Psychonomic Bulletin & Review
2000, 7 (3), 375-402

Exemplar-based accounts of "multiple-system" phenomena in perceptual categorization

ROBERT M. NOSOFSKY and MARK K. JOHANSEN
Indiana University, Bloomington, Indiana

We demonstrate that a wide variety of recently reported "rule-described" and "prototype-described" phenomena in perceptual classification, which have led to the development of a number of multiple-system models, can be given an alternative interpretation in terms of a single-system exemplar-similarity model. The phenomena include various rule- and prototype-described patterns of generalization, dissociations between categorization and similarity judgments, and dissociations between categorization and old–new recognition. The alternative exemplar-based interpretation relies on the idea that similarity is not an invariant relation but a context-dependent one. Similarity relations among exemplars change systematically because of selective attention to dimensions and because of changes in the level of sensitivity relating judged similarity to distance in psychological space. Adaptive learning principles may help explain the systematic influence of the selective attention process and of modulation in sensitivity settings on judged similarity.

Recent theorizing in the field of perceptual classification has seen a proliferation of models that posit the operation of multiple categorization systems (e.g., Ashby, Alfonso-Reese, Turken, & Waldron, 1998; Erickson & Kruschke, 1998; Nosofsky, Palmeri, & McKinley, 1994; Palmeri & Nosofsky, 1995; J. D. Smith & Minda, 1998; Vandierendonck, 1995). Although the models differ in important details, a general theme is that one system computes category summary representations such as rules or prototypes, whereas the second system relies on more specific representations such as stored exemplars or complex, nonverbalizable decision boundaries. The idea that multiple systems serve category learning and representation is highly plausible; however, the adoption of multiple-system models comes at a price. In particular, because such models are so powerful and flexible and typically involve numerous free parameters, they may resist falsification. On grounds of scientific parsimony, an important alternative approach is to consider single-system models with fewer free parameters and attempt to push them as far as they will go. Beyond their potential to provide a conceptually simpler account of classification, a concerted effort at applying such models is highly instructive even if it fails. In particular, the failures of such models may provide firmer grounds on which to base the argument that the more complex multiple-system approaches are needed.

An excellent candidate for a single-system approach to modeling categorization is the class of models known as exemplar models (e.g., Brooks, 1978; Hintzman, 1986; Kruschke, 1992; Medin & Schaffer, 1978; Nosofsky, 1986). According to exemplar models, people represent categories by storing individual exemplars in memory, and classify objects on the basis of their similarity to these stored exemplars. Beyond accounting for diverse categorization phenomena, exemplar models have been successful at explaining relations between categorization and other fundamental cognitive processes, including individual object identification, old–new recognition memory, problem solving, and the development of automaticity in tasks of skilled performance (e.g., Estes, 1986, 1994; Hintzman, 1986, 1988; Logan, 1988; Nosofsky, 1986, 1987, 1988, 1991a; Nosofsky & Palmeri, 1997; Palmeri, 1997; Ross, 1987).

Furthermore, exemplar models have shown success at accounting for phenomena that previous investigators have cited as evidence in favor of prototype abstraction or rule induction. For example, a well-known phenomenon in the categorization literature involves prototype enhancement effects, in which category prototypes that are not experienced during training are nevertheless classified as well as or sometimes better than the old training exemplars (Homa, 1984). Although such results were often taken as evidence of prototype abstraction processes, exemplar theorists have demonstrated repeatedly that prototype enhancement effects are well predicted by pure exemplar models (e.g., Busemeyer, Dewey, & Medin, 1984; Hintzman, 1986; Medin & Schaffer, 1978; Nosofsky, 1988, 1991a; Shin & Nosofsky, 1992). The general idea is that, although not presented during training, the category prototypes are often highly similar to numerous training instances from their own category and tend to be quite dissimilar from the training instances of alternative

This work was supported by Grant R01 MH48494-09 from the National Institute of Mental Health. We thank F. Gregory Ashby, John Wixted, and two anonymous reviewers for their helpful criticisms of earlier versions of this article. We are also indebted to Michael Erickson for providing us with the original program code used to conduct Experiment 1 from Erickson and Kruschke (1998), and to J. David Smith for providing us with the individual subject data from J. D. Smith, Murray, and Minda (1997). Correspondence should be addressed to R. M. Nosofsky, Department of Psychology, Indiana University, Bloomington, IN 47405 (e-mail: nosofsky@indiana.edu).

 Copyright 2000 Psychonomic Society, Inc.

categories. By contrast, any given training exemplar may be highly similar only to itself. The redundancy afforded the prototype often gives it an advantage in tests of classification transfer.

Likewise, Nosofsky (1984, 1986, 1991b; Nosofsky, Gluck, Palmeri, McKinley, & Glauthier, 1994) demonstrated that an exemplar model provided excellent quantitative accounts of a variety of phenomena that seemed interpretable in terms of rule abstraction. For example, the exemplar model provided excellent accounts of Shepard, Hovland, and Jenkins's (1961) classic study of the difficulty of learning classification problems defined by rules of differing complexity (Nosofsky, 1984; Nosofsky, Gluck, et al., 1994). Furthermore, it accounted for typicality judgments and response–speed effects in a series of classification problems described by unidimensional and conjunctive rules (Nosofsky, 1991b). Finally, it provided excellent quantitative accounts of classification choice probabilities during transfer in a wide variety of categorization conditions involving continuous-dimension logical rules (e.g., Nosofsky, 1986, 1987; Nosofsky, Clark, & Shin, 1989).

The purpose of the present article is to consider some of the more modern results in the classification literature that have led investigators away from the single-system, exemplar-model approach in favor of the more complex multiple-system models. In each section of our article, we first review the basic empirical result being considered and explain why the original investigators interpreted the result in terms of rule use, prototype formation, or the operation of multiple systems. We then develop an alternative exemplar-based account of the same phenomenon.

Before proceeding with our investigation, we offer here some important caveats. First, we do not claim that the exemplar model provides a *superior* account of the phenomena than do the multiple-system models. Moreover, we do not claim that other alternative accounts of these phenomena are unavailable. Rather, each of the reported phenomena is a major source of evidence that investigators have used to *challenge* the exemplar model. We argue instead that reasonable applications of a single-system exemplar model may be sufficient to account for these phenomena, thereby bringing into question the need to posit the more complex multiple-system accounts.

Second, we limit consideration to "free-strategy" situations involving perceptual classification in which novel categories are learned via induction over individual training exemplars. We do not doubt, for example, that people can in large measure apply specific rules if provided with explicit instructions to do so. Rather, the key question is whether a single exemplar-based system tends to subserve performance on typical perceptual categorization tasks in which the strategy is left to the option of the observer. The extent to which exemplar models may be applicable to more general situations, such as those involving explicit instructions for alternative strategy use, prior knowledge, abstract/conceptual forms of categorization, or to other cognitive activities related to categorization such as inference and feature prediction, remains an open question. We briefly consider some of these more complex issues in the General Discussion.

OVERVIEW OF THE FORMAL MODEL

The exemplar model that guides the present research effort is the *generalized context model* (GCM; Ashby & Maddox, 1993; Nosofsky, 1984, 1986, 1991a). The GCM generalizes the original version of the context model proposed by Medin and Schaffer (1978) and integrates it with classic theories and ideas proposed in the areas of choice and similarity (Carroll & Wish, 1974; Garner, 1974; Luce, 1963; Shepard, 1958; Shepard & Chang, 1963).

The GCM uses a multidimensional scaling (MDS) approach to modeling similarity. According to the model, exemplars are represented as points in a multidimensional psychological space, and similarity between exemplars is a decreasing function of their distance in the space. Selective attention processes systematically modify the structure of the space in which the exemplars are embedded. An important working hypothesis is that, with experience in a given task, observers often learn to distribute their attention over psychological dimensions in a manner that tends to optimize performance.

These ideas are illustrated schematically in Figure 1. In the top panel, there are two categories, X and O, defined by five exemplars each. The exemplars reside in a two-dimensional space. Exemplars X2 and O4 are close together in the space, and so are highly similar to one another, whereas Exemplars X5 and O2 are far away, and so are dissimilar. Suppose that an observer needs to classify item *i* (illustrated in the space). According to the model, the observer sums the similarity of item *i* to all the X exemplars and to all the O exemplars. The classification decision is based on the relative magnitude of these sums.

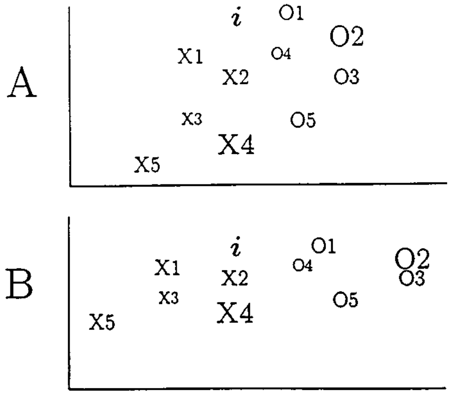

Figure 1. Schematic category structure for illustrating the workings of the generalized context model. X, exemplars of Category X; O, exemplars of Category O. Panel A: Category structure with equal attention to both dimensions. Panel B: Category structure with selective attention to the horizontal dimension.

Specifically, in an experiment involving multiple categories and in which there are no response biases, the probability that item i is classified into Category J is given by

$$P(J \mid i) = \frac{\left(\sum_{j \in J} s_{ij}\right)^{\gamma}}{\left[\sum_{K}\left(\sum_{k \in K} s_{ik}\right)^{\gamma}\right]}, \tag{1}$$

where s_{ij} denotes the similarity of item i to exemplar j and the index $j \in J$ denotes that the sum is over all exemplars j belonging to Category J. The parameter γ, first introduced into the GCM response rule by Ashby and Maddox (1993), is a response-scaling parameter. When $\gamma = 1$, the observer responds by "probability matching" to the relative summed similarities, whereas when γ grows greater than one, the observer responds more deterministically with the category that yields the largest summed similarity (for more detailed discussion and closely related models, see Ashby & Maddox, 1993; Maddox & Ashby, 1993; McKinley & Nosofsky, 1995; Nosofsky, 1991a). A direct process interpretation for the emergence of the summed similarity computation and the γ response-scaling parameter was recently developed by Nosofsky and Palmeri (1997) in terms of their exemplar-based random walk model of categorization.

A critical component assumption in the GCM is that similarity between exemplars is not an invariant relation but a highly context-dependent one. In particular, it is assumed that selective attention processes modify psychological similarity relations among exemplars, usually in a manner that tends to optimize performance for any given task. This selective attention process can have a dramatic influence on the classification predictions that are made by the model.

For example, in the top panel of Figure 1, item i is roughly equally similar to the exemplars of Category X and Category O, and so would be classified in each category with roughly equal probability. Note, however, that the horizontal dimension is far more relevant than is the vertical dimension for discriminating the categories. An experienced observer would presumably learn this aspect of the category structure and would attend selectively to this relevant dimension. This process is represented in the model in terms of a set of selective attention weights that "stretch" the space along attended relevant dimensions and "shrink" it along unattended irrelevant ones, as illustrated in the bottom panel of Figure 1. Note from the illustration that this selective attention process would tend to *optimize* performance because it would separate further in psychological space the two categories that need to be discriminated. In addition, note that whereas item i was roughly equally similar to the X and O exemplars when equal attention was given to the dimensions (top panel), it takes on far greater similarity to the exemplars of Category X following selective attention to the relevant dimension (bottom panel).

These ideas are represented formally in the model as follows. Assume that the exemplars reside in an M-dimensional psychological space, and let x_{im} denote the value of exemplar i on psychological dimension m. (These psychological coordinate values for the exemplars are often derived by conducting a variety of similarity-scaling studies in which MDS solutions for the exemplars are derived; for a review, see Nosofsky, 1992.) The distance between exemplars i and j is computed by using the weighted Minkowski power-model formula,

$$d_{ij} = \left[\sum_{m} w_m \cdot \left| x_{im} - x_{jm} \right|^{r}\right]^{1/r}. \tag{2}$$

In this equation, the value r defines the distance metric of the space. Common values are $r = 1$, which defines a city block distance metric; and $r = 2$, which defines a Euclidean distance metric (see, e.g., Garner, 1974; Shepard, 1964). The city block metric is typically assumed when modeling distances among highly separable dimension stimuli, whereas the Euclidean metric is used to model distances among integral-dimension stimuli (Shepard, 1987). The parameters w_m are the "attention weights" (Carroll & Wish, 1974). The w_m parameters model the degree of attention that an observer gives to each dimension in making psychological distance judgments among exemplars. As illustrated previously, a geometric interpretation for the attention weights is that of stretching and shrinking the psychological space along its dimensions.

On the basis of a large body of research from the field of stimulus generalization (see Shepard, 1987, for a review), the similarity between exemplars i and j (s_{ij}) is assumed to be a nonlinearly decreasing function of their distance (d_{ij}),

$$s_{ij} = \exp\left(-c \cdot d_{ij}^{p}\right), \tag{3}$$

where c is an overall scaling or sensitivity parameter and the value p defines the precise form of the similarity gradient. Common values of the generalization gradient are $p = 1$, which defines an exponential similarity gradient (Shepard, 1958, 1987), and $p = 2$, which defines a Gaussian similarity gradient (Nosofsky, 1985, 1986). The exponential model is favored in situations in which observers learn to classify fairly discriminable, nonconfusable stimuli (Shepard, 1986, 1987), and this model is used exclusively in the remainder of this article.

The sensitivity parameter c in Equation 3 determines the rate at which similarity declines with distance. Its role is illustrated in Figure 2. The top panel shows an example in which the value of c is relatively high. In this case, the generalization gradient relating similarity to distance is steep, so that exemplars that are even a moderate distance apart in the space are judged as very dissimilar. By contrast, as shown in the bottom panel, when c is small, the generalization gradient is shallow, and exemplars that are separated by large distances in the space may still be judged as similar.

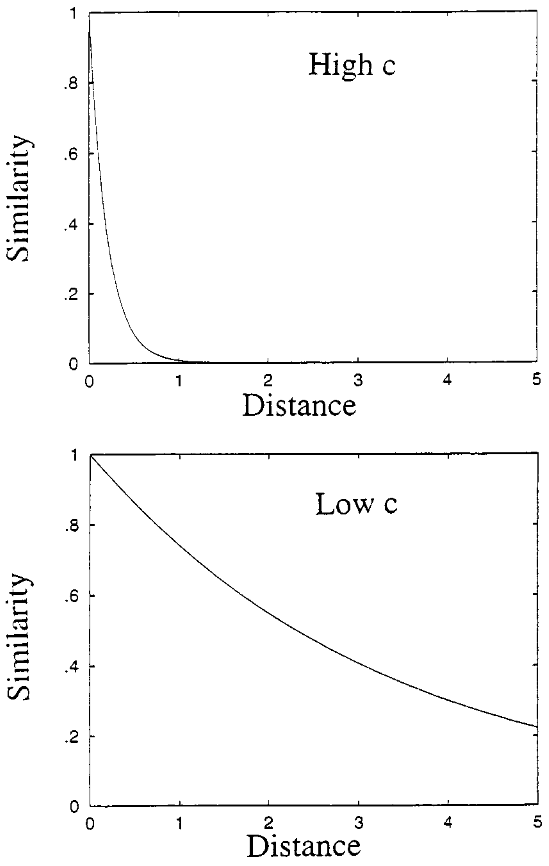

Figure 2. Illustration of exponential gradients relating similarity to distance. Top panel: high value of sensitivity. Bottom panel: low value of sensitivity.

Finally, the s_{ij} values computed from Equations 2 and 3 are substituted into Equation 1 to yield the classification predictions made by the GCM.

In summary, the GCM is defined by the system of Equations 1–3. Its free parameters include the response-scaling parameter γ, the sensitivity parameter c, and the attention weights w_m.

ACCOUNTS OF THE PHENOMENA

With this formal background in mind, we now turn to the recently reported phenomena that have been interpreted as challenging the exemplar model and that have promoted the multiple-system categorization approaches. In many of the cases that we consider, the standard version of the GCM that we have just presented is shown to account for the phenomena with no modifications. In other cases, we introduce extensions to the model that are, in our view, sensible to make in light of the experimental procedures that were used or the goals of the analysis. For example, in some sections we consider the extent to which the exemplar model can account for individual-subject differences in performance. In these sections, we make allowance for the idea that the exemplar-model parameter values may vary across the individual subjects. Other extensions will involve deeper proposed revisions of the standard model, but the essential spirit of the single-system exemplar model is left intact. Although such extensions are admittedly post hoc, we will argue that they are conceptually well grounded. That is, there is a strong sense in which the extensions *should* have been proposed in advance of seeing the data. For example, we introduce the idea that the same principles of parameter optimization that have always been hypothesized to influence the attention weights in the exemplar model may influence the setting of the sensitivity parameter as well. It is important to acknowledge at the outset, however, that the effectiveness of our case will rest on the degree to which our proposed extensions are viewed as reasonable and compelling.

A Bias Toward Verbal Rules

In experiments conducted by Ashby et al. (1998), subjects learned to classify stimuli into two bivariate normally distributed categories. The category structures are illustrated in Figure 3, where "+" and "o" indicate stimuli that belong to Categories A and B, respectively. A key aspect of the design was that the categories were displaced more in the vertical direction than in the horizontal direction, as is illustrated schematically in Figure 4A. Ashby et al. used this design to test a prediction of their newly proposed COVIS (competition between a verbal and an implicit system) model of categorization. According to COVIS, two mental systems compete with each other in making categorization responses. First, there is a nonverbal, implicit system that learns optimal decision boundaries for separating a space into category regions. The optimal decision boundary for the present design is the solid linear decision boundary with a .60 slope, illustrated in Figure 3. To maximize accuracy, an ideal observer would classify all objects falling to the upper left of the boundary into Category A and all objects falling to the lower right of the boundary into Category B. Second, there is an explicit system that learns verbal rules. According to COVIS, this system would develop the boundary represented by the dashed horizontal line in Figure 3. In using this boundary, the subject sets a criterion along the orientation dimension in such a way that any object with orientation greater than this criterion is classified into Category A, whereas any object with orientation less than the criterion is classified into Category B. Finally, according to COVIS, on some proportion of trials the implicit system chooses the response, whereas on the remaining proportion of trials the explicit system chooses the response.

Ashby et al. (1998) demonstrated that if these ideas are correct, then an individual subject's *observable* performance should be well described by a linear decision boundary with a slope that is *shallower* than the optimal bound slope of .60. Intuitively, the slope of the best-fitting linear boundary should be "pulled" in the direction of the horizontal line verbal rule.

Ashby et al. (1998) tested 5 individual subjects in this experimental design and fitted a linear decision-boundary

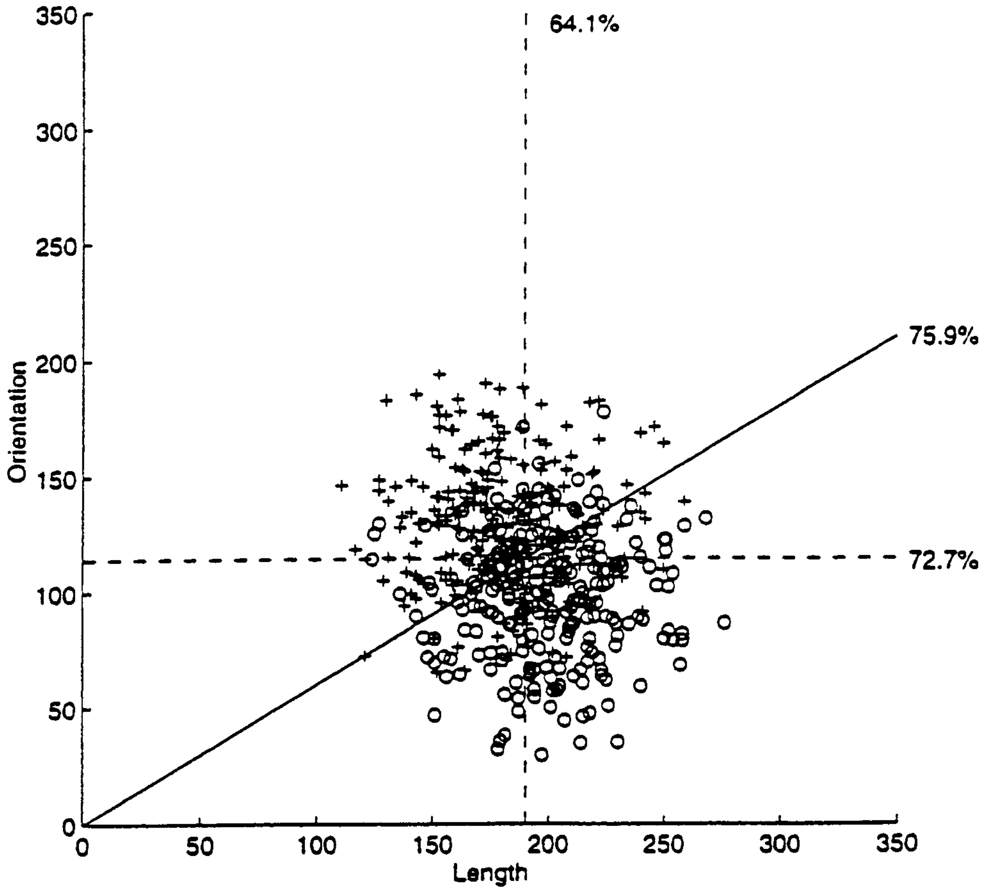

Figure 3. Category structure tested by Ashby, Alfonso-Reese, Turken, and Waldron (1998). +, exemplar from Category A; ○, exemplar from Category B. The optimal population decision boundary is represented by the solid line with a .60 slope. The vertical and horizontal dashed lines are the most accurate single-dimension rule boundaries available on each dimension. The numerical values are the accuracies yielded by each decision boundary. Note—From "A Neuropsychological Theory of Multiple Systems in Category Learning," by F. G. Ashby, L. A. Alfonso-Reese, A. U. Turken, & E. M. Waldron, 1998, *Psychological Review, 105,* p. 467. Copyright 1998 by the American Psychological Association. Reprinted with permission.

model to each of the individual subject's classification responses. In all cases, the best-fitting linear boundary did indeed have a slope that was significantly less than .60, as predicted by COVIS. Beyond concluding that the results were consistent with the COVIS predictions, however, Ashby et al. argued that the results challenged the predictions of the exemplar model. They wrote:

> As a consequence, COVIS predicts that even experienced participants will adopt a bound with a slope of less than 0.6 [i.e., the slope of the optimal bound]. To our knowledge, COVIS is the only model of category learning that predicts (a priori) such an asymptotic bias in categorization performance. For example, consider Nosofsky's (1986) generalized context model, which, arguably, has been the most successful of the formal exemplar models . . . the generalized context model predicts a priori that there will be no systematic bias in this experiment. (pp. 467–468)

Contrary to Ashby et al.'s (1998) claim, we suggest here that the GCM *does* predict a priori this bias in the slope of the best-fitting bound. As described in the Overview of the Formal Model section, one of the central assumptions of the GCM is that people selectively attend to psychological dimensions, and that, with learning, greater attention will be given to relevant dimensions than to irrelevant ones. A good working hypothesis, which has often gained support in past empirical investigations, is that observers learn to distribute attention in a manner that tends to optimize performance. Now, as illustrated in the Figure 4A design, the vertical dimension is more relevant than is the horizontal dimension for purposes of classification, so observers would be expected to give greater attention to the vertical dimension. Thus, the space would be "stretched" along the vertical dimension and "shrunk" along the horizontal dimension, as illustrated in Figure 4B. As can be seen in the figure, if this selective attention strategy operated, the boundary of equal similarity between the exemplars of the two categories would take on an increasingly more shallow slope.

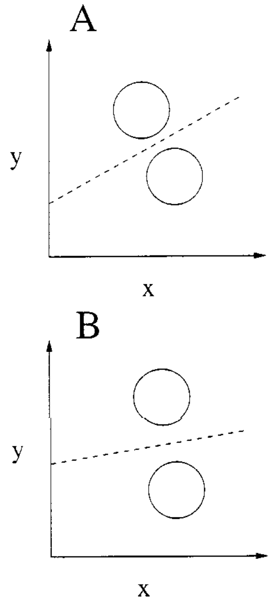

Figure 4. Schematic illustration of the category structure depicted in Figure 3. Panel A: the category structure with equal attention to both dimensions. Panel B: the structure with selective attention to the vertical dimension.

To verify these intuitions, we conducted the following series of computer simulations. First, we generated two bivariate normally distributed categories with 500 exemplars each that satisfied the constraints of Ashby et al.'s (1998) experimental design (for the parameters of these distributions, see Ashby et al., 1998, p. 469). Next, for each of a set of combinations of values of the sensitivity parameter c and the response-scaling parameter γ, we conducted a search for the value of the attention-weight parameter that would optimize performance (i.e., maximize percentage correct) in Ashby et al.'s experimental design. The values of c and γ that were chosen in these investigations are in the range of values that have produced good fits of the GCM to previous data sets obtained in this type of paradigm (McKinley & Nosofsky, 1995, 1996). Furthermore, because the stimuli in these experiments were fairly discriminable and composed of separable dimensions, we assumed a city block metric ($r = 1$ in Equation 2) and an exponential similarity function ($p = 1$ in Equation 3) in applying the GCM in these investigations (Nosofsky, 1998b; Nosofsky et al., 1989; Shepard, 1987).

The results are summarized in Table 1, where w_1 indicates the value of the Dimension 1 attention weight that would optimize performance. In all cases, the value of this Dimension 1 attention weight is less than .50, which indicates, as expected, that it is optimal to give *less* attention to the horizontal dimension than to the vertical dimension. Next, we used the GCM to predict the classification probabilities associated with each stimulus in Ashby et al.'s (1998) design, assuming that a subject used these optimal values of the attention weights. Finally, we conducted modeling analyses in which the linear decision-boundary model was fitted to these GCM-predicted classification probabilities. Representative results are summarized in Table 2. The key point is that whenever the Dimension 1 attention weight is less than .50, the best-fitting linear decision boundary has a slope that is significantly less than .60, in accord with the intuitions developed above. Only when the attention weight is equal to .50 does the slope of the best-fitting linear decision boundary come close to the .60 value.

In summary, assuming that subjects learn to give greater attention to the relevant vertical dimension, which, according to the GCM, is the optimal strategy in this experimental design, the GCM does indeed predict the phenomenon that Ashby et al. (1998) described as a "bias toward the verbal rule" (p. 467).[1] Thus, one of the major phenomena that Ashby et al. used to challenge the exemplar model and to suggest the need for the more complex multiple-system categorization model is apparently not diagnostic. Furthermore, previously reported work (Maddox & Ashby, 1993; McKinley & Nosofsky, 1995) indicates that the GCM (with the γ response-scaling parameter) and the decision-boundary models give virtually identical quantitative fits to the data obtained in this type of classification paradigm when the models' parameters are allowed to vary freely. Thus, this paradigm fails to distinguish the models on quantitative grounds as well.

It might be noted that even a simple prototype model that assumes optimal weighting of dimensions will also predict the bias-toward-verbal-rule phenomenon in this particular paradigm. Again, we remind readers that we are not claiming that the exemplar model provides the only viable account of the present phenomena. Rather, we are defending the exemplar model as also providing a viable account in those situations in which it has been challenged. Results from other closely related paradigms with more diagnostic category structures have strongly favored the predictions of the exemplar model over those of prototype models, however (e.g., Maddox & Ashby, 1993; McKinley & Nosofsky, 1995).

There is a potential limitation associated with our exemplar-based account of Ashby et al.'s (1998) bias-toward-verbal-rule phenomenon. The concern is that, for Ashby et al.'s paradigm, the GCM often predicts only small performance benefits when the optimal attention weights are used as opposed to when equal weighting of the dimensions occurs. For example, in the case in which $c = .05$ and $\gamma = 3$, the GCM predicts .683 correct per-

Table 1
Optimal Dimension 1 Attention Weight (w_1) for the GCM as a Function of c and γ in Ashby et al.'s (1998) Experimental Design

	$\gamma = 1$	$\gamma = 3$	$\gamma = 5$
$c = .02$	.00	.00	.00
$c = .05$	.00	.15	.33
$c = .10$	.20	.38	.45

Note—GCM, generalized context model.

Table 2
Linear Decision Boundaries Fitted to GCM-Predicted Data in Ashby et al.'s (1998) Experimental Design

GCM Parameters	Best-Fitting Linear Boundary
$c = .05, \gamma = 3, w_1 = .15$	$Y = .13X + 89$
$c = .05, \gamma = 5, w_1 = .33$	$Y = .27X + 65$
$c = .05, \gamma = 3, w_1 = .50$	$Y = .56X + 9$

Note—GCM, generalized context model.

formance when the dimensions are weighted equally, whereas it predicts .697 correct performance when the dimensions are weighted optimally. Perhaps this difference is too small to lead observers systematically away from an equal attention weighting toward a more nearly optimal weighting. To explore this issue, we conducted additional simulations using a closely related exemplar-based model known as ALCOVE (Kruschke, 1992; Nosofsky, Kruschke, & McKinley, 1992). This model basically incorporates key principles of the GCM within a connectionist framework. An important advantage of ALCOVE, however, is that it provides an explicit mechanism that learns the attention weights on a trial-by-trial basis. Because the mechanism is error driven, it tends to learn those weights that optimize performance in a given task (e.g., Nosofsky, Gluck, et al., 1994). Although a detailed discussion goes beyond the scope of this article, our simulations with ALCOVE generally corroborated our attention-optimization arguments involving the GCM. In particular, when applied to Ashby et al.'s paradigm, the ALCOVE model learned to place greater attention weight on the more relevant vertical dimension than on the less relevant horizontal dimension, and therefore it too predicted that the best-fitting linear decision boundary would be pulled in the direction of the verbal rule. This result is important because it shows that, even though the predicted performance differences are sometimes small, they are still sufficient to drive an explicit learning mechanism in the direction of the optimal attention weights.

Predicting Distributions of Individual-Subject Generalization Patterns

A second study that suggested a major limitation of exemplar models was one reported by Nosofsky, Palmeri, and McKinley (1994), who advanced an alternative rule-plus-exception (RULEX) model of classification. According to RULEX, people learn to classify objects by forming simple logical rules along single dimensions and then by storing occasional exceptions to those rules. To illustrate the workings of the model, consider the category structure shown in Table 3, which is the classic structure used by Medin and Schaffer (1978) in their seminal studies of category learning. The stimuli vary along four binary-valued dimensions. There are five Category A training exemplars, four Category B training exemplars, and seven transfer stimuli. Logical Value 1 on each dimension tends to indicate Category A, and Logical Value 2 tends to indicate Category B, but there are no singly necessary and jointly sufficient sets of features that define the categories. According to RULEX, by the time the learning process is completed, an individual observer might have stored the following information in memory. First, the observer might store the (imperfect) single-dimension rule that objects with Value 1 on Dimension 1 belong to Category A, and that objects with Value 2 on Dimension 1 belong to Category B (Table 3). We summarize these rules by using the notation 1*** → A, 2*** → B, where the asterisks denote dimension "wild cards" that match any value. Exemplars A5 and B1 are exceptions to this rule, so the observer must store additional information to learn the categories. For example, the observer might store the exceptions 2*11 → A, 1*22 → B (Table 3). Note that, with these rules, the categorization problem is solved, even though no complete exemplars are stored in memory. The learning process in RULEX is stochastic, and a key property of the model is that different observers form alternative rules and exceptions. For example, numerous observers might instead form rules along Dimension 3, **1* → A, **2* → B, and store information to classify the A4 and B2 exceptions, for example, 1*21 → A, 2*12 → B. Averaged classification data are assumed to represent probabilistic mixtures of these idiosyncratic rules and exceptions. An explicit learning process is formalized in the RULEX simulation that incorporates classic principles of hypothesis testing (e.g., Levine, 1975; Trabasso & Bower, 1968) and probabilistic storage of exception information. Thus, although a vast array of different rules and exceptions are involved in predicting the averaged classification data, these rules and exceptions emerge from a probabilistic learning process described by relatively few free parameters.

Nosofsky, Palmeri, and McKinley (1994) demonstrated that RULEX provided excellent quantitative fits to averaged classification data, fits that were essentially the same as those achieved by the exemplar-based GCM. Beyond predicting averaged classification data, however, RULEX also fared well at predicting patterns of performance at the individual observer level. A highly challenging form of data is what Nosofsky, Palmeri, and McKinley referred to as a *distribution of generalizations* (see also Nosofsky et al., 1989; Pavel, Gluck, & Henkle, 1988). Consider the transfer stimuli in Table 3. During test, each transfer stimulus is classified by an individual observer into either Category A or Category B. The specific pattern of clas-

Table 3
Category Structure Tested in Medin and Schaffer's (1978) Experiments

Category A	Category B	Transfer Stimuli
A1 1112	B1 1122	T1 1221
A2 1212	B2 2112	T2 1222
A3 1211	B3 2221	T3 1111
A4 1121	B4 2222	T4 2212
A5 2111		T5 2121
		T6 2211
		T7 2122

Note—A, training exemplar of Category A; B, training exemplar of Category B; T, transfer stimulus.

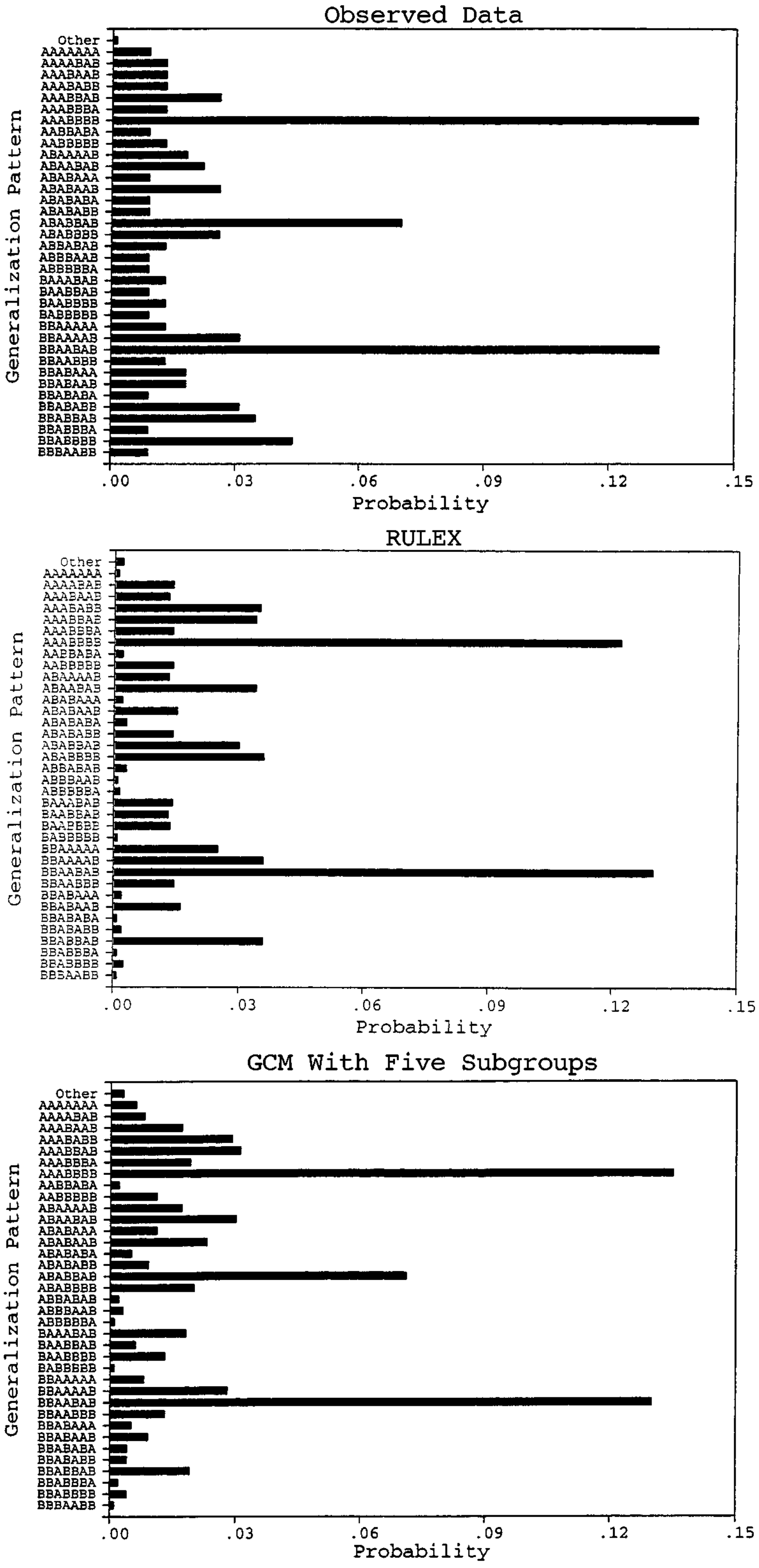

Figure 5. Distribution of generalizations from Nosofsky, Palmeri, and McKinley's (1994) categorization experiment. Top panel: the observed distribution. Middle panel: the distribution predicted by RULEX. Bottom panel: the distribution predicted by the five-subgroup version of the generalized context model.

sification responses given to the transfer stimuli defines a *generalization profile* for an individual observer. For example, an observer classifying T1–T3 into Category A, and T4–T7 into Category B, yields the generalization profile AAABBBB. (Note that this particular profile would be produced, for example, by observers using the RULEX strategy 1*** → A, 2*** → B, with the complete A5 and B1 exemplars learned as exceptions.) The distribution of generalizations is then obtained by computing the frequency of individuals displaying each profile. The top panel of Figure 5 shows the distribution of generalizations observed in Nosofsky, Palmeri, and McKinley's replication of Medin and Schaffer's (1978) experiment. The middle panel shows the distribution predicted by RULEX. As can be seen, RULEX performed reasonably well (although it underestimated the frequency of Profile ABABBAB). Nosofsky, Palmeri, and McKinley deemed this achievement important, because in addition to accounting for the averaged transfer data, RULEX simultaneously characterized the patterns of performance observed at the individual observer level. By contrast, the exemplar-based GCM failed dramatically to predict the distribution-of-generalization data (see Nosofsky, Palmeri, & McKinley, 1994, Figure 11). These model-fitting results were the major lines of evidence that led Nosofsky, Palmeri, and McKinley to argue in favor of RULEX over the exemplar model in this experimental paradigm.

However, we argue here that there was an important lack of comparability between the models that makes it difficult to draw strong conclusions. The RULEX model incorporates a stochastic learning process that leads to heterogeneous individual subject behavior. By the time learning is complete, for example, some subjects may have formed rules along Dimension 1 and stored exceptions to the Dimension 1 rule, whereas other subjects may have formed rules along Dimension 3 and stored exceptions appropriate to this alternative rule. By contrast, the standard version of the GCM that was fitted to the data assumed homogeneous individual-subject behavior: All subjects were assumed to store all exemplars in memory, and similarity comparisons to the exemplars were described by an identical set of free parameters across subjects. Thus, Nosofsky, Palmeri, and McKinley's (1994) model-fitting analyses confounded the issue of whether subjects classify by using rules or exemplars with the issue of whether individual subjects behave identically. Clearly, an important goal is to test versions of the exemplar model that allow for heterogeneity in the behavior of individual subjects.[2]

As a first step toward achieving this goal, we report here a model-fitting analysis involving an extended version of the GCM that includes the idea that there are distinct subgroups of subjects who employ different attention-weight configurations to learn the Table 3 category structure. Our main idea was that observers are likely to start their learning of the category structure by selectively attending to a single dimension and then gradually spreading attention to other dimensions composing the exemplars. We modeled the behavior of Subgroups 1–4 as reflecting such an attentional distribution. Thus, Subgroup 1 was assumed to place attention weight w on Dimension 1 and to divide evenly the remaining attention $(1 - w)$ among Dimensions 2–4. Likewise, Subgroup 2 placed attention weight w on Dimension 2 and divided evenly the remaining attention among Dimensions 1, 3, and 4, and so forth for Subgroups 3 and 4. Finally, we supposed that there was a fifth subgroup of subjects who had learned to divide attention optimally among the four dimensions. (The optimal attention distribution for learning the Table 3 category structure was reported in a previous study by Nosofsky, 1984, p. 113, Figure 4. It places the majority of attention on Dimensions 1 and 3, which are the most diagnostic dimensions; somewhat less attention on Dimension 4; and nearly zero attention on Dimension 2, which is the least diagnostic dimension.) Presumably, with extended learning, numerous subjects might eventually move toward such an optimal attention distribution. In Nosofsky, Palmeri, and McKinley's (1994) study, however, learning took place for a total of only 144 trials, so it seemed reasonable to posit that most subjects were still attending primarily to just a single dimension at the time of classification transfer.

We tested this five-subgroup version of the GCM by fitting it to the data obtained in Nosofsky, Palmeri, and McKinley's (1994) experiment. In this experiment, 227 subjects learned the Table 3 category structure. During the training phase of the experiment, there were 16 blocks of nine trials each, with each training stimulus presented once per block in a unique random order for each subject. Following the training phase, a transfer phase was con-

Table 4
Average Probability With Which Each Stimulus Was Classified into Category A During the Transfer Phase of Nosofsky, Palmeri, and McKinley's (1994) Experiment

Stim.	Obs.	GCM
	Category A	
A1 1112	.770	.779
A2 1212	.780	.854
A3 1211	.830	.882
A4 1121	.640	.588
A5 2111	.610	.590
	Category B	
B1 1122	.390	.462
B2 2112	.410	.465
B3 2221	.210	.186
B4 2222	.150	.119
	Transfer	
T1 1221	.560	.548
T2 1222	.410	.463
T3 1111	.820	.805
T4 2212	.400	.471
T5 2121	.320	.335
T6 2211	.530	.555
T7 2122	.200	.238

Note—Stim., stimulus; Obs., observed probability; GCM, generalized context model predicted probability.

Table 5
Probability of Each Generalization Profile in Nosofsky, Palmeri, and McKinley's (1994) Experiment

Profile	Obs.	GCM
AAAAAAA	.009	.001
AAAABAB	.013	.020
AAABAAB	.013	.017
AAABABB	.013	.035
AAABBAB	.026	.044
AAABBBA	.013	.013
AAABBBB	.141	.130
AABBABA	.009	.001
AABBBBB	.013	.007
ABAAAAB	.018	.018
ABAABAB	.022	.046
ABABAAA	.009	.002
ABABAAB	.026	.017
ABABABA	.009	.002
ABABABB	.009	.011
ABABBAB	.070	.043
ABABBBB	.026	.032
ABBABAB	.013	.003
ABBBAAB	.009	.001
ABBBBBA	.009	.001
BAAABAB	.013	.020
BAABBAB	.009	.017
BAABBBB	.013	.018
BABBBBB	.009	.002
BBAAAAA	.013	.009
BBAAAAB	.031	.038
BBAABAB	.132	.130
BBAABBB	.013	.019
BBABAAA	.018	.002
BBABAAB	.018	.011
BBABABA	.009	.001
BBABABB	.031	.004
BBABBAB	.035	.034
BBABBBA	.009	.002
BBABBBB	.044	.017
BBBAABB	.009	.001
Other	.001	.002

Note—Obs., observed probability; GCM, generalized context model predicted probability; Other, average probability of all other generalization profiles.

ducted in which all 16 stimuli (including the transfer stimuli) were presented three times each, and subjects judged whether each stimulus belonged to Category A or B. The average probability with which each transfer stimulus was classified in Category A is reported in Table 4, whereas the distribution-of-generalization data are reported in Table 5. Recall that the latter data set gives the probability with which each generalization profile was observed across the 227 subjects.

We fitted the five-subgroup version of the GCM simultaneously to the Table 4 and 5 data. As a criterion of fit, we searched for the free parameters that maximized the summed percentage of variance accounted for in each data set. The free parameters were the overall sensitivity parameter c (Equation 3) and the response-scaling parameter γ (Equation 1), the values of which were held fixed across the five subgroups; a common attention weight w (Equation 2) given to the "primary" dimension in each of Subgroups 1–4; and the proportions of subjects p_1 through p_5, making up Subgroups 1–5, respectively. Because the five subgroup proportions are constrained to sum to one, this model makes use of seven free parameters.

The predicted data are shown along with the observed data in Tables 4 and 5, with the predicted distribution-of-generalization results also illustrated in Figure 5 (bottom panel). (The best-fitting free parameters were $c = 8.648$, $\gamma = 1.702$, $w = .949$, $p_1 = .31$, $p_2 = .21$, $p_3 = .30$, $p_4 = .11$, and $p_5 = .08$.) Whereas the standard exemplar model, which did not allow for individual-subject parameter variability, had failed dramatically to account for the distribution-of-generalization data (see Nosofsky, Palmeri, & McKinley, 1994, Figure 11), the present application is quite successful: The model accounts for 95.4% of the variance in the averaged classification transfer data (Table 4), and, more importantly, for 88.0% of the variance in the distribution-of-generalization data (Table 5 and Figure 5).

Interestingly, according to the parameter estimates reported above, it was most common for observers to attend selectively to Dimension 1 or 3 when learning the Table 3 category structure. This result is sensible because these two dimensions are the most diagnostic ones for classifying objects into the two categories (Table 3). We also remark that it is the inclusion of the fifth subgroup in our modeling analysis, namely the one that is assumed to adopt an optimal distribution of attention, that allows the exemplar model to predict fairly well the central peak in the distribution of generalizations (i.e., Profile ABABBAB). If this subgroup is removed (i.e., if the parameter p_5 is held fixed at zero), then the model is still able to achieve fairly good fits (94.5% of the variance in the averaged transfer data, 78.1% of the variance in the distribution-of-generalization data); its main failing is that, like RULEX, it then underestimates the probability of Profile ABABBAB.

To gain converging evidence regarding the plausibility of an exemplar-based account of the distribution-of-generalization data, we also applied Kruschke's (1992) ALCOVE model to this experiment. Whereas our analyses involving the GCM relied on our positing particular patterns of attention weights across five hypothetical subgroups, in our ALCOVE simulations the variability in performance was produced solely by virtue of the model's attention-weight learning mechanism. Specifically, to introduce individual-subject heterogeneity, ALCOVE learned the Table 3 category structure by being trained on a unique random sequence of the stimuli for each of 1,000 individual-subject simulations. Because the attention-weight learning mechanism in the model is highly sensitive to the sequence of stimuli (Lewandowsky, 1995), the attention weights can vary dramatically across individual subjects, especially at early stages of learning. Using the same analytic techniques as already described for the GCM, a standard four-parameter version of ALCOVE (see Kruschke, 1992, for details) accounted for 95.9% of the variance in the averaged classification transfer data

(Table 4) and for 84.3% of the variance in the distribution-of-generalization data (Table 5). Furthermore, the learning mechanism tended to produce attention-weight configurations similar to those posited in our five-subgroup version of the GCM, with many individual subjects attending primarily to a single dimension (particularly Dimensions 1 and 3) and a smaller subgroup distributing attention in a more nearly optimal manner.

In summary, the key point made in this section is that the heterogeneity in the distribution-of-generalization data reported by Nosofsky, Palmeri, and McKinley (1994) is apparently not as diagnostic of rule use and multiple categorization systems as was originally argued. Exemplar models that make allowance for forms of individual-subject variability in attention weighting can account for these data as well.

On Evidence for a Hybrid Connectionist Model of Categorization

Another major source of evidence that has been used to argue in favor of the need for a multiple-system categorization model comes from work reported by Erickson and Kruschke (1998). These investigators developed a hybrid connectionist model known as ATRIUM that incorporates both rule- and exemplar-based representations. Specifically, ATRIUM is composed of a rule module that learns to establish single-dimensional decision boundaries, an exemplar module that learns associations between exemplars and categories, and a competitive gating mechanism that links the two modules together. The system basically learns which of the two modules, the rule module or the exemplar module, is best suited for classifying objects in specific regions of the psychological space. In general, the model learns to rely on the exemplar module when classifying objects that are exceptions to the category rule; otherwise it favors the rule module.

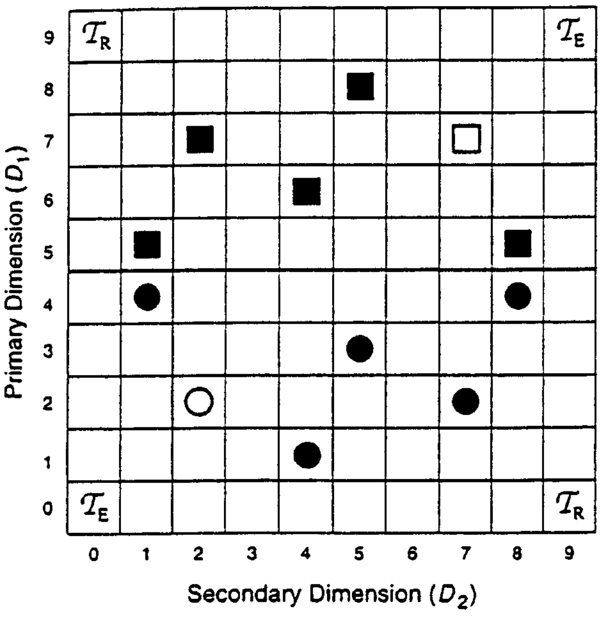

Figure 6. Schematic illustration of the category structure tested by Erickson and Kruschke (1998). Filled squares, training instances from Rule Category 1; filled circles, training instances from Rule Category 2; open circle, training instance for Exception Category 3; open square, training instance for Exception Category 4. T_E, critical transfer stimuli that are near to one of the exception categories; T_R, critical transfer stimuli that are near to the rule categories. Note—From "Rules and Exemplars in Category Learning," by M. A. Erickson and J. K. Kruschke, 1998, *Journal of Experimental Psychology: General, 127*, p. 110. Copyright 1998 by the American Psychological Association. Reprinted with permission.

A key source of evidence in favor of ATRIUM and that challenged a single-system exemplar model was obtained by Erickson and Kruschke (1998) in their Experiment 1. The experiment is illustrated schematically in Figure 6. As shown in the figure, the stimuli varied along two continuous dimensions, one referred to as a primary dimension and the other as secondary. The solid circles and solid squares represent training exemplars assigned to Categories 1 and 2, respectively. The stimulus represented by the open circle was assigned to Category 3, whereas the stimulus represented by the open square was assigned to Category 4. Note that a simple rule, namely the horizontal-line boundary formed along the primary dimension, exists to separate the members of Categories 1 and 2. The members of Categories 3 and 4, however, can be viewed as "exceptions" to this rule. In essence, ATRIUM learns the structure by having the rule module form the single-dimension boundary, while having the exemplar module learn associations for the exception items.

A key prediction from ATRIUM is that the exemplar module will contribute to classification judgments primarily for stimuli that are similar to the learned exceptions; otherwise the rule module will tend to be used. To test this prediction, Erickson and Kruschke (1998) presented a large number of transfer stimuli following the learning phase of their experiment. The transfer stimuli of major interest are those labeled T_R and T_E in Figure 6. T_R and T_E are equidistant from the rule boundary. However, T_E lies closer to one of the exception items than does T_R. Given the properties of this proximity structure, Erickson and Kruschke argued that exemplar models predict that T_E should be classified into the exception category with higher probability than T_R. They argued, however, that T_E was sufficiently far from the exception category so that in ATRIUM the exemplar module would contribute negligibly to performance on this item. Instead, the rule module would be used to classify both T_R and T_E, so both transfer items should be classified into the exception category with low and equal probability. This prediction was strongly supported by Erickson and Kruschke's data: T_R was classified into the nearest-exception category with probability .10 and T_E was classified into the nearest-exception category with probability .11.

Nevertheless, even the standard exemplar model can account for this single result if it is assumed that the observer gives virtually all attention to the primary dimension when making classification judgments. Therefore, to sharply distinguish between ATRIUM and the exemplar model, Erickson and Kruschke (1998) quantitatively fit the models to the complete set of classification data and compared the models on their indices of fit. They found that ATRIUM provided far better fits to the complete set of classification data than did the pure exemplar model.

We believe that a potential limitation of Erickson and Kruschke's (1998) design, however, involves the fact that in the presentations of the stimuli, numeric labels were included to indicate the precise magnitude along each perceptual dimension. This procedure was used to improve subjects' ability to learn this very difficult category structure. An unintended consequence of this procedure, however, is that the psychological representation for the objects may not have corresponded simply to the two-dimensional layout pictured in Figure 6. Instead, the numeric labels may also have been part of the psychological representation and may have altered the presumed similarity relations that hold among the objects.[3]

Indeed, we have formulated extended versions of the exemplar model that make explicit allowance for the idea that the labels themselves are stored in memory as part of the category representation, and these extended exemplar models yield fits to Erickson and Kruschke's (1998) classification transfer data that are nearly as good as those of ATRIUM. Unfortunately, however, these extended exemplar models rely on post hoc assumptions concerning the manner in which the numeric labels may have been represented. We decided that a more compelling approach to investigating this issue was simply to replicate Erickson and Kruschke's original experiment but to withhold the numeric labels when presenting the stimuli. Under these modified experimental conditions, our prediction was that there would be little difference in quantitative fit between the baseline version of the GCM and the more elaborate rule-plus-exemplar ATRIUM model.

Experiment

In this experiment, we replicated Erickson and Kruschke's (1998) Experiment 1, except that we withheld the presentation of the numeric labels in the stimulus displays. Because we used Erickson and Kruschke's original computer program and procedure for controlling the experiment (except for the modification of the program code to withhold the numeric labels), we refer the reader to the original article for details regarding the stimulus materials and procedures. These aspects of the study are simply summarized below.

Method

Subjects. The 156 subjects participated to receive experimental credit for an introductory psychology course at Indiana University.

Stimulus materials. Each stimulus was composed of a rectangle with 1 of 10 possible heights and an interior vertical line segment near the bottom of the rectangle with 1 of 10 possible positions from left to right. Assignment of the physical dimensions to the abstract dimensions of the category structure was counterbalanced across subjects.

Procedure. As was the case in Erickson and Kruschke's (1998) study, the general procedure was that in each block of 14 trials, each of the training stimuli was presented once in a random order except for the exception stimuli, which were presented twice. Each subject was trained on 29 blocks of 14 trials each for a total of 406 training trials. A test phase followed the training phase and consisted of the presentation of 50 transfer stimuli in random order (see Appendix A for details regarding the transfer stimuli that were used). Feedback was withheld during transfer. The only differences between our procedure and the one used by Erickson and Kruschke were as follows: (1) As a source of motivation for good performance, the subjects were instructed that the 3 subjects with the highest average accuracies for the latter part of training would receive a bonus of $25. (2) Subjects were given a maximum of 30 sec rather than 6 sec to respond on each trial. (3) We withheld a warning tone for incorrect responses and presented only visual corrective feedback.

Results

Following Erickson and Kruschke (1998), we were interested in modeling the results for only those observers who learned the category structure fairly accurately. Therefore, we included in the results only those subjects whose average accuracies in the last two blocks of training for the stimuli in the rule categories and exception categories were both at least 60% correct. Ninety-six out of 156 subjects met this criterion. (Erickson and Kruschke used similarly strict criteria involving both accuracy and reported strategy use, which led them to cull 125 out of 187 subjects in their modeling analyses. It is an issue of concern that learning the categories was very difficult for many subjects, but in the present case we were constrained to use the same category structure as in the original study.)

The complete set of transfer data from the subjects who met the learning criterion is reported in our Appendix A. These data consist of the probability with which each of 50 transfer stimuli was classified into Categories 1–4, respectively. An initial qualitative result of interest concerns the classification probabilities for Transfer Stimuli T_E and T_R. Recall that in Erickson and Kruschke's (1998) study, the probability with which these items were classified into the nearest-exception category was virtually identical. By contrast, in our follow-up experiment, T_E was classified into the nearest-exception category with probability .34, whereas T_R was classified into the nearest-exception category with probability only .04. This difference in classification probabilities was statistically significant [$t(95) = 6.12$, $p < .001$]. The finding that T_E was classified into the exception category with higher probability than T_R is consistent with the predictions of the exemplar model and provides initial evidence that the use of the labels in Erickson and Kruschke's original design may indeed have had an important influence on the patterns of generalization.

The key question concerns the ability of the alternative models to quantitatively fit the complete sets of transfer data. To begin, as a source of comparison, we fitted both ATRIUM and the GCM to the classification transfer data from Erickson and Kruschke's (1998) original experiment. We fitted the alternative models to these data by

Table 6
Fits of ATRIUM and the Generalized Context Model (GCM) to the Classification Transfer Data From Erickson and Kruschke's (1998) Experiment 1 Paradigm

	Study					
	Original			Follow-Up		
Model	−ln L	% Var.	SSD	−ln L	%Var.	SSD
ATRIUM	324.0	97.7	0.403	321.0	98.5	0.335
GCM	547.1	89.9	1.810	338.3	97.3	0.600

Note—SSD, sum of squared deviations.

using a maximum-likelihood criterion (see, e.g., Wickens, 1982). The criterion is to minimize the statistic −ln L, given by

$$-\ln \mathrm{L} = \sum \ln N_i! - \sum\sum \ln f_{ij}! + \sum\sum f_{ij} \cdot \ln p_{ij}, \qquad (4)$$

where f_{ij} is the observed frequency with which stimulus i was classified in category j, p_{ij} is the predicted probability from the model with which stimulus i was classified in category j, and N_i is the total frequency with which stimulus i was presented. As auxiliary measures of fit, we report the sum of squared deviations (SSD) between predicted and observed classification probabilities and the percentage of variance in the observed classification probabilities accounted for by each model. Fitting ATRIUM required the estimation of eight free parameters (for details regarding these parameters, see Erickson and Kruschke, 1998, pp. 117–120), whereas fitting the GCM required the estimation of three free parameters (the sensitivity parameter c, the response-scaling parameter γ, and a single attention weight w_1).

The model-fitting results are summarized in the left portion of Table 6. The GCM performs dramatically worse than does ATRIUM in quantitatively fitting Erickson and Kruschke's (1998) classification transfer data. The value of the log-likelihood statistic is more than 50% higher for the exemplar model than for ATRIUM (smaller values of −ln L indicate a better fit for a model), and the SSD for the exemplar model is over four times greater than that of ATRIUM. These dramatically better fits for ATRIUM corroborate Erickson and Kruschke's findings regarding the superiority of the multiple-system categorization model over the exemplar model in their original study.

However, when the models are fitted to the classification transfer data obtained in our follow-up experiment, the results are quite different. As shown in the right portion of Table 6, the GCM now provides an excellent overall fit to the complete set of transfer data, nearly as good as that of the eight-parameter multiple-system ATRIUM model. (The predictions from the GCM are shown along with the observed data in Appendix A.) The best-fitting parameters were $c = .767$, $w_1 = .322$, and $\gamma = 2.140$. The attention-weight parameter estimate indicates that observers gave more weight to Dimension 2 (the primary dimension in Figure 6) than to Dimension 1 (the secondary dimension), which is sensible given that the primary dimension is more diagnostic of category membership.

Overall, then, the results are consistent with our hypothesis that the single-system exemplar model may be sufficient to account for the pattern of data observed in Erickson and Kruschke's (1998) paradigm. In the situation in which labels were presented as part of the stimulus display, they apparently formed part of the exemplar representation and accordingly had a strong influence on the pattern of generalization. A reasonable application of the exemplar model to such a design would therefore need to explicitly incorporate the labels as part of the multidimensional similarity representation for the stimuli. By contrast, in the situation in which the labels were not presented as part of the display, the pattern of generalization is as predicted by the standard exemplar model, and it provides an excellent quantitative account of the data.

Performance on Prototypes Versus Exceptions

Thus far in our article, we have considered cases in which investigators argued in favor of multiple-system models involving rule formation. In this section we consider evidence that has been used to argue for a multiple-system model involving prototypes. Specifically, J. D. Smith, Murray, and Minda (1997; J. D. Smith & Minda, 1998) argued for a mixed prototype-plus-exemplar model of categorization. According to their view, prototypes are often abstracted during early stages of category learning or when categories have highly coherent structures. Exemplar storage is used to supplement prototype abstraction as learning is extended in time. In the J. D. Smith et al. (1997) model, the term *prototype* refers to an idealized object that is composed of the most frequently occurring dimension values of the members of a category.

The main experimental design that J. D. Smith and colleagues used to support their hypothesis of prototype abstraction is shown in Table 7. The stimuli varied along six binary-valued dimensions, with Logical Value 1 tending to indicate Category A and Logical Value 2 tending to indicate Category B. In this design, the prototype of Category A is 111111, and the prototype of Category B is 222222. Note also that each category in J. D. Smith et al.'s design contains an "exception" item. Specifically, 222212 is an exception stimulus that belongs to Category A, whereas 111211 is an exception stimulus that belongs to Category B.

As a representative example of the use of this design, we consider J. D. Smith et al.'s (1997) Experiment 1. In this experiment, J. D. Smith et al. tested 16 individual

Table 7
Nonlinearly Separable Category Structure Tested by J. D. Smith, Murray, and Minda (1997)

Category A	Category B
1. 1 1 1 1 1 1	8. 2 2 2 2 2 2
2. 2 1 1 1 1 1	9. 1 2 2 2 2 2
3. 1 2 1 1 1 1	10. 2 1 2 2 2 2
4. 1 1 2 1 1 1	11. 2 2 1 2 2 2
5. 1 1 1 1 2 1	12. 2 2 2 1 2 2
6. 1 1 1 1 1 2	13. 2 2 2 2 2 1
7. 2 2 2 2 1 2	14. 1 1 1 2 1 1

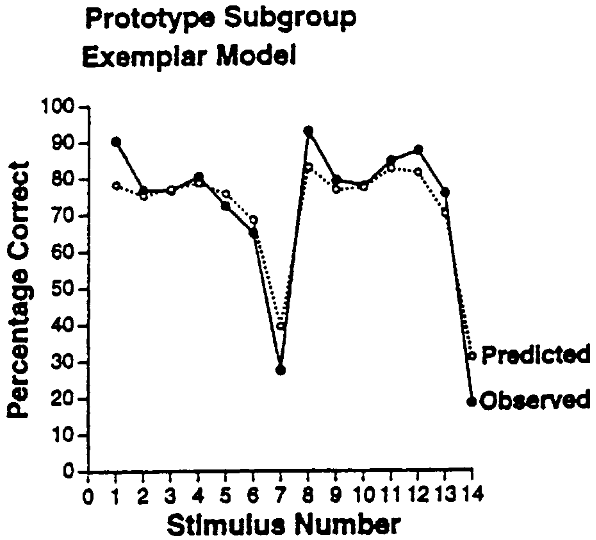

Figure 7. Solid line shows the average observed percentage of correct category decisions displayed by the "prototype subjects" for each of the 14 stimuli in J. D. Smith, Murray, and Minda's (1997) Experiment 1 (nonlinearly separable, easy condition). Dotted line shows the average percentage of correct category decisions for each stimulus predicted by the generalized context model. Note—From "Straight Talk About Linear Separability," by J. D. Smith, M. J. Murray, & J. P. Minda, 1997, *Journal of Experimental Psychology: Learning, Memory, & Cognition, 23,* p. 669. Copyright 1997 by the American Psychological Association. Adapted with permission.

subjects on the Table 7 category structure and compared a prototype model and the GCM on their ability to fit the individual-subject learning data. Although the GCM provided better fits to the data of 8 of the subjects, the prototype model fared better on the remaining 8 subjects. The averaged observed and GCM-predicted data for this latter "prototype subgroup" of subjects are displayed in Figure 7 (from J. D. Smith et al., 1997, Figure 5). As can be seen, the GCM underpredicted the observed performance on the prototype stimuli (Stimulus 1 [111111] and Stimulus 8 [222222]), but overpredicted performance on the exceptions (Stimulus 7 [222212] and Stimulus 14 [111211]).

Furthermore, J. D. Smith et al. (1997) argued that these deviations represented a fundamental qualitative shortcoming of the exemplar model. Figure 8 provides a scatterplot, for each of a set of 16 "prototype subjects" (across two experiments), of the mean percent correct performance observed on the prototypes and exceptions (from J. D. Smith et al., 1997, Figure 10). In general, these subjects are ones who performed extremely accurately on the prototypes but who classified the exceptions with low accuracy. By way of comparison, Figure 9 illustrates the "response surface" of the exemplar model, that is, the set of all possible performances that it can predict on these items under any of its parameter settings (after J. D. Smith et al., 1997, Figure 10). It is easily observed that the behavior of the "prototype subjects" lies in a region of the performance space where the exemplar model cannot go.

However, the version of the exemplar model fitted by J. D. Smith et al. (1997) assumed $\gamma = 1$ in the response-rule equation (Equation 1), which is the value assumed in the original version of the context model of Medin and Schaffer (1978). Although J. D. Smith et al.'s "prototype subjects" effectively rule out this special-case model, we know of no very strong reason why one should assume $\gamma = 1$. In an early article that tested the context model, Medin and Smith (1981) justified use of this particular response rule simply by saying, "The best defense of the response rule is that it is a fair approximation and that it seems to work" (p. 250). Numerous independent lines of evidence since then have indicated that when fitting individual-subject data with the exemplar model, it is important to allow γ to take on values greater than 1 (see, e.g., Maddox & Ashby, 1993; McKinley & Nosofsky, 1995; Nosofsky & Palmeri, 1997). As discussed in our Overview of the Formal Model section, the γ response-scaling parameter describes the extent to which observers use probabilistic as opposed to deterministic response strategies in classification, and values of $\gamma > 1$ allow the exemplar model to account for the types of deterministic responding that are often evidenced at the individual subject level.[4]

In Figure 10, we plot the response surface of the exemplar model with γ allowed to vary freely. Inspection of

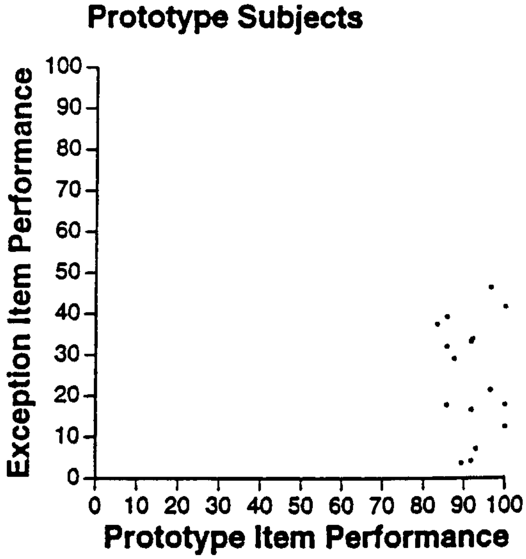

Figure 8. Scatterplot of performances for each of a set of 16 "prototype subjects" (across two experiments) from the study of J. D. Smith, Murray, and Minda (1997). The vertical axis gives the exception-items performance of each subject, and the horizontal axis gives the prototype-items performance. Note—From "Straight Talk About Linear Separability," by J. D. Smith, M. J. Murray, & J. P. Minda, 1997, *Journal of Experimental Psychology: Learning, Memory, & Cognition, 23,* p. 675. Copyright 1997 by the American Psychological Association. Adapted with permission.

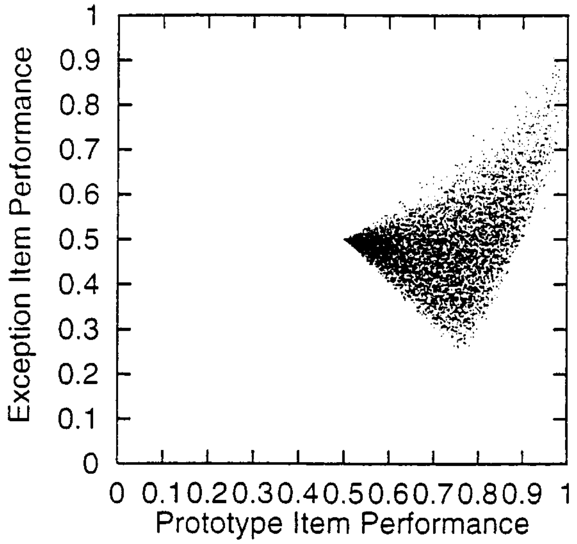

Figure 9. The set of all possible performances that the GCM can jointly predict for the prototype and exception items in J. D. Smith, Murray, and Minda (1997; Experiment 1—nonlinearly separable, easy condition). These predictions assume $\gamma = 1$ in the response-rule equation (Equation 1).

the figure indicates that the exemplar model can easily accommodate the performance profiles displayed by the "prototype subjects" in Figure 8. At the same time, the model can account for the performances of all of the "exemplar subjects" who are *not* illustrated in Figure 8 (but who were just as prevalent as the "prototype subjects"). To provide a complete account of J. D. Smith et al.'s (1997) data, a single-system model is required to account for both types of performance.

The performance profiles of the "prototype subjects" are well predicted by the exemplar model when the overall level of sensitivity (c) in the model is low and when the response-scaling parameter (γ) is at least moderately high. As discussed previously, when sensitivity is low, there is a great deal of stimulus generalization. Because the exception items in the Table 7 structure are highly similar to numerous objects from the contrast category, their summed similarity to members of the contrast category exceeds their summed similarity to members of their own category, so the observer classifies them into the wrong category. Note that it is at early stages of learning that one expects the sensitivity parameter to take on lower values. One of the major learning mechanisms in the exemplar model has always involved the idea that overall sensitivity (ability to discriminate objects in memory) increases with training (Nosofsky, 1987; Nosofsky et al., 1992). By comparison, once sensitivity reaches high levels, as would be expected later in training, the exception items will have low summed similarity to the members of the contrast category; that is, there is little stimulus generalization. The exception items perfectly match their own memory traces, however, so still have high similarity to themselves. Thus, at late stages of learning, the exemplar model predicts correctly that the exception items will be accurately classified. Finally, the role of the γ parameter is simply to scale the extremeness of the response-probability predictions made by the model, that is, the degree to which the probabilities deviate from .50 in the direction of either zero or unity.

To gain further evidence bearing on the adequacy of the exemplar model, we conducted maximum-likelihood fits of the model to J. D. Smith et al.'s (1997, Experiment 1) individual subject data. Figure 11 provides a composite summary of the results of these individual-subject fits, analogous to the composite summary that was displayed previously for the $\gamma = 1$ model. The figure plots the observed and predicted classification probabilities for each of the 14 stimuli, averaged across the 16 individual observers. The figure now reveals no systematic deviations between predicted and observed data values. Indeed, as shown in Table 8, the exemplar model gives a better overall maximum-likelihood fit to J. D. Smith et al.'s individual subject data than does the alternative prototype model, and for none of the individual subjects is there a substantial advantage for the prototype model. (We fitted the same version of the prototype model as was used by J. D. Smith et al., 1997. Details of the model-fitting procedures are provided in our Appendix B.)

We emphasize that we are not claiming here a strong advantage for the exemplar model over the prototype model in this paradigm. Drawing strong conclusions is particularly difficult in view of the fact that the exemplar model makes use of an additional free parameter relative to the prototype model (see Appendix B). Rather, we argue that J. D. Smith et al.'s (1997) strong claim of a qualitative advantage of the prototype model over the ex-

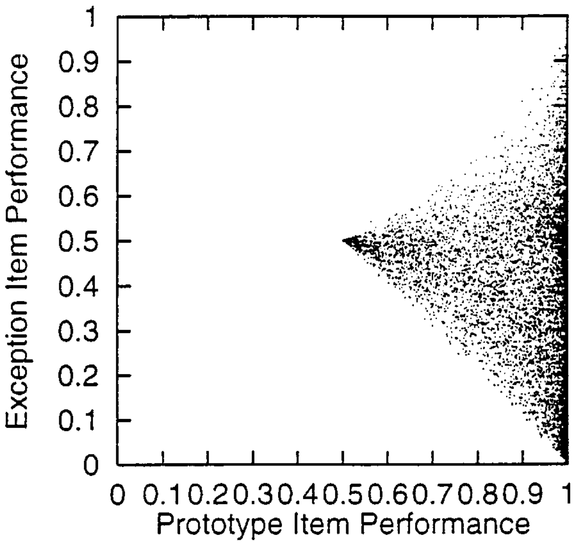

Figure 10. The set of all possible performances that the GCM can jointly predict for the prototype and exception items in J. D. Smith, Murray, and Minda (1997; Experiment 1—nonlinearly separable, easy condition), with the γ response-scaling parameter allowed to vary freely.

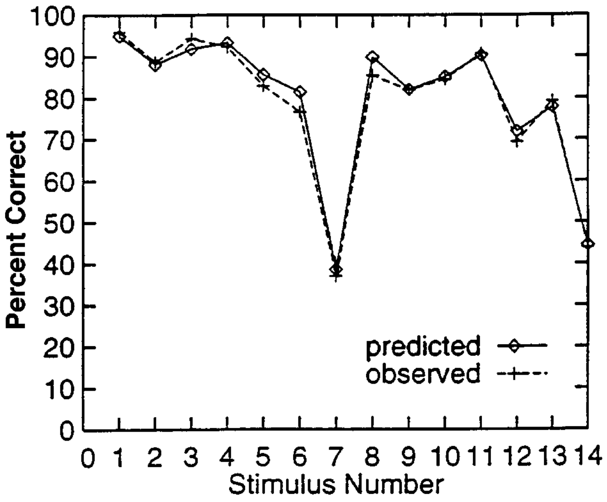

Figure 11. **Solid line, observed percentage of correct category decisions for each of the 14 stimuli in J. D. Smith, Murray, and Minda's (1997) Experiment 1 (nonlinearly separable, easy condition), averaged across all 16 observers. Dotted line, average percentage of correct category decisions for each stimulus predicted by the GCM (with γ allowed to vary freely).**

emplar model is not merited by the data. Apparently, a more diagnostic experimental paradigm will be needed to gain clear evidence of the need to extend the exemplar model with a separate prototype-abstraction system.

Although they did not provide detailed comparisons of fits to the individual subject data, J. D. Smith and Minda (1998, pp. 1430–1432) acknowledged that when the γ parameter is allowed to vary freely, the pure exemplar model could perform well in this paradigm. They argued against the exemplar-model interpretation, however, by suggesting that the response-scaling parameter "can be a prototype in exemplar clothing" (J. D. Smith & Minda, 1998, p. 1431). In other words, their suggestion was that the exemplar model with the response-scaling parameter was simply mimicking what is a simpler description provided by the prototype model. Our view, however, is that simplicity is in the eye of the beholder. It is true that for some of the individual subjects or stages of learning, the exemplar model with an additional free parameter is providing roughly the same fits as is the prototype model. For numerous other subjects or stages of learning, however, the prototype model performs substantially worse than does the exemplar model. The preferred interpretation of J. D. Smith and his colleagues is that subjects are both abstracting prototypes and storing exemplars, with the particular representational system that is dominant varying across individual subjects and stages of learning. By contrast, we are pointing out that a single-system exemplar model can account for *all* of J. D. Smith et al.'s data, without the need to posit the second representational system. In a nutshell, in the comparisons that J. D. Smith and his colleagues have conducted in their paradigm, the "competition" is between an exemplar model with seven free parameters versus a multiple-system model in which the exemplar system uses six free parameters, the prototype system uses six free parameters, and where the assumed system for any individual subject at a given stage of learning is based on post hoc fits to the data. Viewed from this perspective, the exemplar model with the response-scaling parameter seems at least as parsimonious as the alternative multiple-system approach proposed by J. D. Smith and Minda (1998). It is true that our exemplar-based account does not predict in advance which individual subjects will use smaller values of γ and which will use larger ones; but neither does J. D. Smith and Minda's multiple-system account predict in advance which individual subjects will use exemplars and which will use prototypes.

Dissociations Between Categorization and Similarity Judgment

A classic study that challenged the sufficiency of exemplar-similarity accounts of categorization is one reported by Rips (1989; see also Rips & Collins, 1993). In this study, Rips demonstrated certain dissociations between categorization and similarity judgments that were interpreted as evidence for the role of rule-based categorization.[5]

An example of the type of phenomenon documented by Rips (1989) is as follows. Subjects were asked to imagine a circular object with a 3-in. diameter. One group was asked to judge whether the object was more similar to the category QUARTER or to the category PIZZA. (It was previously established in the study that the 3-in. object was intermediate in diameter between the subjects' experiences of the largest quarter and the smallest pizza.) A second group, however, was asked to decide to which of the two categories the object was more likely to belong. Whereas the similarity-judgment group judged the object as more similar to QUARTER, the category-judgment

Table 8
Summary Fits of Exemplar and Prototype Models to the Data From J. D. Smith, Murray, and Minda's (1997) Experiment 1, Nonlinearly Separable (Easy) Condition

	−ln Likelihood	
Subject	Exemplar	Prototype
1	20.4	22.1
2	22.3	21.8
3	23.9	23.6
4	21.2	20.9
5	14.8	15.6
6	26.3	30.5
7	18.1	22.6
8	17.1	23.8
9	10.8	19.8
10	20.0	20.0
11	18.6	18.5
12	19.2	25.2
13	18.0	14.4
14	18.2	23.0
15	14.3	26.5
16	19.5	19.6

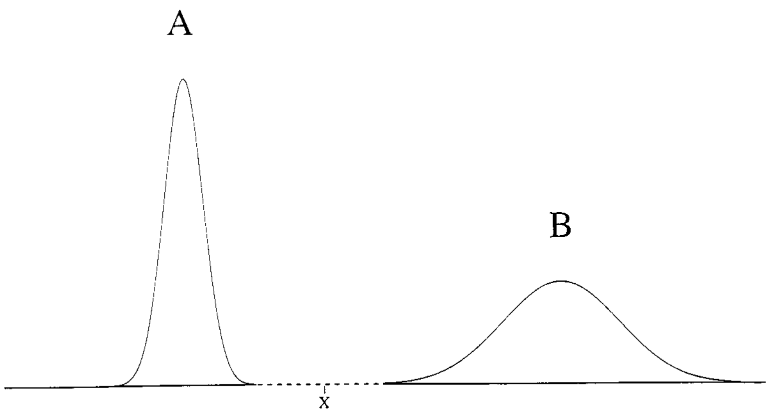

Figure 12. Schematic illustration of the types of category structures queried by Rips (1989) in his studies of effects of variability on categorization and similarity judgment.

group judged the object as more likely to belong to the PIZZA category. Thus, Rips demonstrated a dissociation between similarity and categorization judgments. The key point is that although the 3-in. object is judged more similar to the QUARTER category, it is judged more likely to belong to the PIZZA category because the latter is a high-variability category on the size dimension. By contrast, the category QUARTER has virtually no variability on this dimension. Subjects apparently used their knowledge of the low variability of the QUARTER category to override their similarity judgments. The interpretation is that subjects used a "rule" that an object must have virtually the exact diameter of a QUARTER to be classified in that category. Indeed, E. E. Smith and Sloman (1994) summarized the interpretation of Rips's studies as follows:

> The clearest evidence for rule-based categorization, and for its difference from similarity-based categorization, comes from studies of Rips (1989) . . . Because categorization decisions favored one category but similarity decisions favored the other, the categorization decisions could not have been based on similarity. (p. 378)

Rips's (1989) study does indeed pose an interesting challenge to single-system, exemplar-similarity accounts of categorization. In this section, we speculate on possible approaches to meeting the challenge. To begin, we focus first on the categorization aspect of the study. The structure of the categorization problem is illustrated schematically in Figure 12. Category Distribution A is an extremely low-variability category; Distribution B is the high-variability category; and x represents an object that is equidistant from the highest exemplar of Category A and the lowest exemplar of Category B. Note that, according to the standard GCM, the summed similarity of x to the exemplars of Category A exceeds its summed similarity to the exemplars of Category B (because all exemplars of A are a fairly small distance from x, whereas relatively few exemplars from B are a small distance from x). Yet, the phenomenon observed by Rips was that subjects tended to classify x into the high-variability category.

Our main proposal is that the phenomenon can be understood in terms of the exemplar-similarity model if allowance is made for the idea that separate sensitivity parameters may govern similarity computations to the low- and high-variability categories (see also E. E. Smith & Sloman, 1994, p. 384). Recall that the sensitivity parameter determines the rate at which similarity decreases with distance (Figure 2). Although most applications of the GCM assume a single sensitivity parameter, a reasonable extension of the model makes allowance for the role of category-specific sensitivities. The idea is that, in making a classification judgment, an observer learns that an object should be considered as similar to members of a low-variability category only if it virtually exactly matches those members on the relevant attribute. (We will explain below why we believe that such an assumption is extremely reasonable.) Thus, the observer adopts a high setting of the sensitivity parameter when computing similarities to the exemplars of the low-variability category. By contrast, a looser criterion can be used in assessing the similarity of an object to members of a high-variability category, so here the sensitivity parameter is set at a low value. The result is that the summed similarity of object x to the exemplars of the high-variability category exceeds its summed similarity to the exemplars of the low-variability category, so it is classified into the high-variability category.

Why is the proposal sensible that separate sensitivities may govern similarity comparisons to low- and high-variability categories? We believe that it follows directly from the same considerations of adaptiveness and optimality that have motivated many of the other predictions of the exemplar model. For example, as discussed previously in this article, it has long been hypothesized that there are adaptive learning principles that lead observers to distribute attention across psychological dimensions in a manner that tends to optimize performance (Nosofsky, 1984, 1986). This attention weighting causes systematic changes in the structure of psychological space so as to maximize within-category similarities among exemplars and minimize between-category similarities. Likewise, an observer who wished to maximize his/her percentage of correct classification decisions would adjust the setting of the sensitivity parameter in accord with experienced category variabilities. In Appendix C, we verify that if an observer's goal is to maximize his/her percentage of correct generalizations to new objects generated from the Figure 12 category distributions, then it is indeed optimal to use a lower setting of sensitivity when computing similarities to the high-variability category than to the extremely low-variability one.

Note that previous investigators have argued that observers may adjust the setting of the sensitivity parameter in accord with task demands and have provided empirical evidence in support of this hypothesis (e.g., Estes, 1986; Lamberts, 1994; Nosofsky, 1991a; L. B. Smith, 1989), so the current proposal converges with past ones made in the field. For example, Lamberts reported studies about how the role of background knowledge may influence the types of generalization gradients that observers use when categorizing. In a training phase, subjects experienced a set of schematic faces that were experimentally defined as belonging to one of two families. In a transfer phase, new faces were shown and subjects were required to classify them into the families. The crucial manipulation was that for each transfer face, subjects were informed either that it was a brother of one of the training faces or a cousin of one of the training faces. Owing to subjects' background knowledge that a brother is likely to closely resemble a particular family member, whereas a cousin might only vaguely resemble a number of different family members, it was hypothesized that subjects would use different generalization gradients for classifying the two types of items. Specifically, it was hypothesized that subjects would use a steep generalization gradient (a high level of sensitivity) for classifying brothers, but a shallow gradient (a low level of sensitivity) for classifying cousins. Lamberts obtained consistent evidence in support of this hypothesis across three separate experiments.

Although our hypothesis regarding category-specific sensitivity may allow the exemplar model to account for subjects' classification judgments, we have addressed only half the story. Recall that Rips (1989) also demonstrated that when asked to make direct similarity judgments, subjects judged the 3-in. object as more similar to QUARTER than to PIZZA.

In a nutshell, we believe that it is a mistake to view the similarity-judgment question as a "stand-in" for the exemplar model's predictions of classification judgments. The type of summed similarity computation that we hypothesize is performed by the exemplar model is not something that can be assessed simply by asking subjects to provide a similarity judgment. Even in situations in which observers are asked to provide similarity ratings between individual objects, the nature of such ratings has been treated with extreme caution in the field. It is well documented that such ratings are highly context dependent and are strongly influenced by complex cognitive and decision factors. Although Nosofsky (1991a; Shin & Nosofsky, 1992) has sometimes collected similarity

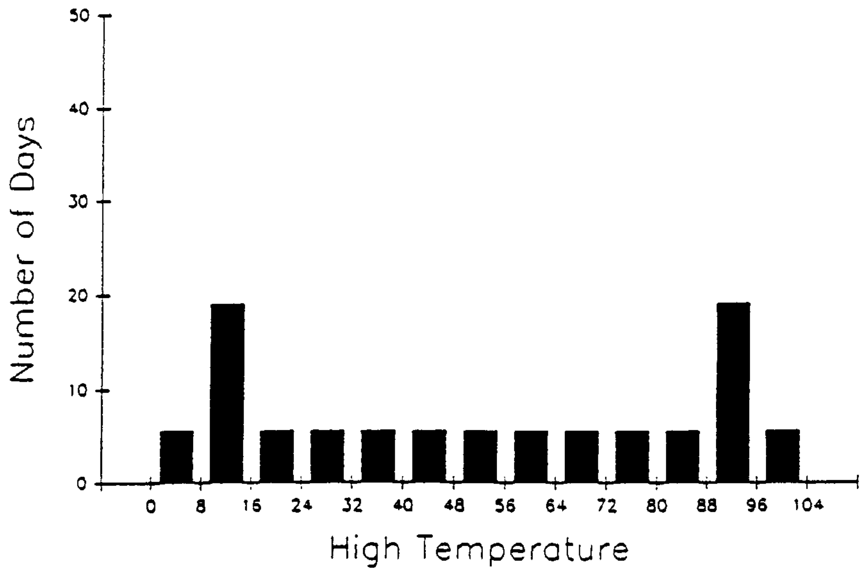

Figure 13. Stimulus histogram from Rips and Collins's (1993) Experiment 1, illustrating a sample of 100 daily high temperatures from February and August. Note—From "Categories and Resemblance," by L. J. Rips and A. Collins, 1993, *Journal of Experimental Psychology: General, 122*, p. 471. Copyright 1993 by the American Psychological Association. Reprinted with permission.

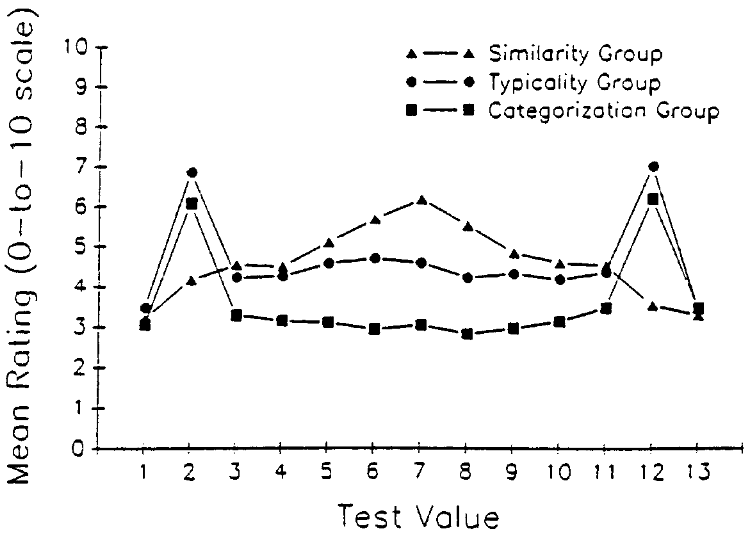

Figure 14. Mean ratings from Rips and Collins's (1993) Experiment 1 for similarity, typicality, and likelihood as a function of the position of the test value within the histogram. Note—From "Categories and Resemblance," by L. J. Rips and A. Collins, 1993, *Journal of Experimental Psychology*: *General, 122,* p. 473. Copyright 1993 by the American Psychological Association. Reprinted with permission.

judgments among objects in studies of categorization, the ratings are never used directly to predict subjects' classification choices. Instead, the ratings are used to construct MDS solutions for the objects, which are then used in combination with the machinery of the formal exemplar model as a means of predicting classification (see Nosofsky, 1992, for a review). Processes such as dimensional attention weighting and modulation of overall sensitivity are presumed to modify the structure of the psychological similarity space in which the exemplars are embedded (Nosofsky, 1986, 1987). The complex cognitive and decision factors that influence similarity judgment are likely to be further compounded in situations such as Rips's (1989) study, in which observers must deal with linguistically difficult and open-ended questions such as how similar a single object is to entire categories of other objects.[6]

Indeed, it is our view that, in a follow-up study, Rips and Collins (1993) themselves actually provided a "demonstration proof" that the direct similarity-judgment question does not serve as a stand-in for the exemplar model's predictions of classification. The design of one of Rips and Collins's experiments is illustrated in Figure 13. On each trial, subjects were presented with a histogram similar to the one in the figure along with a particular test value indicated by an arrow below the histogram's *x*-axis. In the Figure 13 example, the histogram depicts a single "category" of 100 daily high-temperature readings observed during the months of February and August. A *similarity* group rated the similarity of the indicated test value to the depicted category, whereas a *categorization* group rated the likelihood that the test value was a category member. (A third group rated how *typical* the value was of the category.) The main results are shown in Figure 14, which plots mean ratings from the three groups against the various test values in the histogram. It is evident from inspection that whereas the highest similarity ratings occurred for values located centrally in the histogram, the highest categorization ratings occurred for the most frequently occurring values in these bimodal distributions. Rips and Collins argued that the results posed problems for "resemblance-based" models of categorization, such as exemplar-similarity models, because there was again a dissociation between the similarity and categorization judgments.

Because the histogram-judgment experiment does not correspond to a situation in which people have learned categories by induction over individually presented exemplars, we need to introduce some assumptions to try to translate it into a format compatible with the exemplar model. First, we assume that the frequency counts in the histogram translate directly into copies of exemplars stored in subjects' memory; second, that the configuration of exemplars in psychological space corresponds directly to the physical layout pictured in the figure; and third, that the category-likelihood judgment is then monotonically related to the summed similarity of a value to the exemplars pictured in the histogram. Given these assumptions, it is then straightforward to show that the exemplar model predicts that the category-likelihood judgments will follow the bimodal pattern observed in the figure for the category-judgment group.[7] In other words, the most straightforward application of the exemplar model to the histogram-judgment experiment *predicts* the pattern of categorization results that were reported by Rips and Collins (1993). This subset of results, therefore, can hardly be considered as providing evidence against the exemplar model.

Equally relevant in the present context, there are *no* parameter settings for which the exemplar model would yield the highest summed similarities for the central values in the histogram (whereas the subjects in the similarity-judgment group gave these values the highest similarity ratings). Therefore, whatever subjects may be doing when rating how similar an individual value is to the category of values depicted in the histogram, it does *not* correspond to the summed-similarity computation performed by the exemplar model. A reasonable inference, therefore, is that the subjects' similarity judgments in the QUARTER–PIZZA study are also unlikely to correspond to the summed-similarity computations performed by the exemplar model.

In summary, our view is that Rips (1989; Rips & Collins, 1993) has developed an interesting cognitive task, namely one in which subjects must make direct judgments of the similarity of an object to categories of exemplars, and a worthwhile goal is to develop models that can account for such judgments. However, the direct similarity judgment itself cannot be viewed as a "stand-in" for what the exemplar model predicts regarding classification. Rips and Collins's (1993) own data provide a demonstration proof that such judgments often do not correspond to the exemplar model's predictions of classification.

Dissociations Between Categorization and Recognition in Normals and Amnesics

In this final section, we briefly review a recently reported application of the exemplar model (Nosofsky & Zaki, 1998) to another phenomenon that has been interpreted by researchers as providing evidence for the role of multiple cognitive systems in categorization. In a series of experiments, Knowlton, Squire, and their colleagues demonstrated some intriguing dissociations between categorization and recognition in normal and amnesic subjects (Knowlton, Mangels, & Squire, 1996; Knowlton & Squire, 1993; Squire & Knowlton, 1995). For example, Knowlton and Squire tested amnesics and matched normal controls in categorization and recognition tasks involving the classic Posner and Keele (1968) dot-pattern distortion paradigm. In the categorization task, subjects were presented with a list of 40 high distortions of a dot-pattern prototype. Following the presentation of the list, subjects were presented with a test set consisting of the prototype, 20 low distortions of the prototype, 20 new high distortions of the prototype, and 40 random patterns. Subjects were asked to judge whether each pattern belonged to the category defined by the study items. In a recognition task, subjects were presented with a study list of five random dot patterns presented eight times each. They were then presented with a test list consisting of the five study items and five new random dot patterns, and were asked to judge whether each pattern was old or new.

As expected, the amnesics performed significantly worse than did the normals on the old–new recognition task. The interesting result, however, was that the amnesics performed virtually as well as the normals in the categorization task (the differences in performance between the two groups were not statistically significant).

Knowlton and Squire (1993) interpreted this dissociation between categorization and recognition performance as providing evidence of separate memory systems governing the two tasks. According to their interpretation, an explicit memory system based on the storage of individual exemplars underlies recognition, but an implicit system, perhaps based on the formation of prototypes, governs categorization. The explicit memory system of the amnesics is damaged, accounting for their impaired recognition, but the implicit memory system of the amnesics is intact, accounting for their normal categorization performance.

Nosofsky and Zaki (1998), however, demonstrated that a single-system exemplar model, which was essentially the same as the one described in the present article, could account in quantitative detail for the complete set of classification and recognition data reported by Knowlton and Squire (1993). The key idea introduced in their modeling analyses was that both categorization and recognition judgments were based on summing similarities of the test items to exemplars stored in memory; however, the exemplar-based memory of the amnesics was assumed to be impaired. This impairment was modeled in terms of a lower setting of the overall sensitivity parameter in the exemplar model. Although the lowered sensitivity causes severe impairment in the old–new recognition task, Nosofsky and Zaki demonstrated that it is sufficient to support near-normal performance in the categorization task (at least in the paradigms tested by Knowlton and Squire). This result can be described intuitively as follows. To perform well in old–new recognition, it is beneficial to discriminate as sharply as possible between individual old versus new objects. However, to perform well in the categorization task, it is important to take advantage of similarities between new transfer patterns and the old exemplars on which subjects were trained. Making discriminations that are too fine-grained can actually be detrimental to categorization, where an important goal is to generalize appropriately to new objects and to treat distinct objects that belong to the same category as equivalent. Thus, the lower level of memory sensitivity available to the amnesics still enabled them to generalize to the new transfer patterns at near-normal performance levels.

GENERAL DISCUSSION

Summary

In summary, in this article we demonstrated that a single-system, exemplar-similarity model accounts for a wide variety of recently reported perceptual-classification phenomena that investigators have interpreted in terms of rule use or prototype formation. These phenomena have led to the proposal that multiple systems govern perceptual categorization and have stimulated the development of a variety of multiple-system categorization models. The present demonstrations are important because they provide evidence that, at least in free-strategy situations in which people learn categories via induction over training exemplars, the multiple-system models may not be necessary: A single-system, exemplar-similarity approach appears adequate to account for the major phenomena of interest.

We summarize here some of the features of the exemplar model that allow it to account for these rule-described and prototype-described phenomena. The first feature involves the idea that similarity is not an invariant relation. Rather, it is highly context dependent due to the influence of selective attention to the component dimensions that compose the category exemplars. The selective attention process is modeled in terms of a set of dimension weights that stretch and shrink the psychological similarity space in which the exemplars are embedded. An important working hypothesis that has gained support in numerous previous studies is that observers learn to distribute attention across the psychological dimensions in a manner that tends to optimize performance (Nosofsky, 1984, 1986, 1987, 1991a, 1998a; see also Getty, Swets, Swets, & Green, 1979, and Reed, 1972).

In the examples in the present article, the selective attention process was the key to allowing the exemplar model to account for the phenomenon of a "bias toward verbal rules" observed by Ashby et al. (1998) in their studies of probabilistic classification learning. Specifically, according to the model, observers would maximize performance by giving greater attention to the more diagnostic dimension in Ashby et al.'s experimental design. Such an attentional process, in turn, results in the best-fitting linear decision boundary being "pulled away" from the optimal population boundary in the direction of the "verbal rule." Ideas about selective attention were also critical in allowing the model to account for the heterogeneous "distribution of generalizations" observed by Nosofsky, Palmeri, and McKinley (1994) in their studies of rule-plus-exception learning. The exemplar-based account of this phenomenon involved the plausible assumption that individual observers may vary widely in the dimensions to which they selectively attend early in classification training, even though an adaptive learning process may eventually drive many observers in the direction of a nearly optimal attentional distribution later in training. We also found support for the hypothesis that observers may have attended to the numeric labels that were available in Erickson and Kruschke's (1998) experiment, and that some of the evidence for the operation of the rule module in this study may have been reflecting a modified proximity structure instead.

A second important feature of the exemplar model is the role of the overall sensitivity parameter that relates judged similarity to distance in psychological space. When sensitivity is low, there is a great deal of stimulus generalization, in the sense that even objects that are distant in psychological space are judged as fairly similar. Sensitivity is expected to be low at the start of training, before processes of perceptual and memorial differentiation begin to operate (Gibson & Gibson, 1955; Nosofsky, 1987). Under conditions of low sensitivity, the exemplar model predicts large prototype enhancement effects, because prototypes are highly similar to virtually all of the exemplars from their own category. Furthermore, under such conditions, "exception items" will often be classified into the wrong category because they tend to have greater overall similarity to the exemplars of the contrast category than to their own target category. The idea that observers may have low sensitivity, particularly at the start of training, was one of the keys to allowing the exemplar model to account for the poor performance on the exception items observed in the "prototype abstraction" study of J. D. Smith et al. (1997). The sensitivity parameter was also instrumental in allowing the exemplar model to account for the classification–recognition dissociations observed in the studies of Knowlton and Squire (1993). The idea here was that the lowered memory sensitivity of the amnesics was particularly detrimental to their old–new recognition performance, yet sufficient to allow them to make the gross-level discriminations that were needed to achieve near-normal performance on the categorization tasks. Finally, we speculated that strategic adjustments in the sensitivity parameter may be a key idea in allowing the exemplar model to account for patterns of "rule-based" classification judgments observed in the well-known variability studies introduced by Rips (1989).

A third aspect of the exemplar model that plays an important role in its predictions is the operation of the response-scaling parameter (γ). This parameter governs the extent to which observers use probabilistic versus deterministic response strategies in their classification judgments. In the original version of the context model (Medin & Schaffer, 1978; Nosofsky, 1986), the γ parameter was set at 1, in which case the exemplar model tends to predict "probability-matching" behavior (Estes, 1986; Nosofsky et al., 1992). However, there was never any strong motivation for the use of this parameter setting. Our view is that the extent to which observers use probabilistic versus deterministic response strategies is an issue quite separate from the issue of the nature of the underlying category representation. For example, suppose that a "rule boundary" has been established, and an object falls a certain distance from the boundary within Region A of the space. An observer may choose to respond that the object belongs in Category A with probability based on its distance from the boundary, or may choose to respond deterministically that the object belongs in Category A. Analogously, suppose that an object has greater summed similarity to the exemplars of Category A than to the exemplars of Category B. An observer may choose to respond by probability matching to the summed similarities, or may choose to respond deterministically that the object belongs in the category with the greater summed similarity. On the basis of this reasoning, we argue that evidence for rules versus exemplars should be decoupled from the question of whether observers use deterministic or probabilistic response strategies.

In the present investigations, the role of the γ parameter was explored most directly in our exemplar-based accounts of J. D. Smith et al.'s (1997) "prototype abstraction" studies. In particular, our interpretation is that the subset of subjects who, near the start of training, classify the prototype with high accuracy into the target category, but who classify the exception item with high probability into the wrong category, satisfy the following characteristics. First, they tend to have a low value of overall sensitivity; thus, the exception items tend to have greater summed similarity to the members of the contrast category than to their own target category. Second, these observers tend to use a deterministic response rule (values of γ greater than 1), responding with the category that yields the larger summed similarity.

Issues of Model Flexibility

Although the exemplar model accounts well for the numerous multiple-system phenomena that we have considered in this article, we should reemphasize that we are not claiming that it necessarily provides a superior account of the phenomena than do the multiple-system models. Because the exemplar model can be viewed as a single-system model and generally has fewer free parameters than its multiple-system competitors, there is a strong sense in which it provides a more parsimonious account of the full range of data. However, in some of the cases considered, we should admit that the exemplar model did not predict a priori certain key patterns of results; we showed only that it could account for the results with suitable choices of its parameter settings or with certain extensions to the model. A potential danger is that, with enough post hoc extensions, the exemplar model can be made to fit anything and will not be falsifiable. On the other hand, in fairness to the exemplar-based approach, we believe that the types of extensions that were proposed herein were reasonable and sensible. At least at present, we believe that the story told by the exemplar model remains a conceptually simple one, and so the model stands as a viable alternative to its multiple-system competitors.

Obviously, a critical direction for future research is to design new experiments that will distinguish our exemplar-based accounts of the phenomena from the multiple-system accounts. For example, suppose that our parameter-difference explanation of Knowlton and Squire's (1993) categorization–recognition dissociation is correct. Then it ought to be possible to design new, more diagnostic categorization problems in which the lowered memory sensitivity of the amnesics will in fact lead to worse-than-normal performance. By contrast, if Knowlton and Squire's hypothesis of separate memory systems governing categorization and recognition is correct, then amnesics should perform as well as normals regardless of the difficulty of the categorization problem that is involved.

A potential concern is that the reason that the exemplar model accounts so well for the wide and varied phenomena considered in this article is simply that, in a mathematical sense, it is a highly flexible model. In other words, with suitable adjustments of its free parameters, perhaps it can account for virtually any pattern of data. In one previous study designed to get a handle on this issue, Nosofsky et al. (1989, Experiment 2) provided subjects with explicit instructions to use particular rules for classifying a set of objects. Under such conditions, the exemplar model failed dramatically to account for the patterns of data, whereas appropriately formalized rule-based models performed impressively (see Nosofsky et al., 1989, pp. 293–294). By contrast, in free-strategy designs involving the same category structures, the reverse pattern of results was observed. Thus, it was not the case that the exemplar model could act as a "universal data fitter"; rather, at least in those studies, it seemed to be capturing the types of performances that subjects naturally exhibited when learning categories by induction over individual training exemplars.

Nevertheless, it is crucial to gauge the degree of flexibility that a model has when evaluating its goodness of fit, and this issue is currently burgeoning in the field of mathematical psychology (e.g., Grunwald, 2000; Myung, 2000; Myung & Pitt, 1997). New model-evaluation techniques that are currently under development and that consider such flexibility may well hold the key to better assessing the virtues of the single-system versus multiple-system approaches to modeling categorization.

Finally, an important route to reducing the flexibility of a model is to develop theories of its parameters rather than allowing them to vary freely. The attention-optimization hypothesis, which we relied on at several points in this article, provides one example along these lines. However, in its present form, the attention-optimization hypothesis provides only general guidelines and is far from a universal law of human performance. Numerous other factors besides performance optimization are posited to influence the attention weights, including the intrinsic salience of the psychological dimensions, prior learning and beliefs, forms of hypothesis testing behavior, and so forth.

Furthermore, various constraints may prevent an observer from adjusting his/her attention in the theoretically optimal manner. For example, it is well known that it is difficult to attend selectively to individual component dimensions when they combine in an "integral" manner (e.g., Garner, 1974; Shepard, 1964). Thus, whereas Nosofsky, Gluck, et al. (1994) found strong support for the attention-optimization hypothesis in a series of classification learning conditions involving highly separable dimension stimuli, Nosofsky and Palmeri (1996) demonstrated that this attention learning failed to occur in a directly analogous set of classification problems involving integral-dimension stimuli. Likewise, although in this article we posited an extension of the attention-optimization ideas to forms of sensitivity optimization as well, there are clearly limits on the degree to which an observer can adjust his/her sensitivity. Without such limits, the exemplar model would make the absurd prediction that regardless of how similar two exemplars are, they could always be perfectly discriminated in one's memory. Thus, the machinery of the mind must set important constraints on the types of parameter adjustment that can take place. A challenge for future research is to develop models in which all of these multiple influences and constraints can be coordinated within a unified theoretical framework.

Beyond Perceptual Classification?

We reemphasize that in this investigation, we focused primarily on situations in which observers learned categories of perceptual objects by induction over presented exemplars. A natural question is the extent to which the exemplar-similarity principles may be applicable to more general forms of categorization as well as to cognitive tasks related to categorization such as inference and feature prediction. We cannot hope to even scratch the surface of the huge gamut of results pertaining to ideas involving "theory-based" categorization, goal-driven categorization, the role of prior knowledge in categorization, and reasoning-based inference processes (e.g., Barsalou, 1985; Carey, 1985; Gelman & Markman, 1986; Keil, 1989; Murphy & Medin, 1985; Sloman, 1996). Nevertheless, it is worthwhile to examine a few selected results to gain hints on how exemplar-similarity approaches might eventually be extended to these more complex domains.[8]

Among the major phenomena that have been interpreted as providing evidence against similarity-based models are inference studies such as those reported by Gelman and Markman (1986). An example of this type of study is the following. Children are shown a picture of a flamingo and a picture of a bat, and are told that the flamingo is a *bird* and that the bat is a *mammal*. They are also told that the flamingo has one type of internal organ and the bat a second type. Finally, the children are shown a picture of a blackbird, which is perceptually quite similar to the bat and dissimilar to the flamingo, and are told that it is a *bird*. The children are then asked whether the blackbird is more likely to have the same internal organ as the flamingo or the bat. The children tend to guess "flamingo." Thus, with regard to the driving force behind inferences about internal structure, the knowledge that flamingo and blackbird have common membership in the bird category overrides the high perceptual similarity that exists between bat and blackbird. By contrast, when asked to which object the blackbird is more likely to be closer in weight, the inferences begin to be based more on the perceptually visible properties. Thus, the common category label strongly supports certain inferences, but not others.

Such results are often taken as evidence against "similarity-based" models of categorization and in favor of "theory-based" ones (see, e.g., Sloman, 1996, pp. 8–10, for a review). In the above example, children are using an already developed biological knowledge that certain properties are predictable from membership in natural-kind categories whereas others are not. Furthermore, perceptual similarity fails to guide inferences about the internal structure of the objects. Thus, as summarized by Sloman (p. 9), it is argued that "simple similarity structures cannot explain concept use."

We believe, however, that phenomena such as those reported by Gelman and Markman (1986) have a natural interpretation within the perspective of modern exemplar-similarity models. The children in these studies have learned that flamingo and blackbird are both members of the bird category. Thus, these objects are "similar" by virtue of the fact that they share the same category label. Models such as the rational model of Anderson (1991) explicitly include the category label that is learned as one of the many features that an object possesses and along which similarity computations can take place. Furthermore, if one's goal is to predict the internal structure of an organism, people have learned that it is optimal to give a great deal of attention weight to the category-label feature. By contrast, if one's goal is to predict body weight, adaptive learning processes have taught people that visible perceptual features of objects such as overall size are better predictors, so attention would be shifted to these features in this alternative task. In a nutshell, as long as provision is made for the long-held ideas that similarities among objects are determined by selective attention to dimensions, and that the selective attention process is influenced by considerations of adaptiveness and optimality in the pursuit of task goals, there may not be as much difference between "theory-based" and "similarity-based" approaches as is commonly suggested in the literature.

Ultimately, the extent to which exemplar-similarity models may be fruitfully applied in complex conceptual domains will rest on their being endowed with a sufficiently rich theory of similarity. In the present applications to simple perceptual domains, we used classic MDS approaches to modeling similarity, which may be suitable in situations in which objects are composed of a set of well-specified, continuously varying dimensions. Researchers have documented, however, that more complex processes often enter into similarity judgment as well, including processes of analogy, alignment of parts, apprehension of relations, and feature interpretation (e.g., Gentner, 1983; Goldstone, 1994b; Medin, Goldstone, & Gentner, 1990, 1993). A fully specified exemplar-similarity model of categorization will need to incorporate a theory of similarity that reflects such influences. The challenge is whether the richly endowed theory of similarity is still sufficiently constrained so as to yield true predictions of categorization rather than merely circular, post hoc accounts (see Goldstone, 1994a, for extensive discussion bearing on these issues). Note, however, that the same basic challenge faces "theory-based" or "rule-based" models, which also require a groundwork of attributes and relations on which to operate. An advantage of testing between exemplar-similarity and rule-based models in simple perceptual domains, such as the ones considered in this article, is that there is often some consensus regarding the underlying groundwork, so rigorous comparisons between the models can be achieved. By contrast, because the psychological similarity space underlying most conceptual domains is highly multidimensional, complex, and flexible, achieving incisive tests between exemplar-similarity and rule-based models in such domains may prove to be exceedingly difficult.

REFERENCES

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

Anderson, J. R., & Lebiere, C. (1998). *The atomic components of thought*. Mahwah, NJ: Erlbaum.

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

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

Ashby, F. G., & Townsend, J. T. (1986). Varieties of perceptual independence. *Psychological Review*, **93**, 154-179.

Barsalou, L. W. (1985). Ideals, central tendency, and frequency of instantiation as determinants of graded structure in categories. *Journal of Experimental Psychology: Learning, Memory, & Cognition*, **11**, 629-654.

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

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

Carey, S. (1985). *Conceptual change in childhood*. Cambridge, MA: MIT Press.

Carroll, J. D., & Wish, M. (1974). Models and methods for three-way multidimensional scaling. In D. H. Krantz, R. C. Atkinson, R. D. Luce, & P. Suppes (Eds.), *Contemporary developments in mathematical psychology* (Vol. 2, pp. 57-105). San Francisco: W. H. Freeman.

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

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

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

Garner, W. R. (1974). *The processing of information and structure*. New York: Wiley.

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

Gentner, D. (1983). Structure-mapping: A theoretical framework for analogy. *Cognitive Science*, **7**, 155-170.

Getty, D. J., Swets, J. A., Swets, J. B., & Green, D. M. (1979). On the prediction of confusion matrices from similarity judgments. *Perception & Psychophysics*, **26**, 1-19.

Gibson, J. J., & Gibson, E. J. (1955). Perceptual learning: Differentiation or enrichment? *Psychological Review*, **62**, 32-41.

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

Goldstone, R. L. (1994b). Similarity, interactive interaction, and mapping. *Journal of Experimental Psychology: Learning, Memory, & Cognition*, **20**, 3-28.

Grunwald, P. (2000). Model selection based on minimum description length. *Journal of Mathematical Psychology*, **44**, 133-152.

Heit, E. (1994). Models of the effects of prior knowledge on category learning. *Journal of Experimental Psychology: Learning, Memory, & Cognition*, **20**, 1264-1282.

Heit, E. (1997). Knowledge and concept learning. In K. Lamberts & D. Shanks (Eds.), *Knowledge, concepts, and categories* (pp. 7-41). Cambridge, MA: MIT Press.

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

Hintzman, D. L. (1988). Judgments of frequency and recognition memory in a multiple-trace memory model. *Psychological Review*, **95**, 528-551.

Homa, D. (1984). On the nature of categories. *Psychology of Learning & Motivation*, **18**, 49-94.

Keil, F. C. (1989). *Concepts, kinds, and cognitive development*. Cambridge, MA: MIT Press.

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

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

Kruschke, J. 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, & Cognition*, **20**, 1003-1021.

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

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

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

Lovett, M. (1998). Choice. In J. R. Anderson & C. Lebiere (Eds.), *The atomic components of thought* (pp. 255-296). Mahwah, NJ: Erlbaum.

Luce, R. D. (1963). Detection and recognition. In R. D. Luce, R. R. Bush, & E. Galanter (Eds.), *Handbook of mathematical psychology* (pp. 103-189). New York: Wiley.

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

McKinley, S. C., & Nosofsky, R. M. (1995). Investigations of exemplar and decision-bound models in large-size, ill-defined category structures. *Journal of Experimental Psychology: Human Perception & 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 & Performance*, **22**, 294-317.

Medin, D. L., Goldstone, R. L., & Gentner, D. (1990). Similarity involving attributes and relations: Judgments of similarity and differences are not inverses. *Psychological Science*, **1**, 64-69.

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

Medin, D. L., & Schaffer, M. M. (1978). Context theory of classification learning. *Psychological Review*, **5**, 207-238.

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

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

Myung, I. J. (2000). The importance of complexity in model selection. *Journal of Mathematical Psychology*, **44**, 190-204.

Myung, I. J., & Pitt, M. A. (1997). Applying Occam's razor in modeling cognition: A Bayesian approach. *Psychonomic Bulletin & Review*, **4**, 79-95.

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

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

Nosofsky, R. M. (1986). Attention, similarity, and the identification–

categorization relationship. *Journal of Experimental Psychology: General*, **115**, 39-57.
Nosofsky, R. M. (1987). Attention and learning processes in the identification and categorization of integral stimuli. *Journal of Experimental Psychology: Learning, Memory, & Cognition*, **13**, 87-109.
Nosofsky, R. M. (1988). Exemplar-based accounts of relations between classification, recognition, and typicality. *Journal of Experimental Psychology: Learning, Memory, & Cognition*, **14**, 700-708.
Nosofsky, R. M. (1991a). Tests of an exemplar model for relating perceptual classification and recognition memory. *Journal of Experimental Psychology: Human Perception & Performance*, **17**, 3-27.
Nosofsky, R. M. (1991b). Typicality in logically defined categories: Exemplar-similarity versus rule instantiation. *Memory & Cognition*, **19**, 131-150.
Nosofsky, R. M. (1992). Similarity scaling and cognitive process models. *Annual Review of Psychology*, **43**, 25-53.
Nosofsky, R. M. (1998a). Optimal performance and exemplar models of classification. In M. Oaksford & N. Chater (Eds.), *Rational models of cognition* (pp. 218-247). Oxford: Oxford University Press.
Nosofsky, R. M. (1998b). Selective attention and the formation of linear decision boundaries: Reply to Maddox and Ashby (1998). *Journal of Experimental Psychology: Human Perception & Performance*, **24**, 322-339.
Nosofsky, R. M., Clark, S. E., & Shin, H. J. (1989). Rules and exemplars in categorization, identification, and recognition. *Journal of Experimental Psychology: Learning, Memory, & Cognition*, **15**, 282-304.
Nosofsky, R. M., Gluck, M. A., Palmeri, T. J., McKinley, S. C., & Glauthier, P. (1994). Comparing models of rule-based classification learning: A replication and extension of Shepard, Hovland, and Jenkins (1961). *Memory & Cognition*, **22**, 352-369.
Nosofsky, R. M., Kruschke, J. K., & McKinley, S. C. (1992). Combining exemplar-based category representations and connectionist learning rules. *Journal of Experimental Psychology: Learning, Memory, & Cognition*, **18**, 211-233.
Nosofsky, R. M., & Palmeri, T. J. (1996). Learning to classify integral-dimension stimuli. *Psychonomic Bulletin & Review*, **3**, 222-226.
Nosofsky, R. M., & Palmeri, T. J. (1997). An exemplar-based random-walk model of speeded classification. *Psychological Review*, **104**, 266-300.
Nosofsky, R. M., Palmeri, T. J., & McKinley, S. C. (1994). Rule-plus-exception model of classification learning. *Psychological Review*, **101**, 53-79.
Nosofsky, R. M., & Zaki, S. R. (1998). Dissociations between categorization and recognition in amnesic and normal individuals: An exemplar-based interpretation. *Psychological Science*, **9**, 247-255.
Palmeri, T. J. (1997). Exemplar similarity and the development of automaticity. *Journal of Experimental Psychology: Learning, Memory, & Cognition*, **23**, 324-354.
Palmeri, T. J., & Nosofsky, R. M. (1995). Recognition memory for exceptions to the category rule. *Journal of Experimental Psychology: Learning, Memory, & Cognition*, **21**, 548-568.
Pavel, M., Gluck, M. A., & Henkle, V. (1988). Generalization by humans and multi-layer networks. In *Proceedings of the Tenth Annual Conference of the Cognitive Science Society*. Hillsdale, NJ: Erlbaum.
Posner, M. I., & Keele, S. W. (1968). On the genesis of abstract ideas. *Journal of Experimental Psychology*, **77**, 353-363.
Reed, S. K. (1972). Pattern recognition and categorization. *Cognitive Psychology*, **3**, 382-407.
Rips, L. J. (1989). Similarity, typicality, and categorization. In S. Vosniadou & A. Ortony (Eds.), *Similarity and analogical reasoning* (pp. 21-59). Cambridge: Cambridge University Press.
Rips, L. J., & Collins, A. (1993). Categories and resemblance. *Journal of Experimental Psychology: General*, **122**, 468-486.
Ross, B. H. (1987). This is like that: The use of earlier problems and the separation of similarity effects. *Journal of Experimental Psychology: Learning, Memory, & Cognition*, **13**, 629-639.
Shepard, R. N. (1958). Stimulus and response generalization: Tests of a model relating generalization to distance in psychological space. *Journal of Experimental Psychology*, **55**, 509-523.
Shepard, R. N. (1964). Attention and the metric structure of the stimulus space. *Journal of Mathematical Psychology*, **1**, 54-87.
Shepard, R. N. (1986). Discrimination and generalization in identification and classification: Comment on Nosofsky. *Journal of Experimental Psychology: General*, **115**, 58-61.
Shepard, R. N. (1987). Toward a universal law of generalization for psychological science. *Science*, **237**, 1317-1323.
Shepard, R. N., & Chang, J. J. (1963). Stimulus generalization in the learning of classifications. *Journal of Experimental Psychology*, **65**, 94-102.
Shepard, R. N., Hovland, C. I., & Jenkins, H. M. (1961). Learning and memorization of classifications. *Psychological Monographs*, **75** (13, Whole No. 517).
Shin, H. J., & Nosofsky, R. M. (1992). Similarity scaling studies of dot pattern classification and recognition. *Journal of Experimental Psychology: General*, **121**, 278-304.
Sloman, S. A. (1996). The empirical case for two systems of reasoning. *Psychological Bulletin*, **119**, 3-22.
Smith, E. E., & Sloman, S. A. (1994). Similarity- versus rule-based categorization. *Memory & Cognition*, **22**, 377-386.
Smith, J. D., & Minda, J. P. (1998). Prototypes in the mist: The early epochs of category learning. *Journal of Experimental Psychology: Learning, Memory, & Cognition*, **24**, 1411-1436.
Smith, J. D., Murray, M. J., & Minda, J. P. (1997). Straight talk about linear separability. *Journal of Experimental Psychology: Learning, Memory, & Cognition*, **23**, 659-680.
Smith, L. B. (1989). A model of perceptual classification in children and adults. *Psychological Review*, **96**, 125-144.
Squire, L. R., & Knowlton, B. J. (1995). Learning about categories in the absence of memory. *Proceedings of the National Academy of Sciences*, **92**, 12470-12474.
Trabasso, T., & Bower, G. H. (1968). *Attention in learning: Theory and research*. New York: Wiley.
Vandierendonck, A. (1995). A parallel rule activation and rule synthesis model for generalization in category learning. *Psychonomic Bulletin & Review*, **2**, 442-459.
Wickens, T. D. (1982). *Models for behavior: Stochastic processes in psychology*. San Francisco: W. H. Freeman.

NOTES

1. Ashby et al. (1998, p. 468, note 13) justified their claim that the GCM fails to predict the bias toward the verbal rule as follows. The "equivocality contour" in the GCM is defined as the locus of points for which the model predicts that the probability of responding A equals the probability of responding B. On the basis of some previous theorems proved by Ashby and Maddox (1993), Ashby et al. (1998) noted that the equivocality contour in the GCM will be identical to the optimal decision boundary in Ashby and Townsend's (1986) general recognition theory (GRT) when the attention weights in the GCM are equal, $w_1 = w_2$. They then claimed that "predicted accuracy is maximized in the GCM when the equivocality contour equals the optimal decision bound," so the a priori prediction of the model is that subjects will allocate equal attention to each dimension (Ashby et al., 1998, p. 468). The flaw in their argument, we believe, is that there is no justification for the final claim, namely that performance in this paradigm is maximized in the GCM when its equivocality contour equals the optimal decision bound in GRT. Indeed, on the basis of investigations that we have summarized in Table 1, it appears that the claim is incorrect.

2. Because the GCM uses a probabilistic response rule, the model does predict some variability in individual subject performance even if all subjects are assumed to be governed by the same parameter settings. Nosofsky, Palmeri, and McKinley (1994) documented, however, that the probabilistic response-rule mechanism in and of itself was insufficient to account for the degree of heterogeneity in performance observed at the individual subject level.

3. Indeed, Erickson and Kruschke (1998, p. 111) found that in post-experiment interviews, many subjects reported using a salient rule that involved use of the numeric labels for responding to the exception stimuli. Because these stated rules were available in only some of the coun-

terbalanced conditions, Erickson and Kruschke eliminated subjects from these conditions in the main analyses reported in their article. However, it seems plausible that the numeric labels may have entered into other subjects' categorization strategies as well, even in those conditions in which the salient rule was unavailable.

4. We remark as well that a direct process interpretation for the emergence of the γ parameter was provided by Nosofsky and Palmeri (1997) in the form of their exemplar-based random walk (EBRW) model of speeded classification. In that model, γ turns out to correspond to the magnitude of the response criteria in a random walk process in which exemplars are retrieved from memory; that is, it corresponds to the amount of exemplar-based evidence that needs to be obtained before an observer will initiate a response (see Nosofsky & Palmeri, 1997, p. 291). Furthermore, Nosofsky and Palmeri (1997, pp. 270–271) noted that with $\gamma = 1$, the EBRW model made implausible predictions of classification response time, providing still further independent grounds about the importance of using values of $\gamma > 1$ in accounting for perceptual classification. Finally, we note that numerous other models in the field include response-noise parameters that are analogous to γ, including the response-scaling constant ϕ in Kruschke's (1992) ALCOVE model (see Kruschke, 1992, p. 24), the goal-value parameter G and response-noise parameter t in Anderson and Lebiere's (1998) ACT-R theory (see, e.g., Lovett, 1998, pp. 256–257, 276–277), and the criterial-noise parameter σ_c^2 in Ashby and Maddox's (1993) decision-bound theory (see Ashby & Maddox, 1993, pp. 377–378).

5. The experiments did not involve situations in which observers learned categories of perceptual objects via induction over training exemplars, so the study goes beyond the domain of inquiry we established for this article. Nevertheless, the study has had such a major impact on the field that some discussion is merited.

6. Interestingly, Goldstone (1994a, p. 143) provided examples suggesting that with various modifications in the exact wording of the similarity-judgment question, it is likely that different patterns of results may be obtained in these tasks.

7. As pointed out in numerous previous articles, the exemplar model reduces to a pure "likelihood-based" model when there is zero similarity between distinct exemplars—the summed similarity of an item to the category exemplars is given by the frequency with which the item was experienced in the category. When there is nonzero similarity, the model is often referred to as a "similarity-likelihood" model, and it still tends to yield predictions that follow the individual exemplar likelihoods—see Estes (1986), Nosofsky (1998a), and McKinley and Nosofsky (1995) for details.

8. For a much broader consideration of how similarity-based ideas may be applicable in wide and varied situations involving both perceptual and conceptual categorization, the reader is referred to Goldstone (1994a). For ideas regarding how effects of background knowledge in categorization may be modeled in terms of the storage of prior exemplars, the reader is referred to Heit (1994, 1997).

APPENDIX A
Classification Transfer Data From Erickson and Kruschke's (1998) Experimental Paradigm

The classification transfer data obtained in our follow-up of Erickson and Kruschke's (1998) Experiment 1 are reported in Table A1. The table reports the probability with which each transfer stimulus was classified in Categories 1–4, respectively. Recall that the categories are illustrated pictorially in Figure 6 (Category 1 = solid circles, Category 2 = solid squares, Category 3 = open circle, Category 4 = open square). The transfer stimuli consisted of all objects lying along the even-numbered columns of the secondary dimension (D_2) in the figure. The stimuli are numbered moving in an upward direction in each column, with Stimuli 1–10 lying in column 0, Stimuli 11–20 in column 2, Stimuli 21–30 in column 4, Stimuli 31–40 in column 6, and Stimuli 41–50 in column 8. Thus, T_E corresponds to Stimulus 1 and T_R corresponds to Stimulus 10. The "nearest-exception" category for T_E is Category 3 and the "nearest-exception" category for T_R is Category 4.

Table A1
Quantitative Fit of the GCM to Our Follow-up of Erickson and Kruschke’s (1998) Experiment 1

	Category Response					Category Response			
Stimulus	1	2	3	4	Stimulus	1	2	3	4
1	.52	.02	.46	.00	26	.21	.76	.01	.02
1	.58	.07	.34	.00	26	.13	.86	.01	.00
2	.52	.02	.46	.00	27	.06	.89	.00	.05
2	.58	.07	.34	.00	27	.07	.90	.00	.03
3	.33	.03	.64	.00	28	.02	.81	.00	.17
3	.47	.07	.46	.00	28	.04	.85	.00	.10
4	.61	.12	.26	.00	29	.01	.90	.00	.09
4	.76	.07	.17	.00	29	.01	.94	.00	.05
5	.65	.29	.07	.00	30	.01	.90	.00	.09
5	.81	.18	.01	.00	30	.03	.91	.00	.06
6	.19	.78	.02	.00	31	.92	.01	.06	.00
6	.15	.83	.02	.00	31	.92	.02	.06	.00
7	.08	.90	.01	.02	32	.92	.01	.06	.00
7	.05	.94	.01	.00	32	.94	.01	.05	.00
8	.03	.92	.00	.05	33	.87	.02	.11	.00
8	.03	.84	.01	.11	33	.84	.02	.10	.03
9	.02	.94	.00	.04	34	.90	.06	.04	.01
9	.01	.93	.01	.05	34	.90	.05	.01	.04
10	.02	.94	.00	.04	35	.77	.19	.01	.02
10	.01	.93	.02	.04	35	.83	.11	.01	.04
11	.48	.01	.50	.00	36	.28	.65	.00	.07
11	.49	.03	.48	.00	36	.16	.77	.00	.07
12	.48	.01	.50	.00	37	.09	.70	.00	.21
12	.36	.02	.61	.00	37	.04	.64	.01	.31
13	.28	.02	.70	.00	38	.02	.47	.00	.50
13	.16	.03	.81	.00	38	.02	.46	.00	.52
14	.55	.10	.35	.00	39	.01	.70	.00	.29
14	.55	.04	.41	.00	39	.02	.57	.00	.41
15	.59	.29	.11	.00	40	.01	.70	.00	.29
15	.75	.10	.15	.00	40	.02	.60	.00	.38
16	.17	.78	.03	.01	41	.94	.02	.04	.00
16	.13	.85	.01	.01	41	.94	.01	.05	.00
17	.06	.91	.01	.02	42	.94	.02	.04	.00
17	.04	.93	.02	.01	42	.93	.01	.05	.01
18	.02	.91	.00	.06	43	.92	.03	.05	.00
18	.02	.85	.03	.09	43	.93	.00	.06	.01
19	.01	.95	.00	.04	44	.90	.08	.02	.01
19	.00	.93	.01	.06	44	.90	.05	.03	.02
20	.01	.95	.00	.04	45	.78	.19	.00	.02
20	.01	.93	.01	.05	45	.84	.13	.01	.02
21	.86	.01	.13	.00	46	.29	.65	.00	.07
21	.89	.03	.08	.00	46	.14	.78	.01	.07
22	.86	.01	.13	.00	47	.12	.61	.00	.26
22	.84	.01	.15	.00	47	.07	.67	.00	.26
23	.69	.02	.29	.00	48	.03	.33	.00	.64
23	.76	.04	.20	.00	48	.02	.30	.00	.68
24	.82	.08	.10	.00	49	.02	.52	.00	.46
24	.91	.07	.02	.00	49	.03	.43	.00	.54
25	.69	.27	.04	.01	50	.02	.52	.00	.46
25	.81	.16	.01	.02	50	.02	.57	.00	.41

Note—Top line in each pair of rows gives the predicted classification probabilities from the generalized context model (GCM); bottom line gives the observed classification probabilities. Each probability is based on 96 observations.

APPENDIX B
Description of Prototype and Exemplar Models

The prototype model we fitted to J. D. Smith et al.'s (1997) Experiment 1 data is the one described in their article. According to the model, the probability that stimulus i is classified in Category A is given by

$$P(\mathrm{A} \mid i) = (1. - g) \cdot \sum_m w_m \delta_m(i) + g/2, \tag{B1}$$

where g ($0 < g < 1$) is a guessing parameter, w_m ($0 \leq w_m \leq 1$) is the weight given to dimension m, and $\delta_m(i)$ is an indicator variable set equal to 1 if stimulus i matches Prototype A on dimension m, and set equal to 0 if it mismatches. The dimension weights are constrained to sum to one. Thus, because the stimuli are composed of six dimensions, the model uses six free parameters (g and five freely varying weight parameters). Note that without the guessing parameter g, the prototype model predicts that the prototype stimuli, 111111 and 222222, will be classified into their respective categories with perfect accuracy. Thus, this parameter plays a major role in allowing the prototype model to fit the data (because the prototypes are often classified at levels below perfect accuracy). The exemplar model that was fitted to J. D. Smith et al.'s data was the one already described in Equations 1–3. It uses seven free parameters (γ, c, and five freely varying dimension weights). The guessing parameter was not used in fitting the exemplar model. The criterion of fit for both models was to minimize the $-\ln L$ statistic, given by Equation 6 in the text.

APPENDIX C
Optimal Sensitivity When Generalizing to New Instances From Low- and High-Variability Categories

To formalize the type of scenario illustrated in Figure 12, we created two normally distributed categories, each consisting of 100 "training" exemplars and 100 "transfer stimuli." Category A had a mean of 0 and a standard deviation of .01, .02, or .05 (the low-variability category). Category B had a mean of 1 and a standard deviation of 1 (the high-variability category). We then conducted a computer search for the values of c in the exemplar model that maximized its predicted percentage of correct classifications for the transfer stimuli. The upper limit for c was set at 20, and separate c parameters were allowed for each category. The response-scaling parameter in these investigations was set at $\gamma = 3$, which is a value that tends to give good fits of the exemplar model to classification data in a variety of experiments. The results are shown in Table C1, which reports, for each standard deviation value of the low-variability category (.01, .02, and .05), the optimal values of c and the predicted percentage of correct generalizations. The results can be summarized as follows. First, for the low-variability category, it is optimal to set c at its upper limit. However, for the high-variability category, it is optimal to set c at values substantially below its upper limit. This mechanism allows the exemplar model to correctly classify objects lying in the tails of the high-variability category. Such objects are closer in distance to the mean of the low-variability category than to the mean of the high-variability category, even though they were actually generated by the high-variability distribution. Finally, as the standard deviation of the low-variability category increases, the optimal value of c for the high-variability category increases as well, until it is finally optimal to set c at its upper limit.

Table C1
Optimal Values of the Sensitivity Parameter From the GCM for Generalizing to New Instances From Normal-Category Distributions With Differing Variabilities

σ	c_{L}	c_{H}	% Cor.
.01	20.00	2.70	97.6
.02	20.00	5.02	97.2
.05	20.00	7.43	95.8
.10	20.00	20.00	91.3

Note—σ, standard deviation of low-variability distribution; c_{L}, optimal sensitivity parameter for the low-variability distribution; c_{H}, optimal sensitivity parameter for the high-variability distribution; %Cor., predicted percentage of correct generalizations to new transfer stimuli.

(Manuscript received April 22, 1999;
revision accepted for publication December 7, 1999.)