcomes from the modeling could be averaged for the group.

Performance over trial segments. The accuracy data were analyzed using a two-way analysis of variance (ANOVA) with NLS–LS as a between-subjects variable and trial segment as a within-subject variable. This analysis confirmed that significant learning occurred across trial segments, $F(6, 180) = 32.83, p < .05, MSE = 0.005$. By the end of learning, the proportions correct were .81 and .84 for the NLS and LS category structures, respectively.

Guessing and sensitivity parameters over trial segments. Modeling suggested that participants' strategies changed in intuitive ways through time. Guessing, as assessed by the prototype model, steadily declined. When the prototype model was fit to performance on the NLS and LS categories, the correlation of the guessing parameter with trial segment was, $r = -.86, p < .05$, and $r = -.87, p < .05$, respectively. Guessing-rate parameters were lower for the exemplar model overall and were often near zero.

The exemplar model's sensitivity parameter steadily increased over trial segments. When the exemplar model was fit to performance on the NLS and LS categories, the correlation of sensitivity with trial segment was, $r = .97$, $p < .05$, and $r = .93, p < .05$, respectively. High values of the sensitivity parameter for the last trial segment (9.3 and 9.2 for the NLS and LS conditions, respectively) suggest that the exemplars became distinctive and well-individuated for participants and that some processes like exemplar self-retrieval increasingly supported categorization as learning progressed.

The fit of models over trial segments. A key issue in this experiment was whether the basic prototype and exemplar models were each selectively advantaged at different points in learning. Figure 1 shows the fits of the two models (averaged across participants) for each trial segment. An early advantage for the prototype model seemed to wane or reverse as training continued. To assess the significance of this pattern, the fits were entered into a two-way ANOVA with type of model (prototype or exemplar) and trial segment (1 to 7) as within-subject variables. The interaction between type of model and trial segment was significant, both for NLS categories, $F(6, 90) = 3.85, p < .05, MSE = 0.017$, and for LS categories, $F(6, 90) = 5.64, p < .05, MSE = 0.013$.

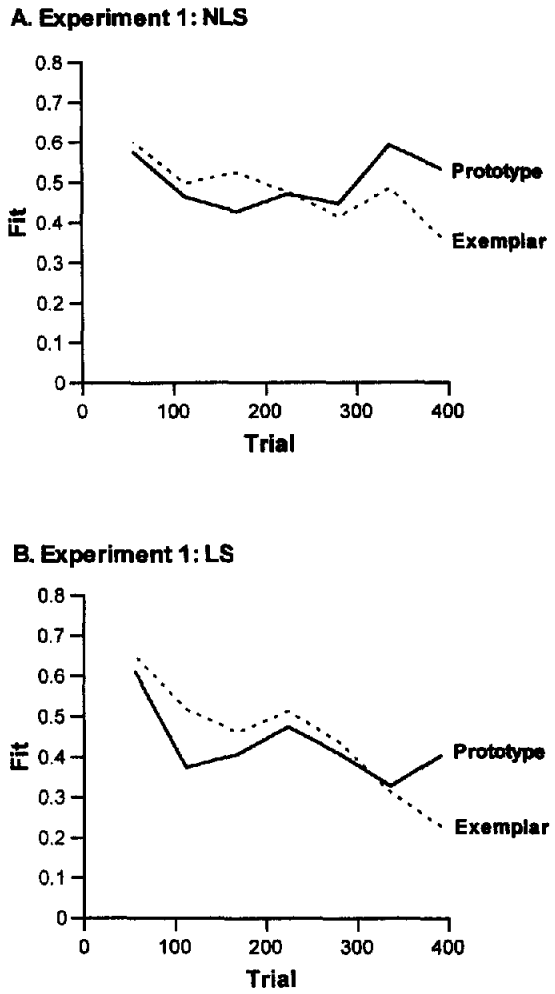


Figure 1. A: The average fit of the prototype and exemplar models at each 56-trial segment to the performance of participants learning not linearly separable (NLS) categories in Experiment 1. B: The average fits for participants learning linearly separable (LS) categories.

To specify the character of this interaction, the ANOVA was repeated, including only the data from the first three trial segments (12 blocks or 168 trials) or the last three trial segments. The prototype model's early advantage over the exemplar model was significant, with average best fits over 48 observations (16 participants at three trial segments) of 0.49 (SD = .250) and 0.54 (SD = .250), respectively, $F(1, 15) = 10.76$, $p < .05$, $MSE = 0.006$, for the NLS categories, and with average best fits of 0.46 (SD = .292) and 0.54 (SD = .261), respectively, $F(1, 15) = 5.99$, $p < .05$, $MSE = 0.025$, for the LS categories.[1]

Moreover, these fit advantages were as large as many of those that have favored exemplar models in past research. For example, Nosofsky (1992) strongly criticized prototype models by summarizing 13 studies in which the exemplar model had a 0.059 average fit advantage over the additive prototype model. Here the prototype model exactly turned the tables, with a 0.065 fit advantage across the two category structures.

The fits for the prototype and exemplar models were statistically equivalent late in training, with average best fits over 48 observations (16 participants at 3 trial segments) of 0.52 (SD = .283) and 0.41 (SD = .250) for the NLS categories, $F(1, 15) = 3.00$, ns, $MSE = 0.084$, and average best fits of 0.38 (SD = .200) and 0.33 (SD = .243) for the LS categories, $F(1, 15) = 1.51$, ns, $MSE = 0.046$. In Trial Segments 5, 6, and 7, respectively, 17 of 32, 16 of 32, and 10 of 32 performance profiles (including both LS and NLS tasks) were fit better by the prototype model. This balanced pattern of best fits replicates exactly the results of Smith et al. (1997), who also modeled performance over Trials 225 to 392 (Trial Segments 5, 6, and 7 here), and who also found that half the profiles were fit better by the prototype model. However, this balance contrasts sharply with the results from many previous studies (Medin & Schaffer, 1978; Medin & Smith, 1981; Nosofsky, 1987, 1992).

Statistically, the equivalence of the models late in the NLS condition occurred because some participants (but not others) passed through a transition that produced excellent fits by the exemplar model but poor fits by the prototype model. These individual differences severely inflated the value of the relevant error term by a factor of 14 (i.e., 0.084 and 0.006 for late and early performance, respectively). Even so, the NLS result approached significance, raising the possibility that extending training would bring the basic exemplar model into favor.

Conceptually, the equivalence between the basic prototype and exemplar models is probably linked to Experiment 1's category structures (see also Smith et al., 1997). Typically studies have used three- or four-dimensional stimuli with about four exemplars per category and with structural ratios of about 1.3. But here the stimulus dimensionality was of a higher order (six dimensions), the categories were better differentiated from one another (structural ratios of about 1.8), and the exemplar pools were larger (seven per category). Any of these factors could have slowed the emergence of strongly differentiated exemplar traces in Experiment 1, or could have set aside the perceived need for those traces, allowing alternative categorization strategies to be more influential. For example, Homa et al. (1981) supported the specific claim that larger exemplar pools foster the emergence of prototype-based categorization strategies.[2]

One sees that the levels of fit obtained by the prototype and exemplar models are higher than typically reported (e.g., Nosofsky, 1992). The reason lies in our modeling of four-block epochs of performance to gain temporal resolution. This modeling strategy means that every observed

---

[1] One might wonder whether the letters combine into integral patterns that would be modeled better by a Euclidean similarity metric, and whether assuming a Gaussian similarity-decay function might let the exemplar model accommodate the present data more comfortably. Accordingly, we modeled each participant's data at each trial segment using all four alternative versions of the context model that are featured in the literature (city-block metric with exponential decay, city-block metric with Gaussian decay, Euclidean metric with exponential decay, Euclidean metric with Gaussian decay; Ashby & Perrin, 1988; Maddox & Ashby, 1993; Nosofsky, 1985, 1989). All versions of the exemplar model produced successions of average fits just like those already shown. All were disadvantaged early on relative to the simple, additive prototype model with only seven parameters.

Note that to the extent one evaluates different Minkowski metrics and decay functions trying to reduce the exemplar model's disadvantage, one starts to grant the exemplar model more parameters, making it complex, unwieldy, and doing it a disservice. It would be helpful if the situation were clarified regarding these alternative versions of the model that may have outlived their usefulness.

[2] We also explored the possibility that rule-based processing, not prototype-based processing, caused the prototype model's advantage. To do so, we first found for each participant at each trial segment the one-dimensional rule (among six) that fit the data best. The rule model was a prototype model, shorn of its guessing parameter (participants using such a simple rule have no need to guess), with all attention placed on one dimension. This model configuration predicts the rule-use pattern of 100% or 0% Category A responses, respectively, as the focal dimension takes on the typical Category A or Category B value. Complete attention to all six single dimensions was modeled, and the smallest fit was assumed to represent the one-dimensional focus closest to participants' actual attentional allocation. The fits achieved by assuming the best one-dimensional rule averaged 1.70 and 1.37 for the NLS and LS category structures, respectively.

We also found for each participant at each trial segment the best-fitting two-dimensional rule. The two-dimensional rule model assumes that participants evenly divided their attention between some combination of two dimensions. This model configuration predicts either 100% or 0% Category A responses, respectively, as the two focal dimensions both take on the typical Category A or Category B values. If the two features disagreed on category membership, participants were predicted to guess given this conflict and to make the Category A response 50% of the time. All 15 two-dimensional rules were evaluated in this way, and the best-fitting combination was assumed to represent the two-dimensional attentional allocation closest to that of the participants. The fits achieved by assuming the best two-dimensional rule averaged 1.14 and 0.88 for the NLS and LS category structures, respectively.

The fits for both one-dimensional and two-dimensional rules were massively worse than the average fits of 0.50 and 0.43 achieved by the prototype model for the NLS and LS category structures. Rule use did not adequately describe performance.

categorization proportion must be either .00, .25, .50, .75, or 1.00, even if the participant's categorization probabilities in some longer run would fall between these steps. Both models must fit more poorly than usual these steplike data, because committing to explain performance on one .25 step makes it more difficult to simultaneously explain performance that sits on 13 other .25 steps. Another way to put this is that some of each performance is error (i.e., a .87 ideal proportion expresses itself as either 1.00 or .75, and neither is exactly right).

To illustrate the effect on fit of modeling 56-trial segments, we turned to a simulation that included 500 simulated exemplar-based categorizers and 500 simulated prototype-based categorizers. Each simulated categorizer was made obedient to a randomly selected configuration of its respective model, with a particular set of attentional weights and particular levels of guessing and sensitivity as applicable. Each categorizer then performed 56 trials (four blocks) in Experiment 1's NLS task, with chance disturbing the categorization probabilities around the long-run values each configuration of the model would produce. That is, if the model predicted 80% Category A responses in the long run, the simulated categorizer had an 80% chance of making that Category A response on each trial, but also a 20% chance of making a B response. The performance for the trial segment was the average of the four independent events associated with each stimulus. Having run each categorizer for 56 trials, we then fit the prototype and exemplar models to each of the 1,000 performances.

Prototype-based categorizers were fit better by the prototype model than by the exemplar model (average fits of 0.49 [SD = .217] and 0.56 [SD = .236], respectively). Exemplar-based categorizers were fit worse by the prototype model than by the exemplar model (average fits of 0.51 [SD = .263] and 0.40 [SD = .236], respectively).

Three points follow from this simulation. First, these fits on known exemplar-based and prototype-based performances are exactly at the level we observed, confirming that our fits are where they should be for modeling with temporal precision. Second, despite the graininess of performance, modeling easily resolves whether the simulated performances were prototype or exemplar based. Third, and most important, the advantage for the prototype model we observed in Experiment 1's early performance (.05 for the NLS category structure, .08 for the LS category structure, and .065 on average) is the same as that found for simulated categorizers that are all perfectly prototype based (0.07). Thus, our observed fit advantages early in performance are as large as they could possibly be in principle, and they are consistent with the possibility that participants' early performances were produced purely in accordance with a simple prototype-based strategy.

The frequent failures of the basic exemplar model, both early in learning and later in learning, illustrate a limitation on that model to which we return. They also urge a broader exploration of different category structures and the different categorization strategies they encourage. However, these frequent failures would be expected if participants were transitioning at different times toward categorization strategies that were more based in the encoding of specific exemplars and that favored an exemplar-based description of categorization performance. This raises the possibility, addressed in Experiment 2, that extending training would produce an advantage for the exemplar model.

Experiment 2

When Experiment 1 ended at Trial Segment 7, it seemed that an advantage might be emerging for the exemplar model. We may have stopped filming our video too soon. By extending training, one might let this advantage express itself fully. Accordingly, Experiment 2 replicated Experiment 1 but gave participants 10 trial segments (560 trials) instead of 7 trial segments (392 trials). Once again we predicted an early advantage for the prototype model, but now we predicted that the often-reported advantage for the exemplar model would assert itself in the end.

Method

Participants. Thirty-two students participated to fulfill a course requirement.

Stimuli and category structures. The stimuli, prototype pairs, and category structures were the same as those of Experiment 1.

Procedure. The procedure was like Experiment 1's except that participants received 40 blocks of the 14 stimuli.

Results

Performance over trial segments. We divided the 560 trials into ten 56-trial segments. Accuracy data were analyzed using a two-way ANOVA with NLS–LS as a between-subjects variable and trial segment as a within-subject variable. This analysis confirmed that significant learning occurred, $F(9, 270) = 46.72$, $p < .05$, $MSE = 0.005$. The task-final proportions correct were .86 and .89 for the NLS and LS category structures, respectively.

Guessing and sensitivity parameters over trial segments. Both the prototype and exemplar models were fit to the data for each 56-trial segment for each participant. Once again the value of the guessing parameter declined over time. When the prototype model was fit to performance on the NLS and LS category structures, the correlations of the guessing parameter with trial segment were −.78 and −.60, respectively. Guessing-rate parameters were lower for the exemplar model overall and soon fell to nearly zero.

The exemplar model's sensitivity parameter steadily increased over trial segments. When the exemplar model was fit to performance on the NLS and LS category structures, the correlations of this parameter with trial segment were .86 and .73, respectively. The task-final sensitivities were 13.50 and 10.05 for the NLS and LS category structures, respectively. This magnification of psychological space is again consistent with participants' coming to hold strongly individuated exemplar traces.

The fit of models over trial segments. Once again we asked whether the two models were each selectively advantaged at different points of learning. Figure 2 shows, for the

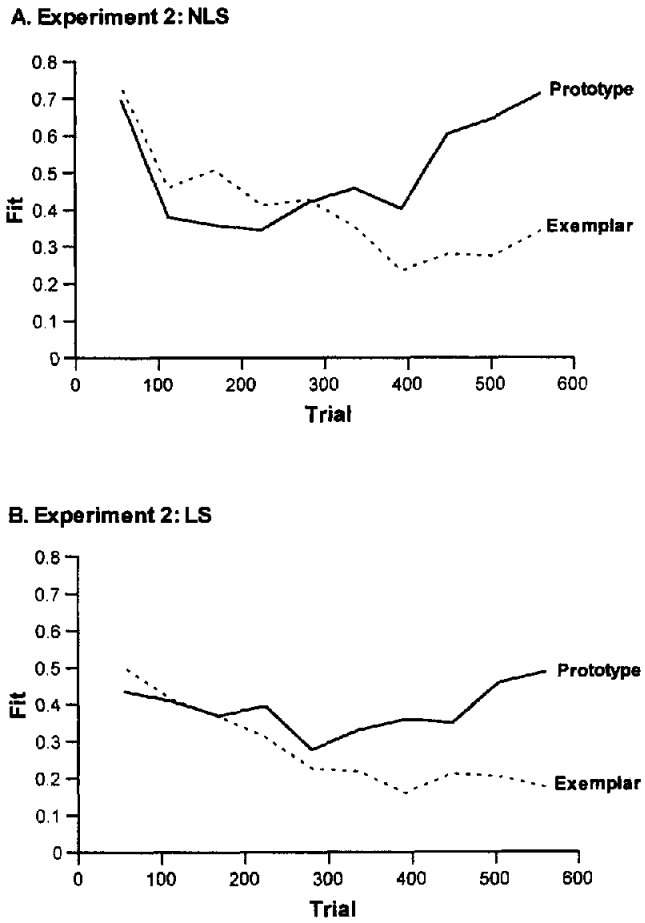


Figure 2. A: The average fit of the prototype and exemplar models at each 56-trial segment to the performance of participants learning not linearly separable (NLS) categories in Experiment 2. B: The average fits for participants learning linearly separable (LS) categories.

NLS and LS category structures, the average fits of the two models for each trial segment. Here there seems to have been an early advantage for the prototype model only for the NLS category structure, and there seems to have been a late advantage for the exemplar model for both category structures.

To evaluate these patterns, the fits were entered into a two-way ANOVA with type of model and trial segment as within-subject variables. There was a main effect for model for the LS category structure only. The average fits were 0.39 and 0.28 for the prototype and exemplar models, respectively, $F(1, 15) = 5.48$, $p < .05$, $MSE = 0.170$. The interaction between type of model and trial segment was significant for NLS categories, $F(9, 135) = 7.83$, $p < .05$, $MSE = 0.041$, and for LS categories, $F(9, 135) = 6.90$, $p < .05$, $MSE = 0.017$.

To interpret this interaction, we repeated these ANOVAs, including first the data from the first three trial segments and then the data from the last three trial segments. The prototype model's early advantage over the exemplar model was significant only for the NLS category structure (average best fits of 0.47 and 0.56, respectively), $F(1, 15) = 10.98$, $p < .05$, $MSE = 0.018$. The exemplar model's late advantage was significant for both category structures. For the NLS category structure, the average best fits for the prototype and exemplar models were 0.65 and 0.30, respectively, $F(1, 15) = 9.40$, $p < .05$, $MSE = 0.323$. For the LS category structure, the average best fits for the prototype and exemplar models were 0.43 and 0.20, respectively, $F(1, 15) = 10.83$, $p < .05$, $MSE = 0.121$.

The character of early performance. Figures 3A and 3B show snapshots that compare observed and predicted performance for both models at Trial Segment 3 (Trials 113–168) of Experiment 2's NLS condition. To make these figures, the 16 observed performances were averaged into the observed composite profile shown. Then each participant's performance during Trial Segment 3 was modeled individually, and the 16 best-fitting predicted profiles were averaged into the predicted composite profile shown. In this way, if each participant's predicted profile matched well his or her observed profile, the observed and predicted composites would match well. If not, the two composites would diverge appropriately.

Figures 3A and 3B make plain that only the prototype model grants the prototype items (Stimuli 1 and 8) their observed performance advantage while simultaneously allowing poor performance on exception items (Stimuli 7 and 14). The basic exemplar model fits less well because it persistently underpredicts and overpredicts performance on the prototypes and exceptions, respectively. It homogenizes performance too much. The NLS category structure diagnoses this failure clearly because it offers participants both prototypes and exceptions. The NLS condition of Experiment 1 produced the same failure by the basic exemplar model, and Experiment 1's LS condition produced this failure too.[3]

[3] LS categories reveal the exemplar model's difficulties less clearly than do NLS categories, because they lack the exceptions that are so diagnostic. This may be why only one of the two LS conditions strongly differentiated the two models early in performance. Even given the exemplar model's failure in the LS condition of Experiment 1, it is more difficult to show the character of that failure than in the NLS case. The problem is that even if some participants regard some LS stimuli as exceptional and perform poorly on them, other participants will regard other stimuli as exceptional. The averaged profiles of observed and predicted performances will wash out these differences and reveal no localized, interpretable failure of the exemplar model.

To preserve participants' individual senses of prototypes and exceptions in the LS category structure, one can rank order performances within subjects and then examine how each model deals with participants' own particular worst, medium, and best category exemplars. This analysis lets one create consensus prototypes and exceptions for the LS category structure, too, by letting each participant define his or her own through his or her performance.

Illustrating this technique with Experiment 1's LS condition, we ranked the observed performances from the third trial segment within subjects and then averaged within ranks across subjects. The prototype model does reach lower and higher than the basic exemplar model does to predict participants' worst and best performances. The significance of this pattern was confirmed with a two-way ANOVA on

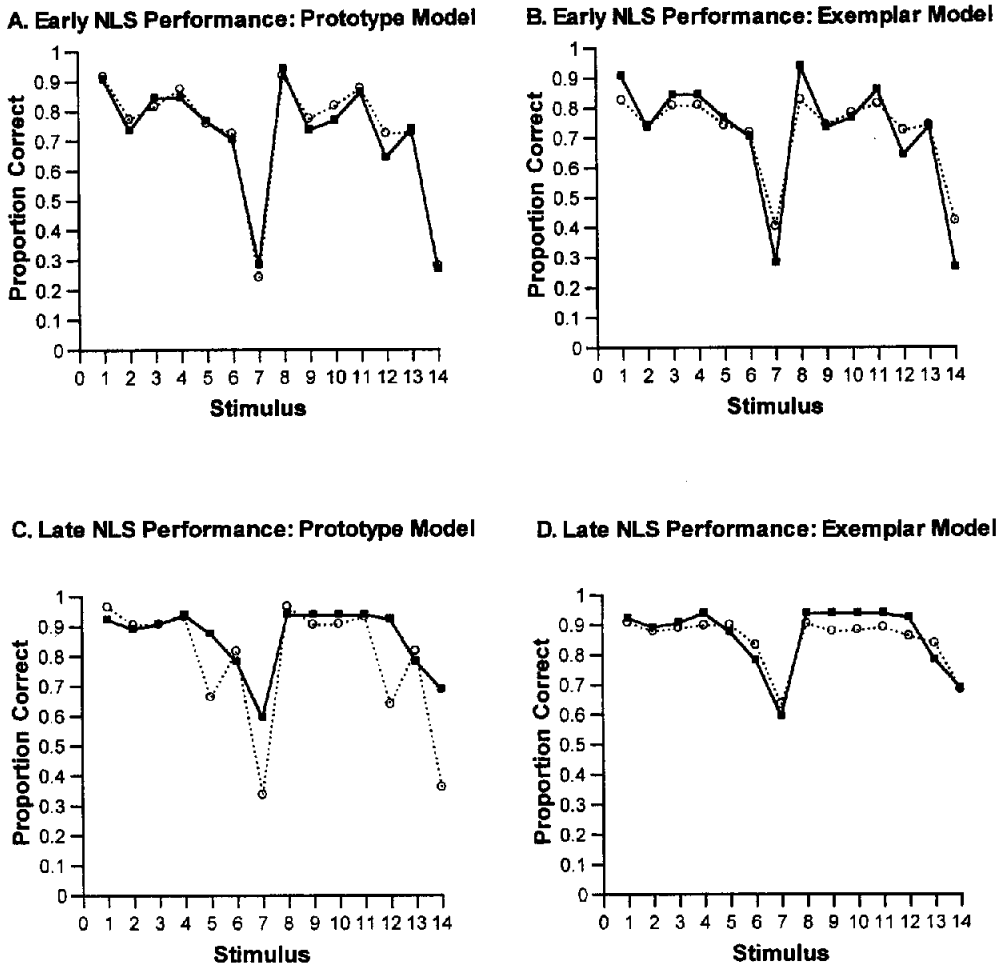


Figure 3. A: The fit of the prototype model to the early performance of participants learning Experiment 2's not linearly separable (NLS) categories. The solid line represents the average observed proportion correct, and the dotted line represents the average predictions of the prototype model. B: The fit of the exemplar model to the same observed performance. C: The fit of the prototype model to the late performance of participants learning Experiment 2's NLS categories. D: The fit of the exemplar model to the same performance.

The character of late performance. Figures 3C and 3D show performance snapshots that compare the models' performances during Trial Segment 10 of Experiment 2's NLS condition (Trials 505–560). The exemplar model fits these performances well. The prototype model cannot. Experiment 2's LS condition produced the same failure by the prototype model. The problem for the prototype model late in learning is that it still must predict that participants classify the stimuli in strict obedience to the typicality gradients in the task. But participants were simply better on the exceptions in the NLS condition and better on all items generally than the basic prototype model can predict.

The results from the later trial segments add a new fact to those established by Smith et al. (1997). Smith et al. (and the present Experiment 1) found that about equal numbers of performance profiles were fit better by the basic prototype and exemplar models for the blocks spanning Trials 225–392. Clearly, though, the basic exemplar model holds a more dominant position when one models performance later in learning (i.e., Trials 505–560). This supports the idea that processing based in specific exemplars comes on more strongly as learning progresses and the training exemplars become highly familiar.

the prediction errors of the two models—that is, with the quantity (observed minus predicted performance) as a dependent measure. Model and stimulus rank were within-subject variables in this analysis. There was a significant interaction, $F(13, 195) = 4.89$, $p < .05$, $MSE = 0.003$, showing that the exemplar model makes larger negative prediction errors for lower performance, but larger positive prediction errors for higher performance. To ground this analysis further, we conducted the identical analysis using Experiment 1's NLS condition and found again a significant interaction, $F(13, 195) = 7.77$, $p < .05$, $MSE = 0.003$. The only difference between the LS and NLS analyses is that the lower NLS ranks are consensually occupied by the true exceptions, whereas the lower LS ranks represent participants' self-defined exceptions.

The trajectory of category learning. Three other views of the data make clear the trajectory that participants traced through the larger performance space as they learned. First, Figure 4A shows, for Experiment 2's NLS condition, the performance variance among 14 stimuli for the composite observed profile for each trial segment. Early in learning (i.e., Trials 1–224) there is a sharp increase in the heterogeneity of observed performance as the task comes into focus for participants. The figure also shows the performance variance among stimuli for the composite predicted profiles of both models at each trial segment. The prototype model predicts correctly the large variance of participants' early observed performances, and this helps the model comfortably fit that early data pattern. At this point, the basic exemplar model fails seriously by predicting only half of the variance that observed performances show. Then, over the next six trial segments, the observed performance variance drops as participants come to perform better on all items. As this reduction occurs, the two models exchange fit advantages (Figure 2A) and the exemplar model comes into its own.

Second, Figures 4B and 4C show the relation between prototype-item and exception-item performance over the 10 trial segments in Experiment 2's NLS condition. Early on there is a profound improvement in performance on the prototypes without any improvement on the exception items. Indeed, through the early trial segments, the exception items essentially are *misclassified* as efficiently as the normal category members are classified. This stage of learning lasts for more than 200 trials—the whole length of many category experiments. Through this misclassification, participants turn the NLS categories into good LS ones, ignoring over 16 presentations of each exception item the corrective feedback and the possibilities for exemplar memorization. It is clear from this reallocation of the exceptions that participants initially make a strong linear separability assumption about the categories they are learning.

The participants' learning trajectory is overlain on the entire constellation of performances that are producible by the basic prototype model (Figure 4B) or the exemplar model (Figure 4C). The early training epochs bring the average performance of the entire sample into a region of performance space that no configuration of the basic exemplar model can occupy, even granted an additional free parameter, but where prototype-based descriptions of performance capture performance well. This is a stronger result than that of Smith et al. (1997), who found that half of their participants occupied this region of performance space. One sees that the basic exemplar model's failure is qualitative at this stage of learning; there simply is no configuration of the model that predicts what the participants average early on. The exemplar model can predict exception performance as low as that shown by participants early on, but then it underpredicts their prototype performance by an average of 19%. The exemplar model can predict prototype performance as high as that shown by participants early on, but then it overpredicts their exception performance by an average of 44%. The problem is that the exemplar model cannot accomplish these two goals simultaneously. The prototype model accommodates this pattern easily. Nonethe-

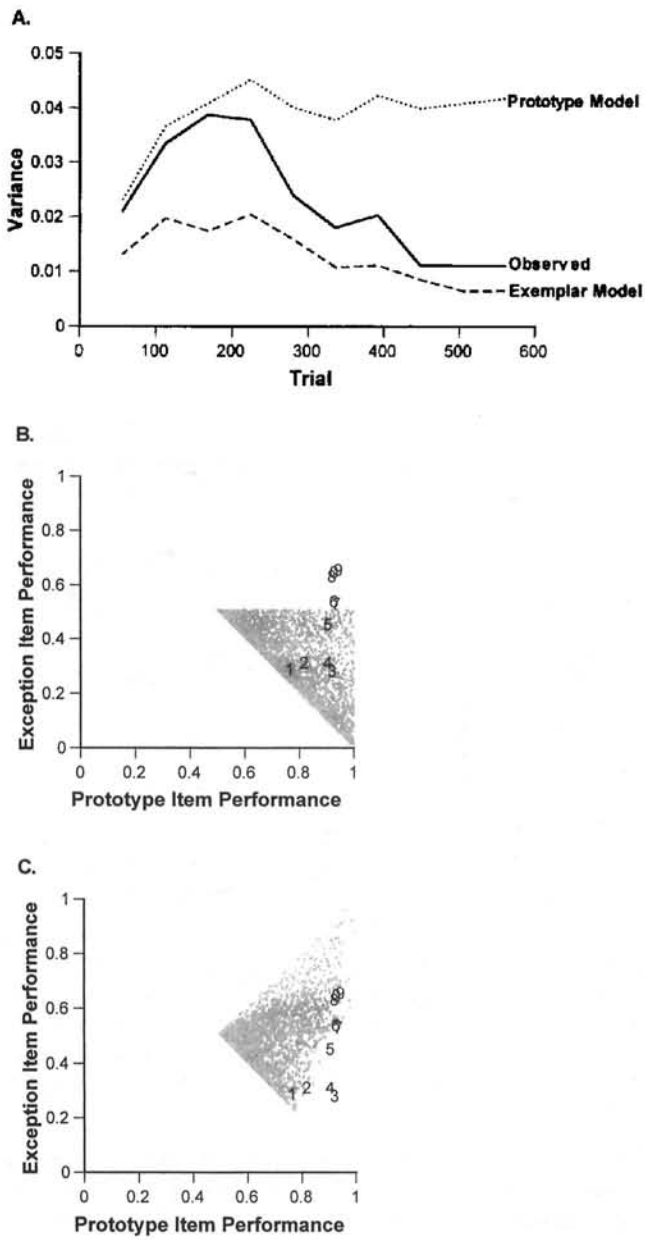


Figure 4. A: The performance variance among 14 stimuli for the composite observed and composite predicted performance profiles at each trial segment (Experiment 2, not linearly separable [NLS] category structure). B: Average prototype-item and exception-item performance by trial segment for participants learning Experiment 2's NLS categories. The 10 trial segments are numbered from 1 to 0. Also shown is the constellation of prototype-exception performances that the seven-parameter additive prototype model can produce. The behavior of the model was found by sampling 3,000 randomly selected configurations of parameter settings. C: The same observed performances together with the prototype-exception performances of 3,000 randomly selected configurations of the eight-parameter exemplar model.

less, by Trial Segment 6, participants enter a region of performance space that no configuration of the prototype model can occupy. They transcend their initial linear separability assumption, using some auxiliary strategies to master the exception items. It is possible that participants memorize the offending members or begin to rely on familiar exemplar traces in some other way. It is possible that participants begin to code multiple stimulus features more configurally or correlationally. Exemplar storage and configural representations are two key features of the exemplar model. The character of later performance replicates many published successes of the basic exemplar model.

Third, many aspects of performance show the same transition from early performances that disfavor the basic exemplar model toward mature performances that favor it. Figures 5A and 5B show the two models' general expectations regarding prototype performance, normal-item performance, exception-item performance, prototype advantage over normal items, and the sum of the squared deviations of the 14 categorization probabilities around their mean (this precursor of the variance is scaled better for inclusion on the graph). To draw the prototype model's behavioral space, we chose 500 randomly selected configurations of the basic prototype model at each value of guessing from 0.00 to 1.00 in .05 steps and found the average performance characteristics at each level of guessing. To draw the exemplar model's behavioral space, we chose 500 randomly selected configurations of the basic exemplar model at each value of sensitivity from 1 to 15 and found the average performance characteristics at each level of sensitivity.

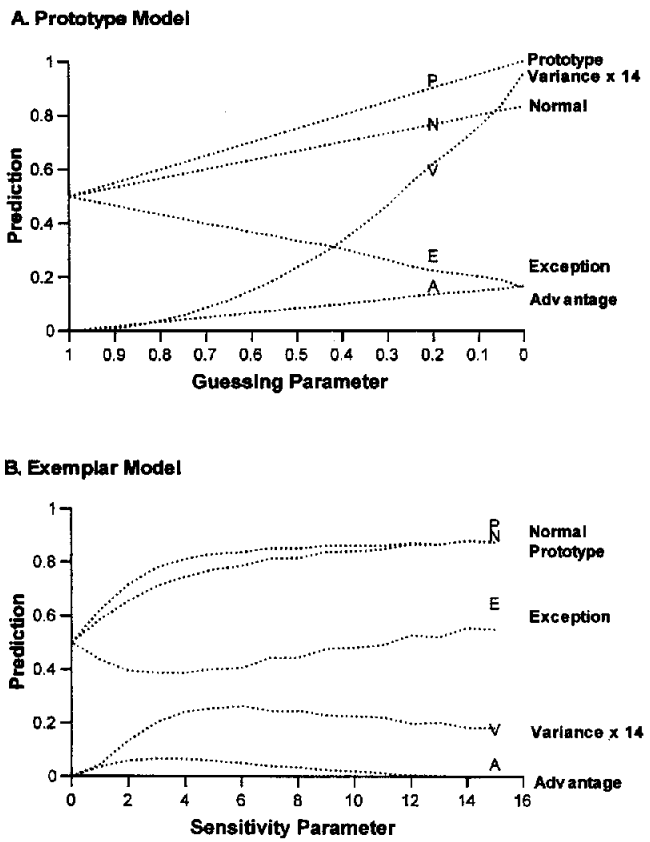


Figure 5. A: The average predictions of the additive prototype model at different levels of guessing (five dotted lines) for five aspects of performance (performance on prototypes, normal items, exception items, the prototype advantage over normal items, and 14 multiplied by the variance among the 14 predicted categorization probabilities [this precursor of the variance is scaled better for inclusion on the graph]). Also shown are the actual performance characteristics (P, N, E, A, and V) observed in Trial Segment 3 of Experiment 2's not linearly separable (NLS) condition. B: The average predictions of the basic exemplar model at different levels of sensitivity for the same five aspects of performance. Also shown are the actual performance characteristics (P, N, E, A, and V) observed in Trial Segment 10 of Experiment 2's NLS condition.

Overlain on Figure 5A are the actual performance characteristics (indicated by capital letters) that participants showed during Trial Segment 3 of Experiment 2's NLS condition. A least-squares procedure allowed us to minimize the five-dimensional error and place the data in their best-fitting spot on this graph. The minimum error distance was 0.0023. It is instructive to imagine sliding these performance characteristics across the behavioral space of the exemplar model shown in Figure 5B. No place along that x-axis offers remotely the right performance characteristics. The minimum error distance (at a sensitivity level of 4.0) was 0.0348, which is 15.1 times as large as that found for the additive prototype model. Thus, participants' early performance characteristics have precisely a prototype-model configuration. The basic exemplar model cannot produce anything like these performance characteristics.

Once again, though, turnabout is fair play (Figure 5B). The performance characteristics at Trial Segment 10 fit somewhat comfortably (minimum error distance of 0.0195 at a sensitivity of 15) on the exemplar model's behavioral space. They fit nowhere in the behavioral space of the prototype model (minimum error distance of 0.1728 at a guessing value of .40, which is 8.9 times the exemplar model's error distance). The stamp of exemplar-based processing may be on these late performances.

These three perspectives all show that the trajectory of category learning through the larger space of categorization strategies is so pronounced that both the basic prototype and exemplar models fail to provide a complete description of all stages of learning. The prototype model fails later on; the exemplar model fails early on. This leads us in the Additional Modeling Perspectives section to consider alternative models that may better capture the whole progression of learning and may illuminate the changing strategies participants choose at different stages. First, though, we demonstrate that different category structures can deflect the progression of learning into very different regions of performance space.

Experiment 3

Experiments 1 and 2 demonstrated that a prototype-based description made a strong showing over the first 200 learning trials for larger, better differentiated categories. In three out of four cases, it fit participants' early performance

profiles better than the basic exemplar model did. In one case, it fit participants' performance profiles as well as the basic exemplar model did. Later in learning, especially in the later trial segments of Experiment 2, the exemplar model assumed the dominant position. To provide a contrast to Experiments 1 and 2, Experiment 3 considered smaller, less differentiated categories. One idea behind the present research is that these categories will encourage exemplar-based strategies to emerge sooner and take hold more strongly. Thus, we hypothesized that humans learning these categories might not perform early on in accordance with a prototype-based description. In fact, we hypothesized that these categories would produce no prototype-model advantage anywhere.

Method

Participants. Thirty-two students participated to fulfill a course requirement.

Stimuli and category structures. The NLS and LS category structures were those used by Medin and Schwanenflugel (1981, Experiment 2). The NLS Category A members were 0 0 0 1, 0 1 0 0, 1 0 1 1, and 0 0 0 0; the Category B exemplars were 1 0 0 0, 1 0 1 0, 1 1 1 1, and 0 1 1 1. The LS Category A members were 1 0 1 0, 0 1 1 0, 0 0 0 1, and 1 1 0 0; the Category B exemplars were 1 1 1 0, 1 0 1 1, 1 1 0 1, and 0 1 1 1. These exemplars share on average only about 2.6 features with their prototypes 0 0 0 0 (Category A) and 1 1 1 1 (Category B), and individual features are poorly predictive (65%) of category membership. Both characteristics reflect the reduced category differentiation in these category structures (structural ratios of about 1.2) compared with those in Experiments 1 and 2 (structural ratios of about 1.8).

The stimuli were derived from the four prototype pairs (*buno–kypa, daki–sego, mufa–vosy, leta–giru*). The first member of each pair was the stimulus 0 0 0 0; the second was the stimulus 1 1 1 1. These prototypes were created subject to the constraints described in Experiment 1.

Procedure. The procedure was like Experiment 2's except that here the 560 trials comprised 70 blocks of the eight stimuli.

Results

Performance over trial segments. We again divided the 560 trials into ten 56-trial segments, each now containing seven blocks of the eight stimuli. Accuracy data were analyzed using a two-way ANOVA with NLS–LS as a between-subjects variable and trial segment as a within-subject variable. Significant learning occurred across trial blocks, $F(9, 270) = 38.26$, $p < .05$, $MSE = 0.012$. The task-final proportions correct were .82 and .93 for the NLS and LS category structures, respectively.

Guessing and sensitivity parameters over trial segments. Both the prototype and exemplar models were fit to the data just as in Experiments 1 and 2. Once again the guessing parameter always declined strongly over trial segments. The estimated guessing rates were substantially higher in Experiment 3 than in Experiments 1 or 2, consistent with these categories' poor differentiation and difficulty.

The exemplar model's sensitivity parameter once again increased strongly over trial segments, up to 10.85 and 20.00 for the NLS and LS category structures, respectively.

The fit of models over trial segments. The fits of the two models for each category structure and each trial segment are shown in Figures 6A and 6B. Once again the fits by participant and trial segment were entered into a two-way ANOVA with type of model and trial segment as within-subject variables. Here we observed a main effect for type of model for the NLS category structure, $F(1, 15) = 21.79$, $p < .05$, $MSE = 0.472$, and for the LS category structure, $F(1, 15) = 29.32$, $p < .05$, $MSE = 0.261$. The interaction between type of model and trial segment was also significant both for NLS categories, $F(9, 135) = 10.56$, $p < .05$, $MSE = 0.045$, and for LS categories, $F(9, 135) = 18.03$, $p < .05$, $MSE = 0.021$.

Early in training (Trial Segments 1–3), the prototype model had no advantage for either category structure. Late in training (Trial Segments 8–10), the exemplar model had the clear advantage. For the NLS category structure, the average best fits for the prototype and exemplar models were 0.73 and 0.13, respectively, $F(1, 15) = 32.84$, $p < .05$, $MSE = 0.265$. For the LS category structure, the average best fits for

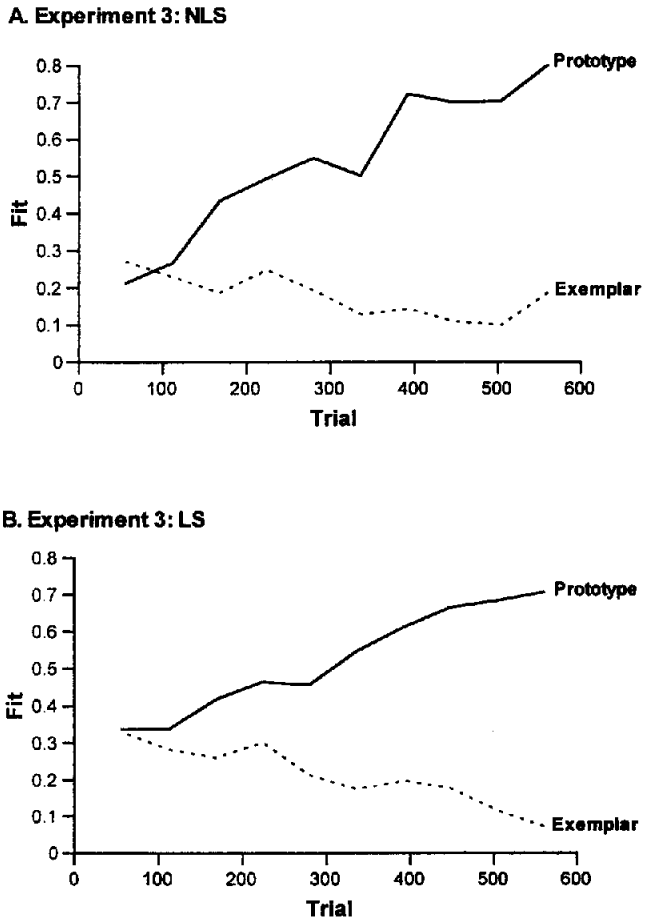


Figure 6. A: The average fit of the prototype and exemplar models at each 56-trial segment to the performance of participants learning not linearly separable (NLS) categories in Experiment 3. B: The average fits for participants learning linearly separable (LS) categories.

the prototype and exemplar models were 0.69 and 0.12, respectively, $F(1, 15) = 57.34, p < .05, MSE = 0.130$.

Moreover, it is clear that participants in Experiment 3 moved very differently through performance space than did those in Experiment 2. For example, Figures 7A and 7B show for each trial segment the average prototype-item performance and exception-item performance by participants in Experiment 3's NLS condition. This trajectory is once again overlain on the entire constellation of performances that are producible by the basic prototype model (Figure 7A) or exemplar model (Figure 7B). Early prototype performance was never so high as in Experiment 2; early exception performance was never so low. Participants always stayed within a region of performance space that the exemplar model can accommodate. They quickly entered a region of performance space that the prototype model cannot accommodate. Their movement up the major diagonal of performance space is consistent with the gradual strengthening of all exemplar traces through time and training epochs. There is none of the backward-L trajectory (Figure 4B) that suggests an early commitment to something like a prototype strategy. Something about Experiment 3's category structure caused participants never to enter this corner of performance space. The result of all these differences is that the prototype model not only never had any advantage in Experiment 3—it never had a chance.

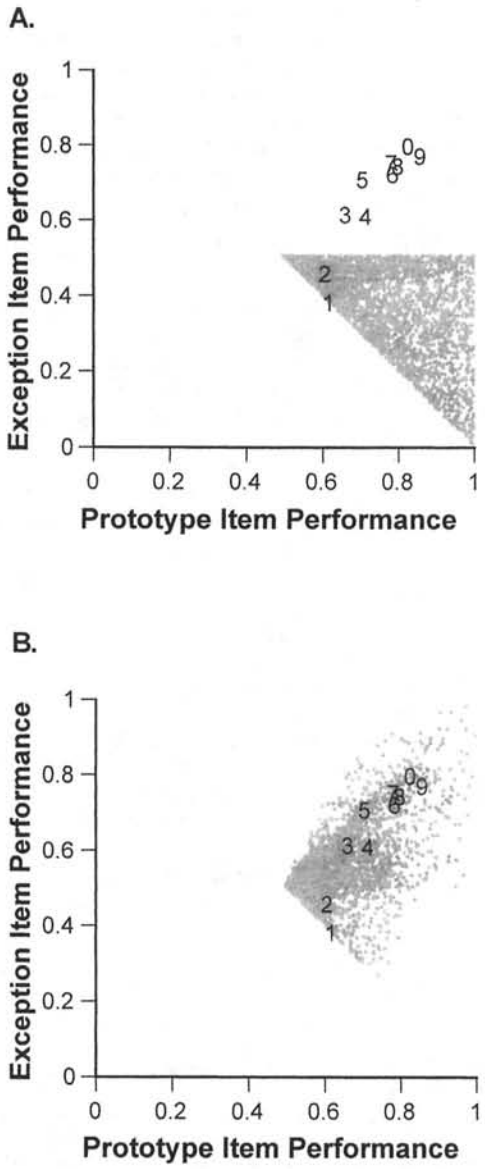


Figure 7. A: Average prototype-item and exception-item performance by trial segment for participants given not linearly separable categories in Experiment 3, together with the prototype-exception performances of 3,000 randomly selected configurations of the five-parameter additive prototype model. B: The same observed performances, together with the prototype-exception performances, of 3,000 randomly selected configurations of the six-parameter exemplar model.

All in all, the results show clearly that the small, poorly differentiated categories of Experiment 3 gave the exemplar-based description of categorization a strong advantage over the prototype-based description, far more so than did the category structures used in Experiments 1 and 2. Of course, many successes of the basic exemplar model have featured small and poorly differentiated categories like those in Experiment 3 (Medin, Altom, & Murphy, 1984; Medin et al., 1983; Medin & Schaffer, 1978; Medin & Schwanenflugel, 1981; Medin & Smith, 1981; Nosofsky, Gluck, et al., 1994; Nosofsky et al., 1992; Nosofsky, Palmeri, & McKinley, 1994; Palmeri & Nosofsky, 1995). Thus, Experiment 3's results confirm all that work.

What is the information-processing basis for these successes? The four-dimensional categories of Experiment 3 featured smaller exemplar pools, more stimulus repetitions, weaker prototypes, and more easily assimilable exceptions. Any of these factors could have discouraged the use of prototypes while encouraging exemplar strategies and reminding participants of them (see also Homa et al., 1981). The six-dimensional category structures of Experiments 1 and 2 featured larger exemplar pools, fewer stimulus repetitions, stronger (more useful) prototypes, and stranger (more disconcerting) exceptions. Any of these factors could have camouflaged and undermined exemplar strategies while encouraging and reinforcing prototype-based strategies. These differences lead us to stress the importance of research programs that explore both general kinds of categories. The reliance on three or four stimulus dimensions has hampered this exploration by constraining the available exemplar pools and levels of category differentiation. It is evident (compare Figures 4B and 7A) that these constraints profoundly alter the course of category learning.

Experiment 4

The importance of these issues led us to replicate and extend our results with different stimulus materials. Therefore, in Experiment 4 participants learned NLS categories with the category structures from Experiments 2 and 3, but with line drawings of bug-like creatures, not nonsense-word stimuli. Once again we predicted that the simple prototype

model would have the fit advantage early in learning over the basic exemplar model when participants learned the large, well-differentiated categories. In contrast, we predicted that the prototype model would be profoundly disadvantaged over the whole course of learning when participants learned the small, poorly differentiated categories.

Method

Participants. Thirty-two students participated in this experiment to fulfill a course requirement.

Stimuli. The stimuli were line-drawn bug-like creatures, shown in profile facing left (Appendix B). These bugs were analogous to the Brunswick faces that have been used extensively in categorization research (Medin & Smith, 1981; Nosofsky, 1991; Reed, 1972; Smith et al., 1993). Different creatures could be created by different combinations of the binary-valued attributes that defined the stimulus space. Other studies have featured similar stimuli such as drawings of starfish (Ahn & Medin, 1994), rocket ships (Palmeri & Nosofsky, 1995), or imaginary creatures (Brooks, 1978; Malt, 1989).

The four binary features used to construct the four-dimensional bugs were as follows: short or long body (a 1.5-cm × 1.0-cm oval or a 2.4-cm × 1.0-cm oval), round or oval head (a 0.8-cm diameter circle or a 0.7-cm × 1.4-cm oval), red open eye or green half-closed eye (a 0.4-cm circle with a 0.2-cm central red dot or a 0.4-cm circle with a green-colored top half), and short or long legs (0.6 cm or 1.0 cm). The six-dimensional bugs also were distinguished by having a curved forward antenna or a straight back antenna (a 1.0-cm forward curve with a purple 0.2-cm terminal dot or a back-angled 0.7-cm straight line with an orange 0.2-cm terminal dot), and by having gray-triangle feet (0.4 cm across the bottom and 0.2 cm high) or blue-semicircle feet (0.4 cm across the bottom and 0.3 cm high).

Pilot experiments. Similarity-scaling experiments were conducted to ensure that the features of the bugs had approximately equal salience. Participants rated pairs of bugs for how alike or different they were. If the average difference between bugs was about the same for all single-feature differences, it would suggest that perceptual salience was approximately balanced across the stimulus dimensions.

Ten participants rated pairs of four-dimensional bugs on a scale of 1 to 5 for how *alike* (1) or *different* (5) they were. Trials were presented in 20 blocks of random permutations of eight trials. There were three possible trial types: bugs that were the same (two trials in each block), bugs that were different on one feature (four trials in each block—one for each feature contrast), bugs that were different on two features (two trials in each block). For the trials containing two-feature differences, the choice of the contrasting features was made randomly. The zero-difference and two-difference trials served only to anchor participants' single-difference judgments, and they were not analyzed further.

The average difference rating was then calculated for each single-feature difference for each participant. Average ratings for the four dimensions ranged from 3.37 to 2.78. These ratings were entered into an ANOVA with feature as the single variable. There was no significant difference in salience over the four dimensions, $F(3, 36) = 1.98$, *ns*, $MSE = 0.396$, suggesting that all of the features had about the same perceptual impact.

Thirteen participants rated pairs of six-dimensional bugs on the same 1–5 scale of increasing difference. Trials were presented in 20 blocks of random permutations of 10 trials. There were three possible trial types: bugs that were the same (2 trials in each block), bugs that were different on one feature (6 trials in each block—1 for each feature contrast), bugs that were different on two features (2 trials in each block).

The average difference rating was then calculated for each single-feature difference for each participant. Average ratings for the six-dimensional bugs ranged from 2.96 to 2.13. These ratings were entered into an ANOVA with features as the single variable. There was no significant difference in salience over the six dimensions, $F(5, 72) = 2.25$, *ns*, $MSE = 0.504$, suggesting again that all of the features had about the same perceptual impact.

Four-dimensional category structure. The four-dimensional NLS categories had the logical structure used in Experiment 3. The logical Category A prototype 0 0 0 0 was made to correspond to a randomly chosen polarity configuration of the binary stimulus features (i.e., 0 referred to a short or a long body for different participants, to short or long legs for different participants, etc.). The Category A stimuli were then generated from that prototype. The logical Category B prototype 1 1 1 1 was always made to correspond to the featural combination that was the opposite of the Category A prototype, and the Category B stimuli were then generated from this Category B prototype. Four random polarity configurations were built for the experiment, and a random quarter of the sample received each configuration.

Six-dimensional category structure. The six-dimensional NLS categories had the logical structure used in Experiment 2. The logical Category A prototype 0 0 0 0 0 0 was made to correspond to a randomly chosen polarity configuration of the binary stimulus features (i.e., 0 referred to a round or oval head for different participants, to a straight-back or curved-forward antenna for different participants, etc.). The Category A stimuli were then generated from that prototype. The logical Category B prototype 1 1 1 1 1 1 always corresponded to the featural combination that was the opposite of the Category A prototype, and the Category B stimuli were then generated from this Category B prototype. Four random polarity configurations were built for the experiment, and a random quarter of the sample received each.

Procedure. Participants were first assigned randomly to one of the two category structures (four dimensional or six dimensional) and to one of the four feature-polarity configurations. Sixteen participants were assigned to learn each category structure. The bug stimuli were presented in blocks of 8 trials (four dimensional) or 14 trials (six dimensional), with each block containing a random permutation of all the stimuli in the experiment. As in Experiments 2 and 3, participants received a total of 560 trials (70 blocks for the four-dimensional categories, 40 blocks for the six-dimensional categories). Trials continued in an unbroken fashion until the 560 trials had been presented.

Participants were tested individually. Each participant was seated at the computer and read the following instructions on the screen:

> In this experiment, you will see a series of line drawings of bugs which can be classified either as "Group 1" bugs or as "Group 2" bugs. Your job is to look carefully at each bug and decide if it belongs to Group 1 or Group 2. Type a "1" on the keyboard if you think it is a Group 1 bug and a "2" if you think it is a Group 2 bug. If you choose correctly, you will hear a "whoop" sound. If you choose incorrectly, you will hear a low buzzing sound. At first, the task will seem quite difficult, but with time and practice, you should be able to answer correctly.

The stimuli were presented on a 11.5-in. (29.5 cm) diagonal computer screen on a white background. Each trial consisted of a drawing of the bug, which appeared slightly to the left of center. Slightly to the right of center, the large numerals *1* and *2* appeared on the screen to remind participants how to respond. Participants had unlimited time to view each stimulus before responding. A correct response was followed by a brief, computer-generated

whooping sound; an error was followed by a 1-s low buzzing sound. The stimulus remained visible after wrong choices during the 1-s error signal.

Results

Performance over trial segments. We again divided the 560 trials into ten 56-trial segments containing four or seven blocks of the 14 or 8 stimuli for the six- and four-dimensional categories, respectively. In both cases, the accuracy data were analyzed with a one-way ANOVA with trial segment as a within-subject variable. In both cases, significant learning occurred across trial blocks: $F(9, 135) = 4.88$, $p < .05$, $MSE = 0.003$, for the six-dimensional categories; $F(9, 135) = 38.14$, $p < .05$, $MSE = 0.006$, for the four-dimensional categories. The task-final proportions correct were .84 and .91 for the six-dimensional and four-dimensional categories, respectively.

Guessing and sensitivity parameters over trial segments. Both the prototype and exemplar models were fit to the data just as in Experiments 1–3. Once again the guessing parameter declined strongly over trial segments. Once again the estimated guessing rates were substantially higher in the four-dimensional condition than in the six-dimensional condition, replicating the contrast between Experiments 3 and 2.

The exemplar model's sensitivity parameter once again increased strongly over trial segments. The task-final sensitivities were 7.65 and 12.30 for the six- and four-dimensional categories, respectively.

The fit of models over trial segments. The fits of the two models through time for the six-dimensional categories are shown in Figure 8A. The fits were entered into a two-way ANOVA with type of model and trial segment as within-subject variables. As in Experiments 1 and 2, there was a significant interaction between type of model and trial segment, $F(9, 135) = 3.83$, $p < .05$, $MSE = 0.015$.

To specify the character of this interaction, the ANOVA was repeated, including only the data from the first three trial segments (12 blocks or 168 trials) or the last three trial segments. The prototype model's early advantage over the exemplar model was significant, with average best fits over 48 observations (16 participants at three trial segments) of 0.32 ($SD = .240$) and 0.37 ($SD = .220$), respectively, $F(1, 15) = 10.92$, $p < .05$, $MSE = 0.005$. The exemplar model's late advantage approached significance, with average best fits for the prototype and exemplar models of 0.43 and 0.30, respectively, $F(1, 15) = 3.93$, $p = .07$, $MSE = 0.100$. The exemplar model failed to show a significant advantage late in Experiment 4 for the same reason it failed to show one late in Experiment 1. Participants were transitioning at different rates toward strategies that favored the exemplar model, and these strategy differences multiplied the relevant error term here by a factor of 20. These strategy transitions are the focus of the present research.

In all respects, the results from the six-dimensional categories replicated those of Experiments 1 and 2 (compare Figures 1A and 2A). Indeed, it appears that the transition to performance strategies that favor exemplar-based descriptions occurred even more gradually with the pictorial stimuli

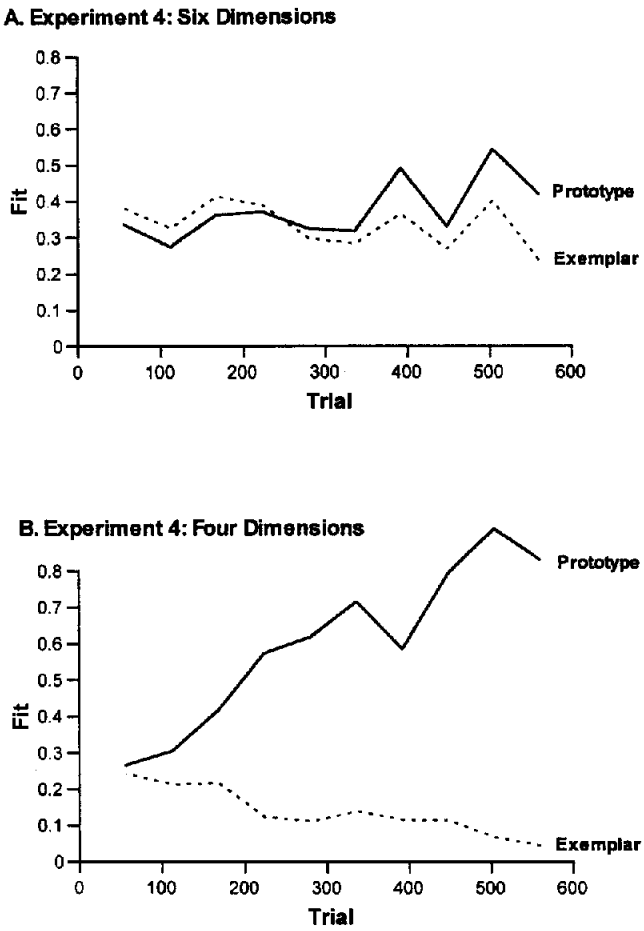


Figure 8. A: The average fit of the prototype and exemplar models at each 56-trial segment to the performance of participants learning six-dimensional categories in Experiment 4. B: The average fits for participants learning four-dimensional categories.

than with the nonsense-word stimuli. Here the standard exemplar model could not even gain a significant fit advantage after 560 trials.

The fits of the two models through time for the four-dimensional categories are shown in Figure 8B. The fits were entered into a two-way ANOVA with type of model and trial segment as within-subject variables. As in Experiment 3, we observed a main effect of model, with the exemplar model fitting far better overall than did the prototype model (average fits of .14 and .60, respectively), $F(9, 135) = 10.56$, $p < .05$, $MSE = 0.045$. The interaction between type of model and trial segment was also significant, $F(9, 135) = 20.29$, $p < .05$, $MSE = 0.031$, indicating that the fit advantage for the exemplar model grew stronger over the course of the experiment. In all respects, the results from the four-dimensional categories replicated Experiment 3 (compare Figure 6A).

Thus, Experiment 4 generalized all the results of Experiments 1–3 from the domain of nonsense-word stimuli to a more pictorial stimulus domain. The very different trajecto-

ries by which participants learn different category structures (including trajectories that favor prototype-based descriptions for large, well-differentiated categories) may be a general phenomenon deserving more research attention.

Summary of Experiments 1–4

In Experiment 3, participants learned small, less differentiated categories. The prototype model was never at an advantage and quickly became seriously disadvantaged. Experiment 4 confirmed this result with different stimulus materials. Both experiments suggest that exemplar processes dominate for the category structures typically used and for the mature performances typically modeled.

In Experiments 1 and 2, participants learned larger, more differentiated categories. The prototype model fit early performance better for both the NLS and LS category structures in Experiment 1 and for the NLS category structure in Experiment 2. Experiment 4 confirmed this result with different stimulus materials. This result illuminates the early stages of category learning and shows the effect of larger, better differentiated categories on learners. Yet there also seemed to be a transition in performance as learning progressed. The basic exemplar model finally gained ascendance in Experiment 2, for both NLS and LS category structures, though more slowly than when participants learned the small, poorly differentiated categories of Experiments 3 and 4. The exemplar model nearly gained ascendance for Experiment 4's large, well-differentiated categories. A measure responsive to this point of ascendance could help assay the relative prominence of exemplar strategies in different categorizing populations facing different category structures.

The prototype model's early advantage is related to the heterogeneity of performance across stimuli. The prototype model performs well when prototype-item performance and exception-item performance diverge strongly in an NLS category structure. This divergence correctly reflects the prototypes' perfect self-similarity and the exceptions' dissimilarity to the prototype of their category. In contrast, the basic exemplar model fits poorly heterogeneous performance profiles.

The exemplar model's late advantage in Experiments 2, 3, and 4 is related to participants' homogeneously good performance across stimuli of differing levels of typicality. Of course this pattern is consistent with performance based in the retrieval of highly familiar patterns. Accordingly, the exemplar model predicts this performance pattern; the prototype model cannot.

Additional Modeling Perspectives

The failure of both models clearly emphasizes the trajectory that participants trace through the larger space of categorization strategies as they learn. Neither model bends flexibly enough to retrace this trajectory. Therefore, we now consider additional models that may better capture the whole progression of learning and that may illuminate further the changing strategies of participants during learning.

Prototypes Combined With Exemplar Memorization: A Mixture Model

First, we examine a simple mixture of prototypes and exemplar memorization. The mixture model adopted here received some early attention (Medin et al., 1983; Medin & Smith, 1981), though it was not used to describe the course of category learning or to trace an increasing reliance on exemplar processes. The mixture model assumes that participants base their classification decisions either on the simple, additive similarity of a given stimulus to the prototype (in which case they obey typicality gradients very strictly), or on random guesses (in which case they place stimuli into Categories A or B haphazardly), or on the recognition of memorized specific exemplars (in which case they definitely classify the item correctly). The key aspect of fitting data with the mixture model is to estimate the balance among these three processes that best accounts for any participant's performance profile.

Note that the exemplar process assumed by the mixture model is simple memorization—that is, individual exemplars are stored, self-retrieved, and self-boosted toward correct categorization. This process is quite different from the context model's exemplar process in which many training exemplars enter the computations that produce a classification decision. In fact, the context model's exemplar-to-exemplar comparison processes have seemed implausible to some (see discussions in Palmeri & Nosofsky, 1995, p. 548; Nosofsky & Palmeri, 1997, p. 292), making it valuable to see whether a simpler exemplar process suffices also.

The mixture model evaluated here contained a guessing parameter that functioned as it did for the other models. It contained an exemplar-memorization parameter that could give all exemplars a performance boost to reflect the successful categorizations that would result from relying on specific, memorized exemplar traces. It contained a prototype-processing parameter that incorporated additive similarity to the prototype into the categorization decision (exactly like the prototype model already described). These three parameters were constrained to sum to 1.0, and the fitting process found the best-fitting mixture of these three alternative processing strategies. The mixture model also contained four or six dimensional weight parameters (for the four- and six-dimensional tasks, respectively) that were constrained to sum to 1.0, just as in the other models. Accordingly, the mixture model, just like the context model, had five or seven free parameters, with two free parameters for the alternative processing strategies, and three or five free parameters for the attentional weights.

The mixture model was fit to the data from Experiment 2's NLS condition for each participant for each trial segment, using the seeding and hill-climbing procedures already described. It re-creates the early advantage of the prototype model over the basic exemplar model (because the specific exemplar parameter can be estimated near zero), and it re-creates the late advantage of the exemplar model over the prototype model (because it can also emulate a strong reliance on memorized exemplars; see Figure 9A). Thus, the mixture model captures performance well throughout learn-

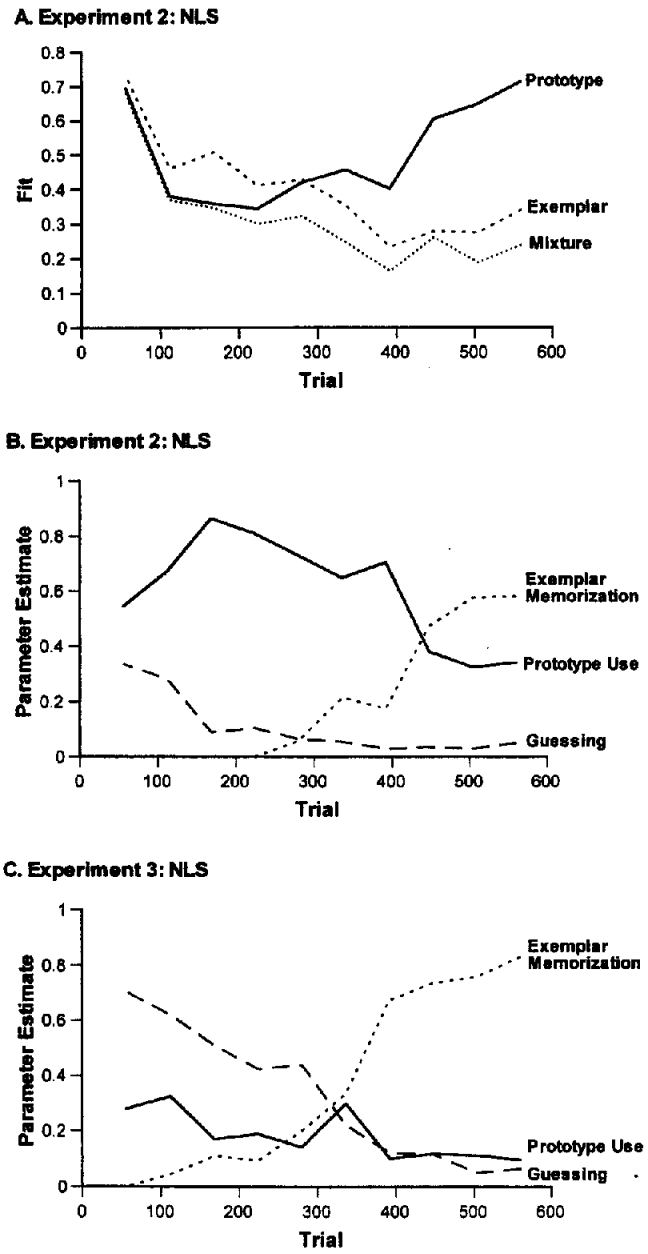


Figure 9. A: The average fit of the prototype, exemplar, and mixture models at each 56-trial segment to the performance of participants learning not linearly separable (NLS) categories in Experiment 2. B: Median parameter estimates across 16 participants of the mixture model fitting these performances. C: Median parameter estimates across 16 participants of the mixture model fitting the performances at each trial segment of participants learning NLS categories in Experiment 3.

ing. It bends more flexibly through strategy space by supposing that two different processes are prominent during different stages of learning.

Figure 9B shows the median estimates of guessing, prototype use, and exemplar memorization across 16 participants when the model tracks these parameter values across the 10 trial segments of Experiment 2's NLS condition. The mixture model's description has guessing tail off early on, yielding to prototype use as the dominant process through 400 trials of the experiment. Gradually, though, estimates of exemplar memorization increase, consistent with an increased reliance on highly familiar traces. This increase begins just where the prototype model falters (Figure 9A) and continues until exemplar memorization finally becomes the dominant categorization process as estimated by the mixture model.

Figure 9C shows the same parameter estimates through time for Experiment 3's NLS condition with small, poorly differentiated categories. This trajectory through parameter space is profoundly different. Higher rates of guessing reflect the slowness with which these categories come into focus for participants. Higher terminal reliances on exemplar memorization eventuate. Most strikingly, the use of prototypes is never the dominant categorization process under the description of the mixture model.

We believe this mixture perspective has many interpretative strengths. First, it makes plain that the trajectory through performance space during category learning is as profound as a transition from strong reliance on prototypes to strong reliance on exemplar memorization. Second, the mixture model points to a face-valid time of guessing during which the task crystallizes for participants. Third, the model isolates a time in learning when a simple prototype model fits the data well, with no guessing and no exemplar process at all (Figure 9B). To our knowledge, this extended stage of learning has not been pointed out before, and we believe it may be an important stage in the learning of many categories. Fourth, the mixture model allows one to consider the possibility that the balance between category representations shifts during learning. Neither prototype nor exemplar models allow this possibility—they are too single-minded. Fifth, the mixture model demonstrates the sufficiency here of a very simple exemplar process (memorization). One may not always need the context model's exemplar-to-exemplar comparisons that have garnered criticism (see discussions in Nosofsky & Palmeri, 1997, p. 292; Palmeri & Nosofsky, 1995, p. 548). Sixth, the mixture model makes plain the different learning trajectories produced by different category structures (Figures 9B and 9C) and lets one consider why different category structures so strongly deflect participants into different regions of performance space. Seventh, the mixture perspective can even help one understand and interpret the parameters of the context model itself. For example, over the last six trial segments of Experiment 2's NLS condition, the correlation between the increasing value of the sensitivity parameter (in the basic exemplar model) and the increasing value of the exemplar memorization parameter (in the mixture model) was $r(94) = .84$, $p < .05$. This close relationship between sensitivity and memorization recommends the consideration of simpler exemplar principles in category research.

By the mixture model's description, something like prototype use precedes something like exemplar memorization in the learning of the six-dimensional categories, whereas exemplar memorization dominates for the four-

dimensional categories. This difference is consonant with the suggestion of Homa et al. (1981) that prototype- and exemplar-based generalization processes, respectively, figure more strongly in the learning of larger and smaller categories.

Even so, it is useful to consider why this sequence (prototypes to exemplars) is the one the mixture model finds (aside from the obvious fact that participants' early and late performance profiles dictate this sequence). Our view is that participants can quickly start to grasp the task's regularities. After only a few trials, they can start to see which features inform well and which should be discounted. In contrast, before specific exemplars can guide categorization, participants must store those exemplars as separate, individuated traces—traces of the sort that might serve the related processes of identification, recognition, and so forth. Then, they must develop categorization cues that are grounded in that separateness and individuation. In our view, these capacities will develop slowly, especially when the categories are large, when the exemplars use the same six binary traits repetitively, when the exemplars are highly confuseable with each other, when the participants are told to categorize the stimuli (not identify or memorize them), and when participants do not even know that the same stimuli will repeat. Remember also that every exemplar can reinforce prototype-level information, whereas in effect participants receive only 1/8 or 1/14 as much specific-exemplar training. For all these reasons, it may be common and natural for regularities (i.e., prototype-level information) to impress themselves on participants earlier than uniquenesses (i.e., exemplar-level information).

Reed (1978) reinforced this idea. He found that category learning proceeded far faster than did item learning. He argued that early in learning (up to about Trial 200), there is not enough "within-categories discrimination of patterns to use a classification rule that requires comparing the distance of a test pattern to the individual exemplars" (p. 617). He argued that participants may not even bother to store the individuated exemplars required by an exemplar model because they have no incentive to under categorization instructions. He concluded that performance during the early and middle stages of category learning was inconsistent with exemplar processes but consistent with prototype abstraction.

For the small, poorly differentiated categories used here, our view is that exemplar-based categorization processes can emerge early and strongly, possibly even turning the categorization task into an identification or item-learning task. Medin and Schwanenflugel (1981, p. 365) expressed this fear. Homa's work (e.g., Homa et al., 1981) also clearly raises this possibility. McKinley and Nosofsky (1995, p. 129) also worried about this possibility. In fact, it is surprising, given this clear concern, that so much research has focused on category structures that favor a priori the exemplar principle.

Recommending research on a range of category structures, Reed (1978, p. 619) said

> As the number of patterns within a category increases, prototype abstraction should improve (Homa et al., 1973), but exemplar learning should decrease. As the variability of patterns within a category is reduced, prototype abstraction should improve (Peterson, Meagher, Chait, & Gillie, 1973), but exemplar learning should decrease.

These factors of category size and within-category coherence are two key factors distinguishing the six- and four-dimensional categories used in all four experiments here. Therefore, the finding that the two category structures produce very different trajectories of learning is just what Reed might have anticipated.

In the next two sections of this article, we consider a complementary modeling approach that may also accommodate successfully the progression of performance profiles produced during learning. To introduce this approach, we first explain why the exemplar model's processing assumptions are the original source of its failure to accommodate early performance. Then, we consider a recent profound modification of the exemplar model that fares better but remains problematic.

Why the Standard Exemplar Model Fails to Describe Early Performance

The source of the standard exemplar model's failure and of the prototype model's success lies in the heterogeneity of early performance—whether indexed by the large prototype advantages, the dismal exception-item performance, or any other assay of performance. Two aspects of the exemplar process assumed by the exemplar model ensure that it will fail to accommodate these kinds of performance profiles.

First, the exemplar model allows exemplar self-retrieval. All the items experienced repeatedly in training—prototypes and exceptions included—can rely on these self-retrieval processes, boosting their performance levels, bringing their performance levels closer together, and homogenizing the overall performance profile. In contrast, the prototype model fits well the strongly heterogenized performances participants produced early in learning, because exceptions are more similar to the prototype of the opposing category (and no exemplar self-retrieval counters this effect), and because the prototype's perfect self-similarity does boost its predicted performance. In fact, the prototype model assumes that the better the prototype is performed, the worse the exceptions will be performed (e.g., across 5,000 random configurations of the prototype model, the correlation between those two performance levels was $-.75$). In contrast, the behavior of the standard exemplar model does not express this relationship (across 5,000 random configurations of that model, the correlation between these performances was .27).

Second, the exemplar model relies on exemplar-to-exemplar comparisons in making category decisions. It is seldom recognized that specific, stored exemplar traces are noisy signals for guiding categorization decisions, because category members are often surprisingly unlike each other. To see this, consider Stimulus A2 in the LS task shown in Appendix A (0 1 0 0 0 0). That stimulus shares 5, 6, 4, 3, 3, 3, and 5 features in common with the Category A stimuli (including itself) and shares 1, 2, 2, 1, 3, 1, and 3 features in common with the Category B stimuli. By a rough calculation (that the exemplar model carries out finely), the evidence

favoring a Category A response for A2 is 29 Category A features shared divided by (29 Category A features shared plus 13 Category B features shared), or about 69%. It is this low because, by storing the specific exemplars, one stores the chaff (the atypical features that individuate the exemplars) along with the wheat (the typical features that define the prototype). As a result, in the comparisons to the exemplars, the evidence base favoring a Category A decision is systematically undercut because, when the exemplars are retrieved, all those featural disagreements are retrieved also and they reduce the surety of a Category A response. Any plausible processing account will realize this reduction, whether participants attend to some dimensions and not others, retrieve some exemplars and not others, or even if they process all dimensions of all exemplars. The opposite effect will occur for exception items with many atypical features. The atypicality signal will be muted on comparison to the exemplars. If strong evidence weakens and weak evidence strengthens, the result will tend to be the homogenized performance profiles that the exemplar model shows in our studies and in others, too.

In contrast, if participants simply compare the stimulus 0 1 0 0 0 0 with the Category A and Category B prototypes, they will find that the evidence favoring a Category A response is roughly 5 Category A features shared divided by (5 Category A features shared plus 1 Category B feature shared), or about 83%. The evidence base favoring the Category A response is stronger than in the case of exemplar storage. This occurs because the prototypes contain only the wheat, not the chaff. As a result, in the comparisons to the prototype, the evidence base favoring a Category A decision is not undercut by any featural disagreements that the other exemplars had with the prototype. Any plausible processing account will realize this strengthening of the evidence base. Once again, exception items will receive the opposite effect (moderate dissimilarity to the exemplars will imply strong dissimilarity to the prototype). If positive evidence and negative evidence both grow more extreme, the result will tend to be the heterogeneous performance profiles that the prototype model shows, and that participants showed early in performance.

Figures 10A and 10B illustrate these contrasting tendencies of the two models by showing the composite behavior of 20,000 random configurations of the prototype and exemplar models, respectively, given the NLS stimulus set of Experiments 1, 2, and 4. Prototype-based processing produces terrible exception-item performance (Stimuli 7 and 14), large prototype advantages (Stimuli 1 and 8), and sprawling performance profiles overall. Its performance profiles look like those that participants produced early in Experiment 2 (Figure 3A). Exemplar-based processing produces much higher exception-item performance, minimal prototype advantages, and cramped performance profiles overall. This statement is not judgmental—in many cases these operating characteristics will be just the ones needed to fit performance. For example, the exemplar model's profiles look exactly like those that participants produced late in Experiment 2 (Figure 3D).

In fact, notice that these cramped kinds of performance profiles will emerge just when one models the aggregate

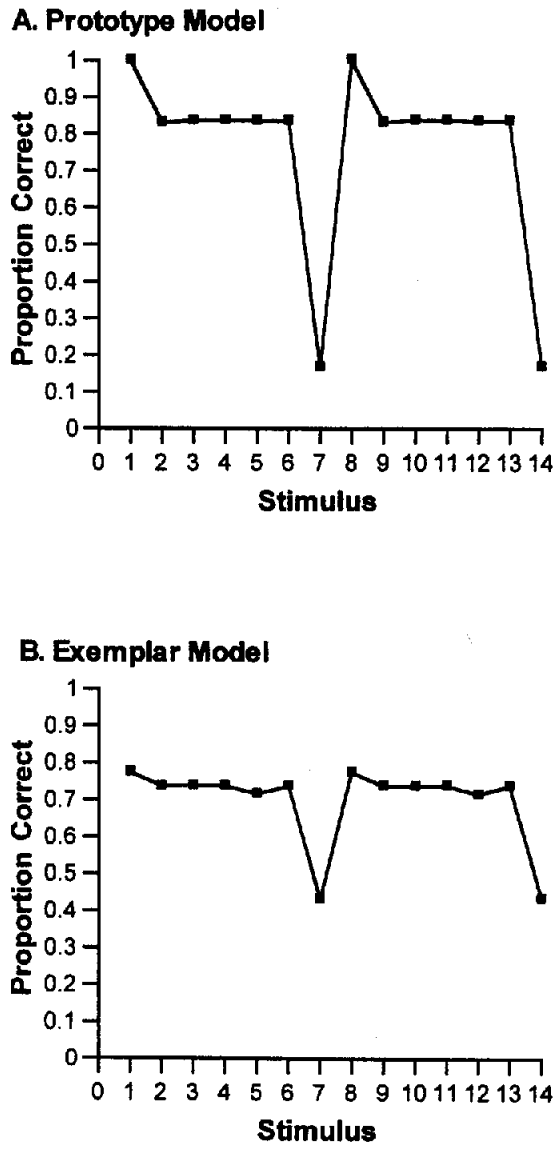


Figure 10. A: The composite performance profile of 20,000 randomly selected configurations of the prototype model on the six-dimensional not linearly separable (NLS) categories. B: The composite performance profile of 20,000 randomly selected configurations of the standard exemplar model on the six-dimensional NLS categories.

performance for a group (and people's idiosyncracies cancel out), when one models task-final performance (with high performance on all stimuli), or when one models performance on small, poorly differentiated categories (and performance on all exemplars improves homogeneously in parallel). That is, the standard exemplar model has had its biggest successes at just the times when methodology guaranteed performance profiles that the model was inherently most comfortable with. This happy confluence of method and model is a possible reason why exemplar models were so successful early on and became so dominant. It is a possible reason to remain cautious about them now.

On the other hand, it is a possibility, not previously considered, that heterogeneous performance profiles like those shown in Figures 10A and 3A may indicate that participants are referring to unitary representations (i.e., prototypes) in the service of categorization, not to the noisy signals sent by specific, stored exemplars.

Gamma

Other research has suggested that exemplar processing, as originally conceived by exemplar theorists and instantiated in 20 years of exemplar models, may be insufficient to explain what individual participants are doing (Ashby & Gott, 1988; Maddox & Ashby, 1993). Therefore, researchers have occasionally modified the context model profoundly by adding on the gamma parameter (Maddox & Ashby, 1993; McKinley & Nosofsky, 1995, 1996). Gamma intervenes by allowing every quantity in the choice rule to be raised to whatever power best recovers participants' actual performance profiles. That is, whereas the choice rule that long served category models was

$$P(R_A|S_i) = \frac{\sum_{j \in C_A} \eta_{ij}}{\sum_{j \in C_A} \eta_{ij} + \sum_{j \in C_B} \eta_{ij}},$$

the augmented version is

$$P(R_A|S_i) = \frac{\left(\sum_{j \in C_A} \eta_{ij}\right)^{\gamma}}{\left(\sum_{j \in C_A} \prod \eta_{ij}\right)^{\gamma} + \left(\sum_{j \in C_B} \eta_{ij}\right)^{\gamma}}.$$

The need for gamma to supplement exemplar processing has strong resonances with the utility of prototype models demonstrated here and in Smith et al. (1997). For example, Smith et al. showed that good prototype-model fits remain hidden while one models group performance. The requirement for gamma also remained hidden while researchers modeled group performance (Maddox & Ashby, 1993). In contrast, the present results show the usefulness of prototype models for modeling individual profiles. This is when gamma is necessary and why gamma was invented.

Moreover, we have shown that prototype descriptions are useful because they naturally produce more heterogeneous performance profiles (Figure 10A) than those of the standard exemplar model (Figure 10B). Gamma produces heterogeneity too. It systematically undoes the homogenizing effect of exemplar–exemplar comparisons and restores the heterogeneity that participants actually show, by raising the choice rule to the 1.8th power, the 4.6th power, or the 9.7th power.

In fact, one can show that gamma has representational entanglements with prototypy that remain to be explored. We discussed earlier how prototypes provide clearer signals than exemplars for guiding categorization. The scatter plot in Figure 11 gives precise mathematical shape to this idea. To make this scatter plot, we chose randomly 1,000 attentional configurations. For each, we found the Category A response proportions predicted by the additive prototype model and the standard exemplar model for the 14 stimuli of the six-dimensional NLS category structure. The scatter plot shows the median Category A response prediction of the prototype model at each level of Category A response prediction of the exemplar model. That is, if a stimulus would prompt the exemplar model to predict that there would be about 40% or 60% Category A responses, the stronger, clearer classificatory signal from the prototype would predict about 20% or 80% Category A responses, respectively.

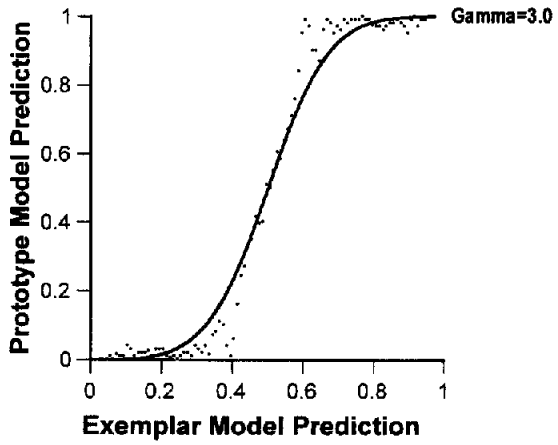


Figure 11. The relationship between the prototype and gamma models. The points represent the median Category A response prediction of the prototype model at each level of Category A response prediction of the exemplar model for the six-dimensional not linearly separable (NLS) category structure (see text for details). The line shows the Category A response prediction of the gamma model ($\gamma = 3$) at each level of Category A response prediction of the exemplar model for the six-dimensional NLS category structure.

Overlain on these points is the Category A response prediction of the gamma model ($\gamma = 3$) at each level of Category A response prediction of the exemplar model. This line shows how gamma-based categorization intrinsically relates to exemplar-based categorization. The gamma line explains 97.9% of the variance in the relationship between prototype-based and exemplar-based similarity. Thus, gamma can arrange a precise mathematical conversion from a pattern of responding consistent with exemplar processing to one consistent with prototype processing.

Given this fact, a variety of criteria, like simplicity and the intuitive framing of psychological questions, encourage a significant role for prototype-based descriptions of data like those from the early trial segments that show clearly every expected feature of prototype processing (Figure 5A). On the other side, advocates of the gamma parameter as an adjunct to exemplar processing must note carefully that in many cases gamma has exactly the effect on categorization profiles that is created by processes that have been the historical and theoretical antithesis of exemplar processing. Gamma must be used and interpreted with caution, for

gamma can be a prototype in exemplar clothing (for a related discussion, see Ashby & Maddox, 1993).

The relation of gamma to prototypy was especially clear for participants' early performances that the exemplar model failed to accommodate (Figure 12A). In such cases, a large value of gamma let the gamma model use more parameters to emulate the behavior of an additive prototype model and better fit the performances during the early trial epochs (Figure 12B).

In doing so, the gamma model chose just the parameters that allowed it to imitate best the simple prototype model. Figure 13 shows the results of the careful fitting procedure that confirmed this. To make the curve for varying levels of gamma, we took 1,000 random attentional configurations, and for each calculated the gamma model's Category A predictions (at 91 levels of gamma from 1 to 10 in steps of .10, with sensitivity always given a random value between 1

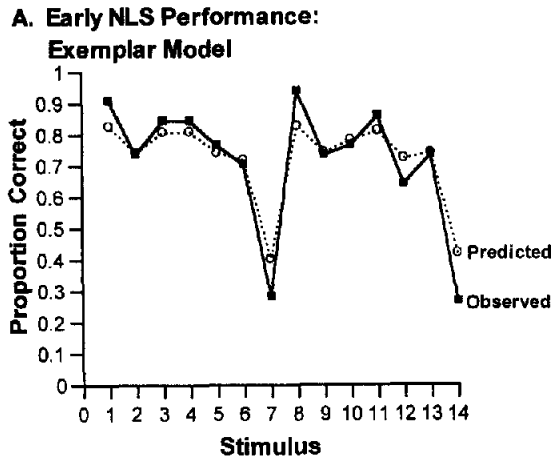


Figure 12. The fit of the exemplar model to the early performance of participants learning Experiment 2's not linearly separable (NLS) categories. B: The composite predicted performance profiles of the gamma and prototype models when they fit the early performances of participants learning Experiment 2's NLS categories.

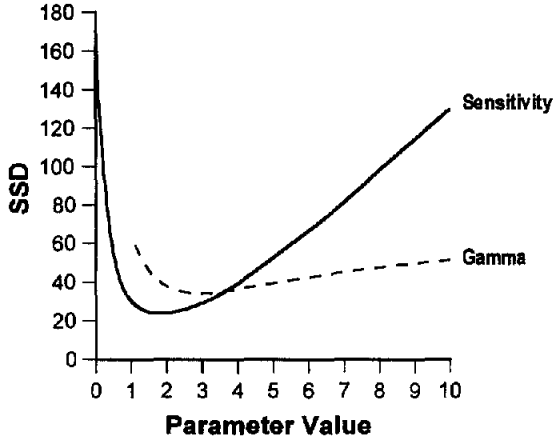


Figure 13. The fit (sum of the squared deviations [SSD]) of the gamma model to the simple prototype model for different levels of gamma and sensitivity (see text for details).

and 10) and the prototype model's Category A predictions, for each of the 14 stimuli in our six-dimensional NLS category structure. We then calculated the sum of the squared deviations (SSDs) between the 14,000 prototype-based predictions and the 14,000 gamma-based predictions at each level of gamma. The line in the figure plots these SSDs by levels of gamma and shows that the fit function was minimized for gammas near 3. For our real participants, the average median gamma estimated over the first three trial segments of Experiment 2's NLS condition was 3.4, nearly the gamma that lets the context model act most prototype based.

By a parallel analysis, we found the levels of sensitivity that let the context model imitate best the prototype model. To make the curve for varying levels of sensitivity, we took 1,000 random attentional configurations, and for each calculated the gamma model's Category A predictions (at 101 levels of sensitivity from 0.0 to 10.0 in steps of .10, with gamma always given a random value between 1 and 10) and the prototype model's Category A predictions, for each of the 14 stimuli in our six-dimensional NLS category structure. We then calculated the SSDs between the 14,000 prototype-based predictions and the 14,000 predictions of the gamma model at each sensitivity. The line plots these SSDs by levels of sensitivity and shows that the fit function was minimized for sensitivities near 2. The average median sensitivity for our real participants over the early trial segments was 2.0, just the sensitivity that lets the context model act most prototype based. Therefore, of the hundreds of trillions of configurations that the gamma model can take on, it takes on, to fit the early performances, just the terribly narrow range of configurations that allows it to imitate exactly a simple prototype model.

Now, this is no coincidence because participants at this stage of learning do occupy a spot in the larger space of categorization strategies that is describable by a simple prototype model. In a sense, the empirical story here is that participants do remain for an extended period in this singular place in the huge space of possibilities. One can then describe this place as one where performance is consistent

with the pure use of additive prototypes. Or one can describe it as the place where gamma equals 3 and sensitivity equals 2 (while wondering what gamma equals 3 and sensitivity equals 2 means, and while being helped by knowing that gamma equals 3 and sensitivity equals 2 means that performance is consistent with the pure use of additive prototypes). We believe the prototype description is simply more illuminating than the gamma description. For example, only the prototype description makes plain the singularity of this place in the larger space of categorization strategies and encourages appropriately sharp research attention to it. Only the prototype description encourages one to ask what kinds of (exemplar?) processes participants add to the mix to move out of this place as learning matures. Only the prototype description emphasizes appropriately the dramatic trajectory that participants trace through the larger space of categorization strategies. The trajectory they trace is as profound as that from performance based in simple prototypes (early in learning) to performance based in memorized exemplars (late in learning). Our view is that learning trajectories of this kind deserve sharper research attention too, because they may reveal participants' early default strategies as they enter category tasks, the singular positions in the larger performance space they sometimes occupy, the succession of strategies that they adopt through learning, and the different successions of strategies fostered by different category structures. Our view is that gamma does not highlight any of these issues as well as do the prototype or mixture perspectives.

General Discussion

Humans' Ultimate Capacities and Natural Predispositions Regarding Categorization

Whatever parameterization of the formal modeling space one chooses, our results are about the trajectory through the large space of categorization strategies that participants trace during the learning of many categories. Early on, participants' performance profiles are consistent with processing based in prototypes. They show large prototype effects, strictly obey intuitive typicality gradients, show dismal exception-item performance, and heterogeneous performance profiles overall. Moreover, their strong reassignment of the exception items shows that they initially make a strong linear separability assumption about category structure—they naturally assume, despite extensive evidence and feedback to the contrary over 200 trials, that the exemplars cluster in similarity space near or around the prototypes.

Later on, all aspects of performance are consistent with a gradual transition to strategies that feature some kind of exemplar processing given highly familiar training exemplars. Many parameterizations of the formal space of categorization strategies would describe later performance in this way, though it remains an open question whether those exemplar processes are to be understood as pure exemplar memorization (as in the mixture model) or as exemplar-to-exemplar comparison processes (as in the context model).

In any case, the demonstration of this trajectory of learning joins results from others (Nosofsky, Gluck, et al., 1994; Nosofsky, Palmeri, & McKinley, 1994; Smith et al., 1997) to assert the importance of organizing principles in early category learning that may not be exemplar based and that are very different from those at the end of extensive category training. Therefore, we believe that it is useful to distinguish two research questions regarding human categorization that the literature has not distinguished sufficiently.

First, there is the question of humans' ultimate categorization capacities, at the end of extended training, when they have developed an asymptotic strategy for coping with difficult category structures. Illustrating research on this question, McKinley and Nosofsky (1995) gave participants 4,000 trials (the amount of training our participants would have had if we had trained them every day for a week) on category tasks that were so difficult that asymptotic performances were still only 81% and 68% in two experiments. Then they modeled the last 300 trials to assess ultimate performance. This research tradition raises important issues. Can humans master exceptions? Can they learn NLS categories completely? Can they defend nonlinear decision boundaries? These issues have been prominent for 20 years (McKinley & Nosofsky, 1995; Medin & Schaffer, 1978; Medin & Schwanenflugel, 1981; Nosofsky, 1987). Experiments like those by McKinley and Nosofsky have shown that humans can do all these things and have shown that exemplar models describe these performances best. However, these experiments cannot reveal humans' natural default assumptions about category tasks or their natural default strategies when entering category tasks. Instead, these experiments bear on humans' ultimate capacities in categorization and on the kinds of category structures they can eventually transcend and learn.

Consequently, there remains the second question about humans' natural, initial default strategies and their assumptions on entering a category task. Do humans naturally assume LS categories and linear decision boundaries? Do they sometimes begin with strategies that are captured especially well by prototype-based algorithms? Do they systematically redefine exceptions into the opposing category in an expression of these assumptions and strategies? The present data suggest that humans do these things too early in category learning, and this may, too, be why prototype models describe these performances best. This question is about humans' defaults and predispositions. We hope that more research will examine them.

Converging Perspectives on Categorization

This examination is beginning. For example, Nosofsky, Palmeri, and McKinley (1994) suggested that participants learn categories by first using a rule to make a straight-edged cut through the stimulus space (see also Ahn & Medin, 1992; Medin et al., 1987; Regehr & Brooks, 1995). They suggested that participants then invoke special learning algorithms (exemplar memorization or configurally coding featural complexes) to master the problematic exceptions.

Nosofsky and his colleagues (Nosofsky, Gluck, et al., 1994; Nosofsky, Palmeri, & McKinley, 1994) have developed these ideas by using the rule-plus-exception (RULEX) model. RULEX supposes that participants initially seek

rules that handle many items in a task. Then, they encode more holistically or configurally the rule's exceptions and remember their appropriate category labels. RULEX successfully predicts various aspects of participants' performance given three- and four-dimensional category structures like those adopted in Experiments 3 and 4. RULEX is even able to track through the epochs of learning the performance levels on prototype and exception items, providing a formal video of performance as we have done here. Nosofsky's participants, like ours in Experiments 3 and 4, show strong improvement on both central and exceptional category members. Both his and our participants move up the major diagonal of the performance spaces shown in Figure 7 (see Figures 6–8 in Nosofsky, Palmeri, & McKinley, 1994).

In contrast, the data from our six-dimensional category task provide a good target for Nosofsky, Palmeri, & McKinley's (1994) model to shoot at and could represent an exception to RULEX. For example, in Experiment 2 participants improved dramatically on prototypes over 224 trials while showing no improvement on exceptions. It would be interesting to know whether the RULEX model can separate so cleanly in time the early improvement on prototypes and the later cognitive work that masters exceptions. Here too, larger, well-differentiated categories could shed a new light on category learning.

It also may be important that our participants' early performances seem to be based in something closer to prototypes than to rules. However, we note that rule-based and prototype-based performance descriptions have some commonalities. For example, both descriptions predict the strong reassignment of exception items; both embody a strong linear separability assumption about category structure and predict performance patterns that reflect that assumption.

Generally speaking, then, the ideas formalized in RULEX converge with our own. RULEX endorses that different epochs of learning involve different categorization strategies and varieties of stimulus coding. It endorses that early learning epochs reveal a strong linear separability assumption. It places exemplar-based processing in later epochs and predicts the exemplar model's advantage there, because it also links exemplar processes to the eventual-transcendence strategies that conquer exceptions. Thus, our research shares with RULEX a perspective that blends what human adults naturally assume and ultimately can do. This kind of perspective could eventually provide a fuller description of human categorization than has been prevalent.

Continuities and Discontinuities With Other Categorizers

Moreover, a focus on epochs of learning, and the different processing strategies that attend them, could also inform the literatures on developmental change in categorization and species differences in categorization. For example, young children may lack the deliberate cognitive set that adults bring to category tasks or the abstraction/rule-use ethic that formal education fosters (Kemler Nelson, 1984; Smith, 1989; Smith & Kemler Nelson, 1984; Smith et al., 1993; Smith & Shapiro, 1989). Young children may also lack the sophisticated metacognitive capacities that let adults monitor error signals sensitively, target problematic exemplars, and enact secondary exemplar-based strategies to master them (Baker, 1985; Brown, Bransford, Ferrara, & Campione, 1983; Nelson, 1992, 1996). One could explore these age differences in category learning by considering the plot line of the 4-year-old's categorization video. This could show the initial commitments children make to either prototypes or exemplars and their ultimate capacities to transcend obstacles and exceptions in category tasks. This research would make constructive contact with existing developmental research, in which young children's task-final categorization processes have been variously described as holistically based in exemplar processes (Kemler Nelson, 1984), holistically based in prototypes (Kemler Nelson, 1984, 1988, 1989; Smith, 1989; Smith & Kemler Nelson, 1984; Smith & Shapiro, 1989), or based in narrow and rule-like attention to single dimensions (Ward, 1988; Ward & Scott, 1987). This research would also link current formal approaches and current debates about exemplars and prototypes to earlier, elegant research exploring young children's progression of hypothesis-testing activities during discrimination learning (e.g., Kemler, 1978).

One could also use the category structures of Experiments 1 and 2 to compare the performances of human and nonhuman animals, to see when humans' unique cognitive capacities come most into play. The human edge might be sharpest at the start of the categorization video (for the early epochs of learning), in which case animals would produce a weaker initial advantage for prototype-based descriptions than humans do. Or the human edge might be sharpest at the end of the video (for the final epochs of learning), in which case animals might stay welded to a primitive linear separability constraint and assumption for longer than humans. Either pattern of results would be interesting, for either pattern would establish both continuities and discontinuities between human and animal species, either in their initial approaches to categorization tasks or in their flexibility at shoring up those approaches given special circumstances.

The crucial point for us is that both patterns of results emphasize the value of granting a time depth to analyses of categorization and to our formal models of those processes, and emphasize the value of understanding the trajectory of the learning processes and the strategy transitions that occur. The moving picture, and not the frozen still, may well provide the richer description of categorizers great and small.

References

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

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. (1993). Relations between prototype, exemplar, and decision bound models of categorization. *Journal of Mathematical Psychology, 37,* 372–400.

Ashby, F. G., & Perrin, N. A. (1988). Toward a unified theory of similarity and recognition. *Psychological Review, 95,* 124–150.

Baker, L. (1985). How do we know when we don't understand? Standards for evaluating text comprehension. In D. L. Forrest-

Pressley, G. E. MacKinnon, & T. G. Waller (Eds.), *Metacognition, cognition, and human performance* (pp. 155–205). New York: Academic Press.

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

Brown, A. L., Bransford, J. D., Ferrara, R. A., & Campione, J. C. (1983). Learning, remembering, and understanding. In J. H. Flavell & E. M. Markman (Eds.), *Handbook of child psychology* (Vol. 3, pp. 77–164). New York: Wiley.

Estes, W. K. (1986a). Array models for category learning. *Cognitive Psychology, 18,* 500–549.

Estes, W. K. (1986b). Memory storage and retrieval processes in category learning. *Journal of Experimental Psychology: General, 112,* 155–174.

Estes, W. K., Campbell, J. A., Hatsopoulos, N., & Hurwitz, J. B. (1989). Base-rate effects in category learning: A comparison of parallel network and memory storage-retrieval models. *Journal of Experimental Psychology: Learning, Memory, and Cognition, 15,* 556–571.

Homa, D., & Chambliss, D. (1975). The relative contributions of common and distinctive information on the abstraction from ill-defined categories. *Journal of Experimental Psychology: Human Learning and Memory, 10,* 351–359.

Homa, D., Cross, J., Cornell, D., Goldman, D., & Shwartz, S. (1973). Prototype abstraction and classification of new instances as a function of number of instances defining the prototype. *Journal of Experimental Psychology, 101,* 116–122.

Homa, D., Dunbar, S., & Nohre, L. (1991). Instance frequency, categorization, and the modulating effect of experience. *Journal of Experimental Psychology: Learning, Memory, and Cognition, 17,* 444–458.

Homa, D., Rhoads, D., & Chambliss, D. (1979). Evolution of conceptual structure. *Journal of Experimental Psychology: Human Learning and Memory, 5,* 11–23.

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

Jacoby, L. L., & Brooks, L. R. (1984). Nonanalytic cognition: Memory, perception, and concept formation. *Psychology of Learning and Motivation, 18,* 1–47.

Kemler, D. G. (1978). Patterns of hypothesis testing in children's discriminative learning: A study of the development of problem solving strategies. *Developmental Psychology, 14,* 653–673.

Kemler Nelson, D. G. (1984). The effect of intention on what concepts are acquired. *Journal of Verbal Learning and Verbal Behavior, 23,* 734–759.

Kemler Nelson, D. G. (1988). When category learning is holistic: A reply to Ward and Scott. *Memory & Cognition, 16,* 79–84.

Kemler Nelson, D. G. (1989). The nature and occurrence of holistic processing. In B. E. Shepp & S. Ballesteros (Eds.), *Object perception: Structure and processes* (pp. 357–386). Hillsdale, NJ: Erlbaum.

Lamberts, K. (1994). Flexible tuning of similarity in exemplar-based categorization. *Journal of Experimental Psychology: Learning, Memory, and Cognition, 20,* 1003–1021.

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

Maddox, W. T., & Ashby, G. (1993). Comparing decision bound and exemplar models of categorization. *Perception and Psychophysics, 53,* 49–70.

Malt, B. C. (1989). An on-line investigation of prototype and exemplar strategies of classification. *Journal of Experimental Psychology: Human Learning and Memory, 15,* 539–555.

McKinley, S. C., & Nosofsky, R. M. (1995). Investigations of exemplar and decision bound models in large, ill-defined category structures. *Journal of Experimental Psychology: Human Perception and Performance, 21,* 128–148.

McKinley, S. C., & Nosofsky, R. M. (1996). Selective attention and the formation of linear decision boundaries. *Journal of Experimental Psychology: Human Perception and Performance, 22,* 294–317.

Medin, D. L. (1975). A theory of context in discrimination learning. In G. H. Bower (Ed.), *The Psychology of Learning and Motivation* (Vol. 9, pp. 269–315). New York: Academic Press.

Medin, D. L., Altom, M. W., & Murphy, T. D. (1984). Given versus induced category representations: Use of prototype and exemplar information in classification. *Journal of Experimental Psychology: Learning, Memory, and Cognition, 10,* 333–352.

Medin, D. L., Dewey, G. I., & Murphy, T. D. (1983). Relationships between item and category learning: Evidence that abstraction is not automatic. *Journal of Experimental Psychology: Learning, Memory, and Cognition, 9,* 607–625.

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

Medin, D. L., & Schwanenflugel, P. J. (1981). Linear separability in classification learning. *Journal of Experimental Psychology: Human Learning and Memory, 7,* 355–368.

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

Medin, D. L., Wattenmaker, W. D., & Hampson, S. E. (1987). Family resemblance, conceptual cohesiveness, and category construction. *Cognitive Psychology, 19,* 242–279.

Mervis, C. B., & Rosch, E. (1981). Categorization of natural objects. In M. R. Rosenzweig & L. W. Porter (Eds.), *Annual Review of Psychology, 32,* 89–115.

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

Nelson, T. O. (Ed.). (1992). *Metacognition: Core readings.* Needham Heights, MA: Allyn & Bacon.

Nelson, T. O. (1996). Metacognition and consciousness. *American Psychologist, 51,* 102–116.

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. (1985). Overall similarity and the identification of separable-dimension stimuli: A choice model analysis. *Perception and Psychophysics, 38,* 415–432.

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). Similarity, frequency and category representations. *Journal of Experimental Psychology: Learning, Memory, and Cognition, 14,* 54–65.

Nosofsky, R. M. (1989). Further tests of an exemplar-similarity approach to relating identification and categorization. *Perception and Psychophysics, 45,* 279–290.

Nosofsky, R. M. (1991). Tests of an exemplar model for relating perceptual classification and recognition memory. *Journal of Experimental Psychology: Human Perception and Performance, 17,* 3–27.

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

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., Kruschke, J. K., & McKinley, S. C. (1992). Combining exemplar-based category representations and connectionist learning rules. *Journal of Experimental Psychology: Learning, Memory, and Cognition, 18,* 211–233.

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., & 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.

Posner, M. I., & Keele, S. W. (1970). Retention of abstract ideas. *Journal of Experimental Psychology, 83,* 304–308.

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

Reed, S. K. (1978). Category vs. item learning: Implications for categorization models. *Memory & Cognition, 6,* 612–621.

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

Rosch, E. (1973). On the internal structure of perceptual and semantic categories. In T. E. Moore (Ed.), *Cognitive development and the acquisition of language* (pp. 111–144). New York: Academic Press.

Rosch, E. (1975). Cognitive reference points. *Cognitive Psychology, 7,* 192–238.

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

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

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

Smith, J. D. (1989). Analytic and holistic processes in categorization. In B. E. Shepp & S. Ballesteros (Eds.), *Object perception: Structure and processes* (pp. 297–323). Hillsdale, NJ: Erlbaum.

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

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

Smith, J. D., & Shapiro, J. H. (1989). The occurrence of holistic categorization. *Journal of Memory and Language, 28,* 386–399.

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

Ward, T. B. (1988). When is category learning holistic? A reply to Kemler Nelson. *Memory & Cognition, 16,* 85–89.

Ward, T. B., & Scott, J. (1987). Analytic and holistic modes of learning family resemblance concepts. *Memory & Cognition, 15,* 42–52.

Whittlesea, B. W. A. (1987). Preservation of specific experiences in the representation of general knowledge. *Journal of Experimental Psychology: Learning, Memory, and Cognition, 13,* 3–17.

Whittlesea, B. W. A., Brooks, L., & Westcott, C. (1994). After the learning is over: Factors controlling the selective application of general and particular knowledge. *Journal of Experimental Psychology: Learning, Memory, and Cognition, 20,* 259–274.

Appendix A

Category Structures Used in Experiments 1 and 2

Category A		Category B	
Structure	Stimuli	Structure	Stimuli
Categories linearly separable			
000000	banuly	111111	kepiro
010000	benuly	111101	kepilo
100000	kanuly	110111	keniro
000101	banilo	101110	kapiry
100001	kanulo	011110	bepiry
001010	bapury	101011	kapuro
011000	bepuly	010111	beniro
Categories not linearly separable			
000000	gafuzi	111111	wysero
100000	wafuzi	011111	gysero
010000	gyfuzi	101111	wasero
001000	gasuzi	110111	wyfero
000010	gafuri	111011	wysuro
000001	gafuzo	111110	wyseri
111101	wysezo	000100	gafezi

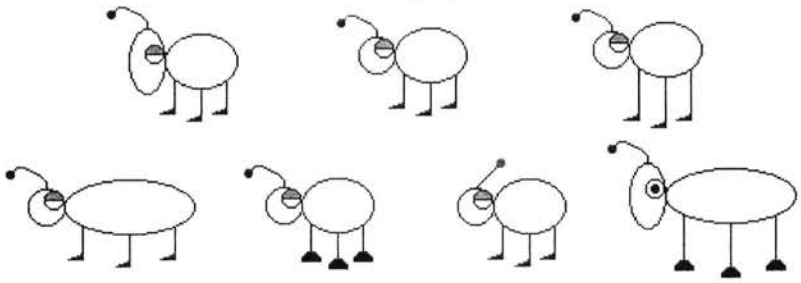
(Appendix B follows on next page)

Appendix B

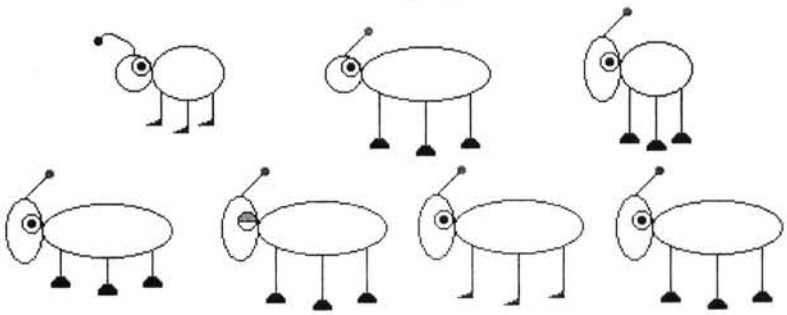
Stimulus Materials Used in Experiment 4

Six-Dimensional Bugs

Category A

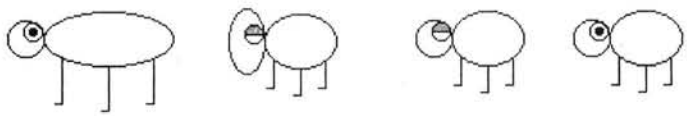


Category B

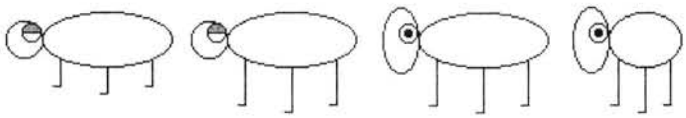


Four-Dimensional Bugs

Category A



Category B



Received December 16, 1996
Revision received May 18, 1998
Accepted May 25, 1998 ■