Memory & Cognition
2004, 32 (3), 379-389

Measures of similarity in models of categorization

TOM VERGUTS
Ghent University, Ghent, Belgium

and

EEF AMEEL and GERT STORMS
University of Leuven, Leuven, Belgium

This paper concerns the use of similarities based on geometric distance in models of categorization. Two problematic implications of such similarities are outlined. First, in a comparison between two stimuli, geometric distance implies that matching features are not taken into account. Second, missing features are assumed not to exist. Only nonmatching features enter into calculations of similarity. A new model is constructed that is based on the ALCOVE model (Kruschke, 1992), but it uses a feature-matching similarity measure (see, e.g., Tversky, 1977) rather than a geometric one. It is an on-line model in the sense that both dimensions and exemplars are constructed during the categorization process. The model accounts better than ALCOVE does for data with missing features (Experiments 1 and 2) and at least as well as ALCOVE for a data set without missing features (Nosofsky, Kruschke, & McKinley, 1992). This suggests that, at least for some stimulus materials, similarity in categorization is more akin to a feature-matching procedure than to geometric distance calculation.

Having expectations about the behavior of living things and about the use of nonliving things is essential in daily life. Because experiences never recur exactly, a record of encountered experiences alone would be of little help to us in dealing with new problems or in anticipating the future. This is why *categorization*—the process that allows us to partition animals, plants, objects, experiences, and so on into groups—forms the essence of memory organization. It is also basic to all of our intellectual activities (Estes, 1994).

In recent years, a large number of formal models of categorization have been proposed. Many of these (e.g., Johansen & Palmeri, 2002; Kruschke, 1992; Lee & Navarro, 2002; Love, Medin, & Gureckis, 2004; Nosofsky, 1984; Nosofsky & Palmeri, 1997; Rosseel, 2002; J. D. Smith & Minda, 2000) derive from the exemplar model described by Medin and Schaffer (1978). Essentially, all these models construe categorization as a two-step process. First, the similarity of a newly presented stimulus to a number of stored exemplars is computed. (For clarity, from now on we will refer to the internally stored reference stimuli as *exemplars,* although in some of these models prototypes are used rather than exemplars.) Second, on the basis of these similarities, a decision is made as to which category the new stimulus belongs to. Op de Beeck, Wagemans, and Vogels (2001) recently advanced neurophysiological evidence for such a two-step procedure of categorization.

The similarity measure that is used in most of the formal categorization models is based on a *geometric distance* between the stimulus and the exemplar, in a *D*-dimensional coordinate space. Several criticisms have been raised against such spatial representations in the literature on similarity judgments (e.g., Sattath & Tversky, 1987; Tversky, 1977; Tversky & Gati, 1982). Tversky (1977) presented the well-known *contrast model* of similarity. In the contrast model, similarity is an additive function of the common and distinctive features of the two stimuli under investigation. Tversky also described the less well-known *ratio model*, in which similarity equals a function of the number of common features divided by a function of all features (both common and distinctive). This way of conceptualizing similarity as a feature-matching process has distinct advantages over the traditional geometric distance approach in many categorization situations. For the present purposes, the geometric distance approach contends with two problematic aspects. First, the geometric distance approach assumes that for all stimuli in a categorization context, all relevant dimension values are known; otherwise, the distance function cannot be calculated. However, unknown dimension values are rather common in real-life situations in which, for instance, perceptually presented stimuli can be seen from only one particular angle, or when limited information is given about verbally described stimuli. This will be called the *missing features problem*.

Part of this work was performed while T.V. was a postdoctoral researcher from the Fund for Scientific Research (Flanders) at the University of Leuven. E.A. is a research assistant of the Fund for Scientific Research. The contribution of G.S. was partly supported by Grants G.0266.02 from the Fund for Scientific Research and by Grants OT/01/15 and ZKB1578 from the Research Council of the University of Leuven. We thank Siegfried Dewitte, Paul De Boeck, and Koen Lamberts for their comments on the topic of this paper, and Robert Nosofsky for the use of his data. Correspondence concerning this article should be addressed to T. Verguts, Department of Experimental Psychology, Ghent University, H. Dunantlaan 2, 9000 Ghent, Belgium (e-mail: tom.verguts@ugent.be).

Second, if the feature values of two stimuli under comparison are equal, the difference in feature values is zero and, hence, does not influence the (additive) distance function that forms the basis of geometric similarity. Young and Wasserman (2002) have shown that this assumption is problematic. They showed that in a discrimination task, a simple discrimination (A– vs. B+) is easier than a discrimination with one common feature added to the two stimuli (XA– vs. XB+). The authors concluded that common features add to similarity and hence to categorization. This could be called the *matching features problem*. To handle both problems, we will present a model of categorization that is based on a feature-matching approach to similarity, akin to the ratio model of similarity mentioned above. Due to this feature-matching assumption, the two problems—of missing and matching features, respectively—disappear in a natural fashion.

The outline of this paper is as follows: First, we describe ALCOVE in detail because the new model we will propose is inspired by ALCOVE. However, our discussion of missing and matching features applies more broadly to all categorization models incorporating the critical geometric distance assumption (see, e.g., Nosofsky, 1984). Second, the two problematic implications of the geometric distance assumption are outlined in more detail. In the same section, a slight extension of ALCOVE is proposed, which deals with the first problem (but not the second). Then, a new model called "ADDCOVE" is described, which dispenses with the geometric distance assumption and avoids both problems in a straightforward manner. We discuss two experiments in which the problematic aspects of ALCOVE are highlighted and the validity of ADDCOVE is illustrated. Finally, we show that the new model also performs well with a more traditional category-learning experiment (Nosofsky, Kruschke, & McKinley, 1992, their Experiment 2).

ALCOVE

ALCOVE is a three-layer connectionist network model with input (i.e., feature) nodes, exemplar nodes, and output (i.e., category) nodes. Suppose a stimulus is presented in a categorization situation. The distance between this stimulus and a set of stored exemplars is then calculated. The stimulus is coded as $\mathbf{x} = (x_1, \ldots, x_D)$, which is represented in the input layer. An exemplar indexed by j is coded as $\mathbf{h}_j = (h_{j1}, \ldots, h_{jD})$, and

$$\text{dis} = \left[\sum_{d=1}^{D} \alpha_d \, | \, x_d - h_{jd} \, |^r \right]^{1/r}, \qquad (1)$$

where dis is the distance between the stimulus and the exemplar and the values α_d are attention weights and represent the amount of attention that is assigned to dimension d (or feature d: The terms *feature* and *dimension* will be used interchangeably). Furthermore, the parameter r determines the metric used in the distance calculation (e.g., $r = 1$: city-block metric; $r = 2$: Euclidean metric). The similarity between stimulus and exemplar is a decreasing function of this distance. Specifically, the similarity is equal to $\exp(-p \times \text{dis})$, where p is a scaling parameter. These similarities constitute a pattern of activation in the exemplar layer.

Once the relevant stimulus–exemplar similarities have been calculated, a category response is chosen on the basis of these similarities in the following way. Suppose there are two relevant categories of choice, A and B. If we denote the similarity between the stimulus and the exemplar j by $A_j^{\text{ex}}(\mathbf{x})$, then the evidence in favor of Category A is a linear function of activations A_j^{ex}:

$$A_{\text{A}}^{\text{cat}}(\text{x}) = \sum_{j=1}^{J} A_j^{\text{ex}}(\text{x}) w_{j\text{A}}, \qquad (2)$$

where it is assumed that there are J exemplars. The parameter $w_{j\text{A}}$ is the connection between exemplar j and Category A − response node in the output layer. These weights are adapted with the Widrow–Hoff learning rule. An equation analogous to Equation 2 holds for Category B, resulting in an output activation value $A_{\text{B}}^{\text{cat}}$. Finally, the probability of choosing Category A is obtained from a weighting of these two output values:

$$P(\text{A}) = \frac{\exp\left(\varphi A_{\text{A}}^{\text{cat}}\right)}{\exp\left(\varphi A_{\text{A}}^{\text{cat}}\right) + \exp\left(\varphi A_{\text{B}}^{\text{cat}}\right)}. \qquad (3)$$

The parameter φ scales the effect of the difference between the two category activation values $A_{\text{A}}^{\text{cat}}$ and $A_{\text{B}}^{\text{cat}}$: If φ is close to zero, probability $P(\text{A})$ will be close to .5; larger values of φ bring $P(\text{A})$ closer to one or to zero, depending on whether $A_{\text{A}}^{\text{cat}}$ or $A_{\text{B}}^{\text{cat}}$ is larger.

Two different versions of ALCOVE can be distinguished, depending on how exemplars are added to the network (Kruschke, 1992). In the first approach, a number of exemplars is chosen a priori that cover the input space (the *covering map* version). In the second approach, each new stimulus that is presented is added as a new exemplar. We will concentrate on the second version because, in the case in which participants do not practice the task (which is what usually happens in real life, as well as in our experiments), the covering map version seems implausible.

MISSING FEATURES, MATCHING FEATURES, AND GEOMETRIC DISTANCE

We consider the case of *additive features* (Tversky, 1977). These are binary-valued features of which the two values indicate presence or absence of a property—for example, presence or absence of headache in a medical context. The opposite of additive features are *substitutive features*, in which case there are at least two possible feature values (e.g., feature "length," with values *short* and *long*). As will be discussed later, substitutive binary features and continuous features can be recoded

as additive features, so there is no loss of generality if we restrict attention to the latter.

In addition to being *present* or *absent*, a feature can also be *missing*, in the sense that the value of that feature is not known. Consequently, there are three possible correspondences for the values of a particular feature in the comparison of two stimuli: (1)Two values are *matching* if both stimuli have the same feature value; (2) two values are *nonmatching* if the feature value is different for the two stimuli; and (3) one of the two values may be *missing*. Note that the term *missing* refers both to the value of one feature and to a correspondence between two features.

As was noted in the introduction, a first problem with the geometric distance measure is that it assumes that the values on all relevant dimensions are known, so there is no mechanism for handling missing features (in which case one or both of the feature values are unknown). This problem has already been ascertained by Estes, Campbell, Hatsopoulis, and Hurwitz (1989) and by Nosofsky et al. (1992) in the context of the ALCOVE model. These authors suggested that a dimension in Equation 1 should simply be dropped if values are missing on either the stimulus or the exemplar for that dimension. Since this is the only version of ALCOVE that has been described in the literature that can in principle cope with missing dimensions, we will focus on this version of the model in the following.

Ignoring a missing dimension is equivalent to assuming that a best possible match occurs on that dimension,[1] since then the term $\alpha_d |x_d - h_{jd}|^r$ in Equation 1 is zero for the dropped dimension, or $x_d = h_{jd}$. However, this leads to a problematic implication. Suppose there are two categories and we abbreviate the values of *presence*, *absence*, and *missing* as 1, 0, and ?, respectively. With features X_1 and X_2, stimuli such as ($X_1 = 1$, $X_2 = ?$), ($X_1 = ?$, $X_2 = 1$), and ($X_1 = 1$, $X_2 = 1$) cannot then be discriminated. This is because, on each dimension, each pair of stimuli either matches or has a missing value for one of the stimuli, and a missing value is equivalent to a match ($x_d = h_d$) when missing dimensions are dropped. Therefore, if at least one of the three stimuli belongs to a different category, high accuracy cannot be achieved on all three stimuli. In Experiment 1, we show that this prediction is contrary to empirical data. Therefore, Estes et al.'s (1989) and Nosofsky et al.'s (1992) solution to the problem of missing dimensions introduces new problems and so this solution is not optimal.

Instead of dropping a dimension when one or more features is missing, one could assume that a missingness parameter s is inserted if either the stimulus or the exemplar (but not both) have a missing value on a dimension.[2] The ALCOVE model extended with this assumption will be called "AlcoveMD" (i.e., ALCOVE with missing dimensions). It can be shown that AlcoveMD adequately solves the problem of lack of discriminability between the stimuli outlined in the previous paragraph.

The second problem to be discussed here is that, with a geometric-distance-based similarity measure, matching features are not taken into account. Indeed, from Equation 1 it is clear that matching features do not influence distance and, hence, do not influence the similarity between two stimuli. For example, if the stimulus and the stored exemplar match on dimension i, then $x_d = h_{jd}$ and so $|x_d - h_{jd}| = 0$, and the dimension might just as well be removed from the equation.

In a way, missing and matching dimensions are handled very similarly in ALCOVE (but not in AlcoveMD), since neither influences the similarity between stimulus and exemplar. In this sense, geometric distance deals only with nonmatching features. Recently, Lee and Navarro (2002) also criticized the geometric distance assumption in ALCOVE. They proposed changing the distance as calculated in Equation 1 with a distance based on binary features that are obtained from an independent additive clustering algorithm. Although this algorithm uses a common-features-based similarity, the relevant features were then used in a distinctive-feature distance measure and plugged into the standard ALCOVE model (see Lee & Navarro, 2002, for the rationale of this approach). Their nonspatial version of ALCOVE has its distinct merits, but from the present point of view, it also focuses on nonmatching features in categorization only.

To sum up, the assumption of existing models is that categorization works with nonmatching dimensions only (exceptions are addressed in the General Discussion). Whereas ALCOVE can be patched up to deal with missing dimensions (the resulting model being AlcoveMD), taking matching dimensions into account is less straightforward. In the next section, a new model is described that deals with all three feature correspondences (matching, nonmatching, and missing) in a natural fashion. Since the new model is inspired by ALCOVE, it is called "ADDCOVE."

ADDCOVE: ADDITIVE ALCOVE

Basic Principles

Like ALCOVE, the ADDCOVE model can be described as a three-layer connectionist network model with an input layer, an exemplar layer, and an output layer. In this model, a stimulus is processed as follows: The stimulus is first encoded as a list of additive (present/absent) features. In many categorization experiments, this assumption is quite natural, since the stimuli are shown to the participant in this format. If a feature appears that was not mentioned earlier, two extra input nodes (in the input layer) are created in the model: The first node becomes activated if the feature is present, and the second if it is absent (Gluck & Bower, 1988). If a particular feature is missing but nodes were already constructed earlier for that feature, neither of the two nodes is activated. Alternatively, an extra node could be created that codes for missingness of the feature. This approach is not pursued here, because it would lead to the implausible prediction that if two stimuli have a missing value on the same di-

mension, this would count as a match and hence increase similarity between the two stimuli.

The stimulus is thus coded as a pattern of binary values in the input layer. This pattern is then compared in parallel with all feature patterns that were already presented earlier. These earlier feature patterns are stored as separate nodes in the second, or exemplar layer of the model. If the new stimulus pattern is "sufficiently different" from all previous patterns, a new node is created in the exemplar layer that codes for this new pattern. That is, this new node becomes maximally activated when the new stimulus pattern is presented at the input layer. In any case, whether a new node is created or not, the stimulus at hand leads to a pattern of activation over the nodes in the exemplar layer. The activation value of each exemplar node indicates the similarity of the new stimulus to the exemplar (feature pattern) corresponding to that node. Here, the feature-matching approach to similarity that was referred to earlier is used. After this comparison process, the values in the exemplar layer are transformed such that the highest values (those indicating highest similarity to the stimulus) are relatively enhanced and the others are relatively diminished. This transformation could be considered a high-level description of a competition process between the different exemplar nodes (Love et al., 2004). Then this pattern of activation over the exemplar nodes is sent to the output layer in the same way as in ALCOVE (see Equation 2), and, finally, the probability of choosing one of the categories is determined in the same way as in ALCOVE (Equation 3).

The description in the previous paragraph shows that the model is an on-line model in the sense that nothing is assumed to be known prior to exposure to the task except, of course, the relevant categories. The input dimensions and exemplars are built from the ground up during the task. This is yet another advantage that is made possible through the use of a feature-matching procedure; indeed, with a geometric distance approach, dimensions cannot be added during the task because the distance function requires all relevant input dimensions for the calculation of similarity.

Model Equations

We now describe the activation equations in more detail. Let a stimulus be coded by a pattern of activation $\mathbf{x}^{\text{in}} = (x_1^{\text{in}}, \ldots, x_I^{\text{in}})$ over the input layer, where I refers to the number of input nodes. Activation A_j^{ex} of an exemplar node is then equal to

$$A_j^{\text{ex}} = \sum_{i=1}^{I} x_i^{\text{in}} w_{ij}^{\text{in}}, \tag{4}$$

where $j = 1, \ldots, J$ and J denotes the number of exemplar nodes. The set of parameters $\mathbf{w}_j^{\text{in}} = (w_j^{\text{in}}, \ldots, w_{Ij}^{\text{in}})$ is comprised of the weights of nodes feeding into exemplar node j. Input connections to an exemplar node are normalized in the sense that $\|\mathbf{w}_j^{\text{in}}\| = 1$, where $\|.\|$ denotes the (Euclidean) length function (see Equation 5 below). The normalization $\|\mathbf{w}_j^{\text{in}}\| = 1$ ensures that exemplar node j with vector $\mathbf{w}_j^{\text{in}}$ closest to the input pattern, relative to other exemplar nodes, reacts most strongly. Equation 4 is the reason that the model is called "additive ALCOVE," because similarity is computed from a sum rather than being based on a distance measure, as in Equation 1.

Learning in the input-to-exemplar mapping proceeds as follows: A newly created exemplar node j receives connections

$$w_{ij}^{\text{in}} = \frac{x_i^{\text{in}}}{\left\| \mathbf{x}^{\text{in}} \right\|}. \tag{5}$$

Hence, the length of $\mathbf{w}_j^{\text{in}}$ (i.e., $\|\mathbf{w}_j^{\text{in}}\|$) equals 1, as was claimed above. Furthermore, it follows that this exemplar node j will respond most strongly (over all exemplar nodes) to pattern $\mathbf{x}^{\text{in}}$ in the future, and its response will be equal to $\|\mathbf{x}^{\text{in}}\|$.

There is an additional node in the exemplar layer (indicated as node $J + 1$) with activation according to

$$A_{J+1}^{\text{ex}} = G\left\| \mathbf{x}^{\text{in}} \right\|. \tag{6}$$

The node $J + 1$ functions as a *novelty detector*: If the parameter G is close to 1, the activation of this node will be stronger than all other nodes for *novel* patterns. Indeed, the maximal response of an exemplar node is $\|\mathbf{x}^{\text{in}}\|$, and this occurs only if an earlier stimulus is shown again. Hence, if the activation of the novelty detector is larger than the activation of all exemplar nodes (Nodes $1, \ldots, J$), the pattern is judged to be a new one. For old patterns, its activation will be weaker than at least one of the other nodes. If G decreases toward zero, the model becomes "sloppy" and mistakenly judges new stimuli to be equal to old ones (which were already coded as exemplars).

The novelty detector can be thought of as expressing the amount of "surprise" experienced by the network. If the amount of surprise is sufficiently large, a new exemplar node is added. Specifically, a new exemplar node is added if $A_{J+1}^{\text{ex}} > A_j^{\text{ex}}$ for $j = 1, \ldots, J$. Hence, the number of exemplars increases throughout learning in an adaptive manner—that is, the number of nodes only increases if it is judged necessary (see also Love et al., 2004): Only "sufficiently different" patterns (depending on the value of G, the input pattern, and the old patterns) receive their own exemplar node. New input patterns that are close to old patterns will not receive a new exemplar node.

In the present study, the parameter G was set quite large (close to 1), so that a new exemplar is added for each distinct stimulus. In this way, the procedure may be considered a network implementation of the principle used in ALCOVE, in which a new exemplar node is recruited for each different stimulus. At the same time, it can be used as a generalization of the ALCOVE procedure if G is not very close to 1. Since G was set very large in the present study, it may be asked why it cannot be removed altogether and why we did not simply add a

new exemplar for each different stimulus. However, the inclusion of the novelty detector has an extra advantage, which will be described in the Implications section below.

The responses of the exemplar nodes (A_j^{ex}) are then normalized as follows:

$$x_j^{ex} = \frac{\left(A_j^{ex}\right)^{\varphi_1}}{\sum_{k=1}^{J+1}\left(A_k^{ex}\right)^{\varphi_1}} \tag{7}$$

for exemplar nodes $j = 1, \ldots, J + 1$. This is the transformation in the exemplar layer referred to above. These values x^{ex} represent the calculated similarities in ADDCOVE.

The parameter φ_1 has a positive value, and it either increases or decreases the differences between activations of the different exemplar nodes. In a sense, this parameter is similar to the specificity parameter (c) in ALCOVE (see Equation 2). In the limiting case, if $\varphi_1 = 0$, all exemplar node activations equal $1/(J + 1)$, so the differences are reduced to 0. At the other extreme, if φ_1 tends to infinity, the activation x^{ex} of the strongest node tends to 1 and the activation of all other nodes tends to 0, so the differences between exemplar node activations are increased. In this sense, the parameter φ_1 controls the amount of differentiation between (normalized) exemplar node activation values.

Once the similarities x^{ex} of Equation 7 are calculated, ADDCOVE is identical to the original ALCOVE model. The activations of category layer nodes A_A^{cat} or A_B^{cat} are a linear combination of the normalized exemplar responses x_j^{ex} as in Equation 2 above, resulting in activation values A^{cat}. These are normalized to yield response probabilities as follows:

$$P(\mathrm{A}|x) = \frac{\exp\left(\varphi_2 A_A^{cat}\right)}{\exp\left(\varphi_2 A_A^{cat}\right) + \exp\left(\varphi_2 A_B^{cat}\right)}. \tag{8}$$

The parameter φ_2 works similarly to φ_1 and corresponds to the parameter φ in ALCOVE. Learning in the exemplar-to-category mapping proceeds by the standard Widrow–Hoff (delta) rule, by which the weight change is proportional to the difference between the target (or required) value on an output node and the actual output value. The target values, representing the feedback provided by the experimenter, are of the "humble teacher" type described by Kruschke (1992). This means that activation values "better than necessary" are considered correct. For example, with targets 0 or 1, activation 1.2 is better than necessary when the target is 1.

In total, the model has four parameters: G, φ_1, and φ_2, described above, and β, which is a learning parameter for the exemplar-to-category weights used in the Widrow–Hoff rule. The parameter G is in all applications set to the constant $G = 0.9999$. This ensures, at least in the experiments reported in this paper, that for each distinct stimulus a new exemplar is created, as in ALCOVE. Therefore, there are only three effective parameters here: the two normalization parameters φ_1 and φ_2, and the learning rate parameter β.

Implications

We now consider the implications of the model equations of the previous paragraphs. Consider what happens with exemplar node $\mathbf{h}_j$ if a pattern $\mathbf{x}^{in}$ is presented. If we plug Equation 5 into Equation 4 (and replace $\mathbf{x}^{in}$ in Equation 5 with $\mathbf{h}$), the result is

$$A_j^{ex} = \sum_i x_i^{in} h_{ji} \,/\, \|\mathbf{h}_j\| = \frac{n \text{ match}_j}{\sqrt{n \text{ features in } \mathbf{h}_j}},$$

where the numerator is the number of matches between $\mathbf{x}^{in}$ and $\mathbf{h}_j$ and the denominator is the square root of the number of features in $\mathbf{h}_j$. When this is plugged into Equation 7, activation of the exemplar node j becomes

$$x_j^{ex} = \frac{\left(\dfrac{n \text{ match}_j}{\sqrt{n \text{ features in } \mathbf{h}_j}}\right)^{\varphi_1}}{\sum_k \left(\dfrac{n \text{ match}_k}{\sqrt{n \text{ features in } \mathbf{h}_k}}\right)^{\varphi_1} + \left(G\sqrt{n \text{ features in } \mathbf{x}}\right)^{\varphi_1}}. \tag{9}$$

Equation 9 shows that the new model is akin to Tversky's (1977) ratio model described above, since similarity is a function of the number of common features (n match) divided by a function of all features. Three properties are of interest here. First, *matching* features contribute to similarity in the model, due to the n match terms in Equation 9. Second, suppose that, in an existing configuration of stimulus, dimensions, and exemplars, a *nonmatching* feature is added to an exemplar $\mathbf{h}_j$ and the current stimulus $\mathbf{x}$. This will increase both the terms $G\sqrt{}$(n features in $\mathbf{x}$) and $\sqrt{}$(n features in $\mathbf{h}_j$), and all other terms are unaffected. Therefore, the overall similarity defined by Equation 9 will decrease. Third, consider what happens if the same particular feature is added to both the stimulus $\mathbf{x}$ and the exemplar $\mathbf{h}_j$, and a *missing* value for that particular feature is inserted for either the stimulus or the exemplar, but not for both. This will cause an increase in either the term $G\sqrt{}$(n features in $\mathbf{x}$) or the term $\sqrt{}$(n features in $\mathbf{h}_j$), and again similarity will decrease. Note that the novelty detector is needed for this to occur, which reveals the extra advantage of this node referred to above. If the value of this feature is missing for both stimulus and exemplar, no terms will change and similarity is then, of course, also unchanged. Note that missing values do not require an extra parameter in ADDCOVE (in contrast to AlcoveMD), but simply follow from the conceptualization of similarity as a feature-matching process. Hence, matching, nonmatching, and missing dimensions are dealt with in a unified manner in ADDCOVE.

From Equation 9, it also follows that if the stimulus and the exemplar are equal ($\mathbf{x} = \mathbf{h}$), x^{ex} can be brought arbitrarily close to 1 (the maximum value) by choosing a sufficiently high value of φ_1. This simple observation solves a problem with which earlier "additive" models of categorization have been confronted (e.g., Hayes-Roth & Hayes-Roth, 1977; Reed, 1972). In these models, an additive calculation of similarity was problematic in that if a Category A was required on presentation of features x and y together but not on presentation of one of the features separately, feature values f_x and f_y were required to be high in order to evoke the Category A response on simultaneous presentation (Medin & Schaffer, 1978). However, in a simple additive scheme this would also evoke a strong response on presentation of only one of the features. In Medin and Schaffer's seminal paper, the additive form was replaced by a multiplicative form, which solved the problem, and the multiplicative form is still the form chosen for most current models of categorization (e.g., ALCOVE). However, the problem confronted by additive similarity does not necessitate a multiplicative approach to similarity: The model expressed in Equation 9, in which exemplar strengths are normalized and multiplied by a factor φ_1, yields another solution. Other recent models that work with an additive similarity but avoid the problem of earlier additive models appear in the SUSTAIN model (Love et al., 2004) and the APPLE model (Kruschke, 1993).

EXPERIMENT 1

Experiment 1 provides a straightforward test of the missing-dimensions coding scheme proposed by Estes et al. (1989) and Nosofsky et al. (1992) for ALCOVE. Recall that they proposed that missing features are simply ignored in similarity calculation. This implies that stimuli such as ($X_1 = 1, X_2 = ?$), ($X_1 = ?, X_2 = 1$), and ($X_1 = 1, X_2 = 1$) will not be discriminated, since for each pair of stimuli each dimension is either matched or ignored, which amounts to the same thing in this coding scheme. It follows that if at least one of the three stimuli is assigned to a different category than the other two, high accuracy cannot be achieved on all three stimuli. This implication does not follow for either AlcoveMD or ADDCOVE: Both models can predict high accuracy on all three stimuli with appropriate parameter settings.

Method

Participants. Twenty-seven students from the University of Leuven participated for course credit.

Procedure. The experiment was set up as a medical patient diagnosis task, and on each trial the name of a fictitious patient was shown together with his/her symptom pattern, which could be ($X_1 = 1, X_2 = ?$), ($X_1 = ?, X_2 = 1$), or ($X_1 = 1, X_2 = 1$). The first two patterns belonged to the first category, and the third to the second. The symptoms *earache* and *dizziness* were used. For example, in one trial the stimulus could be "Mr. Jones has an earache and it is not known whether he feels dizzy," which corresponds to ($X_1 = 1, X_2 = ?$). Missing dimensions were mentioned explicitly because otherwise a missing symptom could be confused with absence of the symptom. A list of 24 patients (and their symptoms) was presented in a random order, but the order was the same for all the participants. The categories were two fictitious diseases called "agilentia" and "bogumitis," which were randomized over the two categories. The randomization factor had no effect and will not be discussed. The experiment lasted about 5 min for each participant.

Results and Discussion

Already in the first part of the test (Items 1–12), accuracy was high (probability correct = .74), and in the second part (Items 13–24) it was almost perfect (probability correct = .96). This high accuracy cannot be explained by ALCOVE, which cannot distinguish the three patterns, but it can be captured by AlcoveMD and ADDCOVE with appropriate parameter settings. Hence, it cannot be claimed that missing features were simply ignored.

EXPERIMENT 2

In Experiment 2, we intended to investigate how well the three models can cope with items of different dimensionality. A set of nine stimuli, each belonging to one of two categories, was shown repeatedly in the experiment, but the number of features shown on each trial varied from one to four. In a first block, only the first feature of each stimulus was shown, in a second block only the first two features were shown, and similarly for Blocks 3 and 4, in which three and four features were shown, respectively. The blocked presentation of dimensions (first one dimension, then two dimensions, and so on) was used to ensure that the low-dimensional stimuli were actually treated as such. Indeed, suppose items of different dimensionalities were presented in a random order and a four-dimensional item is presented before a particular one-dimensional item. Since the participants already know the four relevant dimensions at the point at which they are confronted with the one-dimensional item, one cannot be sure how the one-dimensional item is encoded. For example, the dimensions that are not shown for this item could be treated as missing, absent, or simply ignored. This ambiguity does not arise if all relevant dimensions are not yet known while the one-dimensional item is being classified.

To check the validity of the respective models, a cross-validation approach was taken. This allows the complexity of each model to be taken into account (Pitt, Myung, & Zhang, 2002). Parameters were estimated on the data of the first part of the test, and on the basis of these parameters categorization probabilities were predicted on the second part. Conversely, parameters were estimated on the second part and used for predicting probabilities in the first part. Pitt et al. have shown that cross-validation works very well in empirically distinguishing between different models with data sets as large as those in the present experiment (see their Table 5).

Method

Stimuli. We used the *5–4 stimulus structure*, which has often been used in the categorization literature (e.g., Medin & Schaffer,

Table 1
The 5–4 Data Structure

Pattern	Category	Dimension 1	Dimension 2	Dimension 3	Dimension 4
1	A	0	0	0	1
2	A	0	1	0	1
3	A	0	1	0	0
4	A	0	0	1	0
5	A	1	0	0	0
6	B	0	0	1	1
7	B	1	0	0	1
8	B	1	1	1	0
9	B	1	1	1	1

1978; J. D. Smith & Minda, 2000). The different possible data patterns of this structure with four binary-valued features are shown in Table 1. Patterns 1–5 belong to Category A, and Patterns 6–9 belong to Category B (hence the name "5–4 stimulus structure"). There is usually also a set of transfer patterns, but none was used here.

We introduced some changes relative to previous studies of the 5–4 structure. First, additive features were used rather than substitutive features. The second and more important change was that a different number of features was shown in different blocks of the experiment: In the first block, only the first feature was presented; in the second block the first two features; in the third block the first three; and in the last block all four features were shown. Hence, the first part of the experiment (Blocks 1 and 2) consisted of low-dimensional stimuli only (one or two dimensions), whereas the second part (Blocks 3 and 4) consisted of high-dimensional stimuli only (three or four dimensions). Following the optimal strategy, the expected probability of success is 78% (7 out of 9 items correct) in Blocks 1, 2, and 3 and 100% in Block 4.

The entire stimulus sequence consists of 252 items. Each of the four blocks contained 63 items; in each block, each of the nine stimuli (see Table 1) was presented seven times. Each participant received the same stimulus sequence. A similar procedure was used by Nosofsky et al. (1992, their Experiment 2). In fact, the abstract stimulus sequence that we used was equal to theirs, except for the two differences mentioned in the previous paragraph. As in Experiment 1, the category labels were the two fictitious diseases "agilentia" and "bogumitis," and the feature names were four familiar symptoms (stomachache, headache, earache, and muscular pain).

Participants. Twenty-three students of the University of Leuven (at Kortrijk) participated for course credit.

Procedure. The test was administered collectively with each participant seated in front of his or her own computer. All 252 stimulus patterns [e.g., (stomachache, no headache) in Block 2] were shown consecutively without breaks. However, the participants were allowed to rest between stimuli.

Three parameters were estimated for ALCOVE: the specificity parameter c, the learning rate for the weights λ_w, and the normalization constant φ. A previous analysis had shown that the learning rate parameter for the attention values λ_α should be restricted in order to obtain an identifiable model (Verguts & Storms, 2004) and without loss of generality we set this parameter to 1. For AlcoveMD, the missingness parameter s was also estimated. For ADDCOVE, the relevant parameters were φ_1, φ_2, and β. These parameters were estimated by the maximum likelihood method. Specifically, each trial was treated as a binomial experiment with 23 observations (the number of participants) with the two probabilities of choosing either of the two diseases. These probabilities were a function of the parameters of the model that was estimated. Note that this procedure makes abstraction of possible individual differences. To obtain the overall likelihood function, the product was taken over the binomial likelihoods for the 252 individual trials. As was noted earlier, the parameter G was set to 0.9999, so a new exemplar was created for each distinct feature pattern in ADDCOVE. To ensure comparability, also in ALCOVE and AlcoveMD a new exemplar was created on each trial in which a new feature pattern was shown.

The parameters of the three models were estimated on the first part (Blocks 1 and 2, Items 1–126) and the second part (Blocks 3 and 4, Items 127–252) separately, resulting in six sets of parameters. We calculated the root mean squared error (RMSE), which is the square root of the mean squared deviation between observed and predicted data (Pitt et al., 2002). This measure was computed over observed and expected response probabilities on the nine different stimuli in each part separately. Predicted response probabilities were calculated on the basis of the parameters estimated on the other part of the test. For example, when investigating how well ALCOVE accounted for the data in Part 1, we calculated RMSE over the probabilities on the nine stimuli in Part 1 but used the parameters obtained from Part 2 for that purpose. In addition to the cross-validation procedure, we also estimated parameters on the complete data set (Blocks 1–4) for each model and calculated RMSE for the complete data set on the basis of these parameters.

Results and Discussion

The overall mean accuracy over persons and items equaled .70, with a standard deviation of 0.09. For the different blocks, the accuracies were .63, .69, .71, and .76, respectively.

When calculated on the complete data set, RMSE values are about equal for the three models (.050, .045, and .048 for ALCOVE, AlcoveMD, and ADDCOVE, respectively). However, these values should be interpreted with caution because the complexity of models (e.g., number of parameters) is not taken into account.

The top panel of Figure 1 shows observed and (cross-validated) predicted probabilities of success over participants in the first two blocks for the three models separately. Table 2 shows the corresponding RMSE measures and parameter estimates. The probability of success is above .5 for Stimuli 1, 2, 3, 4, 7, 8, and 9, because they have the prototypical (i.e., modal) values of their categories on the first dimension (as can be seen in Table 1), and this is the only dimension that is shown in Items 1–63 and one of the two dimensions that is shown in Items 64–126. On the other hand, Stimuli 5 and 6 have a nonprototypical value on the first dimension (see Table 1), so their probability of success is below .5. As can be seen in the top panel of Figure 1, ALCOVE and AlcoveMD underestimate the amount of differentiation in Blocks 1

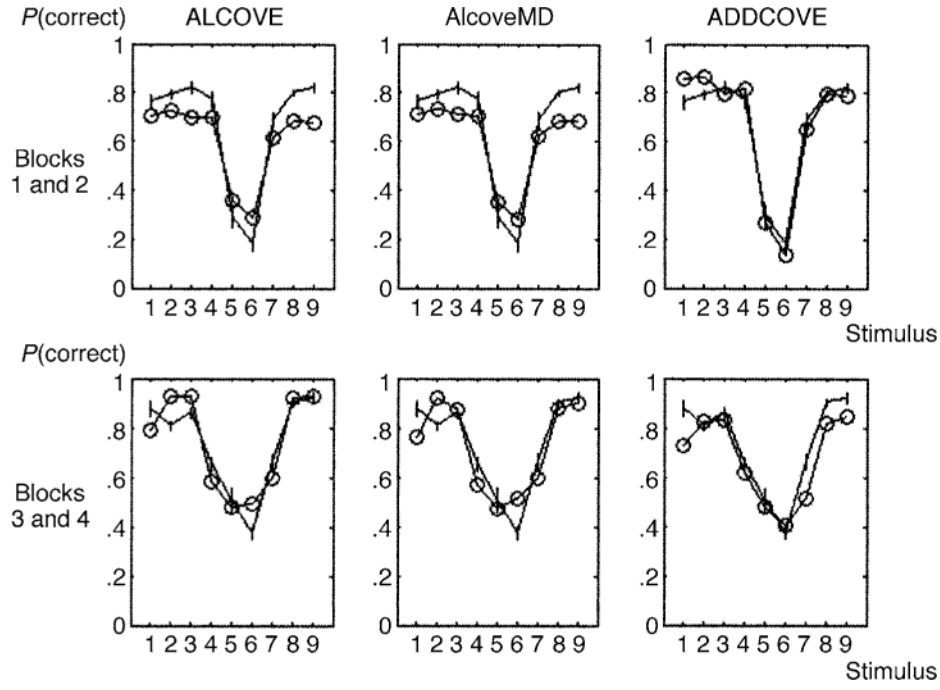

Figure 1. Lines with error bars (±1 standard error of measurement) denote observed probability of success for the different stimuli (abscissa) for the three models. Top panel, Blocks 1 and 2; bottom panel, Blocks 3 and 4. Open circles denote predicted probabilities.

and 2, whereas ADDCOVE does not. Correspondingly, the RMSE of ADDCOVE is almost half as small as that for the other two models (see Table 2). Also, both ALCOVE and AlcoveMD have very high values of the specificity parameter *c*, so only exemplars (feature patterns) that exactly match the current stimulus influence the response probability. Hence, both models produce an almost degenerate solution, in the sense that an essential aspect of ALCOVE—the influence of partly matching exemplars on the stimulus—does not play a role. In Blocks 3 and 4, the three models perform more or less the same (bottom panel of Figure 1); RMSEs for the three models are approximately equal (see Table 2).

All three models are able to account for the categorization of the high-dimensional stimuli (i.e., those with three or four features). However, ALCOVE and AlcoveMD can do so only at the cost of losing an essential aspect of that model, which is the fact that categorization is the result of a balance between different partly matching exemplars. For the low-dimensional stimuli (with one or two features), on the other hand, only ADDCOVE appears to account well for the data. The fact that the global RMSE (without cross-validation) is satisfactory and about equal for the three models shows that ALCOVE and AlcoveMD can be successful in finding a "compromise" in parameter settings between low- and high-dimensional stimuli. However, parameters estimated on high-dimensional items are not satisfactory for low-dimensional items.

ANALYSIS OF THE 5–4 DATA OF NOSOFSKY ET AL. (1992)

In the previous experiments, many of the stimuli had missing values on at least one dimension. It might be suspected that ADDCOVE shows an advantage in this regard because it has been constructed with these situations in mind, and that ADDCOVE might not account for more traditional data showing all features. In the last experiment, we investigate the behavior of ALCOVE and ADDCOVE in such a situation. Note that AlcoveMD reduces to ALCOVE in this case and will therefore not be discussed.

We used the data of the learning phase in Study 2 of Nosofsky et al. (1992). Forty participants categorized the 252 items, consisting of Patterns 2–9 shown in Table 1. In this case, substitutive (e.g., "color of hair") features were used, whereas ADDCOVE is designed to handle *additive* features (e.g., presence or absence of symptoms; Tversky, 1977). However, it is easy to recode substitutive features as additive features if, for each value of a substitutive feature, an input node is created in the network. For example, if the feature is "color of hair" with possible values *red* and *black*, a first node is active if hair color is red, and a second node is active if hair color is black.

The same parameter estimation and model evaluation methods were used as in Experiment 2. RMSE values calculated on the complete data set were .076 and .094

for ALCOVE and ADDCOVE, respectively. However, note again that model complexity is not taken into account in these measures. Cross-validated RMSEs and parameter estimates are shown in Table 2. The evidence here is not clearly in favor of either of the two models: In Part 1, ADDCOVE performs better, but in Part 2 ALCOVE performs better. Importantly, this shows that ADDCOVE (and its new similarity measure) is also a worthy competitor in accounting for data in which there is no missing stimulus information.

GENERAL DISCUSSION

This paper is concerned with similarity measures used in models of categorization. Although we have focused on ALCOVE for definiteness, our general conclusion that the geometric distance measure should not be taken for granted applies more generally. The contributions made in this paper are the following: First, it was shown that ALCOVE and its adaptations by Estes et al. (1989) and Nosofsky et al. (1992) are problematic with respect to missing dimensions. These authors deal with missing dimensions by ignoring them, and, hence, stimulus patterns such as ($X_1 = 1, X_2 = ?$) and ($X_1 = 1, X_2 = 1$) cannot be distinguished. Second, we have changed their assumption of ignoring missing dimensions and introduced a straightforward extension of ALCOVE (AlcoveMD). This model avoids the problem of indiscriminability between stimuli. Third, a new model, ADDCOVE, was presented, which handles matching, nonmatching, and missing features in a unified fashion. Finally, the empirical validity of the ADDCOVE model was tested in two new experiments and in Experiment 2 of Nosofsky et al. In the remainder of this paper, we briefly discuss how continuous dimensions can potentially be modeled in ADDCOVE. Then, we relate our model to a number of competing models in the literature. One original motivation for our work was to create a model that starts from scratch and gradually builds relevant knowledge as it is presented during a categorization experience. At the end of the paper, we return to this issue.

Continuous Dimensions

One concern about the rationale spelled out in this paper is that only additive features can be represented in ADDCOVE. However, this is not as limited as it may appear at first. Substitutive features can be recoded as additive features simply by creating a sufficient number of input units, one for each possible value of the (substitutive) feature, as we did for the Nosofsky et al. (1992) data. If this reasoning is taken to the extreme, continuous features can be handled as well (see also E. E. Smith & Medin, 1981). For example, if length is to be coded, one feature could code for the interval [0, 1), a second for the interval [1, 2), and so on. This procedure is called *place coding* (see, e.g., Kruschke, 1993), and it is used in some other models for continuous-valued dimensions (Kruschke, 1993; Love et al., 2004). Another scheme for recoding continuous features based on inclusion sets was suggested by Tversky and Gati (1982).

Related Models

ADDCOVE has both matching features and nonmatching features with a number of other models proposed in the literature. Another model that, like ADDCOVE, computes similarity in an additive manner is Kruschke's (1993) APPLE. Nevertheless, this model works very similarly to ALCOVE: In fact, weights between the input and exemplar layers are chosen such that ALCOVE is mimicked as closely as possible. Hence, the same problems as those discussed for the latter model apply here. Another model with an additive similarity function is SUSTAIN, developed by Love et al. (2004). This model has many similarities with ADDCOVE: For example, a competitive mechanism in the exemplar layer enhances differences between exemplar nodes (cf. Equation 7). Also, exemplar nodes are added only to the extent that they are necessary. However, there are also important differences. The exemplar nodes are situated in a multidimensional stimulus space, and their positions are adjusted during learning by a Kohonen unsupervised learning procedure. Hence, just as for other geometric distance measures, extra assumptions are needed to deal with missing dimensions or dimensions that are added only in a later phase; it remains to be seen whether SUSTAIN can be extended to cope with this.

Yet another model that bears similarities to ADDCOVE is the RULEX model of Nosofsky, Palmeri, and McKinley (1994). The assumption behind this model is that rules that selectively attend to just a single dimension are initially stored. For instance, a participant can store the rule that a particular value (e.g., *present*) on a particular dimension suggests membership in Category A, regardless of the values on the other dimensions. If one-dimensional rules prove unsuccessful, the model continues to look for good two-dimensional rules. If this is not successful either, complete or incomplete exemplars may be stored. This model has an obvious similarity to ADDCOVE: It stores complete or incomplete patterns and associates them with categories. There are also differences, however. First, whereas in RULEX rules and exemplars are treated as two different entities that are learned in different phases, the distinction between rules and exemplars is blurred in ADDCOVE. Indeed, suppose stimuli are four-dimensional: In this case, an "exemplar" node coding for a pattern such as ($X_1 = 1, X_2 = 0, X_3 = 0$) could be called an incomplete exemplar node because it represents an exemplar of which only three features are specified. Alternatively, it could be called a complex rule node because it specifies the values of three features (rather than those of just one feature, as in a simple rule such as "If X, then Y"). In this way, we feel that ADDCOVE can propose a unified view on the storage of rules and exemplars. Also, RULEX is formulated as an elaborate hypothesis-testing process, whereas ADDCOVE is formulated as a connectionist model. For this

Table 2
Cross-Validation Root Mean Squared Error (RMSE) Values and Parameter Estimates Used for Each RMSE Measure

Source of Data	Model	RMSE	Parameter Estimates Used in RMSE
Experiment 2, Part 1			
	ALCOVE	.097	$c = 14.716$, $\lambda_w = 0.034$, $\lambda_\alpha = 1$, $\varphi = 1.163$
	AlcoveMD	.091	$c = 5.527$, $\lambda_w = 0.035$, $\lambda_\alpha = 1$, $\varphi = 1.242$, $s = 0.013$
	ADDCOVE	.050	$\varphi_1 = 2.182$, $\varphi_2 = 1.774$, $\beta = 0.680$
Experiment 2, Part 2			
	ALCOVE	.076	$c = 8.213$, $\lambda_w = 0.093$, $\lambda_\alpha = 1$, $\varphi = 1.234$
	AlcoveMD	.082	$c = 6.117$, $\lambda_w = 0.080$, $\lambda_\alpha = 1$, $\varphi = 1.417$, $s = 0.030$
	ADDCOVE	.084	$\varphi_1 = 2.728$, $\varphi_2 = 1.325$, $\beta = 0.660$
Nosofsky et al., Part 1			
	ALCOVE	.196	$c = 1.654$, $\lambda_w = 0.099$, $\lambda_\alpha = 0.1$, $\varphi = 1.795$
	ADDCOVE	.106	$\varphi_1 = 1.135$, $\varphi_2 = 1.497$, $\beta = 3.589$
Nosofsky et al., Part 2			
	ALCOVE	.079	$c = 4.098$, $\lambda_w = 0.198$, $\lambda_\alpha = 0.1$, $\varphi = 0.880$
	ADDCOVE	.113	$\varphi_1 = 3.923$, $\varphi_2 = 0.695$, $\beta = 2.210$

Note —The parameter λ_α was always restricted. This parameter was fixed at a different value in the two experiments, however. The fact that a parameter is unidentified does not imply that every value of that parameter is suitable for each data set; in this case, solutions with $\lambda_\alpha = 1$ did not converge for the Nosofsky et al. (1992) data, so the parameter was set at 0.1.

reason, ADDCOVE is more in the spirit of other categorization models, such as ALCOVE and the generalized context model, which apply not just to stimuli composed of a small list of binary features, but also to stimuli that consist of a possibly large number of integral or separable continuous dimensions. Hence, we feel that ADDCOVE also holds more promise for a unifying processing framework of these very different kinds of stimuli.

Finally, the model that arguably is most similar to ADDCOVE is the configural cue model (CCM) of Pearce (1994). In his article, Pearce describes a number of "elemental" theories which hold that individual stimulus elements become connected to a response (e.g., the stimulus sampling theory, and the Rescorla–Wagner model) and shows that these elemental theories cannot account for a large number of findings in the (animal) conditioning literature. The most important problem is that these theories do not have a mechanism for computing "similarity" between different stimuli. He then proposes the CCM, which holds that the cluster of features on a particular trial is stored separately and is connected to a response. As in ADDCOVE, weights from the input layer feeding into such a constellation are normalized, so stimuli (in this case, situations) of different dimensionalities can be distinguished. There are also differences between the two models, however. First, the CCM assumes that the input is normalized, although this is not necessary. Also, the model does not impose the normalization in the exemplar layer (Equations 7 and 9), so it does not take into account nonmatching features but only matching features. Recently, Young and Wasserman (2002) have criticized the similarity measures used in both ALCOVE and CCM because that of the former uses only nonmatching features and that of CCM uses only matching features, whereas Young and Wasserman showed that both influence categorization performance. To account for their results, Young and Wasserman proposed normalizing attention in ALCOVE (cf. Kruschke & Johansen, 1999), so that if there are more dimensions, less attention is assigned to each particular dimension and the categorization task will become more difficult as a result. If one wants also to account for missing features, however, this option is not as straightforward as it may seem. For example, if a two-dimensional exemplar and a four-dimensional stimulus are compared (e.g., as in Experiment 2), how should attention be normalized? Different, equally plausible possibilities are to normalize over the common dimensions, over all dimensions, over the dimensions of the stimulus, or over the dimensions of the exemplar. In the context of Experiment 2, these reduce to only two possibilities, but in general they are distinct. Also, it is not clear whether the gradient descent rule of attention learning (Kruschke, 1992; Kruschke & Johansen, 1999) would still be valid. To sum up, we think ADDCOVE fills an unexplored and interesting niche in the space of categorization models that is worthy of further attention.

On-line Learning

An important facet of ADDCOVE is that it is an on-line model in the sense that both dimensions and exemplars are added during category learning. Most models of categorization suppose that participants start out with a great amount of information about the experiment. For example, in ALCOVE all relevant dimensions are assumed to be known; in the covering map version of ALCOVE, the relevant parts of the input space are known also because the input space is optimally covered with exemplars that are used for categorization. In contrast, ADDCOVE assumes only knowledge of the relevant categories. Input dimensions and exemplars are constructed during the task. In the experiments described in this paper, the novelty detector parameter G was set close to 1 to ensure that a new node was recruited for each different

stimulus pattern. However, choosing different values of G allows the model to add exemplars in an adaptive manner (see also Love et al., 2004; Rosseel, 2002), in the sense that exemplars are added only to the extent that they are sufficiently different from old stimuli. Furthermore, the stimuli that are added to the network are not always complete exemplars (as in ALCOVE) but may correspond to lower-dimensional stimuli if such stimuli are presented by the experimenter. Modeling the process of category learning with ADDCOVE remains an ambition to be fully worked out. However, acknowledging the existence and influence of matching, nonmatching, and missing features is a first step in that direction.

REFERENCES

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

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

Gluck, M. A., & Bower, G. H. (1988). From conditioning to category learning: An adaptive network model. *Journal of Experimental Psychology: General*, **117**, 227-247.

Hayes-Roth, B., & Hayes-Roth, F. (1977). Concept learning and the recognition and classification of exemplars. *Journal of Verbal Learning & Verbal Behavior*, **16**, 321-338.

Johansen, M. K., & Palmeri, T. J. (2002). Are there representational shifts during category learning? *Cognitive Psychology*, **45**, 482-553.

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

Kruschke, J. K. (1993). Human category learning: Implications for backpropagation models. *Connection Science*, **5**, 3-36.

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

Lee, M. D., & Navarro, D. J. (2002). Extending the ALCOVE model of category to featural stimulus domains. *Psychonomic Bulletin & Review*, **9**, 43-58.

Love, B. C., Medin, D. L., & Gureckis, T. M. (2004). SUSTAIN: A network model of category learning. *Psychological Review*, **111**, 309-332.

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

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

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. (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 category learning. *Psychological Review*, **101**, 53-79.

Op de Beeck, H., Wagemans, J., & Vogels, R. (2001). Inferotemporal neurons represent low-dimensional configurations of parameterized shapes. *Nature Neuroscience*, **4**, 1244-1252.

Pearce, J. M. (1994). Similarity and discrimination: A selective review and a connectionist model. *Psychological Review*, **101**, 587-607.

Pitt, M. A., Myung, I. J., & Zhang, S. (2002). Toward a method of selecting among computational models of cognition. *Psychological Review*, **109**, 472-191.

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

Rosseel, Y. (2002). Mixture models of categorization. *Journal of Mathematical Psychology*, **46**, 178-210.

Sattath, S., & Tversky, A. (1987). On the relation between common and distinctive feature models. *Psychological Review*, **94**, 16-22.

Smith, E. E., & Medin, D. L. (1981). *Categories and concepts.* Cambridge, MA: Harvard University Press.

Smith, J. D., & Minda, J. P. (2000). Thirty categorization results in search of a model. *Journal of Experimental Psychology: Learning, Memory, & Cognition*, **26**, 3-27.

Tversky, A. (1977). Features of similarity. *Psychological Review*, **84**, 327-352.

Tversky, A., & Gati, I. (1982). Similarity, separability, and the triangle inequality. *Psychological Review*, **89**, 123-154.

Verguts, T., & Storms, G. (2004). Assessing the informational value of parameter estimates in cognitive models. *Behavior Research Methods, Instruments, & Computers*, **36**, 1-10.

Young, M. E., & Wasserman, E. A. (2002). Limited attention and cue order consistency effects affect predictive learning: A test of similarity models. *Journal of Experimental Psychology: Learning, Memory, & Cognition*, **28**, 484-496.

NOTES

1. The attentional value α_d cannot be zero, since it would imply that the dimension is ignored altogether, in other stimuli as well.

2. We thank an anonymous reviewer for this suggestion.

(Manuscript received November 29, 2002;
revision accepted for publication October 22, 2003.)