Journal of Experimental Psychology:
Learning, Memory, and Cognition
2000, Vol. 26, No. 1, 3–27

0278-7393/00/$5.00 DOI: 10.1037//0278-7393.26.1.3

Thirty Categorization Results in Search of a Model

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

One category structure dominated in the shift toward exemplar-based theories of categorization. Given the theoretical burden on this category structure, the authors reanalyzed 30 of its uses over 20 years in 8 articles. The authors suggest 4 conclusions. (a) This category structure may encourage exemplar-memorization processes because of its poor structure, the learning difficulties it causes, and its small, memorizable exemplar sets. Its results may only generalize narrowly. (b) Exemplar models have an advantage in fitting these 30 data sets only because they reproduce a performance advantage for training items. Other models fit equally well if granted this capacity. (c) A simpler exemplar process than assumed by exemplar models suffices to explain these data sets. (d) An important qualitative result predicted by exemplar theory is not found overall and possibly should not even be expected. The authors conclude that the data produced by this category structure do not clearly support exemplar theory.

Categorizing objects into psychological equivalence classes is a basic cognitive task. Historically, some descriptions of categorization were prototype-based—humans were supposed to average their exemplar experience into a category prototype, compare new items to it, and accept the new items as category members if similar enough (Homa, 1984; Homa, Rhoads, & Chambliss, 1979; Homa, Sterling, & Trepel, 1981; Mervis & Rosch, 1981; Posner & Keele, 1968, 1970; Rosch, 1973, 1975; Rosch & Mervis, 1975).

However, 20 years ago some evidence began to suggest that prototype theory might not account completely for humans' categorization processes. In influential articles, Medin and his colleagues argued that prototypes are an insufficient organizing principle for categories, that prototype models sometimes offer poor descriptions of humans' categorization performance, and that humans learn categories that would be unlearnable if categorization depended on prototypes (Medin, Dewey, & Murphy, 1983; Medin & Schaffer, 1978; Medin & Schwanenflugel, 1981; Medin & Smith, 1981).

Thus, Medin and his coworkers proposed instead that exemplar-based processes underlie categorization and that exemplar-based models of categorization describe categorization performance better. These models assume that stored memories of the specific exemplars encountered in training form the representational core of a category. These models assume that these stored memories become the comparative-reference standard for categories, so that new tokens are placed into the category with the most similar stored exemplars. By making these assumptions, Medin's exemplar model—the context model—granted participants exemplar-based strategies that were powerful enough to master poorly structured categories and to learn exceptional items within categories. Nosofsky generalized the context model in a series of articles, and it has had a profound influence on categorization theory and research (Lamberts, 1994, 1995; Medin, 1975; Medin et al., 1983; Medin & Schaffer, 1978; Medin & Smith, 1981; Medin, Altom, & Murphy, 1984; Nosofsky, 1984, 1987; Palmeri & Nosofsky, 1995; Smith & Minda, 1998; Smith, Murray, & Minda, 1997).

One category structure dominated the science that motivated the shift toward exemplar-based descriptions of categorization. Table 1 shows this category structure. It contains five A exemplars and four B exemplars that are used in training. For this reason, we refer to it as the 5–4 category structure. Seven additional items are reserved for measuring transfer performance. The logical structure shown in Table 1 has been instantiated by using geometric designs, Brunswick faces, yearbook photographs, and line-drawn rocketships. This stimulus set was featured in the article that introduced the context model to human psychology (Medin & Schaffer, 1978) following Medin's (1975) important comparative monograph. It dominated the early papers and findings that favored exemplar-based categorization (Medin et al., 1984; Medin et al., 1983; Medin & Smith, 1981). It was crucial in Nosofsky's (1992) critique of prototype models and in his further explorations of the context model (Nosofsky, Kruschke, & McKinley, 1992; Nosofsky, Palmeri, & McKinley, 1994; Palmeri & Nosofsky, 1995).

Given the critical role this category structure has played in asserting exemplar theory in the categorization literature, it is important to interpret carefully the results it yields and to be certain that the assumptions of exemplar theory are warranted and necessary regarding it. In that way, one may establish whether the 5–4 category structure bears well the burden of supporting exemplar theory, or whether that burden should be placed on other category structures and category tasks that have seemed to show exemplar theory's worth.

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, Buffalo, New York 14260. Electronic mail may be sent to psysmith@acsu.buffalo.edu.

Table 1
The 5–4 Category Structure

Type and stimulus	Dimension (D)			
	D1	D2	D3	D4
Category A				
A1	1	1	1	0
A2	1	0	1	0
A3	1	0	1	1
A4	1	1	0	1
A5	0	1	1	1
Category B				
B6	1	1	0	0
B7	0	1	1	0
B8	0	0	0	1
B9	0	0	0	0
Transfer (T)				
T10	1	0	0	1
T11	1	0	0	0
T12	1	1	1	1
T13	0	0	1	0
T14	0	1	0	1
T15	0	0	1	1
T16	0	1	0	0

Accordingly, the present article revisits the large body of data on the 5–4 category structure, including 30 data sets reported in eight articles (Lamberts, 1995; Medin et al., 1984; Medin et al., 1983; Medin & Schaffer, 1978; Medin & Smith, 1981; Nosofsky et al., 1992; Nosofsky et al., 1994; Palmeri & Nosofsky, 1995). These 30 data sets represent essentially all the results using this category structure, making this a comprehensive survey and a deliberately nonselective one. Using this comprehensive set of results with the most influential category structure in the literature, we ask whether these data encourage the general assumption that humans categorize by using systematic exemplar-to-exemplar comparisons based in specific exemplar traces that form the representational cores of categories.

To do so, we explore the fits of many models that have been used in the literature, and consider the reasons that some fail to fit the 5–4 data whereas others succeed. We explore the convergences among the successful models to see if they have implications for interpreting performance in the 5–4 category task. We consider the present status of the primary result within the 5–4 data that has suggested that exemplar theory's assumptions are qualitatively correct—not just quantitatively better. Finally, we consider the place of the 5–4 category task and related tasks in the larger space of category structures and categorization phenomena.

Method

Category Structure, Stimuli, and Data Sets

Category structure. Table 1 shows the category structure used to collect the 30 5–4 data sets. The stimuli are derived from the category prototypes 1 1 1 1 and 0 0 0 0. Category A has four exemplars that share three features with the Category A prototype and one exemplar that shares two features. Thus, Category A has no exceptional exemplars (sharing more features in common with the opposing prototype) but has one ambiguous exemplar (sharing features equally with both prototypes). Category A exemplars share an average of 2.8 features out of 4 with their prototype; they share an average of 2.4 features with each other (including their perfect self-identities). The features in Category A are .70 predictive of category membership on average. This means that a rule based on any one feature would work only about 70% of the time.

The Category B exemplars share 2, 2, 3, and 4 features, respectively, with their prototype. Thus Category B also has no exceptional exemplars, but half its exemplars are ambiguous. Category B exemplars also share an average of 2.8 features out of 4 with their prototype and 2.4 features with each other (including their perfect self-identities). The features in Category B are .69 predictive of category membership on average.

Across categories, the four features are .77, .55, .77, and .66 predictive of category membership. No feature is perfectly diagnostic, but all four carry at least minimal information. The second dimension does carry minimal information, for it can only categorize correctly five of the nine training exemplars. An adaptive participant might learn to ignore this confusing stimulus dimension and allocate more attention to Dimension 4 (with .66 predictiveness) and especially to Dimensions 1 and 3 (with .77 predictiveness).

To derive an overall index of within-category coherence and between-category differentiation, one can divide within-category similarity by between-category similarity to find the structural ratio (Homa et al., 1979, pp. 13–14; Smith et al., 1997). Given that the exemplars share 2.4 features with each other within category (including their self-identities) and 1.6 features across categories, one can calculate that the structural ratio for this category structure is quite low—1.5. A structural ratio of 1.0 would imply no differentiation—that is, a complete overlap of the categories in multidimensional perceptual space. Structural ratios as high as 3.0 are easy to arrange.

This low index of category differentiation correctly reflects that the individual features are only 70% diagnostic, that exemplars are nearly as similar across categories (sharing 1.6 features) as within categories (sharing 1.9 features if one excludes self-identities), and that three of the nine items are ambiguous because they share features equally with both prototypes. Thus the two categories within themselves are poor assemblages with a weak family resemblance, and they are poorly differentiated from each other. There were constructive methodological reasons for creating categories like these. Still, the fact of poor differentiation may constrain the interpretation of the results they produce. We discuss this problem and the psychological impact of poor category structure below.

However, the 5–4 categories shown in Table 1 do have the characteristic of being linearly separable (LS). LS categories are those that can be partitioned by a linear discriminant function, and for which one can simply sum the evidence offered separately by each feature of an item and use that sum to decide correctly category membership. For LS categories, there is a way to allocate limited attention across the four dimensions that lets one categorize accurately all the training exemplars. For example, if one allocated 40%, 0%, 40%, and 20% of one's attention to Dimensions 1, 2, 3, and 4, respectively, the evidence favoring a Category A response would be, for training items A1 to B9 as shown in Table 1, 0.80, 0.80, 1.00, 0.60, 0.60, 0.40, 0.40, 0.20, and 0.00. That this evidence base is greater than 0.50 for the five Category A exemplars and less than 0.50 for the four Category B exemplars signifies that this attentional allocation would correctly categorize all the stimuli. The fact of linear separability bears on what follows because it means that a prototype-based strategy applied appropriately could categorize correctly all the training exemplars. That is, this category structure does not force participants to desert a prototype-based approach and adopt an exemplar-based approach instead. It leaves both strategies viable.

Stimuli. The 5–4 category structure has been instantiated in a variety of specific stimulus domains—geometric forms (Medin et al., 1984; Medin & Schaffer, 1978; Nosofsky et al., 1992); Brunswick faces (Lamberts, 1995; Medin & Schaffer, 1978; Medin & Smith, 1981); yearbook photos (Medin et al., 1983); and line-drawn rocketships (Nosofsky et al., 1994; Palmeri & Nosofsky, 1995).

Data sets. Appendix A describes various aspects of the 30 data sets (source, experimental or training condition, stimulus materials, and so forth). Appendix B provides a resource for modelers in this area by summarizing the 30 performance profiles. All 30 profiles are reported as Category A response probabilities in the stimulus order of Table 1 (A1 to T16). Authors in the original sources adopted a variety of reporting techniques and stimulus orders.

Formal Modeling Procedures

The context model. In evaluating the exemplar model, we focused 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, 1987). In the context model, the to-be-classified item in the 5–4 category structure is compared to the five A exemplars (including itself if it is an A) and to the four B exemplars (including itself if it is a B), yielding the overall similarity of the item to Category A exemplars and Category B exemplars. Dividing overall A similarity by the sum of overall A and B similarity essentially yields the probability of a Category A response.

We calculated the similarity between the to-be-categorized item and any exemplar in three steps as follows. First, we compared the values (1 or 0) of the item and the exemplar along all four dimensions. 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 their dimension carried. In the present model, each dimensional weight ranged from 0.0 to 1.0, and the four weights were constrained to sum to 1.0.

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

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

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.

Third, we calculated the similarity η_{ij} between the item and exemplar by taking $\eta_{ij} = e^{-d_{ij}}$, with d_{ij} the scaled psychological distance between the stimuli.

We repeated these three steps to calculate the psychological similarity between a to-be-categorized item and each A and B exemplar. Then, summing across the Category A and Category B exemplars, we calculated the total similarity of the item to Category A and to Category B members. These quantities can be used to produce directly the probability of a Category A response (R_A) for stimulus i (S_i) by taking

$$P(R_A|S_i) = \frac{\sum_{j \in C_A} \eta_{ij}}{\sum_{j \in C_A} \eta_{ij} + \sum_{j \in C_B} \eta_{ij}}. \tag{2}$$

This equation means that one sums up the similarity the to-be-categorized item has to each member of the A category and divides this by the similarity the to-be-categorized item has to all the members of both the A and B categories. Repeating this process for each of the 16 items, one would derive the performance profile predicted by the model. However, for reasons we describe later, we added an additional guessing-rate parameter to the context model as follows. We assumed that some proportion of the time participants simply guessed A or B haphazardly and that otherwise participants used exemplar-based categorization in the way just described. Others have proceeded similarly in granting the context model a guessing parameter (Lamberts, 1994, 1995; Smith & Minda, 1998). With that parameter added, the context model had six parameters (five free parameters)—four dimensional weights constrained to sum to 1.0, a sensitivity parameter, and a guessing parameter.

To find the best-fitting parameter settings of the context model, we seeded the space with a single parameter configuration and calculated predicted categorization probabilities for the 16 stimuli according to that configuration. The measure of fit was the sum of the squared deviations between the 16 predicted probabilities and the 16 observed categorization probabilities of some study's performance profile. We minimized this measure 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). To ensure that local minima were not a serious problem in the present parameter spaces, we repeated this procedure 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 we found were close to global ones.

The additive prototype model. We followed the influential research of Medin and his colleagues (Medin & Schaffer, 1978; Medin & Smith, 1981) by evaluating the simple additive prototype model that has been so prominent. That is, we supposed that each to-be-categorized item would be compared to the category prototype along the four stimulus dimensions and that matching features would simply add to prototype similarity in the amount of their dimension's weight. In the simplest case, the item's prototype similarity could be taken to be the probability of a correct categorization and its complement, the probability of an error. However, we added an additional guessing parameter to the prototype model as we did for the context model and for every model considered in this article. Thus we assumed that some portion of the time participants simply guessed A or B haphazardly and used additive prototype-based similarity otherwise (see also Medin & Smith, 1981). The guessing parameter is especially important for modeling participants' sometimes poor performance given the poorly differentiated 5–4 category structure. Without it, for example, the category prototypes (Stimuli B9 and T12), which of course have perfect prototype similarity, would be predicted to be categorized perfectly. With the guessing parameter added, the additive prototype model had five parameters (four free parameters)—one guessing parameter and four dimensional weights constrained to sum to 1.0. For the additive prototype model, and for all the models described below, we hill-climbed, found best-fitting configurations, and avoided local minima by using the procedures already described for the context model.

The multiplicative prototype model. The additive prototype model is at an inherent fitting disadvantage because its additive similarity calculations are so simple. It lacks any capacity to allow psychological similarity to decrease exponentially (not linearly) with increasing distance between stimuli. It lacks any sensitivity parameter that might be appropriate for capturing knowledge gains and performance improvements during learning. This is particularly important when considering the 30 5–4 data sets because researchers have often only considered mature, task-final performance. Consequently, the prototype model's failures might not always be due to participants' reliance on exemplar-based processes instead of prototype-based processes. They might sometimes be due to the less delicate and sensitive nature of the prototype model's calculations.

Accordingly, Nosofsky (1987, 1992) described a prototype model that incorporated exponentially decreasing similarity functions and a sensitivity parameter. In this prototype model, increased sensitivity exaggerates the closeness of category members to their own prototype relative to the opposing prototype, strengthens the evidence base supporting a correct categorization, and increases the estimated percentage correct.

In the multiplicative prototype model, the to-be-classified item in the present tasks would be compared only to the A and B prototypes to yield the overall similarity of the item to Category A and Category B. We calculated the similarity between the to-be-categorized item and a prototype in three steps as described for the context model. First, we determined the psychological distance between the item and prototype by summing the weights of the mismatching features between the two. We then scaled this raw psychological distance between item and prototype by using a sensitivity parameter that could vary from 0.0 to 20.0. Formally, then, the scaled psychological distance between the to-be-classified item i and the prototype was given by

$$d_{iP} = c\left[\sum_{k=1}^{N} w_k|x_{ik} - P_k|\right]. \tag{3}$$

Here, x_{ik} and P_k are the values of the to-be-classified item and the prototype on dimension k, w_k is the attentional weight granted dimension k, and c is the freely estimated sensitivity parameter. Third, we calculated the similarity η_{iP} between the item and prototype by taking $\eta_{iP} = e^{-d_{iP}}$, with d_{iP} representing the scaled psychological distance between the stimuli.

Dividing Category A similarity by the sum of Category A and Category B similarity, one could essentially derive the multiplicative prototype model's predicted probability of a Category A response for stimulus i by taking

$$P(R_A|S_i) = \frac{\eta_{iP_A}}{\eta_{iP_A} + \eta_{iP_B}}, \tag{4}$$

except that once again we added a guessing parameter to the model. Thus the multiplicative prototype model had six parameters—a guessing parameter, a sensitivity parameter, and four dimensional weights constrained to sum to 1.0.

The gamma model. Research has recently 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; Smith & Minda, 1998). Therefore, researchers have occasionally modified the standard context model profoundly by adding on the gamma parameter (Maddox & Ashby, 1993; McKinley & Nosofsky, 1995). Gamma intervenes by allowing the quantities 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

$$P(R_A|S_i) = \frac{\sum_{j\epsilon C_A} \eta_{ij}}{\sum_{j\epsilon C_A} \eta_{ij} + \sum_{j\epsilon C_B} \eta_{ij}}, \tag{5}$$

the augmented version is

$$P(R_A|S_i) = \frac{\left[\sum_{j\epsilon C_A} \eta_{ij}\right]^{\gamma}}{\left[\sum_{j\epsilon C_A} \eta_{ij}\right]^{\gamma} + \left[\sum_{j\epsilon C_B} \eta_{ij}\right]^{\gamma}}. \tag{6}$$

In this article we address the gamma model only briefly. We show that it represents no fitting advantage over the standard context model in fitting the 30 5–4 performance profiles that represent the aggregate performance profiles for whole groups. Probably the reason the gamma model offers no fit advantage is that aggregating performance systematically washes out the categorization idiosyncrasies of the individual participants and creates homogeneous performance profiles that the standard context model (without gamma) fits comfortably (Maddox & Ashby, 1993; Smith & Minda, 1998; Smith et al., 1997). The gamma parameter appears to be necessary if one wishes to model individual performance profiles instead, and this is why gamma was invented. In fact, one possible reason the context model seemed so apt early on is that it was often used to model the aggregate performances that suited it best (Smith & Minda, 1998). Smith and Minda discussed other theoretical concerns about gamma.

With the guessing parameter added, the gamma model had seven parameters—a guessing parameter, a gamma parameter, a sensitivity parameter, and four dimensional weights constrained to sum to 1.0.

The multiplicative prototype model (twin sensitivities). One instructive approach to allowing a prototype model to cope with the 5–4 data sets is to grant the possibility that participants process old, training items more easily or more fluently than they do new, transfer items. Intuitively, this might mean that practice lets the prototype-comparison processes run more smoothly, or the prototype-based algorithm be applied more skillfully, for old items than for new ones. It might mean that practice creates stronger connections between old items and their prototypes, whereas new items are less strongly connected to their prototype. This global old–new processing difference can be incorporated into a prototype model by assuming that participants respond to old items with a higher level of sensitivity and to new items with a lower level of sensitivity. Formally, one accomplishes this by simply granting the model two parameters, c_o and c_n, that apply to old items and new items in the distance equation of the prototype model as follows:

$$\text{Old items:} \quad d_{iP} = c_o\left[\sum_{k=1}^{N} w_k|x_{ik} - P_k|\right]. \tag{7}$$

$$\text{New items:} \quad d_{iP} = c_n\left[\sum_{k=1}^{N} w_k|x_{ik} - P_k|\right]. \tag{8}$$

In all other respects, this model is identical to the multiplicative prototype model (i.e., it has the identical distance-to-similarity transformation and choice rule).

With the guessing parameter added, this model had seven parameters—a guessing parameter, two sensitivity parameters and four dimensional weights constrained to sum to 1.0.

Prototypes combined with exemplar memorization: A mixture model. Another instructive approach to allowing a prototype model to cope with the 5–4 data is to grant the possibility that participants partially memorize the old, training exemplars after all the practice with them. This possibility can be explored by using a model that mixes prototype-based processing and exemplar memorization. The mixture model evaluated here received some early attention (Medin et al., 1983; Medin & Smith, 1981). It assumes that participants base their classification decisions either on the simple, additive similarity of a given stimulus to the prototype (in which case they will strictly obey typicality gradients), on random guesses (in which case they will place stimuli into Categories A or B haphazardly), or on the recognition of memorized specific exemplars (in which case they will definitely categorize the item correctly). The key aspect of fitting data using the mixture model is to estimate the balance among these three processes that best accounts for any performance profile.

Note that the mixture model's exemplar process is simple memorization—that is, individual training exemplars are stored, self-retrieved, and self-boosted toward correct categorization. In a sense, each old item is compared only to itself in the exemplar process that aids a correct categorization decision. This process is quite different from the context model's exemplar process. The context model assumes that a to-be-categorized item is compared to all the training items of both categories on the way to a categorization decision. In fact, the context model's systematic exemplar-to-exemplar comparison processes have seemed implausible to some (for a discussion, see Palmeri & Nosofsky, 1995, p. 548), making it useful to see whether other kinds of exemplar processes suffice, too.

The mixture model evaluated in the present article contained five free parameters—a guessing parameter, an exemplar-memorization parameter, and a prototype-processing parameter constrained to sum to 1.0, and four dimensional weights also constrained to sum to 1.0.

Results

Model 1: The Trouble With Prototypes

Figure 1A shows the crucial phenomenon—the failure of a simple prototype model to capture performance on the 5–4 category structure. To create this figure, we first averaged the 30 observed performances on each of 16 stimuli into the composite observed performance profile that is shown in Figure 1A and in all succeeding similar figures as the solid line. The performance profiles are given here and throughout in the stimulus order shown in Table 1 and in Appendix B—that is, Stimuli 1 to 5, 6 to 9, and 10 to 16 refer, respectively, to the Category A training exemplars, the Category B training exemplars, and the new, transfer items not seen in training. Categorization performances are given throughout as proportions of Category A responses. Category B's training exemplars appropriately elicit low rates of Category A responses because they are mostly called Bs.

Next, we modeled each individual data set using the additive prototype model that figured so heavily in the early demonstrations of the prototype model's failure and the exemplar model's superiority (Medin et al., 1983; Medin & Schaffer, 1978; Medin & Smith, 1981). For every data set, we found the best-fitting configuration of the prototype model, and this best-fitting configuration implied a predicted profile of 16 response proportions. Averaging the 30 predicted profiles produced the composite predicted profile shown by the dotted line in Figure 1A. (Succeeding similar figures contain composite predicted profiles that were produced and are displayed in the same way.) In this way, if the individual best-fitting profiles fit their individual observed profiles well, the predicted and observed composites will also fit well. But, as the individual best-fitting profiles fit less well, the predicted and observed composites may begin to diverge in a consistent way that can be interpreted meaningfully.

This divergence is clearly shown in Figure 1A. The additive prototype model fails to recover the data well. One useful measure of fit is the average absolute deviation (AAD), which summarizes how much on average the predicted response proportions diverge from the observed response proportions for each of the 16 stimuli in the task. On average, in fitting the 30 data sets, the prototype model erred by .091 per stimulus. A second useful measure of fit is the sum of the squared deviations (SSD) of predicted response proportions from observed response proportions. On average, in fitting the 30 data sets, the SSD over 16 stimuli was .224 for the additive prototype model. A third useful measure of fit is the percentage of variation in the observed profile that was accounted for by variation in the predicted profile (PVA). On average, in fitting the 30 data sets, the additive prototype model accounted for 83.8% of the variance in each study's observed performance profile. Table 2 provides these fit measures for the models considered in the present article. These fit measures for the pure prototype model are only moderately good, and the reader should note this carefully. In the present article, we focus on the relation of the 30 data sets to exemplar theory. As we do so, one must not forget that pure prototype models deserve definite criticism, for they behave poorly according to these criteria of fit.

Beyond the quantitative criteria of fit, the character of the prototype model's failure is also clear. The observed and predicted profiles diverge in a theoretically suggestive way. Time after time (i.e., data set after data set), the prototype model predicts that participants will do less well than they do on the Category A and B training items (Stimuli 1–5 and 6–9, respectively). Remember that for the Category B items the low observed Category A response rates as graphed imply high rates of correct B responses—higher than the prototype model can predict. In contrast, the prototype model predicts participants' performance on the transfer items (Stimuli 10–16) well. To confirm this, we calculated fit measures separately for the old and new items and found that the prototype model erred two or three times as much in predicting old-item performance as it did in predicting new-item performance, because it cannot handle the levels of old-item performance that participants actually show. Something about those old, familiar traces leads to their good performance and that something cannot be simple

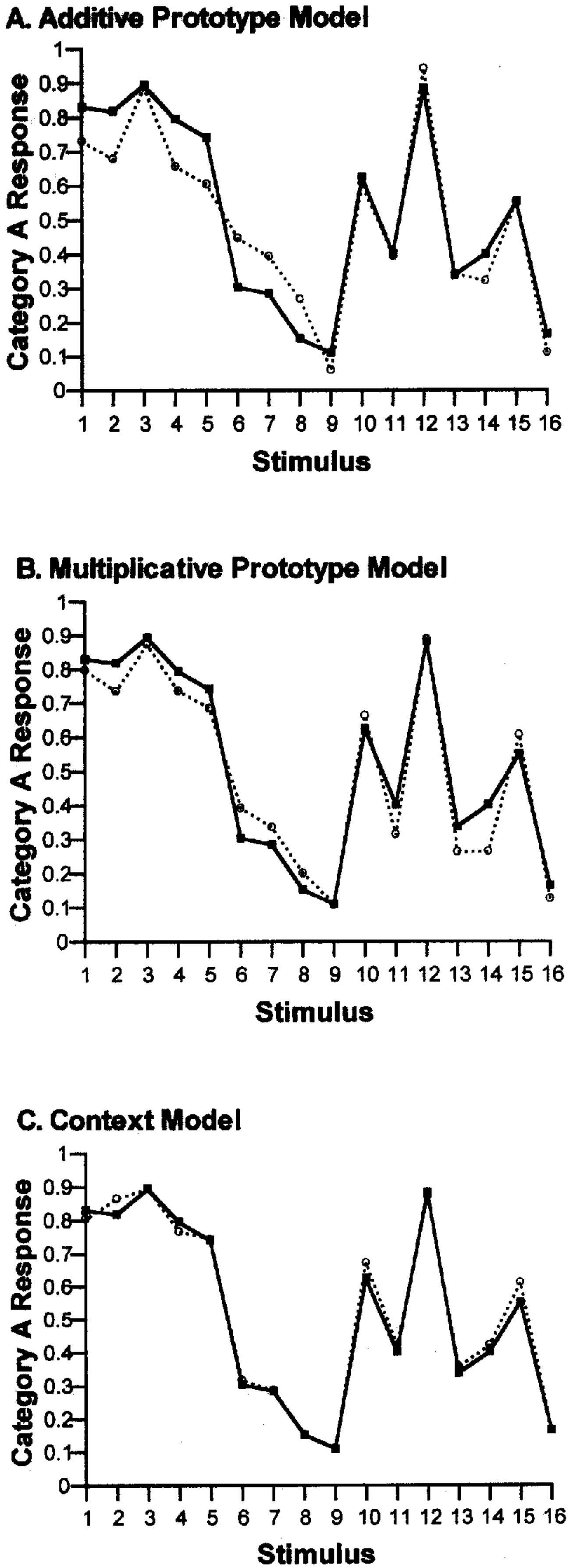


Figure 1 (left). A: The composite observed performance profile produced by averaging the 30 data sets (solid line). Stimuli 1–5, 6–9, and 10–16 denote the training exemplars of Category A, the training exemplars of Category B, and the transfer items, respectively. Also shown is the average of the best-fitting predicted performance profiles found when the 30 data sets were fit individually using the additive prototype model (dotted line). B: The same observed profile shown with the composite predicted profile of the multiplicative prototype model. C: The same observed profile shown with the composite predicted profile of the context model.

Table 2
Measures of Fit

Model	AAD	SSD	PVA
Additive prototype	0.091	0.224	0.838
Multiplicative prototype	0.069	0.137	0.890
Context	0.047	0.062	0.941
Additive exemplar	0.144	0.490	0.664
Fixed low sensitivity	0.149	0.529	0.637
Gamma	0.045	0.055	0.944
Twin sensitivity	0.043	0.054	0.946
Mixture	0.046	0.052	0.944

Note. AAD = average absolute deviation; SSD = sum of squared deviations; PVA = percentage of variance accounted for.

resemblance to the prototype. Remember that we are illustrating this effect with only one category structure. However, we believe that this effect would also occur with many other category structures and transfer sets that have been used in the literature.

The prototype model's problem springs from its cognitive psychology (i.e., the processes and representations it assumes). By assuming that all items (old and new) are referred to the category prototypes, it assumes that all items will equivalently obey the typicality gradients in the task. It has no way to treat training exemplars specially by according them any processing fluency or performance advantage. Clearly, the model with the right cognitive psychology (i.e., with the right assumed processes and representations to fit human psychology and human performance) will have a way to reproduce the observed old-item advantage.

Model 2: A Sophisticated Prototype Model Cannot Help

The sophisticated multiplicative prototype model has the same problematic cognitive psychology. Table 2 shows that including multiplicative similarity computations and a sensitivity parameter did help this model fit seemingly better than the additive prototype model. But this improvement is cosmetic, as one sees on examining the character of the multiplicative prototype model's fit (Figure 1B). Remember that the additive prototype model fit old-item performance poorly, but new-item performance well. Here the availability of the sensitivity parameter lets the multiplicative prototype model reach somewhat higher and lower to predict good performance on Category A and Category B training exemplars, respectively. But it also starts to miss the transfer items more by overreaching them (i.e., by predicting they will be performed more extremely in both directions than they are). To confirm this, we calculated fit measures separately for the old and new items and found that the multiplicative prototype model erred about the same amount in predicting performance on both. The multiplicative prototype model spreads its error evenly over the training and transfer items. Still, the fundamental problem of a prototype psychology remains. Whatever sensitivity this model assumes, it treats all stimuli with that sensitivity. It can no more treat old and new items differentially than can the additive prototype model. It can split the difference between the old and new items, but it cannot predict that difference. The appropriate model must be able to.

Model 3: The Solution Provided by the Context Model

The solution Medin and his coworkers found to this problem was to assume that the primary representations underlying Categories A and B were the old, training exemplars themselves, rather than prototypes that had been abstracted from them (Medin & Schaffer, 1978). By making this theoretical choice, Medin introduced the important possibilities that categorization can be based on exemplar storage, not prototype abstraction, and on item-exemplar comparisons, not item–prototype comparisons. The categorization literature has not been the same since.

Figure 1C shows why. The context model predicts well every general aspect of performance—performance on both categories' training exemplars, performance on transfer items, the old–new performance gap, and so forth. Table 2 shows that in all quantitative respects the context model did a better job fitting the 30 data sets than did either prototype model. Indeed, for all three fit indices, the context model fit significantly better than did the additive prototype model: $t(29) = 6.48$, $p < .05$, for AAD; $t(29) = 4.98$, $p < .05$, for SSD; and $t(29) = -5.45$, $p < .05$, for PVA.[1]

Nosofsky (1992) carried out a survey similar to the one in the present article. He compared the success of various models in fitting some of the 5–4 data sets that were available then. He also included the additive prototype model, the multiplicative prototype model, and the context model. He also found both prototype models wanting by criteria of fit. To this point, his smaller survey of the literature converges with ours. In particular, the reader should note the context model's excellent indices of fit. For as we focus on the relation of these 30 data sets to exemplar theory, one must not forget how well the exemplar model performs in fitting them.

One reason the context model fits these data patterns so well is that it assumes that the old training exemplars are stored in memory as the representational cores of the two categories. On their reappearance as stimuli in the transfer phase of the experiment, they naturally trigger themselves in memory and receive by virtue of this identity match a strong pull or resonance from self-retrieval that causes them to be categorized highly accurately. The transfer items receive only the weaker pull exerted by training items of both categories that they may be similar to but never identical to. As a consequence, their performance disadvantage is predicted.

In fact, one can show intuitively that the context model, by positing exemplar storage and exemplar-to-exemplar comparisons, will be able to produce the required old–new performance advantage. The probability of a Category A response for a transfer item is closely related to the summed similarity of the to-be-categorized item to the A exemplars divided by the summed similarity of the to-be-categorized item to the exemplars of both categories. Thus, the probability of a Category A response for transfer Stimulus i is closely related to the following quantity:

$$(\mathrm{Sim}_{iA1} + \mathrm{Sim}_{iA2} + \mathrm{Sim}_{iA3} + \mathrm{Sim}_{iA4} + \mathrm{Sim}_{iA5}) / ([\mathrm{Sim}_{iA1} + \mathrm{Sim}_{iA2} + \mathrm{Sim}_{iA3} + \mathrm{Sim}_{iA4} + \mathrm{Sim}_{iA5}] + [\mathrm{Sim}_{iB6} + \mathrm{Sim}_{iB7} + \mathrm{Sim}_{iB8} + \mathrm{Sim}_{iB9}]). \tag{9}$$

In contrast, when calculating the Category A response probability for a Category A training item, one of these imperfect similarities will always be replaced by a perfect-match identity. This will increase the numerator of the decision rule proportionally more than the denominator and will predict higher performance for old items than for new ones. How much higher is a decision and a calculation that the context model makes finely as it fits an observed performance profile.

The central point, though, is the success of the context model in predicting the old–new performance gap that prototype models in principle cannot predict. This success in fitting an early group of data sets brought exemplar theory into sharp focus in the literature and brought prototype theory into disfavor. Figure 1C shows that this success is generally repeated across the 30 data sets. This success has been repeated using other category structures, too.

Yet the context model makes several important assumptions to reproduce the performance differential between training and transfer items. It assumes, as other exemplar models do, that specific exemplar traces (not prototypes) form the representational cores of categories. It assumes, as other exemplar models do, that token-exemplar comparisons, not token-prototype comparisons, are the basis of category decisions. It even assumes that humans compare a to-be-categorized item to all the stored members of relevant categories in reaching a categorization decision. This assump-

[1] To ensure that our fitting procedures were stable across different minimization criteria, we also fit all eight models to all 30 data sets using log-likelihood as the criterion for goodness of fit, not the sum of the squared deviations (SSD). The resulting best fitting solutions were practically identical in the two cases. For example, for the context model fitting the 30 data sets, the 480 predicted performance levels (16 stimuli × 30 data sets) found by the SSD minimization criterion and by the log-likelihood minimization criterion correlated at .999. For the additive prototype model, this correlation was .998. The best fitting parameter estimates found by the two minimization criteria correlated extremely highly, too. Given this close correspondence, we emphasize the SSD minimization criterion in this article because it is more intuitive, because it was generally used by others in modeling the 5–4 performance profiles, and because it is even discernible in the graphs comparing observed and predicted performance.

tion, that humans categorize a new dog by comparing it to all the specific dogs they know, has not seemed natural to everyone (see Palmeri & Nosofsky, 1995, for discussion).

The context model also assumes that the subjective psychological space within which a category task is represented and performed can be expanded through a wide range of magnification factors. This magnification is the role of the sensitivity parameter *c*. Higher sensitivities magnify psychological space by increasing the differentiation among stimuli. The context model also assumes that psychological similarity decays exponentially as featural mismatches accumulate between stimuli. This means that even the first featural mismatch can create a large dissimilarity between stimuli. This means that the context model can emphasize strongly an exact match between stimuli, especially when its sensitivity parameter is set at a high level.

As we show now, this whole package of assumptions—exemplar representations, systematic exemplar-to-exemplar comparison processes, exponentially decaying similarity, and the magnification factor expanding psychological space—operates synergistically in allowing the context model to fit successfully the data from the 30 5–4 performance profiles. All these assumptions are needed.

Model 4: Exemplar Processes Alone Do Not Work—An Additive Exemplar Model

Assuming exemplar representations and systematic exemplar-to-exemplar comparisons is not enough to account for the 5–4 performance profiles. To show this, we fit a simple, additive exemplar model to the 30 data sets. This model is the twin of the additive prototype model, with the same lack of exponential similarity decay and the same lack of a sensitivity parameter. The models only differ in their assumptions about the representational cores of categories and about the comparison processes that underlie categorization decisions. Figure 2A shows that this exemplar model fits these data poorly, and Table 2 summarizes its poor fit indices. It falls far short of what the full-fledged context model accomplishes. It even falls short of what the additive prototype model accomplishes (Figure 1A). It is generally acknowledged that exemplar processes are not the sole source of the answer the full context model provides (Medin & Schaffer, 1978; Nosofsky, 1992).

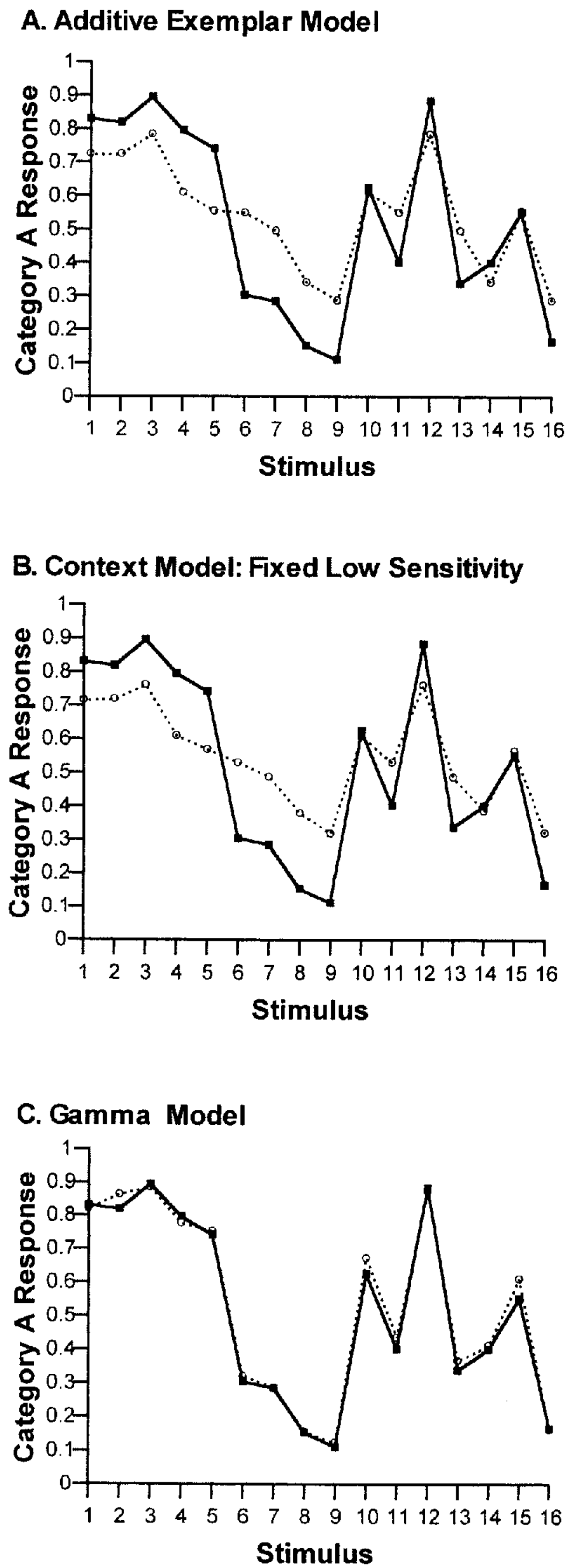


Figure 2 (right). A: The composite observed performance profile produced by averaging the 30 data sets (solid line). Also shown is the average of the best-fitting predicted performance profiles found when the 30 data sets were fit individually using an additive exemplar model (dotted line). B: The same observed profile shown with the composite predicted profile of the context model in low-sensitivity configurations. To make this predicted profile, we fit each of the 30 data sets with 21 versions of the context model in which sensitivity was held at low levels (from 1.0 to 3.0 inclusively, in steps of .10). The predicted profile shown is the average of these 630 individual predicted profiles. C: The same observed profile shown with the composite predicted profile of the gamma model.

Model 5: The Context Model Denied Magnification Power—Fixed Low Sensitivity

Even combining exemplar representations with exponentially decaying similarity is not enough to account for the 5–4 performance profiles. The context model must have extremely high sensitivities—that is, it must be allowed to magnify psychological space enormously. Remember that higher sensitivities magnify psychological space by increasing the psychological differentiation among stimuli in an uneven way that places more emphasis on exact identity between a to-be-categorized item and an old exemplar. To show the context model's dependence on high sensitivities in fitting the 30 data sets, we examined its performance when sensitivity was kept low. To do so, noting that the sensitivity parameter's full range goes from zero to infinity, we fit each of the 30 data sets with 21 versions of the context model that had sensitivity fixed at all values from 1.0 to 3.0, in steps of .10.

Figure 2B shows the predictions of the context model averaged over 21 levels of sensitivity as it tried to fit the 30 data sets. It fits these data very poorly, and Table 2 confirms this fact. It is obvious that the context model needs the flexibility to choose much higher powers of magnification for psychological space to account for these performance profiles and probably others, too. In fact, over the 30 data sets, the unconstrained context model estimated sensitivity to be 8.19 (SD = 2.79). As we consider now the meaning of high sensitivity and its role in fitting 5–4 data, remember that this high sensitivity value was estimated only for the 30 5–4 data sets, that other category structures have produced lower sensitivity estimates in the literature (Nosofsky, 1986, 1987, 1988; McKinley & Nosofsky, 1996), and that in these cases the implications of high sensitivity might apply less strongly.

To explain the meaning of high sensitivity in the context model, Figure 3 shows the relationship between shared features and psychological similarity for two stimuli when the exemplar-comparison system has 8.2 sensitivity. Whereas identical stimuli share 100% similarity, the expansion of psychological space caused by high sensitivity leaves even two items that share all but one feature with only 13% similarity instead. In this psychological space, only exact identity produces strong similarity. The meaning of high sensitivity can also be illustrated by using the 30 5–4 data sets. To do so, we fixed the dimensional weights in the context model at the values obtained on average when the context model fit the 30 data sets. We fixed sensitivity at 8.2. Under the context model's description, same-category exemplars in the 5–4 task are only 5% similar to each other.

This makes one wonder how, if the context model's description is right, participants glue objects that are 95% dissimilar into categories, or whether they even do. Given nine such disparate stimuli, all distant neighbors in psychological space, participants might just pursue an exemplar-memorization strategy that attaches the correct A or B label to each unique exemplar. Old items would then be categorized just by being recognized, not by being systematically compared to all the training items. Notice that categorization by exemplar recognition would involve 100% similarity, not 5% similarity. By the context model's own description of processing in the 5–4 task, exemplar-memorization events have 20 times the impact on the system that exemplar-comparison events have.

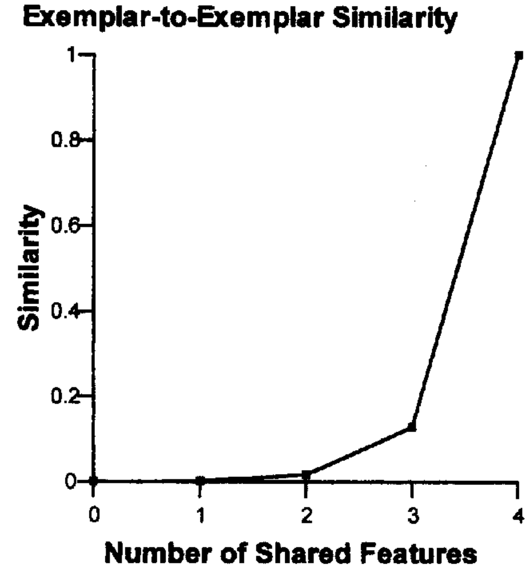


Figure 3. The relationship between similarity and the number of shared features between exemplars when we set the context model's sensitivity parameter at 8.2, assumed that attention was distributed homogeneously, and calculated similarity between two four-dimensional exemplars.

In fact, in the original experiments that motivated exemplar theory, Medin and his coworkers considered the possibility that participants might adopt just such an exemplar-memorization approach (Medin & Schwanenflugel, 1981, p. 365). If so, the categorization task would degenerate into an identification task in which participants would rotely associate whole instances and their labels but would have no sense of organized categories as they applied the labels. Below, we consider further the important effects and implications of high sensitivity in the context model.

In any case, Model 5 joins Model 4 in showing that exemplar processing alone is insufficient to give a model the flexibility it needs to fit the 5–4 data sets well. Instead, the context model needs to be granted all of its assumptions—exemplar storage, systematic exemplar-to-exemplar comparisons, exponential similarity decay, and the extreme magnification of psychological space. These assumptions change traditional thought about the representations and processes underlying humans' categorization, and this is why it is important to keep evaluating those assumptions.

In fact, it is useful to keep in mind that the comparison between the context model and the prototype model was not historically only about one contrast (exemplar-based category representations for the context model; prototype-based representations for the prototype model). Rather, the comparison was about multiple contrasts (exemplar storage, systematic exemplar-to-exemplar comparisons, the magnification of psychological space, exponentially decaying similarity for the context model, prototype storage, item-prototype comparisons, additive similarity, and linearly decaying similarity for the prototype model). Although the term exemplar theory is ubiquitous now, it selectively emphasizes some of the context model's assumptions and not others, and therefore it potentially misleads. In particular, we have just seen that the extreme magnification required by the context model to fit the 30 5–4 data sets could speak against the exemplar-comparison processes of the context model and for a simpler exemplar-memorization process. This recommends the consideration of alternative exemplar processes in category research, and we return to this issue below.

Model 6: The Gamma Model—Too Much Firepower for the Job

The standard context model sometimes fits group performance when it fails to fit individual performances (Ashby & Gott, 1988; Maddox & Ashby, 1993; Smith et al., 1997). This failure has implications for theory in categorization research (Smith & Minda, 1998), and it has led the context model to be augmented by the gamma parameter (Maddox & Ashby, 1993; McKinley & Nosofsky, 1995). Here, because the 30 data sets do represent composite group performances, it seemed likely that gamma would not be required. In fact, the gamma model fit the 30 data sets just like the standard context model does—quantitatively (Table 2) and qualitatively (compare Figures 1C and 2C). The average best-fitting gamma over the 30 data sets was only 1.75, underscoring the minimal requirement for it.

Model 7: Prototypes, But Extra Old-Item Sensitivity—A Twin-Sensitivity Model

Even without gamma, the context model makes four assumptions to capture the 5–4 performance profiles. We now ask whether these assumptions are necessary to describe the 5–4 performances or whether there is a viable set of assumptions that preserves prototypes as the representational cores of categories. Models 7 and 8 address these questions.

One approach is to extend slightly the multiplicative prototype model (Model 2), by assuming that participants apply prototype-based algorithms to the old items with special ease and fluency because practice and repetition has left them skilled at doing so. Accordingly, we fit to all 30 data sets a multiplicative prototype model that featured one level of sensitivity for the old, training items (acknowledging that these might be processed more fluently) and one level of sensitivity for the new, transfer items (acknowledging that these might be processed less fluently).

This twin-sensitivity model fits the 30 performance profiles well (Figure 4A) and slightly better than the context model (Table 2) does. Perhaps exemplar-to-exemplar comparisons are not the key to explaining the 5–4 performances. Perhaps a variety of models can explain the 5–4 performances successfully if only they are granted the capacity to show a special fluency regarding old items.

Model 8: Combining Prototypes and Exemplar Memorization—A Mixture Model

Another approach is to assume that the 5–4 performance profiles reflect prototype-based processing augmented by the partial memorization of the training items and their correct category labels after all the practice. Accordingly, we examined a simple intermixture of prototypes and exemplar memorization that was adopted in some early studies (Medin et al., 1983; Medin & Smith, 1981). The mixture model's exemplar process gives a simple performance boost to all the old items as if they had been partially memorized. Smith and Minda (1998) showed that this model offers insights even in cases where pure prototype models and pure exemplar models fail.

Figure 4B shows the composite observed and predicted performance profiles when this mixture model was fit

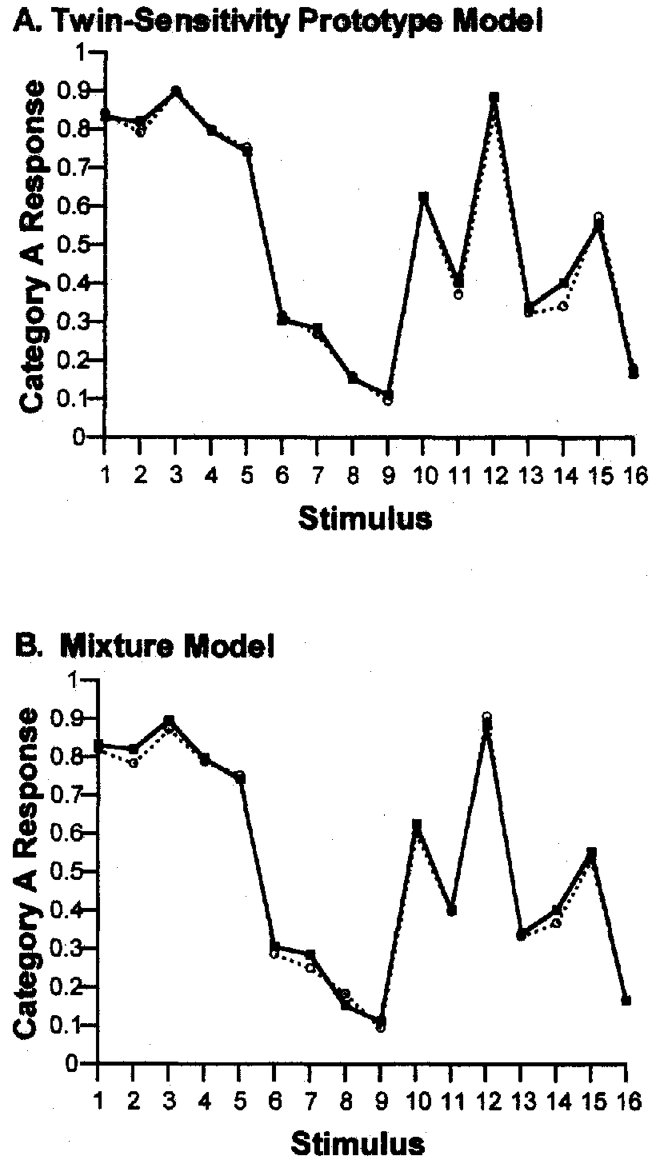


Figure 4. A: The composite observed performance profile (solid line) shown with the composite predicted performance profile (dotted line) of a multiplicative prototype model that allowed differential sensitivity to old and new items. B: The composite observed profile shown with the composite predicted profile of a model that combined prototype-based processing with exemplar memorization.

individually to all 30 data sets. It also fits the data well and slightly better than the context model does (Table 2).

Which Data Sets Should Be Modeled Comparatively?

In the preparation of this article, a reviewer raised a concern about which, if not all, of the 5–4 data sets it was appropriate for us to include in our analyses, and about which, if not all, of the 5–4 data sets the exemplar model might be expected to fit successfully. This concern has implications for interpreting the 5–4 data and for theory in the field.

The reviewer's concern was that the exemplar model cannot be fairly expected to fit comfortably the data produced when participants are operating under a deadline (as in Lamberts, 1995), or the data produced at earlier stages of learning (as in Nosofsky et al., 1992), or the data produced under rule-plus-exception instructions (as in the conditions of Medin & Smith, 1981, and Palmeri & Nosofsky, 1995, that encouraged participants to use one feature as a rule to categorize the stimuli and to memorize the exceptions to that rule), or the data produced under prototype-based instructions (as in the conditions of Medin & Smith, 1981, that encouraged participants to develop a general idea about the categories). If not, then it might not be appropriate for us to compare the fit of the exemplar model to the fit of other models regarding these data sets.

Before addressing this concern, we point out that theoretical implications would attend the need to exclude data sets to isolate the situations that the exemplar model describes well. This need would suggest that the exemplar model may not handle well different deadline conditions (Lamberts, 1995), different stages of learning (Nosofsky et al., 1992), or different instructional sets (Medin & Smith, 1981; Medin et al., 1983; Medin et al., 1984; Palmeri & Nosofsky, 1995). This would narrow exemplar theory's influence within the categorization literature, leaving it applicable only to some of the possible experimental paradigms. This would narrow exemplar theory's influence in the study of real-world categorization, which naturally features different deadline conditions, stages of learning, and instructional sets. In both respects, the need to restrict data would limit the comprehensiveness of exemplar theory, require additional theories to handle additional data, and recommend a broader theory that could span procedural and temporal variations.

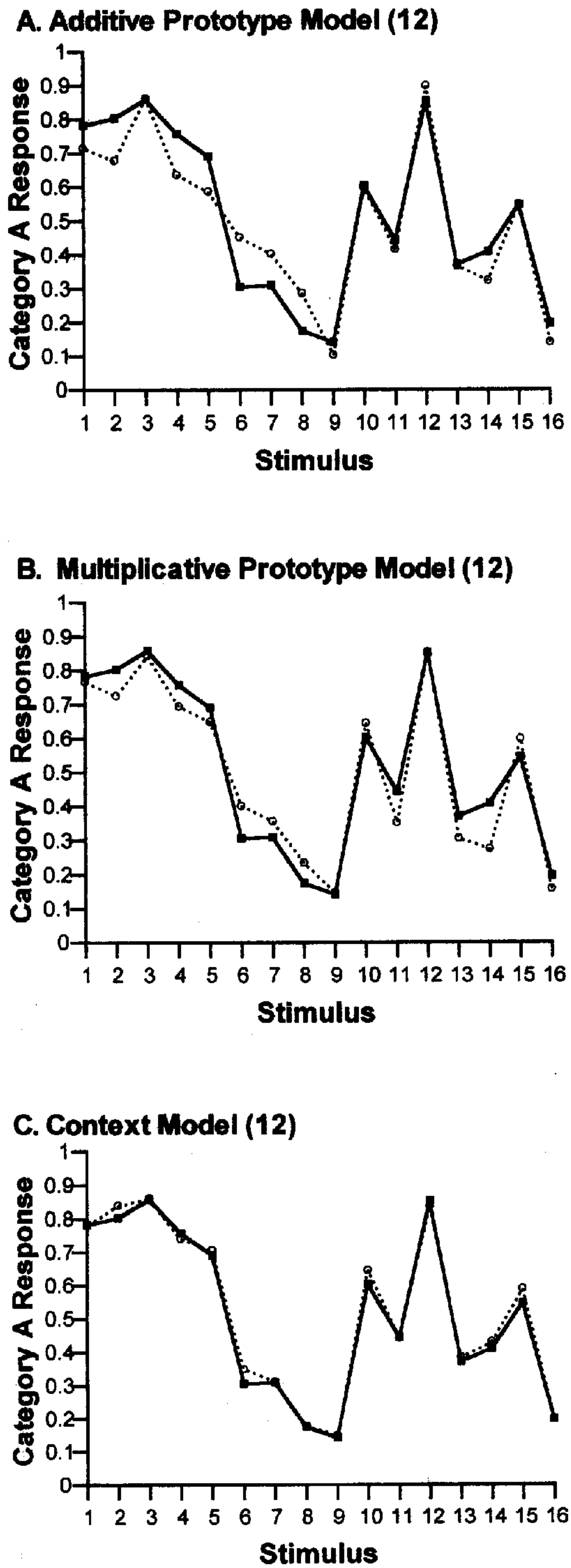
So it seemed important to address the concern about which data sets count. To do so, we temporarily excluded from consideration the three deadline conditions of Lamberts (1995), the early stages of learning from Nosofsky et al. (1992), the seven conditions that used prototype-based instructions, and the two conditions that used rule-plus-exception conditions. We also excluded the data set that sampled exemplars from an infinite pool without replacement, because this left the old items presented at transfer perceptually different from the old items presented during training. At the same time, we addressed the additional concern that we had somehow weighted some articles in the literature artificially heavily, for example, by including multiple data sets from them. In total, 18 of the 30 data sets—3, 4, 5, 6, 11, 12, 14, 15, 17, 18, 19, 20, 21, 22, 25, 27, 28, and 29—were set aside, leaving only the 12 that most heavily favored exemplar processing because they featured highly repetitive training on the same nine exemplars and because they assessed performance at task's end when specific exemplar traces most dominate categorization (Smith & Minda, 1998). Just as before, we averaged the 12 observed profiles into a composite observed profile and the 12 best-fitting profiles of each model into a composite predicted profile. Figure 5 duplicates Figure 1, but with just 12 data sets included, showing the relationship between observed and predicted performance for the additive prototype model, the multiplicative prototype model, and the exemplar model. Figure 6 duplicates Figure 4, but with just 12 data sets included, showing the relationship between observed and predicted performance for the twin-sensitivity prototype model and the mixture model.

All the analyses based on the exclusive group of 12 data sets are identical to those based on the inclusive group of 30. The two observed profiles are nearly identical. The additive prototype model fails just as before. The multiplicative prototype model fails again by trying to split the difference between performance on the old and new items. The exemplar model, the twin-sensitivity prototype model, and the mixture model all fit well again. In fact, all the fits and misfits are so similar to those based on 30 data sets that careful scrutiny is required to see the differences.

In our view, it is instructive and constructive that the 12 data sets that most favor exemplar theory behave identically to the 30 data sets that sample the conditions of categorization more broadly. This means that the 30 data sets do form a coherent body of data. This means that one can treat the data inclusively and work toward an overall description of them.

Interpreting Performance in the 5–4 Task

To work toward this overall description, we now consider the convergences among the three equivalently successful models in this article—the mixture model, the twin-sensitivity model, and the exemplar model—to see if these convergences have implications for interpreting performance in the 5–4 task. In this section, we show that the 5–4 data pose one principal empirical, formal, and psychological problem—the performance advantage of old, training items over new, transfer items. We show that each successful model includes a mechanism that grants old items this simple, global performance boost, but that these mechanisms rest on different representational and processing assumptions. The mixture model assumes prototype-based processing supplemented by partial old-item memorization. The twin-sensitivity model assumes prototype-assimilation processes of greater fluency and skill for the old items. The exemplar model assumes categories that are prototypeless collections of instances and assumes that to-be-classified items are systematically compared to the stored exemplar traces that form the two categories. That these different mechanisms describe equally well 5–4 data has an important implication. It means that interpreting the 5–4 data sets does not require the purely exemplar-based representations and the systematic exemplar-to-exemplar comparison processes of exemplar theory. It means that the 30 5–4 data sets do not



selectively support the assumptions of exemplar theory. Because this section raises controversial issues, we stress that it does not represent a criticism of exemplar theory per se. Indeed, exemplar theory remains one of the several possible formal descriptions of performance in the 5–4 task. Rather, this section shows only that the body of data that most strongly motivated exemplar theory does not particularly do so. But other successes of exemplar theory might sustain it even without the support of these 30 results.

The six panels of Figure 7 allow these points to be made directly. Consider first Figure 7A. It shows the failure of the mixture model when we capped the exemplar-memorization parameter at 0.0 and denied the model any special way to cope with the practice effect on old items. (This constrained version of the mixture model is formally identical to Model 1, the additive prototype model.) This unsuccessful model fits new-item performance (Stimuli 10–16) but undershoots old-item performance (Stimuli 1–9). Reaching up to fit old-item performance is the principal fitting problem regarding the 5–4 data sets. So this model needs a way to selectively target old-item performance and raise it. The exemplar-memorization parameter provides this (Figure 7B). Over 30 data sets its average value was .50 when the mixture model was unconstrained.

Formally, this parameter simply acknowledges a practice effect by selectively increasing old-item performance. This selective increase is clear in the graphs and it is guaranteed because the parameter only applies to the old items. Psychologically, this parameter might reflect that participants partially memorize old items and use recognition-based categorization processes to supplement prototype-based categorization processes. So this model does see a role in categorization for exemplar representations, but its assumptions still differ profoundly from the assumptions of exemplar theory and the context model. The mixture model still grounds categories in prototypes. Prototype-based processing still governs categorization much of the time. The mixture model does not assume, as exemplar theory does, that categories are prototypeless collections of instances. Nor does it assume, as the context model does, that systematic exemplar-to-exemplar comparison processes underlie categorization. Instead, the mixture model just embodies the intuitive idea that people know generally what dogs are, but know specifically their own dogs best.

Consider next Figure 7C. It shows the failure of the twin-sensitivity prototype model when we capped sensitivity at a level that let the model fit new-item performance well. (We chose this level of sensitivity, 1.5, to let the constrained twin-sensitivity model fit new-item performance exactly as well as the constrained mixture model did.) This

---

Figure 5 (left). A: The composite observed performance profile produced by averaging a restricted set of 12 of the 5–4 data sets (solid line). Also shown is the average of the best-fitting predicted performance profiles found when the 12 data sets were fit individually using the additive prototype model (dotted line). B: The same observed profile shown with the composite predicted profile of the multiplicative prototype model. C: The same observed profile shown with the composite predicted profile of the context model.

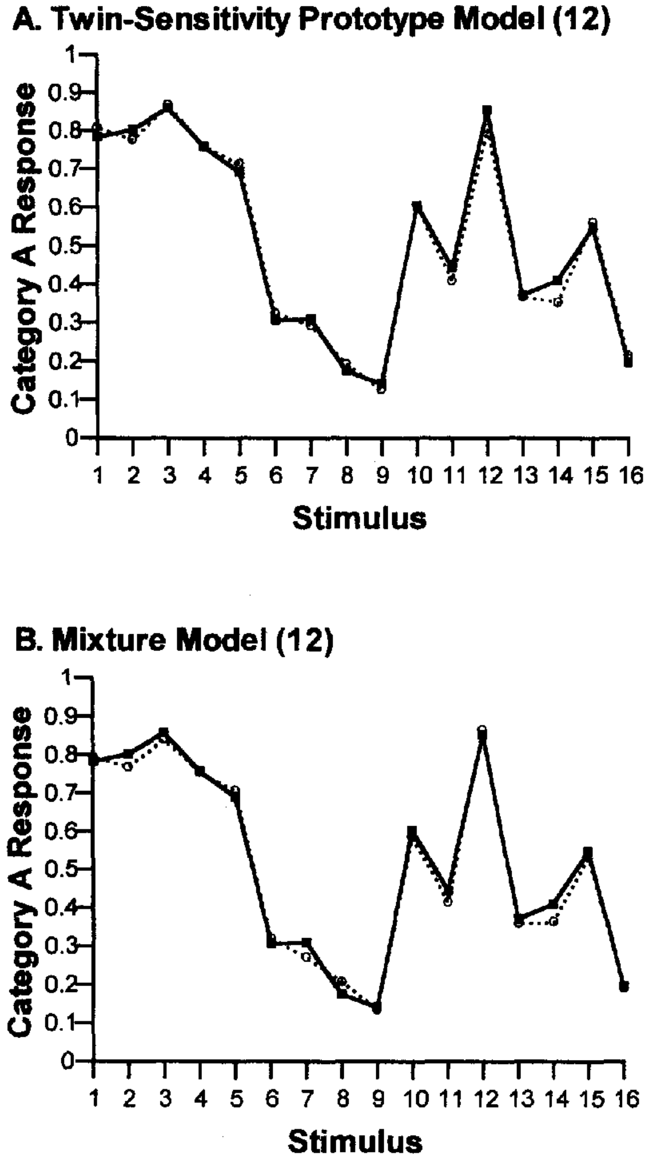


Figure 6. A: The composite observed performance profile produced by averaging a restricted set of 12 of the 5–4 data sets (solid line) shown with the composite predicted performance profile (dotted line) of the twin-sensitivity prototype model. B: The same observed profile shown with the composite predicted profile of the mixture model.

cap let us examine the behavior of the twin-sensitivity model when it is denied any special way to cope with the principal formal problem posed by the 5–4 data sets—the practice effect on old items. This unsuccessful model fits new-item performance but undershoots old-item performance. It too needs a way to selectively target old-item performance and raise it. The old-item sensitivity parameter provides this (Figure 7D). Over the 30 data sets, its average value was 7.09 when the twin-sensitivity model was unconstrained.

Formally, this parameter just selectively increases old-item performance. This selective increase is clear in the graphs, and it is guaranteed because this parameter applies only to the old items. Psychologically, this parameter might mean that practice increases the skill with which old items are assimilated to their appropriate prototype, leading to higher levels of performance on them, whereas the new items are assimilated to prototypes less skillfully, leading to lower levels of performance on them. It seems intuitive that often-repeated comparisons to a prototype would run off more smoothly than first-time comparisons to a prototype. But once again this model grounds categories in prototypes and categorization in prototype-based processes. All comparisons are made to the prototype, some are just made more fluently. This model clearly does not assume, as exemplar theory does, that categories are prototypeless collections of instances or that systematic exemplar-to-exemplar comparison processes underlie categorization.

There are important similarities between the mixture model and the twin-sensitivity model. Both acknowledge that the old exemplars have a special status as to-be-categorized items. (They must, because the old-item performance advantage is the main empirical fact of the 5–4 data sets.) Now one can attribute this processing advantage (as in the twin-sensitivity model) to fluency, priming, the strength of connection to the prototype, or the ease and skill of assimilation to it. Or one can attribute this processing advantage (as in the mixture model) to the item's familiarity, memorization, or explicit recognition. But because there are close connections in the psychological literature between these two sets of variables, one can choose whether to highlight or downplay the processing distinction between the models. One principled distinction could be that under a memorization interpretation, the exemplars would be stored with their category labels, so that self-retrieval directly supported a correct categorization decision that could supplement other categorization strategies. In contrast, under a fluency interpretation, the well-practiced old exemplars would be processed fluently through to a connection with a prototype that itself had a category label—here the exemplars would not be stored with their category labels. For us, though, the critical point is that both models are similar for acknowledging prototype representations and prototype-based comparisons, for restricting the effect of any exemplar process to the old items only, and for contrasting sharply with the purely exemplar-based representations of exemplar theory and the systematic and comprehensive exemplar-to-exemplar comparisons of the context model.

Finally, consider Figure 7E. It shows the failure of the exemplar model when it is granted only the sensitivity that lets it fit new-item performance well. (We chose this level of sensitivity, 4.0, to let the constrained context model fit new-item performance exactly as well as the constrained mixture model and constrained twin-sensitivity model did.) This cap allowed us to examine the behavior of the exemplar model when it is denied any special way to cope with the practice effect on old items. It fits new-item performance but undershoots old-item performance—just as the other models do. It also needs an extra adjustment to selectively target old-item performance and raise it. High values of the sensitivity parameter provide this (see Figure 7F). Over the 30 data sets, its average value was 8.19 when the context model was unconstrained.

One can see from Figures 7E and 7F that the success of the exemplar model is not assured by its exemplar represen-

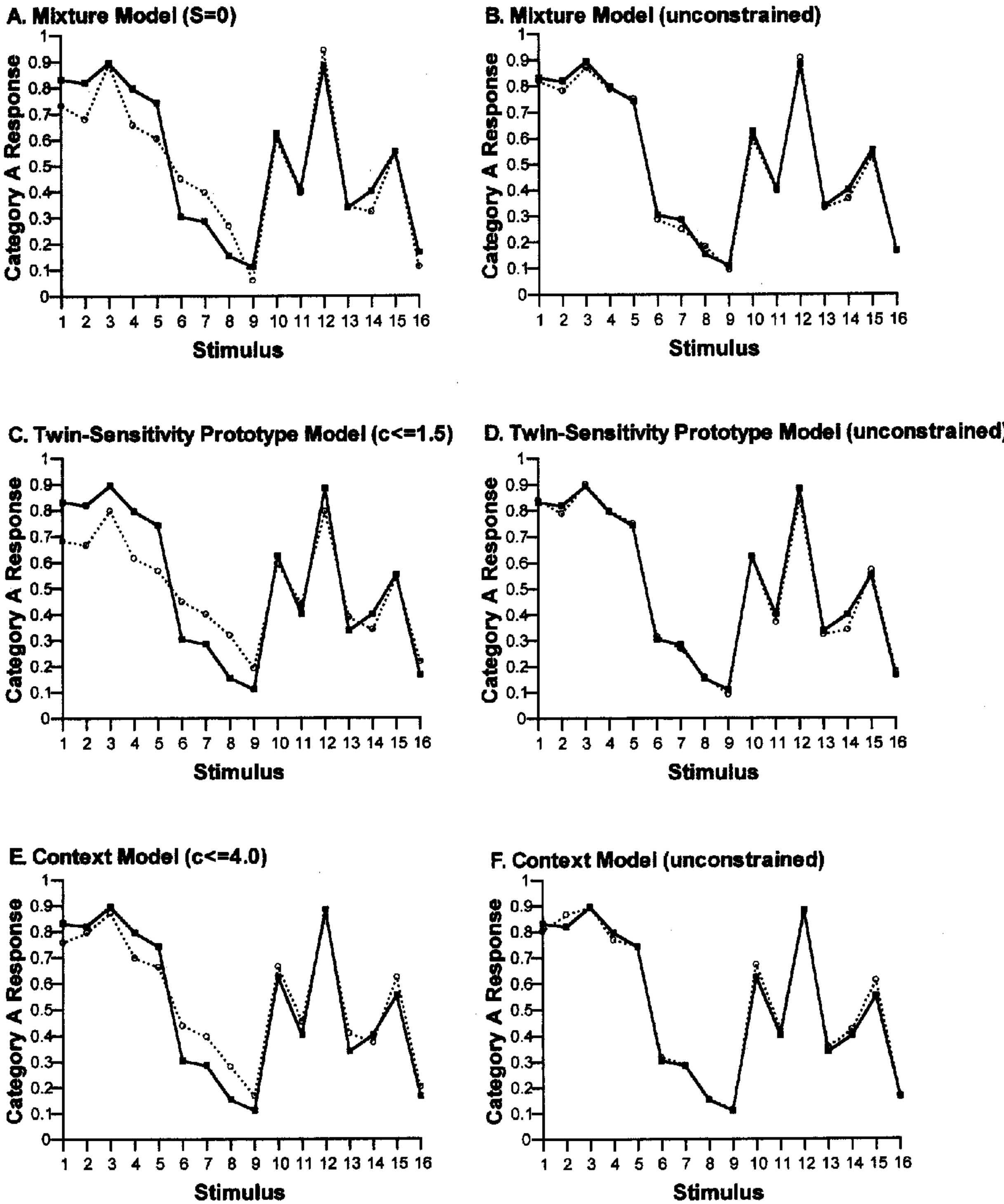


Figure 7. A: The composite observed performance profile from the 30 data sets (solid line) shown with the composite predicted performance profile of the mixture model (dotted line) when the value of the exemplar-memorization parameter (S) was constrained to be 0.0. B: The same observed profile shown with the composite predicted profile of the mixture model when the value of the exemplar-memorization parameter was unconstrained. C: The same observed profile shown with the composite predicted performance profile of the twin-sensitivity prototype model when we constrained both of its sensitivity parameters to be less than or equal to 1.5. D: The same observed profile shown with the composite predicted performance profile of the twin-sensitivity protototype model when the value of both sensitivity parameters was unconstrained. E: The same observed profile shown with the composite predicted performance profile of the context model when we constrained the value of its sensitivity parameter to be less than or equal to 4.0. F: The same observed profile shown with the composite predicted performance profile of the context model when the value of its sensitivity parameter was unconstrained.

tations or its systematic exemplar-to-exemplar comparisons. Rather, its success is assured by the formal mechanics of high sensitivity. High sensitivity just selectively increases old-item performance. This selective increase is clear in the graphs, and one can also show formally that increases in sensitivity beyond 4.0 (in the case of 5–4 performances) act nearly exclusively to raise old-item performance. In essence, the context model has a second sensitivity that applies only to old items, just as the twin-sensitivity model does. In essence, the context model has an old-item parameter, just as the mixture model does.

Taken together, the six graphs of Figure 7 show that performance in the 5–4 task, whether on the new, transfer items or the old, training items, really has no particular representational or process implications. All three models, whether they assume prototype or exemplar representations, explain new-item performance equivalently well and easily. All three confront the selective training boost to old items and incorporate a mechanism that reproduces that boost. But the boost—whether it is modeled in a way that is grounded in prototypes or exemplars, and whether it is attributed to memorization, skilled prototype assimilation, or high sensitivity—only acknowledges that participants perform better on old items. It does not confirm the purely exemplar-based categories of exemplar theory. It does not confirm the systematic exemplar-to-exemplar comparisons of the context model. In fact, those exemplar representations and exemplar processes are demonstrably as insufficient (Figure 7E) as are the assumptions of the prototype models (Figures 7A and 7C) unless they are supplemented (Figure 7F) by the same old-item boost that the prototype models need (Figures 7B and 7D). In all three cases, the boost is the thing; the underlying representation and process remains undetermined and unknown. Therefore, one sees that the 30 5–4 data sets, when described by formal models, are silent on the matter of whether categories are represented in a way that is based on prototypes or in a way that is based on exemplars. As a result, these 30 data sets cannot be judged to selectively support exemplar theory. To the contrary, assuming prototoype-based representational cores is still perfectly appropriate regarding the 30 5–4 performance profiles, if only one assumes that participants also memorize or gain fluency with the training items.

In fact, one might even prefer the prototype-based description, because both memorization and practiced skill ground the old-item advantage in intuitive psychological processes. All of us know that we memorize things and get better with repetition and practice. In contrast, the exemplar model's global sensitivity parameter is less grounded in intuitive psychological processes like these. In fact, high sensitivity in the exemplar model is even potentially misleading regarding the 5–4 data sets because it seems to apply to all the items but really applies selectively to the old items—just as an old-exemplar parameter does. Maybe the simplest interpretation of the 5–4 performances is that something changes in the processing of the old items to grant them their performance advantage. Saying that old items become practiced or memorized addresses this possibility more clearly than does saying that sensitivity increases globally from four to eight.

But even stopping short of any preference for a prototype-based description, one sees that the profound theoretical shift from prototype-based to exemplar-based representations that the 5–4 category structure motivated was not necessary to explain these data. Furthermore, regarding the 30 5–4 performance profiles, one need not assume the systematic exemplar-to-exemplar comparisons that the context model does, in which a to-be-categorized item is compared systematically to all the Category A exemplars and to all the Category B exemplars, and then placed into the category with the more similar stored exemplar traces. A simpler exemplar process suffices—old-item fluency or memorization or recognition. That is, it is sufficient to say that old items simply self-retrieve and boost their own performance. There is no reason to say that any item is ever compared to many exemplars in the processing that produces a categorization decision. In fact, even the context model's description makes it clear that exemplar-memorization events have 20 times the impact on the system that exemplar-comparison events do. Even the context model's description makes it seem implausible that participants would try to scrape together the 5% similarities of exemplar comparisons when they could use instead the 100% similarities of self-retrieval.

If the good fit of the exemplar model in these 30 cases is just about a practice effect, if it is just about memorization or old-item fluency, then it would be useful for theory to say so. For then the good fit is not about systematic exemplar-to-exemplar comparisons, and the context model's description of humans' processing in the 5–4 task is incorrect. There are important psychological differences between these two process interpretations. For example, the memorization interpretation leaves open the theoretical possibility that categories are grounded in prototypes. Put another way, the simpler exemplar process avoids making unnecessary representational commitments about the cores of categories. The context model's exemplar process makes representational commitments that are unnecessary regarding the 5–4 data.

This discussion raises a theoretical concern about exemplar theory. The concern is that higher and higher sensitivity in the context model can quietly change its fundamental character from one that features exemplar-to-exemplar comparisons to one that features exemplar memorization. High sensitivity can stretch psychological space until the exemplars become such distant neighbors that of course they do not contribute to each other's categorization, but only to their own. It is well recognized that high sensitivity stretches psychological space and tightens the circle of exemplar generalization so that fewer exemplars contribute to processing (e.g., Kruschke, 1992; Lamberts, 1994). It is insufficiently recognized that the quantitative stretch of high sensitivity may finally change the exemplar model's qualitative character—turning an exemplar-generalization model into an exemplar-memorization model.

However, as we raise this concern, we also note that there are limits on its present extension. We do not know how broadly the idea of prototypes combined with exemplar memorization will be applicable, because here we are exploring only one category structure, albeit an important one, and 30 influential data sets within the literature. The context model has had significant successes regarding other category structures, too. Regarding one of these successes, Smith and Minda (1998) showed that the idea of prototypes combined with old-item fluency has utility in describing performance on the class of nonlinearly separable category structures that had seemed to favor the context model. Minda and Smith (1999) showed that the idea of prototypes combined with the simpler exemplar process described data from many different category structures that varied widely in structural ratio, in stimulus complexity, and in category size. Even so, future research must still establish how far this idea extends.

We also point out that our claims about simplicity and complexity are only about the exemplar processes in the mixture model and in the context model. Exemplar memorization is simpler than the systematic exemplar-to-exemplar comparisons assumed by the context model. In fact, an emphasis on old-item fluency could help exemplar theory temper its most controversial assumption (that all exemplars are remembered and used in category decisions). However, viewed more broadly, the mixture model's description of processing in a category task is not simpler than the context model's description. Formally, the mixture model and the context model have the same number of free parameters; the twin-sensitivity model has one more. Psychologically, the mixture and twin-sensitivity models could even be viewed as being more complex than the context model because they both invoke prototypes and a simple exemplar process, whereas the context model uses a single representational system more complexly. Future research will have to establish whether and when it is useful to include both prototypes and simple exemplar processes in an overall theory of category learning. For example, in Smith and Minda (1998) there appeared to be an interesting psychological transition during category learning—from prototype-based processing to prototype-based processing combined with memorization or fluency—that a purely exemplar-based model could not emulate and could never in principle illuminate.

The A2 Advantage

Is there an A2 advantage? One result from the 5–4 category structure deserves special attention, because it might provide crucial support for exemplar theory and for the context model's systematic exemplar-to-exemplar comparisons (Medin et al., 1984; Medin & Schaffer, 1978; Medin & Smith, 1981; Nosofsky, 1992; Nosofsky et al., 1992; Nosofsky et al., 1994). So we now consider the relative performance of participants on Stimuli A1 (1 1 1 0) and A2 (1 0 1 0). This stimulus pair is important because A1 seems more prototypical than A2. It is more similar to the Category A prototype (1 1 1 1) than is A2, which shares features equally with both prototypes. In fact, A1 is identical to A2 except for its additional prototypical feature in Dimension 2. This is why prototype models generally predict an A1 performance advantage in the 5–4 task.

However, because of Medin's careful choice of the other stimuli, the context model predicts a performance advantage for Stimulus A2. This is why. Stimulus A2 (1 0 1 0) is highly similar (3 features shared) to two other Category A exemplars (A1, 1 1 1 0; A3, 1 0 1 1) and no Category B exemplars. If participants really do make the systematic exemplar-to-exemplar comparisons assumed by exemplar theory, A2 will seem to be a strong Category A exemplar and will seldom be miscategorized into Category B. Stimulus A1 (1 1 1 0) is highly similar to only one other Category A exemplar (A2, 1 0 1 0) but two Category B exemplars (B6, 1 1 0 0; B7, 0 1 1 0). If participants really do make the systematic exemplar-to-exemplar comparisons assumed by exemplar theory, A1 will seem to be a weak Category A exemplar and will often be miscategorized into Category B. The constraints involved in contrasting prototype-based and exemplar-based similarity explain the 5–4 category structure's poor differentiation and ambiguous items.

To illustrate this prediction of the context model, we found the average A1 and A2 performances of 2,500 simulated samples of 16 exemplar-based processors (Figure 8A). Appendix C contains the details of this simulation. On average, samples of exemplar-based processors show an 8% A2 advantage. They lie above the line of equal performance on A1 and A2.

In contrast, Figure 8B shows the 30 existing A1–A2 results. The overall result from the 30 data sets seems to be that there are A1 advantages and A2 advantages that are scattered around equivalent A1–A2 performance. In fact, real participants have shown no overall performance advantage on A2. A1 and A2 were performed equivalently over the 30 data sets (83.0% correct for A1, $SD = 0.106$; 81.8% correct for A2, $SD = 0.128$).

Which data sets should show an A2 advantage? But which data sets can fairly be included in a search for A2 advantages? Echoing the concern raised earlier, if exemplar theory cannot be expected to describe the data produced under deadline conditions, early in learning, or under prototype-based or rule-plus-exception instructions, then one might not expect A2 advantages under those conditions. In fact, the concern is especially serious here, because the A2 advantage has been so critical in motivating exemplar theory and has offered the main demonstration that exemplar theory's predictions are qualitatively correct, not just quantitatively better.

Accordingly, we first excluded 10 of the performance profiles that might have been produced under instructional sets that were less than ideal for fostering pure exemplar processing (e.g., specific instructions to use a prototype-based or rule-based strategy). On this basis, we set aside data sets 4, 5, 6, 11, 12, 14, 15, 17, 18, and 25. A1 and A2 were still performed statistically equivalently (80.8% for A1, $SD = 0.117$; 82.3% for A2, $SD = 0.128$), $t(19) = -0.57$, *ns*.

Next, we excluded eight data sets that related to early stages of learning, to deadline conditions, or that might have caused us to weight individual articles too heavily (for example, by including multiple data sets from them). On this basis, we set aside data sets 3, 19, 20, 21, 22, 27, 28, and 29. A1 and A2 were still performed statistically equivalently (82.3% for A1, $SD = 0.095$; 80.5% for A2, $SD = 0.135$), $t(21) = 0.67$, *ns*.

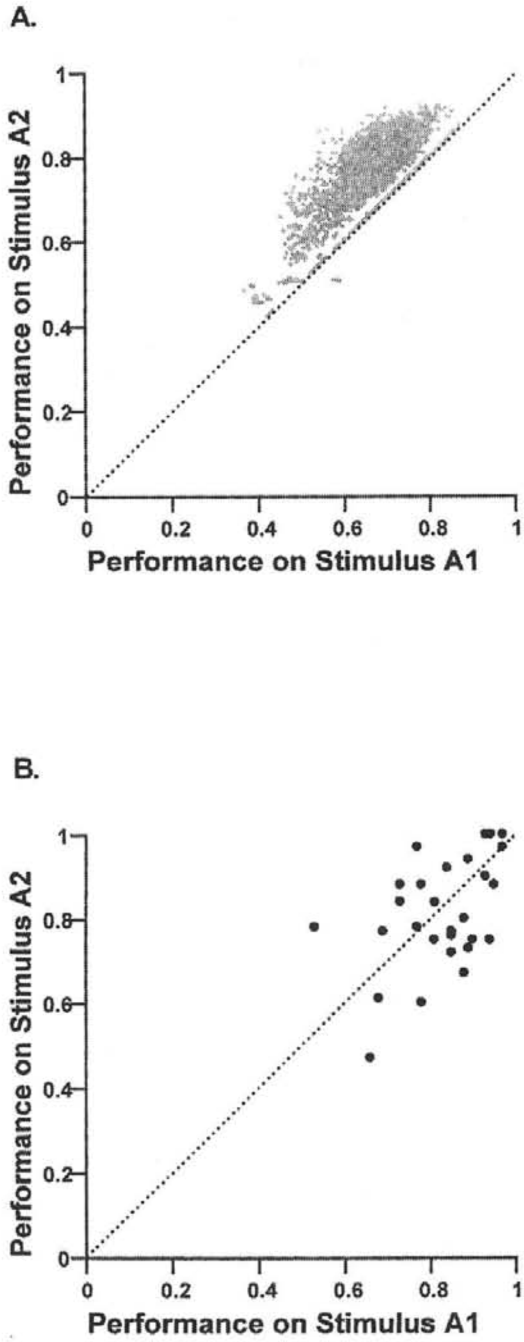


Figure 8. A: The A1–A2 performances of 2,500 simulated samples each containing 16 configurations of the context model (gray dots). See Appendix C for details. B: The observed A1–A2 performances from the 30 data sets (black dots).

To address both concerns simultaneously, we excluded all 18 data sets and focused on the 12 remaining, just as we did in the remodeling that led to Figures 5 and 6. On this basis, we set aside data sets 3, 4, 5, 6, 11, 12, 14, 15, 17, 18, 19, 20, 21, 22, 25, 27, 28, and 29. A1 and A2 were still performed statistically equivalently (78.2% for A1, $SD = 0.097$; 80.3% for A2, $SD = 0.140$), $t(11) = -0.67$, *ns*.

We point out again that theoretical implications would attend the need to exclude existing data to show the effect that exemplar theory predicts. This need would narrow the scope of exemplar theory—in its research applications, in its real-world applications, and in its comprehensiveness in both spheres. However, the A2 analyses on 12 and 30 data sets suggest identical conclusions. The earlier modeling on 12 and 30 data sets suggested identical conclusions. Accordingly, it seems possible to treat the 5–4 data inclusively and to try to explain them broadly. Equivalent performance on A1 and A2 is part of that broader data pattern.

This equivalence is important because the A2 result has long been claimed to provide strong and definite support for the systematic exemplar-to-exemplar comparisons assumed by exemplar theory. But the result is not consistent, robust, present overall, or even present within the 12 data sets that should most favor exemplar theory. We do not rule out that further research will occasionally find the A2 advantage. It will. We do not rule out that further research could establish the conditions that produce the A2 advantage robustly and consistently. It may. This section only considers the present status of the result. At present, the A2 advantage should probably not be granted a place in the lore of the literature, and exemplar theory should be defended on other grounds. For example, one can rightly note that A1 and A2 are only 2 of the 16 stimuli in the 5–4 task and that overall the exemplar model does an excellent job of fitting the 5–4 data sets. Here, though, the problem of the previous section arises again—other models with different psychologies behind them fit as well. The A2 advantage was so crucial theoretically because it qualitatively favored exemplar theory's assumptions alone. Letting that result go for the present is empirically correct but has important theoretical implications.

Interpreting A2 advantages. Given the importance of this issue, we consider specifically the six data sets that have shown clear A2 advantages and show that these data sets create theoretical tensions and fit uncomfortably even within exemplar theory's own framework.

Research using the context model has emphasized that participants attend adaptively to the different features in a task and that the dimensional weights of the context model reflect their attentional strategies (Nosofsky, 1984, 1986, 1991; Nosofsky et al., 1994). In fact, Nosofsky (1987) demonstrated this in an elegant analysis using six different category structures. Nosofsky (1984) also showed this using a 5–4 data set (Data Set 6 in this article). Lamberts (1995) also showed this using his 5–4 data (Data Set 30 in this article). It is important if participants "distribute attention among the component dimensions in a way that tends to optimize performance" (Nosofsky, 1984, p. 109). It is important if the context model successfully describes the stretching and shrinking of psychological space that underlies these attentional shifts.

Figure 9A shows optimal attention in the 5–4 task, as estimated by Lamberts (1995) and Nosofsky (1984). It makes sense that participants should focus their attention on Dimensions 1 and 3, for these features have the highest diagnosticities (.77) and offer the best information for solving the 5–4 category set. It makes sense that participants

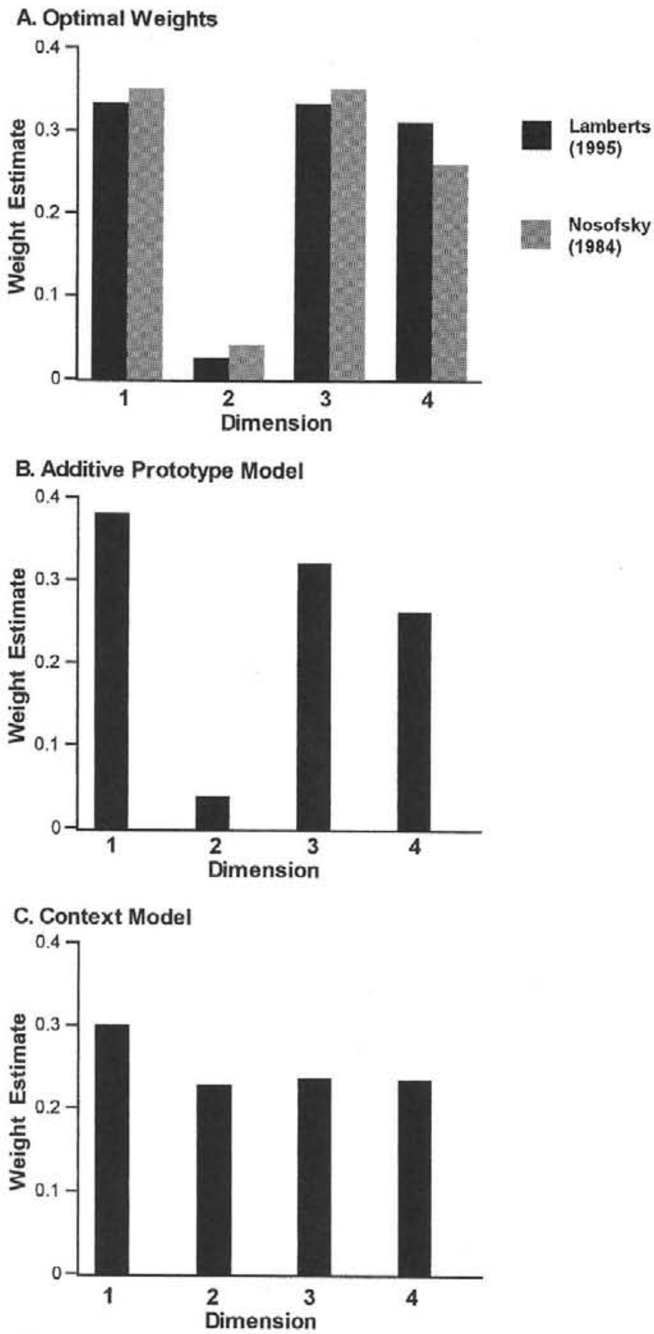


Figure 9. A. The dimensional weights of the context model that optimize performance in the 5–4 task, as reported by Lamberts (1995) and Nosofsky (1984). B. The prototype model's average dimensional weights as it fit the data from the six 5–4 data sets that showed A2 advantages larger than 10%. C. The context model's average dimensional weights as it fit exactly the same data.

should ignore Dimension 2, for this feature has the lowest diagnosticity (.55) and offers almost useless information for solving the 5–4 category set.

But adaptive attention has a crucial performance implication. If Dimension 2 is ignored, Stimuli A1 (1 1 1 0) and A2 (1 0 1 0) become functionally or formally equivalent because they are identical in all other respects. Accordingly, adaptive attention produces equivalent A1 and A2 performance—that is—no A2 advantage.

This makes it clear why A2 advantages fit uncomfortably within the literature on exemplar theory. The literature makes one expect adaptive attention and no A2 advantages, but there they are. The A2 advantages go against the literature's claims of adaptive attention, but there they are. Without judging how to resolve this tension, we point out that A2 advantages create the tension and, instead of cleanly supporting exemplar theory and the context model, raise definite questions.

In particular, one wonders how participants are attending when they show occasional A2 advantages. Figure 9B shows the average dimensional weights offered by the prototype model as it fit the six data sets that showed an A2 advantage that was greater than 10%. Figure 9C shows the average dimensional weights offered by the exemplar model as it fit exactly the same data. These figures underscore the theoretical trouble that A2 advantages cause. If one accepts exemplar theory's expectation of adaptive attention, one disbelieves the exemplar model's attentional description (Figure 9C) and believes instead the prototype model's attentional description (Figure 9B). If one accepts that A2 advantages are the real and critical result, then one accepts the maladaptive attentional policy (Figure 9C) but places a problematic result amidst the exemplar model's most elegant demonstrations.

Figures 9B and 9C also raise an interesting possibility about future research in this area. The attentional descriptions offered by the exemplar and prototype models for the very same data differ in striking ways. One of these descriptions is probably wrong, and this offers a possible way to ask whether prototypes or exemplars ground performance in the 5–4 task.

Specialized Commitments of the 5–4 Category Structure

In the final section of this article, we suggest that the 5–4 category structure samples only one of the interesting regions in the universe of category tasks. It is a specialized structure that for three reasons might favor a particular class of information-processing strategies.

Poor category differentiation. We have already discussed that the 5–4 category structure contains features that are poorly diagnostic of category membership, families of exemplars that have a weak family-resemblance, and categories that overlap substantially in multidimensional perceptual space. Figure 10A shows the relatively impoverished structure offered by the 5–4 task. The figure summarizes a systematic search through the space of category structures based on binary features to illustrate the range of category differentiation that is available to category researchers. Appendix C describes this simulation. The star marks the spot of the 5–4 category structure in this structural-ratio space. The 5–4 task presents to participants a categorization problem with nearly the minimum structure within catego-

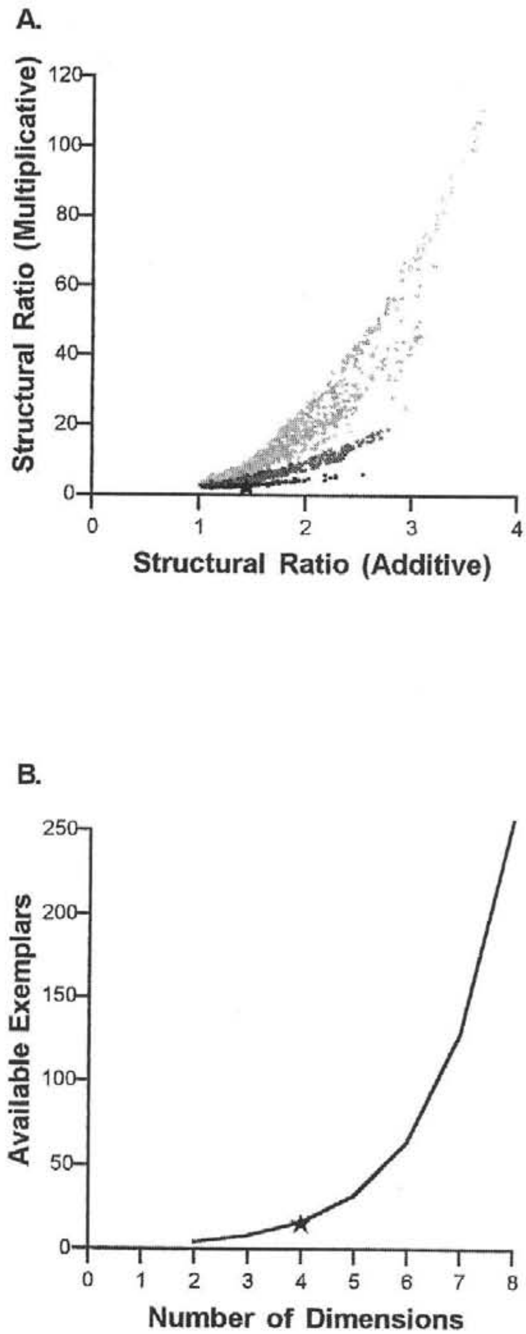


Figure 10. A: The range of category differentiation potentially offered by four-dimensional category structures (dark gray dots), six-dimensional category structures (medium gray dots), and eight-dimensional category structures (light gray dots). See Appendix C for details. The star shows the position of the 5–4 category structure in this larger space of category differentiation. B: An illustration of the range of exemplar availability offered by category structures made from stimuli of different dimensionalities. The star shows the position of the 5–4 category structure in this larger space of category sizes.

ries and differentiation between categories. This is true not only of the 5–4 category structure but also of many related category structures that have been used in the literature.

This impoverished structure could lead participants to focus on one dimension that categorizes most of the stimuli (it probably would not be Dimension 2) and to try to memorize the exceptions created by that focus. Poor structure could lead them to try to memorize all nine training exemplars because strategies based in family resemblance (whether grounded in exemplars or prototypes) will work poorly. Thus, even though the impoverished category structures exemplified by the 5–4 category structure represent one theoretically interesting kind of category structure, they are not theoretically neutral: They may encourage some information-processing strategies and defeat others. Thus, even if the strategies elicited by the 5–4 task are partially based in exemplar memorization, these strategies must be generalized carefully and narrowly—perhaps only to other poorly differentiated category structures.

Accordingly, we hope that existing research will be complemented by research on regions of structural-ratio space that contain categories with stronger family resemblance within categories and more differentiation between categories (Smith & Minda, 1998). In pursuing this goal, it may be useful for researchers to study stimulus spaces of higher dimensionalities (Smith & Minda, 1998; Smith et al., 1997). Figure 10A shows that four-dimensional stimuli generally constrain the researcher to focus on poorly differentiated category structures, whereas six- or eight-dimensional stimuli let the researcher sample other regions of the large space of possible category tasks.

Poor learnability. Poor differentiation may well have a psychological impact on participants. The 5–4 category structure is difficult to learn; participants often fail to reach the preset learning criterion. For example, in Medin and Schaffer (1978, Experiments 2 and 3) only 33 of 64 participants ever achieved one errorless run through the nine training stimuli. In Medin and Smith (1981), only 36 of 96 participants ever achieved one errorless run. In contrast, participants in other studies have met a criterion of 36 errorless trials (Hartley & Homa, 1981), 70 trials (Homa et al., 1981), or even 90 trials (Homa et al., 1979). The 5–4 category structure presents to participants a very difficult categorization problem.

This difficulty could also encourage special categorization strategies, possibly even exemplar-memorization strategies once again. If so, it is interesting that participants have these strategies available and turn to them when the category going gets tough. But even if they do, it is important to realize that these strategies may only generalize to equally difficult tasks that provide equally constant error messages.

Small exemplar sets. The 5–4 category structure is also specialized because it has only 9 training exemplars. Indeed, its whole stimulus set contains only 16 stimuli because just four binary features are used. (This is why the 5–4 task has 7 transfer items.) As one increases the dimensionality of the stimuli to six dimensions, to eight dimensions, and on, the size of the available stimulus population and the size of the possible categories grow exponentially, too. Figure 10B illustrates this point: The star marks the spot of the 5–4 category structure in the larger space of possibilities. The 5–4 structure, like many related category structures that have been used in the literature, presents to participants a categorization problem with nearly the minimum category size and nearly the maximum exemplar repetition during learning. This repetition could encourage and support exemplar-memorization strategies, but once again these supporting conditions and the resulting exemplar-memorization strategies would only generalize narrowly to other small-category tasks. We emphasize that these tasks, these conditions, and these strategies are no less interesting or important than any others. However, a better research balance could be achieved in the literature if there were more research involving larger categories.

Smith and Minda (1998) illustrated the special strategies elicited by small, poorly differentiated categories like those of the 5–4 task. They studied the learning of these categories and of larger, better differentiated categories. They fit a mixture model (similar to Model 8 here) to the data at 10 stages of learning. The mixture model is useful in this case because it combines prototype-based processing and exemplar memorization in the same model, and it gives them both the chance to dominate if their operating characteristics fit performance better. One can then watch the changing strengths of the two processes as learning progresses and see their different strengths for different category structures.

Figure 11A shows the mixture model's estimates of the levels of guessing, prototype use, and exemplar memorization when the model tracked these parameter values through the learning of larger, well-differentiated categories. The mixture model's description has prototype use quickly emerging as the dominant process and staying dominant through 400 trials of the experiment. Estimates of exemplar memorization appear late in learning and never reach a very high level.

Figure 11B shows the same parameter estimates through time as participants learned small, poorly differentiated categories like those in the 5–4 category structure. This trajectory through parameter space is totally different. Higher rates of guessing reflect the slowness with which these poorly structured categories come into focus for participants. Higher reliances on exemplar memorization eventuate. Most striking is that prototype processing is never the dominant categorization process under the description of the mixture model.

Thus, the mixture model makes plain that different category structures deflect participants into different regions of strategy space and alter profoundly the course of category learning. Homa (Homa et al., 1979; Homa et al., 1981; Homa & Chambliss, 1975) endorsed the claim that larger exemplar pools foster the emergence of prototype-based categorization strategies. Reed (1978) believed that both larger categories and better differentiated categories would have this effect (see also Smith & Minda, 1998). Likewise, Medin knew that categories like those in the 5–4 category structure could create a specific task psychology and elicit specialized strategies (Medin & Schaffer, 1978; Medin & Schwanenflugel, 1981). They might even turn a categorization task into an identification–memorization task in which participants pair associatively whole exemplars and their category labels but have no sense of coherent categories in doing so. McKinley and Nosofsky (1995, p. 129) echoed this possibility. Again, the suggestion is that the 5–4 category structure might elicit an exemplar-memorization process (in possible contrast to the systematic exemplar-to-exemplar comparison process assumed by exemplar theory).

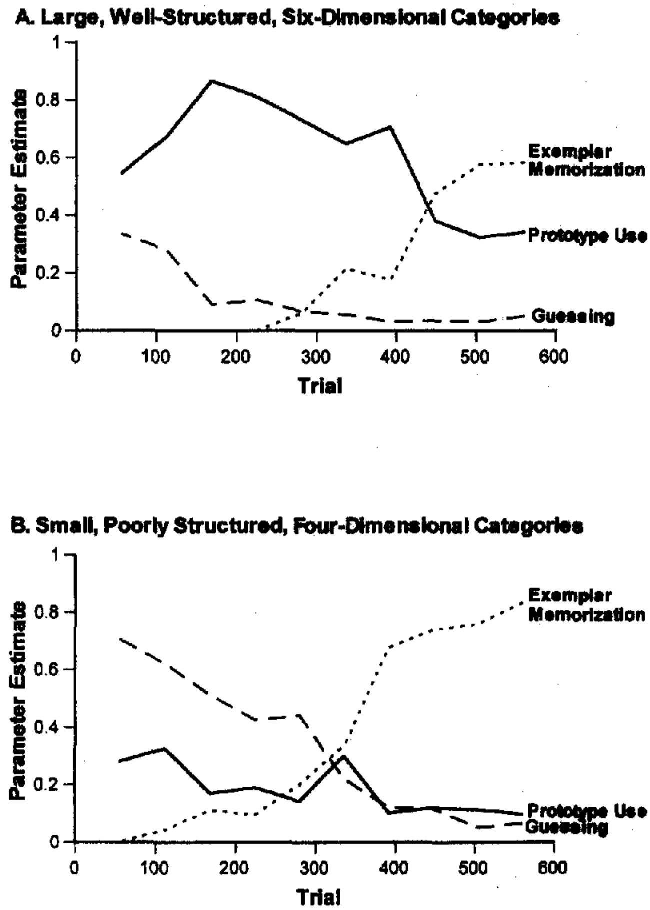


Figure 11. A: Median parameter estimates of the mixture model fitting the performance of participants learning large, well-differentiated category structures (Smith & Minda, 1998, Experiment 2). B: The same for participants learning small, poorly differentiated category structures (Smith & Minda, 1998, Experiment 3).

By all accounts, then, the 5–4 category structure is specialized, because it lies in the sparse, difficult, and poorly differentiated corner of the larger space of category structures and because it likely elicits a particular class of information-processing strategies (some of these possibly based in exemplar memorization). That this category structure has been dominant in asserting exemplar theory means that the data supporting that theory are narrower than has been realized and may unintentionally exaggerate the importance of exemplar processes in categorization.

Conclusions

So what do the 30 5–4 categorization results say in the end about exemplar theory's category representations and comparison processes? The formal problem facing all the models is to explain the old-item practice effect in the 5–4 data. This problem exists apart from how intuitive that practice effect is. The only advantage the context model has in fitting the 30 data sets, compared with the simplest prototype model, is that it accommodates this practice effect. However, several models, making different process and representational assumptions, accommodate this practice effect, too. Thus one need not assume the prototypeless category representations of exemplar theory to explain the 5–4 data. Prototypes combined with a practice effect suffices. One need not assume the systematic exemplar-to-exemplar comparisons of the context model. A simpler process—exemplar memorization—suffices.

This equivalence of prototype and exemplar representations, and of simple and complex exemplar processes, explains the critical importance of the A2 result to exemplar theory. For this result could qualitatively favor exemplar representations and a complex exemplar process. But the A2 advantage is neither present overall nor in the data sets that should most support exemplar theory.

In the end, it is clear that the 5–4 data sets do not offer unambiguous support for exemplar theory. This raises theoretical concern because the 5–4 category structure was so prominent in fostering exemplar theory and because the 5–4 categories represent so well the difficult, sparse, and poorly differentiated category structures that should most favor exemplar theory. If questions about exemplar theory arise here, then questions arise even more definitely about exemplar theory's breadth and extension.

Unfortunately, one cannot presently judge that breadth and extension. For the literature has focused sharply on categories like the 5–4 structure that are sparse, undifferentiated, and difficult. It will be constructive in the future to complement this research with research on more learnable and larger categories. This research could test the extension of exemplar theory, prototype theory, and other theories. It could suggest a broader theory of human categorization that emphasizes the richness and range of humans' approaches to category tasks (Homa, 1984; Reed, 1978; Smith & Minda, 1998). This research could even double back in the end and finally help these 30 categorization results in their search for a model.

References

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.

Hartley, J., & Homa, D. (1981). Abstraction of stylistic concepts. *Journal of Experimental Psychology: Learning, Memory, and Cognition, 7,* 33–46.

Homa, D. (1984). On the nature of categories. *The Psychology of Learning and Motivation, 18,* 49–94.

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, 1,* 351–359.

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.

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

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 & Psychophysics, 53,* 49–70.

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.

Mervis, C. B., & Rosch, E. (1981). Categorization of natural objects. *Annual Review of Psychology, 32,* 89–115.

Minda, J. P., & Smith, J. D. (1999). *Prototypes in category learning.* Manuscript submitted for publication.

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

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

Nosofsky, R. M. (1987). Attention and learning processes in the identification and categorization of integral stimuli. *Journal of Experimental Psychology: Learning, Memory, and Cognition, 13,* 87–108.

Nosofsky, R. M. (1988). Similarity, frequency, and category representations. *Journal of Experimental Psychology: Learning, Memory, and Cognition, 14,* 54–65.

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

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.

Smith, J. D., & Minda, J. P. (1998). Prototypes in the mist: The early epochs of category learning. *Journal of Experimental Psychology: Learning, Memory, and Cognition, 24,* 1411–1430.

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.

Appendix A

Data Sets

Data set	Reference	Experiment	Physical stimuli	Instruction–condition
1	Medin & Schaffer (1978)	Experiment 2	Geometric shapes	Neutral
2		Experiment 3	Brunswick faces	Neutral
3	Medin & Smith (1981)	Standard	Brunswick faces	Neutral
4		Rule plus exception	Brunswick faces	Rule plus exception
5		Prototype	Brunswick faces	Prototype instructions
6	Medin, Dewey, & Murphy (1983)	Last name infinite	Yearbook photos	Neutral
7		Last name only	Yearbook photos	Neutral
8		First name–last name	Yearbook photos	Learn first–last name
9		First name only	Yearbook photos	Learn first name
10	Medin, Altom, & Murphy (1984)	Experiment 1 examples only	Geometric shapes	Neutral
11		Experiment 1 concurrent	Geometric shapes	Prototype facts given concurrently
12		Experiment 1 prototypes first	Geometric shapes	Prototype facts given first
13		Experiment 2 examples only	Geometric shapes	Neutral
14		Experiment 2 concurrent	Geometric shapes	Prototype facts given concurrently
15		Experiment 2 prototypes first	Geometric shapes	Prototype facts given first
16		Experiment 2 examples only (verbal)	Verbal descriptions	Neutral
17		Experiment 2 concurrent (verbal)	Verbal descriptions	Prototype facts given concurrently
18		Experiment 2 prototypes first (verbal)	Verbal descriptions	Prototype facts given first
19	Nosofsky, Kruschke, & McKinley (1992)	Experiment 2 all	Geometric shapes	Neutral
20		Experiment 2 Block 1	Geometric shapes	Neutral
21		Experiment 2 Block 2	Geometric shapes	Neutral
22		Experiment 2 Block 3	Geometric shapes	Neutral
23		Experiment 2 Block 4	Geometric shapes	Neutral
24	Nosofsky, Palmeri, & McKinley (1994)	Experiment 1	Rocketships	Neutral
25	Palmeri & Nosofsky (1995)	Experiment 1	Rocketships	Rule plus exception
26		Experiment 2	Rocketships	Neutral
27	Lamberts (1995)	Experiment 2, 600 ms	Brunswick faces	Neutral–speeded
28		Experiment 2, 1,100 ms	Brunswick faces	Neutral–speeded
29		Experiment 2, 1,600 ms	Brunswick faces	Neutral–speeded
30		Experiment 2, no deadline	Brunswick faces	Neutral

Appendix B

Category A Response Probabilities for 16 Stimuli

Data set	Category A training stimuli					Category B training stimuli				Transfer stimuli						
	1110	1010	1011	1101	0111	1100	0110	0001	0000	1001	1000	1111	0010	0101	0011	0100
1	0.78	0.88	0.81	0.88	0.81	0.16	0.16	0.12	0.03	0.59	0.31	0.94	0.34	0.50	0.62	0.16
2	0.97	0.97	0.92	0.81	0.72	0.33	0.28	0.03	0.05	0.72	0.56	0.98	0.23	0.27	0.39	0.09
3	0.97	0.97	0.92	0.81	0.72	0.33	0.28	0.03	0.05	0.72	0.56	0.98	0.23	0.27	0.39	0.09
4	0.89	0.94	0.94	0.72	0.78	0.27	0.30	0.09	0.05	0.45	0.20	0.88	0.58	0.08	0.75	0.12
5	0.77	0.97	0.98	0.70	0.60	0.55	0.28	0.17	0.13	0.73	0.65	0.87	0.22	0.28	0.52	0.12
6	0.81	0.75	0.95	0.77	0.80	0.42	0.30	0.25	0.11	0.62	0.31	0.89	0.34	0.31	0.62	0.20
7	0.69	0.77	0.92	0.50	0.77	0.36	0.48	0.22	0.09	0.59	0.41	0.87	0.49	0.30	0.57	0.20
8	0.66	0.47	0.56	0.50	0.31	0.47	0.47	0.31	0.41	0.58	0.55	0.69	0.41	0.52	0.50	0.31
9	0.68	0.61	0.74	0.77	0.35	0.68	0.35	0.16	0.36	0.72	0.66	0.76	0.16	0.31	0.35	0.32
10	0.73	0.88	0.95	0.77	0.73	0.25	0.20	0.23	0.06	0.62	0.50	0.86	0.34	0.42	0.59	0.06
11	0.95	0.88	0.98	0.94	0.92	0.28	0.23	0.08	0.05	0.61	0.28	0.98	0.14	0.36	0.61	0.09
12	0.88	0.80	0.95	0.81	0.84	0.31	0.34	0.16	0.02	0.75	0.34	0.94	0.20	0.42	0.67	0.06
13	0.73	0.84	0.89	0.75	0.70	0.25	0.31	0.19	0.16	0.55	0.41	0.80	0.47	0.39	0.61	0.22
14	0.94	0.75	0.91	0.91	0.86	0.31	0.34	0.09	0.06	0.50	0.09	0.92	0.31	0.55	0.50	0.16
15	0.89	0.73	0.89	0.73	0.70	0.30	0.19	0.17	0.02	0.67	0.34	0.92	0.27	0.42	0.56	0.05
16	0.73	0.84	0.84	0.81	0.73	0.22	0.38	0.22	0.17	0.42	0.47	0.77	0.48	0.52	0.58	0.34
17	0.88	0.67	0.81	0.86	0.88	0.34	0.25	0.09	0.11	0.50	0.17	0.95	0.27	0.44	0.55	0.14
18	0.78	0.60	0.84	0.75	0.75	0.21	0.22	0.28	0.12	0.75	0.23	0.86	0.23	0.34	0.59	0.17
19	0.84	0.92	0.93	0.91	0.78	0.13	0.21	0.08	0.11	0.64	0.45	0.83	0.48	0.56	0.56	0.18
20	0.53	0.78	0.75	0.82	0.70	0.35	0.35	0.20	0.25	0.62	0.53	0.70	0.45	0.75	0.53	0.23
21	0.93	0.90	0.95	0.85	0.60	0.05	0.25	0.07	0.12	0.65	0.42	0.82	0.45	0.40	0.45	0.17
22	0.97	1.00	1.00	0.95	0.88	0.18	0.10	0.03	0.00	0.62	0.42	0.90	0.55	0.50	0.65	0.20
23	0.93	1.00	1.00	1.00	0.93	0.00	0.15	0.00	0.05	0.65	0.42	0.90	0.45	0.60	0.62	0.12
24	0.77	0.78	0.83	0.64	0.61	0.39	0.41	0.21	0.15	0.56	0.41	0.82	0.40	0.32	0.53	0.20
25	0.94	1.00	0.97	0.98	0.92	0.13	0.06	0.02	0.02	0.94	0.69	0.94	0.03	0.14	0.32	0.08
26	0.81	0.84	0.86	0.70	0.72	0.32	0.31	0.20	0.11	0.63	0.38	0.85	0.34	0.32	0.59	0.19
27	0.85	0.76	0.85	0.62	0.72	0.50	0.51	0.27	0.21	0.53	0.46	0.89	0.37	0.43	0.57	0.27
28	0.85	0.77	0.94	0.77	0.72	0.43	0.33	0.20	0.11	0.58	0.34	0.99	0.28	0.37	0.58	0.17
29	0.85	0.72	0.96	0.86	0.77	0.35	0.29	0.21	0.08	0.62	0.24	0.99	0.29	0.50	0.59	0.11
30	0.90	0.75	0.97	0.95	0.90	0.23	0.20	0.19	0.04	0.59	0.23	0.99	0.33	0.43	0.60	0.14

Appendix C

Details of Simulations

The Context Model's Prediction of an A2 Advantage

This simulation's goal was to create samples of 16 participants who obeyed the assumptions of the context model and to use these samples to show the range of A1–A2 performance profiles that the model predicts. To build each simulated group, we chose first an initial random configuration of the context model (i.e., a randomly chosen sensitivity and guessing rate and a random set of attentional weights that summed to 1.0). Then, we created 15 other simulated performers by treating the initial parameter setting as the mean of a Gaussian distribution with standard deviations of 5.0 for sensitivity, 0.10 for guessing, and 0.20 for the attentional weights, and by choosing 15 other configurations (each with its own sensitivity, guessing rate, and attentional weights summing to 1.0) in a way that obeyed the probability-density functions around these means. This ensured that each simulated sample contained 16 simulated performers who had a kind of family resemblance to their performance strategy because they shared roughly equivalent levels of guessing and sensitivity, and because they shared a similar attentional strategy.

Structural-Ratio Space

This simulation's goal was to illustrate the range of category structure that is attainable using binary features as in the 5–4 task. To be fair to current ideas about similarity, and to the models considered in this article, we calculated the measure of category structure (structural ratio) additively and multiplicatively. We calculated the structural ratio using the similarity of exemplars to exemplars, both within-category (including exemplar self-identities) and between-category. In every case we assumed homogeneous attention across the dimensions in the task. We constructed categories using four, six, or eight binary dimensions, and with four, seven, or nine exemplars, respectively. Category members were always derived from the nominal category prototypes of 1 1 1 1 and 0 0 0 0, 1 1 1 1 1 1 and 0 0 0 0 0 0, or 1 1 1 1 1 1 1 1 and 0 0 0 0 0 0 0 0.

Searching these structural-ratio spaces exhaustively is not feasible; for example, the eight-dimensional space contains about 10^{19} Category A sets. Searching these spaces by choosing category structures randomly is misleading because chance will not produce the well- and poorly structured categories that a researcher might construct. Thus we chose to search structural-ratio space systematically, but to sample only a limited number of representative category tasks from each region of the space.

To conduct this search, we varied the typicality of the exemplars included in categories. For example, the stimuli 1 1 1 1, 1 1 1 0, 1 1 0 1, and 0 1 1 1 are typical stimuli: Their inclusion will generally produce higher structural ratios for a category task. The stimuli 1 0 1 0, 0 0 1 1, 1 0 0 1, and 1 1 0 0 are less typical stimuli: Their inclusion will generally produce lower structural ratios for a category task. By marrying the different levels of typicality into category tasks, by using different mixes, and by sampling all possible mixes, we made the search systematic but kept it limited, because a given mix will produce about the same structural ratio even if the specific stimuli change.

To illustrate our technique for the case of eight dimensions, we divided the eight-dimensional stimuli into high-typicality items (those with 8 or 7 prototypical features), mid-typicality items (those with 6 prototypical features), and low-typicality items (those with 5 or 4 prototypical features). Then we generated 54 selection procedures—8 1 0, 8 0 1, 7 2 0, . . . , and 0 0 9—that specified how many items at each level of typicality would be included in each category. The first of these (8 1 0) included eight high-typicality items and one mid-typicality item to produce categories with high structural ratios. The last of these (0 0 9) included nine low-typicality items to produce categories with low structural ratios. We examined 20 stimulus sets for each mix, for a total of 1,080 stimulus sets that illustrated the range of structure available in eight-dimensional tasks.

Similarly, we divided the six-dimensional stimuli into high-typicality items (with 6 or 5 prototypical features), mid-typicality items (with 4 prototypical features), and low-typicality items (with 3 prototypical features), generated 35 selection procedures—6 1 0, 6 0 1, 5 2 0, 5 1 1, 5 0 2, . . . , and 0 0 7—that specified the mix of different typicalities, and examined 30 stimulus sets for each mix for a total of 1,050 stimulus sets.

Similarly, we divided the four-dimensional stimuli into high-typicality items (with 4 or 3 prototypical features) and low-typicality items (with 2 prototypical features). We then generated three selection procedures—3 1, 2 2, 1 3—that specified the mix of different typicalities and examined 400 stimulus sets for each mix for a total of 1,200 stimulus sets.

Received May 18, 1998
Revision received April 8, 1999
Accepted April 22, 1999 ■