Psychonomic Bulletin & Review
2007, 14 (4), 560-576

The divergent autoencoder (DIVA) model of category learning

KENNETH J. KURTZ
Binghamton University, Binghamton, New York

A novel theoretical approach to human category learning is proposed in which categories are represented as coordinated statistical models of the properties of the members. Key elements of the account are learning to recode inputs as task-constrained principle components and evaluating category membership in terms of model fit—that is, the fidelity of the reconstruction after recoding and decoding the stimulus. The approach is implemented as a computational model called DIVA (for DIVergent Autoencoder), an artificial neural network that uses reconstructive learning to solve *N*-way classification tasks. DIVA shows good qualitative fits to benchmark human learning data and provides a compelling theoretical alternative to established models.

A focal question in the study of cognition is how people learn and apply categories in order to understand and organize their experience. Despite considerable advances, there is no clear consensus among researchers on the psychological nature of categories—that is, how they are structured, how they function, and how they are acquired (Murphy, 2002). Theorists have variously proposed that categories are best understood as all-or-none rules, or as fuzzy resemblances; as abstracted associations from features to categories, or as stored individual exemplars; as descriptive relationships grounded in data, or as explanatory relationships grounded in theory-like knowledge (for reviews, see Goldstone & Kersten, 2003; Murphy, 2002). Models of category learning have gravitated toward the use of hybrid mechanisms in order to successfully accommodate a wide range of behavioral findings (for a proposed taxonomy of formal accounts of category learning, see Kruschke, 2005). The goal of this article is to introduce a novel theoretical framework and a process-level computational model that can capture key aspects of human category learning in a parsimonious fashion. The guiding notion is that categories are represented in terms of sophisticated yet economical task-constrained models of the statistical properties of their members.

Background on Human Category Learning

In currently favored accounts in the field, category representations consist of one or more *reference points* (Matsuka, 2004) that take the form of stored exemplars, summaries of the central tendency of a set of exemplars, or rule-like definitions. A reference point is a stored set of values for some or all of the features used to encode stimuli. An exemplar is a reference point because it is the set of feature values for one particular example; a prototype is a reference point because it is a summary of central feature values across a set of examples; and a rule can be a reference point if it specifies one or more feature values required for membership. These knowledge structures have in common their status as reference points because (1) they explicitly encode a set of feature values against which incoming stimuli are evaluated and (2) the inputs and the knowledge structures are encoded using the same feature vocabulary. The process of categorizing consists of activating or selecting reference points similar to the input. The reference points are directly or indirectly associated with categories, so that the relationship of a stimulus to the reference points determines the relationship of the stimulus to the categories.

In the rule-based, or *classical*, view (Smith & Medin, 1981), categories are definitions or sets of necessary and sufficient features that must be matched by potential members. The classical view is now largely historical (Medin, 1989; Murphy, 2002), but various contemporary approaches use logical rules (Erickson & Kruschke, 1998; Nosofsky, Palmeri, & McKinley, 1994), causal properties (Ahn, Kim, Lassaline, & Dennis, 2000; Rehder, 2003), or multivariate decision boundaries (Ashby & Maddox, 1993) to offer compelling accounts of selected category learning phenomena.

According to the *probabilistic* view, inputs are compared with either stored exemplars (Brooks, 1978; Medin & Schaffer, 1978; Nosofsky, 1986) or prototypes that capture the central tendency or feature likelihoods across category members (Hampton, 1979; Minda & J. D. Smith, 2001, 2002; Posner & Keele, 1968; Reed, 1972; Rosch & Mervis, 1975; J. D. Smith & Minda, 2000). The core mechanism is an evaluation of the match between an input and acquired reference points associated with category labels. Adaptive network models have been used to implement the probabilistic view of categorization at the

K. J. Kurtz, kkurtz@binghamton.edu

process level. These models operate by activating internal nodes that code for individual exemplars (or sets of exemplars) in accord with their attentionally weighted similarity to the input. The best-known reference point models use quasilocal internal nodes that operate as a specialized type of radial basis function (Kruschke, 1993; Poggio & Girosi, 1990). Other related accounts include the configural cue model (Gluck & Bower, 1988)—an error-driven, two-layer neural network that learns by directly associating input features and precomputed pairwise correlations of input features to category labels—and Anderson's (1991) rational model, which uses Bayesian optimization to establish an underlying clustering of examples that best predicts unobserved feature values. Models with localist internal nodes and dimensional selective attention have produced superior fits to human learning data (e.g., Kruschke, 1992; Nosofsky, Gluck, Palmeri, & Glauthier, 1994; Palmeri, 1999).

The best known reference point model is ALCOVE (Kruschke, 1992), a process-level implementation of the exemplar-based generalized context model (GCM—Nosofsky, 1986). ALCOVE extends the exemplar-based account by showing how dimensional attention strengths can be derived using an incremental learning procedure rather than post hoc parameter fitting. ALCOVE employs error-driven learning to update the dimensional selective attention weights and the association weights between the internal nodes (exemplars) and output nodes (categories). The success of ALCOVE in modeling the course of human learning has given considerable support to the point of view that attentionally weighted similarity to item-specific representations is the best available account (Kruschke, 2005). One criticism of ALCOVE is its demanding storage requirement of a dedicated hidden node for every input example. In addition, the modeler must create the internal representational space using advance knowledge of the details of the training set, or by implementing a covering map that explodes exponentially in size with increasing stimulus dimensionality (Kruschke, 1993).

Another reference point model, SUSTAIN (Love, Medin, & Gureckis, 2004), makes substantial advances in the explanatory range of the approach. While maintaining quasilocal encoding, dimensional selective attention, and error-driven learning, SUSTAIN addresses the need for flexibility in the structural and functional aspects of categories. The flexibility of SUSTAIN to accommodate a range of different learning modes derives from the use of a network architecture designed with the capacity to learn from and predict both features and categories. The structural clusters formed by SUSTAIN are dynamically constructed configurations that constitute a mix of specific and general forms of representation. SUSTAIN employs competitive reference point nodes that can code for prototypes, subprototypes, rules, or individual exemplars. The particular representational configurations that emerge are the result of online increases in structural complexity via recruitment of new internal nodes in response to trial failures.

SUSTAIN implements a multiplicity of representational constructs within a single unified mechanism. There has been a recent trend toward hybrid models that have a more stitched-together nature. ATRIUM (Erickson & Kruschke, 1998) extends ALCOVE by combining attention-mediated exemplar- and rule-based representations to account for a wider range of empirical results. RULEX (Nosofsky, Palmeri, & McKinley, 1994) employs a low-complexity to high-complexity search of the space of possible rules plus, as needed, a mechanism for memorization of exceptions. RULEX has been extended to handle continuous-valued stimulus dimensions (Nosofsky & Palmeri, 1998) but is limited to learning mutually exclusive, two-choice classifications (Love et al., 2004).

Three process models (ALCOVE, RULEX, and SUSTAIN) stand out for the quality and breadth of their fits to behavioral benchmarks. The explanatory core shared across these models (and absent in less successful ones) is computation of similarity to reference points coding for specific items, dimensional selective attention, and learning driven by the error between a category guess (response) and the correct category (target). The GCM (Nosofsky, 1986) and ALCOVE differ primarily in implementational terms, but a notable difference is that the GCM does not employ error-driven learning.

Despite unmatched success in fitting human data, reference point models are open to some criticism. One concern is the relatively high levels of model complexity. Another potential criticism of the network models (ALCOVE and SUSTAIN) is their use of quasilocal representations (Kruschke, 1992), as opposed to the distributed representations central to the "brain-style" cognitive architecture posited in the connectionist approach (Rumelhart, 1990; Rumelhart & McClelland, 1986). From the point of view of researchers interested in natural categories and concepts, reference point models may lack explanatory value beyond the domain of artificial classification learning experiments (Murphy, 2003, 2005). The *theory* or *knowledge* view of categorization (Murphy & Medin, 1985) offers a critique of the probabilistic view (and associated reference point models) as insufficiently constrained in its constructs and/or insufficiently powerful to account for a broad range of properties and uses of the conceptual system (see Goldstone, 1994). The claim is that concepts are organized, at least in part, in terms of top-down explanatory knowledge of *why* it is appropriate to assign equivalence to particular sets of examples or to compute similarity on the basis of a particular set of *respects* (i.e., features and weights). This approach questions the assumption that category representations or input representations are sufficiently constrained by the environment itself (Medin, 1989; Medin, Goldstone, & Gentner, 1993; Murphy & Medin, 1985; Wisniewski & Medin, 1994). Instead, categories and categorizations are considered to emerge not only from statistical regularities (i.e., similarity in the input space), but also from constraints inherent in the nature of the learner and the learning task.

Along these lines, a criticism of available models is that the psychological representations of experienced stimuli are fixed and established independently of the categorization process (Kurtz & Dietrich, 2007; Schyns, Goldstone, & Thibaut, 1998). An alternative idea found in the theory view and in perceptual learning perspectives on catego-

rization is that learning how to encode stimuli is deeply integrated into learning how to categorize them. ALCOVE (like the GCM) approximates a psychological encoding of stimulus items using a multidimensional scaling procedure. This does not actually explain the psychological mechanism for item encoding (Goldstone & Kersten, 2003); instead, it maintains a firm separation between item understanding and categorization. Addressing this issue, Kruschke (1992) emphasizes that ALCOVE is a model of classification performance, not of representation building, and implies that the two are incompatible. This separation is a potentially serious limitation because, on the exemplar view, concept representations consist of the stored instances themselves. Therefore, if the representation of the instances is unconstrained, so is the account of conceptual structure. The one basis for mediating stimulus encoding in reference point models is dimensional weighting, but this mechanism cannot construct item encodings; it can only assign levels of importance to available features. An alternative approach would be a mechanism that recodes inputs in an internal representational space as part of the category learning process.

A recent set of behavioral evidence presents an additional challenge to reference point models: Classification learning performance appears to be equally successful whether participants generate a classification response on each trial or simply study a correct category label provided with each stimulus presentation (observational supervised learning). Specifically, for unidimensional rule-based (Ashby, Maddox, & Bohil, 2002) and resemblance-based (Kurtz & Beck, 2007) categories, generating a response does not appear to be a consequential factor in the ease of category learning. On the other hand, Ashby et al. (2002) found that there was a difference using implicit category structures requiring information integration. In most reference point models, learning is driven entirely by the error between a response and the feedback, so if there is no response, there is no error signal to drive learning. An exception is the GCM (Nosofsky, 1986), which uses Hebbian rather than error-corrective learning, but does not set its attention weights through a learning process. In sum, a mechanism of error-corrective learning that generates an error signal based on something other than the classification response could serve as an important advance.

A New Approach

The focal idea of this article is that categories are sophisticated models of statistical regularities mediated by the structure of the environment and by the conditions of learning. The approach is motivated in part by a straightforward consideration: It would seem that recording only a summary of central tendency is too extreme in terms of information loss, and that recording each and every individual example independently is too extreme in terms of information retention. An example of an intermediate approach that uses a more sophisticated summary with less information loss is Fried and Holyoak's (1984) category density model, which uses both central tendency and variability information about category instances in order to evaluate relative likelihoods of category membership. The present goal is to develop an approach that generates sophisticated statistical abstractions, but does so in a manner that remains closely tied to specific exemplars (see Medin & Ross, 1989).

Unlike accounts that rely on the memorization or induction of reference points articulated in the input feature space, the present proposal is to transform inputs into a reduced-dimensionality representational space as part of learning statistical models of the categories. Principle component analysis (PCA) is a sophisticated technique for statistical analysis based on constructing low-dimensional recodings of the examples in a data set with minimal information loss (Joliffe, 1986). PCA offers impressive utility for data compression in terms of constructed variables that account for a large amount of variance. PCA is often used as a preprocessing stage to perform feature extraction on large inputs prior to applying a learning system that maps the recoded inputs to their class labels (Becker & Plumbley, 1996; Edelman & Intrator, 1997; Zhang, 2000). However, traditional PCA has not been seen as suitable on its own as a basis for classification (Duda & Hart, 1973). The problem is that the information maximization procedure is not constrained by category-level organization and can be substantially at odds with its preservation. Chen and Sun (2005) have suggested one way to address the failure of standard PCA to take advantage of class label information by adding the category label to the input vector for each training example.

Another approach to applying the powerful pattern recognition capability of PCA to classification is to build on the functionality of recoding and decoding examples. Specifically, for each trial the outcome of the recoding/decoding procedure on a particular item can be compared to the original input form of the item. If an input item is sufficiently well recovered after recoding and decoding, it can be considered a good member of the category defined by all items in the data set. Classification tasks that take the form of member/nonmember judgments can be effectively solved in this manner (Japkowicz, Myers, & Gluck, 1995; Oja, 1983). Oja (1989) developed a more general procedure for *N*-way classification by applying PCA to each subclass of a classification problem using a set of independent processing modules. Classification decisions are generated by determining which module produces the most accurate projection of a test input.

This recoding/decoding mechanism of PCA holds promise as an account of human category learning. The basic claim is that people judge category membership by determining how well the statistical model underlying a category accounts for the data. For example, a dog stimulus will be well accounted for (i.e., recoded and decoded with minimal information loss) by the dog category, but not by the chair category. Although it is likely that the dog stimulus is similar to previously experienced dogs, the degree of featural match is not the basis for the membership decision. Instead, the statistical model underlying the dog category is somewhat akin to an implicit theory about dogs, although it is a theory expressed entirely in the language of data.

Medin's (1989) framework for specifying accounts of categorization provides a useful way to clarify the psychological nature of the proposal that categories function

as PCA-driven statistical models. The *concept representation* is a task-constrained statistical model that implements principal component analysis to optimally preserve the form of the data. The *categorization basis* is the relative degree of success in reconstructing the stimulus when submitted to the recoding/decoding procedure—that is, the ability of the model to account for the stimulus. The *unit of analysis* is the stimulus feature, although the core action of the categorization process is to recode the input in a derived multidimensional space. The *weighting of attributes* is a natural consequence of PCA, but, importantly, there is no dimensional constraint on the attentional mechanism. The issues of *interconceptual structure* and *conceptual development* are beyond the scope of the present discussion, but extending the present approach to these topics is a longer term goal.

This explanatory framework is implemented within the connectionist tradition of brain-style computation. *Brain-style* refers to the use of a collection of simple, connected, neuron-like nodes that encode content in a distributed fashion and represent knowledge in the weighted connections between nodes. Brain-style computation showed initial promise as an account of human category learning and associative memory (e.g., Gluck & Bower, 1988; Knapp & Anderson, 1984; McClelland & Rumelhart, 1985; Shanks, 1991), but these models have not fared well on benchmark tests and fail to handle nonlinearly separable classification problems.

Autoencoders are a class of artificial neural networks that function as powerful self-supervised learning devices (McClelland & Rumelhart, 1986). Such networks are trained *autoassociatively* (Anderson, Silverstein, Ritz, & Jones, 1977; Kohonen, 1977), using a bottleneck hidden layer with lower dimensionality than the input and output layers to rerepresent and then reconstruct the input information at the output layer. Learning of this type yields an impressive range of psychologically relevant behaviors, including recognition, recall, generalization, inference, and distortion (Rumelhart, 1989). As a subclass of multilayer, feed-forward neural networks, autoencoders are trained using the backpropagation learning algorithm (Rumelhart, Durbin, Golden, & Chauvin, 1995; Rumelhart, Hinton, & Williams, 1986). Autoencoders have been used for data compression (e.g., Cottrell, Munro, & Zipser, 1988) and have been applied by psychologists to model correlational sensitivity (Mareschal & French, 2000) and asymmetric sequential learning (Mareschal, Quinn, & French, 2002) in human infants. Several models exist (Gluck & Myers, 1993; Intrator & Edelman, 1997; Kurtz & Smith, 2007) that combine an autoencoder with a *heteroassociative* learning module that performs a mapping to output nodes predicting variables other than the input features (i.e., classes).

The autoencoder is an excellent choice for constructing a statistical model of a set of training examples, such as the members of a single category. In specific terms, an autoencoder with a linear activation rule is formally equivalent to PCA (Baldi & Hornik, 1989). In addition to implementing PCA with trial-based learning in a brain-style manner, autoencoders naturally produce the recoding/decoding functionality described above. The activations at the output layer can be used to evaluate the goodness of fit of a test item to the collective training set. When the training set consists of members of a category, the quality of the reconstruction effectively evaluates category membership.

The applied utility of this approach has been demonstrated using a two-choice supervised classification task. Japkowicz (2001) trained an autoencoder network on instances of a single category. Inputs were classified by evaluating the reconstructive success of the autoencoder: Successful reconstructions were classified as members of the training category, whereas poor reconstructions were designated as members of the alternative category. The success of Japkowicz's approach in a machine learning context is encouraging, but the following problems exist in applying this formulation to cognitive modeling: (1) It is not extensible to *N*-way classification; (2) it requires mutual exclusivity of classes; (3) it is not clear how to decide which category to define positively and which negatively (nor whether this is psychologically appropriate); and (4) it assumes independent rather than interdependent categorization. To amplify this last point, a standard autoencoder will learn exactly the same representation of the positively defined category, regardless of the nature of the contrast category.

Until now, the autoencoder has not been theoretically related to human category learning. This is likely a consequence of the seeming mismatch between the tasks of *feature* prediction and *category* prediction. Traditionally, self-supervised autoassociative learning architectures are used to construct statistical models of a training set; by contrast, externally supervised, heteroassociative learning architectures are used to perform classification and regression tasks that predict a variable outside of the input feature space. However, this distinction begins to break down under the view of categories as statistical models and the idea that an autoencoder naturally evaluates the goodness of fit of an input relative to the category it represents.

An additional factor in considering the autoencoder as an account of human category learning is the use of the backpropagation learning algorithm. Backpropagation networks have shown an impressive range of explanatory power in cognitive psychology (Ellis & Humphreys, 1999; McClelland & Rumelhart, 1986; Rogers & McClelland, 2004). The algorithm consists of a generalized delta rule that solves the credit-assignment problem with multilayer architectures. Output error is propagated backward through the network to incrementally adjust weights and perform gradient descent through error space. The essential feature of the learning algorithm is the construction of internal representations—task-driven recodings of inputs—that allow backpropagation networks to act as universal function approximators. A widely raised critique of backpropagation is biological implausibility. Although this has not traditionally been a priority in formal modeling, it is likely to be increasingly considered (e.g., Ashby & Maddox, 2005). It is therefore worth noting that functional equivalents to backpropagation have been developed which do not contradict current neurobiological understanding (O'Reilly, 1998; Xie & Seung, 2003).

A more focal concern is that backpropagation has, to this point, given an extremely poor account of human category learning (Kruschke, 1992, 1993; Love et al., 2004; Palmeri & Noelle, 2002). Although ALCOVE and SUSTAIN also perform error reduction using the delta rule and a feedforward network architecture, these models are pointedly distinguished from backpropagation, primarily in terms of the activation function of the hidden nodes. As will be seen below, the psychologically implausible computational properties of backpropagation turn out to be confined to its use within the traditional multilayer perceptron (MLP) architecture, in which classification is implemented as the heteroassociative learning of a mapping from input features to category outputs. The term *standard backpropagation* will be used in the following to refer specifically to networks with an MLP architecture and linear-logistic hidden nodes.

Architecture and Design Principles of DIVA

DIVA (Kurtz, 2005) is a process model of human category learning that uses the core innovation of *divergent autoencoding* as a basis for applying reconstructive learning to any classification problem.[1] The theoretical claim that categories are task-constrained statistical models is implemented in DIVA. Although the theoretical framework is not inextricably tied to the specific implementation, they are tightly linked in the present discussion. It is possible that the computational level of the account can be effectively realized in alternate terms at the algorithmic level (Marr, 1982).

DIVA is a fully connected, feed-forward connectionist model that uses backpropagation to perform error-driven learning. As an extension of the standard autoencoder, DIVA includes an input layer for stimulus features, a lower dimensional hidden layer, and an output layer at which the input is reconstructed on the basis of a target signal identical to the input. The computational advance in DIVA is a way of addressing any supervised classification problem in terms of reconstructive learning, and of doing so under the mediation of task constraints. Modeling categories as independent autoencoders (along the lines of Oja, 1989) will not account for human learning. The form of a category representation must be mediated by the conditions of learning—that is, the task, the learning mode, and the nature and number of contrasting categories. This is accomplished using multitask learning with a single shared hidden layer (Ben-David & Schuller, 2003; Caruana, 1995). In divergent autoencoding, one reconstructive learning channel is dedicated to each category in an *N*-way classification problem. The category channels are integrated by a shared hidden layer for recoding the input. Accordingly, the architecture of a DIVA network (see Figure 1) consists of a single input layer, a single shared hidden layer, and a set of *N* autoassociative output banks. The coordinated process of conducting PCA in parallel on each category with a shared recoding space means that an additional set of constraints is enforced during the computation; that is, DIVA does not identify principle components for all of the training examples as if they were in a single category, nor does DIVA find the principle components for each category separately. Instead, DIVA finds a set of weights to *recode* all members of the training set (across all categories) in a form that can be *decoded* by category-specific channels realized at the hidden-to-output connections. The quality of the reconstruction (the recoding and decoding) serves as the basis for error-driven learning and for making a classification choice. As a result, the error signal that drives learning is not dependent on the classification choice. The error is based on the quality of the feature reconstruction on the correct category channel, not on the difference between the guessed category and the correct category.

In keeping with traditional backpropagation networks (Rumelhart et al., 1986), the activation rule for the hidden and output nodes is a linear-logistic (sigmoid) function. This nonlinearity yields a useful generalization of standard PCA. Although Boulard and Kamp (1988) concluded that the use of nonlinear hidden nodes neither enhances nor diminishes the computational power of the autoencoder, Japkowicz, Hanson, and Gluck (2000) found advantages of nonlinear nodes relative to linear nodes (standard PCA) in the learning of nonlinear functions and in the degree of sensitivity to complex statistical regularities. Japkowicz et al. (2000) observed this advantage with either linear or nonlinear output nodes, as long as the hidden nodes were nonlinear. The autoencoder with nonlinear nodes is also related to sandglass-style autoencoders, which employ multiple hidden layers to formally implement the statistical technique of nonlinear PCA (NLPCA). Such systems have performed effectively on applied problems (Kramer, 1991; Saegusa, Sakano, & Hashimoto, 2004).

To summarize, DIVA uses divergent autoencoding to extend the autoencoder implementation of PCA to the general form of supervised classification problems. The model conforms to a theoretical view of category learning as the formation of coordinated statistical models that operate by recoding and decoding inputs. Unlike models that compute average feature values for each category or acquire feature-to-category associations, DIVA builds a sophisticated statistical model of each category via recoding examples in terms of principle components. Interdependence arises from the coordinated statistical learning, because the representation of each category is mediated by the alternative categories in the learning task.

By way of comparison to reference point models, DIVA addresses category learning tasks by learning distributed internal representations of stimulus items and does not maintain a strict representational commitment to fixed, externally derived stimulus encodings. DIVA has the ability to weight diagnostic predictors from the input, but does

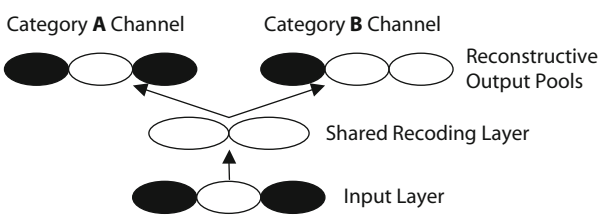

Figure 1. The DIVA model.

so without an explicit, dimensionally constrained mechanism of selective attention. (To clarify this point, reference point models such as ALCOVE employ a single, shared attention weight for all of the outgoing connections from a particular input node; by contrast, the input-to-hidden weights in standard backpropagation and DIVA vary independently, and their function is to perform a recoding of the input in a constructed representational space.) DIVA is sensitive to specific individual exemplars, but without internal nodes that code for particular items in terms of sets of values on input features. DIVA is resemblance sensitive, but never performs an explicit similarity computation between the input and stored reference points. Finally, DIVA is error driven, but the error is generated by comparing the reconstruction of the input to the original, not by comparing a category guess to the correct label.

Training and Testing DIVA

As with most network models, each stimulus is encoded in terms of activation levels on a set of input nodes representing feature values. The input activations can code for discrete or continuous-valued features, although the present research focuses on simulating learning problems with binary features. In DIVA, full feed-forward connectivity is used, so each hidden node receives a signal from each input node, and each output node of each channel receives a signal from each hidden node. The linear-logistic activation function of the output nodes is scaled to match the appropriate target values for the learning task (-1 to $+1$). The input values serve as the targets for error-driven reconstruction, but they are applied only along the channel of the *correct* category. The correct category label (information included as feedback in every trial of supervised learning) is used to determine which category channel to update. This differs from the conventional feedback mechanism used in systems with output nodes representing categories. First and foremost, the error signal is computed strictly in terms of reconstructive success, not classification success. Additionally, the error at each output node is not always applied to generate weight change; this occurs only along the one selected channel. The weight updates incrementally improve the ability of the channel representing the correct category to accurately reconstruct the current training example. Therefore, instead of optimizing a function for mapping features to categories, the supervised classification learning task is modeled by collectively (because of the shared hidden layer) building statistical models for each category. The category representations are inherent in the set of weights that recode and decode inputs.

Although the processing that occurs at the level of the individual node is exactly the same between standard backpropagation and DIVA, the performance characteristics of the divergent autoencoding architecture differ dramatically from a traditional multilayer perceptron. Specifically, DIVA does not transform the inputs into a linearly separable space in the service of category nodes at the output layer. Instead, DIVA learns a set of recodings that are optimized for correct reconstruction of features in accord with each category channel. In a standard autoencoder, the outcome of this process is a set of recodings that achieve *maximal interstimulus separation*, or optimal discriminability of each training example (Harnad, Hanson, & Lubin, 1995). Initial item representations cluster in the center of the recoding space and progressively disperse as singletons toward the boundaries of the representational space. In divergent autoencoding, a multiplicity of such discrimination spaces are superimposed one upon another, since each of the category-specific transformations is supported by the same set of input-to-hidden weights. The ease with which an N-way classification is learned is a direct consequence of the ease of generating N satisfactory statistical models under these constraints. The ease with which particular items are learned, and the accuracy with which test items are classified, can be predicted by the degree of difficulty in accommodating the items within the statistical model of each category.

Backpropagation networks traditionally use a learning rate parameter (a multiplicative factor of 0.1, 0.01, etc.) to ensure the stability of gradient descent and to lower the risk of getting stuck in local minima (Rumelhart et al., 1986). From the standpoint of cognitive modeling, it is often considered a weakness of the connectionist approach that a single learning trial or stimulus exposure in the real world is simulated in terms of many incremental training trials for the network. Unlike standard backpropagation networks, DIVA simulations are conducted (using a learning rate of 1.0), with each real-world trial corresponding to a single training trial for the network.

In order to generate a classification response on the basis of the output node activations, DIVA selects the best reconstruction—that is, the lowest sum-squared error (SSE)—across the set of category channels. No threshold parameter is required. Luce's (1963) choice rule is used to generate response probabilities with one important modification to suit the DIVA approach: Rather than using the activation level of the category nodes (e.g., Kruschke, 1992), the current choice rule uses the inverse of the SSE on each channel. Therefore, DIVA predicts category membership by selecting for low reconstructive error on a category channel rather than for high activation on a category node. The probability of selecting category K from among N choices is given by Equation 1:

$$Pr(K) = [1 / SSE(K)] / \sum_{k=1}^{N} [1 / SSE(k)]. \qquad (1)$$

To illustrate the functioning of the model, consider a two-choice (A/B) classification task. One channel is assigned to reconstruct the inputs labeled A, and a second channel is assigned to reconstruct the inputs labeled B. It is helpful in describing particular DIVA networks to use a summary shorthand: the designation ($3 - 2 - 3 \times 2$) indicates, from left to right, a DIVA network with three input features, two hidden nodes, and three output features (mirroring the inputs) reconstructed along two channels (as in Figure 1). As described above, the target signal (the corrective category feedback in the learning task) is used to select which category channel to train on a particular trial. Weight update is driven only by the error signal gener-

ated on the selected bank of output units. Accordingly, the shared input-to-hidden weights are collectively adjusted on every trial, whereas the hidden-to-output weights are adjusted independently for each category channel. Two implications of this design are that DIVA makes error-driven updates along the correct category channel even when it has achieved the correct category response (unless the reconstruction was perfect), and that DIVA never makes a change along an incorrect category channel, even when an incorrect category response has been selected.

These two characteristics of the model each generate an interesting psychological hypothesis. First, it should be possible for a learner to acquire a classification scheme without making any classification errors along the way, because category formation takes place on each trial regardless of the response outcome. On the basis of informal observation, this phenomenon is not so uncommon, particularly when human learners are taught simple categories and make a lucky guess or two on the initial trials. Second, the learning that arises from an incorrect trial should result in greater improvement in the correct category than in the incorrectly guessed category. By contrast, traditional error-corrective models tend to make adjustments such that the incorrect response becomes less likely and the correct response more likely. These predictions will be addressed in future research.

Modeling Practices and Model Complexity

In conducting simulation experiments with DIVA, the goals are to evaluate qualitative fit to a range of benchmark findings, to generate novel predictions, and to strive for minimalism in the use of free parameters and ad hoc assumptions (see Goldstone & Kersten, 2003; Love et al., 2004; Roberts & Paschler, 2000; Rodgers & Rowe, 2002). There are, in fact, few choices left to the modeler in conducting a DIVA simulation. The architecture of the network is fully determined by the task except for the number of hidden nodes. The number of hidden nodes does have a large influence on the behavior of the model; there must be enough hidden nodes to effectively reduce the reconstructive error, and there must be compression at the hidden layer in order to implement PCA. For the type of learning problems nearly always used in laboratory studies (three to four features, 8–12 training items, and two categories), the use of two hidden nodes has been consistently appropriate. Specifically, this is the smallest number of hidden units that routinely allows minimization of error across the learning conditions of interest (without considering the fit to behavioral data). Therefore, although the use of two hidden nodes is not strictly a fixed property of the model, the current form of the DIVA account is that two hidden nodes are used to simulate traditional category learning experiments.

For purposes of clarity, a distinction is made here between *fixed* parameters and *free* parameters. Fixed parameters are available to the modeler as degrees of freedom in the model formulation that can be called upon to extend the range or precision of model performance. The term *free parameter* is used to refer to settings with which the modeler can tune the quality of the data fit. The fixed parameters of DIVA are the number of hidden units and the learning rate for the weights and biases. Although backpropagation networks can be varied in many other ways (Rumelhart et al., 1995), no such architectural or procedural variations are presently employed in the DIVA framework. For present purposes, the fixed parameters have been assigned default values that reflect preliminary testing but not optimization. The rationale for the use of two hidden nodes is described above; however, for learning problems on a different scale than those used in traditional classification learning studies, this parameter may need to be revisited. The learning rate for all of the weights in the system are set to a default value of 1.

In backpropagation networks, the weights are initially assigned to near-zero values by applying a small amount of random variation away from zero. This technique is used to break symmetry in the initial weights (Rumelhart et al., 1986). A parameter is used to set the range of this random variation. To be clear, the initial weights are always randomly generated—this parameter determines how far the initial random weights are allowed to vary from zero. The random initialization range is not usually considered to be a critical setting, but it can have a significant impact on network performance (Kolen & Pollack, 1990). Since networks based on backpropagation have never come close to capturing the time course of human category learning, there is little precedent regarding the impact of this parameter in this domain. Sensitivity testing reveals that in some cases DIVA shows qualitatively different performance depending on the order of magnitude of the range of initial weight randomization (± 0.5, 0.05, 0.005, 0.0005). Given the observed impact on model performance, the range of random weight initialization is considered a free parameter of DIVA. In the present simulation experiments, the random initialization range is set to a commonly used default value of ± 0.5 (Kolen & Pollack, 1990). When sensitivity testing reveals a qualitative impact of the random initialization range on model performance, it is discussed in detail.

It is also useful to consider the range of random weight initialization as a psychologically meaningful variable. A larger range value (i.e., ± 0.5) means that the initial weights are allowed to deviate substantially from zero. Because of the properties of gradient descent, the initial location in weight space can be consequential. The initial weights, especially at larger values, represent a bias toward particular solution paths. When the initial weights are tightly constrained around zero, there is less likelihood of converging on a suboptimal solution (i.e., a local minimum). The best way to think about this may be in terms of the *flexibility* of the system to seek out the best solution, as opposed to maintaining a commitment to a good, but not optimal, solution. The degree of flexibility shown by a learner can be influenced by characteristics of the learning task and may also arise from individual differences.

SIMULATIONS OF BENCHMARK HUMAN LEARNING DATA

DIVA was tested on three of the best known data sets in the psychological literature on human category learn-

ing. Shepard, Hovland, and Jenkins's (1961) classic study of the ease of acquisition of elemental category structures, the 5–4 problem introduced by Medin and Schaffer (1978), and the comparison of linearly separable versus nonlinearly separable classification learning (Medin & Schwanenflugel, 1981) were chosen on the basis of their being highly consequential and informative results in the empirical literature that have been frequently used for model comparison.

Modeling the Ease of Acquisition of Elemental Category Structures

A foundational study by Shepard et al. (1961), replicated by Nosofsky et al. (1994), has served as something of a litmus test for computational models of human category learning. Only three models (ALCOVE, RULEX, and SUSTAIN) are considered to provide satisfactory fits. Although the articles that describe each of these models were published on the basis of successful accounts of multiple behavioral results, the Shepard et al. data set is the only simulation result common to the three models. Competing models that fail to capture this particular pattern of results have been noticeably marginalized.

Shepard et al. (1961) compared learning performance on the six two-way classifications possible over a set of binary-valued, 3-D stimuli. The six types include three category structures that hold particular psychological interest: Type I, a unidimensional rule (UNI); Type II, an exclusive-or (XOR) rule with an added irrelevant dimension; and Type IV, which corresponds to a family resemblance (FR) as well as a rule-plus-exception (RULE+) structure. Viewed as an FR structure, there are two inverse-valued prototypes, and each example can be correctly categorized according to the prototype with which it shares more features. All of the features are partially predictive, but none are fully predictive. As a RULE+ structure, a unidimensional rule successfully categorizes six out of the eight examples and the remaining cases must be memorized as exceptions to the rule. Type III and Type V also conform to a RULE+ structure, but these category structures do not show an FR organization.

The critical finding is the ordering of the ease of acquisition of the problem types: Type I (UNI) is the easiest to learn; Type II (XOR) is somewhat harder; Types III, IV (FR/RULE+), and V are harder yet and roughly equivalent; and Type VI is the most difficult. It is the rapid learning of the nonlinear, rule-like Type II (XOR), and the relatively slow learning of the linearly separable Type IV (FR/RULE+) in human performance, that tend to foil models that do not conform to the reference point framework of localist encoding and dimensional selective attention. The backpropagation learning algorithm (in a standard multilayer perceptron architecture) is noted for performing successful nonlinear function approximation, but it acquires XOR too slowly and FR/RULE+ too quickly (Kruschke, 1992). Localist encoding and dimensional selective attention are the core design principles thought to contribute to the success of reference point models in capturing the order of acquisition.

Table 1
Fitting DIVA to Shepard et al. (1961)

Type	DIVA Cumulative Error Rate*	Human Cumulative Error Rate
I	0.53	lowest
II	0.95	low
III	1.22	intermediate
IV	1.25	intermediate
V	1.17	intermediate
VI	4.38	highest

*Cumulative error rate measured as the sum of the average probability of an error across 10 sampling points (every five passes through the training set).

The relative ease of acquisition of the six problem types was tested across 60 random initializations of a $3 - 2 - 3 \times 2$ DIVA network (as shown in Figure 1). The default values of 1.0 for learning rate and ± 0.5 for random initialization range were used. The eight training patterns were represented using input and target values of ± 1, and the sigmoidal activation range for the output nodes was scaled accordingly. Weight change was performed using online, trial-by-trial updating. In order to simulate the dependent measure of cumulative errors employed by Shepard et al. (1961), the model was tested on all eight patterns after every five passes through the training set. The aggregate mean probability of an error over a total of 10 sampling points produced the cumulative error. DIVA showed a good fit to the human learning data (see Table 1) by producing the correct ordering of ease of acquisition. Sensitivity testing showed qualitatively consistent performance at varied learning rates. For purposes of direct comparison, a standard $3-2-1$ backpropagation network was tested under matching conditions (using a learning rate of 0.25). As expected, Type IV (FR/RULE+) was learned more easily than Type II (XOR) and nearly as easily as Type I (UNI). Additionally, independent autoencoders assigned to each category failed to fit the data.

Nosofsky et al. (1994) charted the time course of human learning of the six types and generated a richer set of qualitative and quantitative results for model comparison. One finding was that the same general ordering occurred consistently from early to late in learning. DIVA was also fairly consistent across the time course of learning (see Figure 2), but some aspects of the time course fit did not accord completely with the aggregate human learning data. First, the more accurate performance for Type II relative to Types III–V does not emerge from the very beginning of learning, as it does in the human data (Nosofsky et al., 1994). In simulations by Kruschke (1992), ALCOVE produced a similar pattern: The Type II advantage emerged after about 20 passes through the training set. Nosofsky et al. (1994) reported learning curves for ALCOVE in which the Type II advantage begins almost immediately, but under these parameter settings ALCOVE appears to learn Type II too easily—almost as easily as Type I.

The time course data for DIVA reveals a deep curvature in Type V learning. This results in a brief period during which Type V performance is closer to that of Type II than

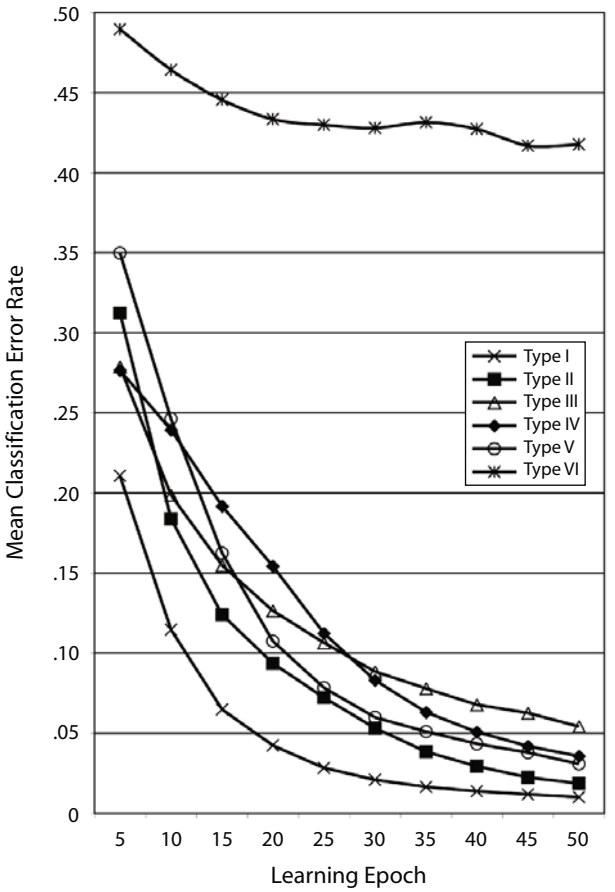

Figure 2. Time course of classification learning by DIVA on the six types of problems (Shepard et al., 1961). Each learning epoch consists of one pass through the training set.

it is to those of Types III and V. This slight inconsistency relative to human learning turns out to be an indicator of an interesting pattern. Sensitivity testing on the range of random weight initialization showed little effect on the relative status of the six types, except for Type V. Specifically, Type V was learned at the same rate as Types III and IV (matching human learning) at less restrictive ranges, but showed faster acquisition under smaller, more restrictive ranges. There are no prior examples of Type V being learned faster than Types III and IV, but there have been no prior attempts to manipulate the extent to which a learner seeks an optimal solution (or is encouraged to resist making an early commitment to a suboptimal solution). The influence of the range of random weight initialization on Type V will be revisited in the discussion below.

The largest departure from the human learning data is the minimal progress shown by DIVA on Type VI. Previous modeling efforts have, in fact, shown the opposite problem: reaching criterion in Type VI learning too easily. Shepard et al. (1961) and Nosofsky et al. (1994) found a dramatic gap in human performance separating Type VI from the other five types. Specifically, Type VI learners still made errors more than 10% of the time after 32 training passes (Nosofsky et al., 1994); at this point, the learning of the other five category structures was essentially perfect. This is not surprising, because there is nothing at the category level to learn in Type VI. Given that no rules or resemblances characterize the categories, the only way to succeed is to memorize the eight individual associations. Such exemplar memorization comes naturally within the reference point framework, and the reported best fits for ALCOVE and SUSTAIN reach asymptote on Type VI learning shortly after Types III–V. DIVA effectively captures the qualitative segregation of Type VI as markedly more difficult than the other types, but the model errs in the opposite direction and seriously overestimates the difficulty of Type VI relative to human performance. DIVA is capable of achieving successful learning of Type VI with lower learning rates or more hidden units, but it remains a goal for future work to provide a full account of how human learners eventually reach criterion when there are no statistical regularities underlying the category structure.

In further research based on Shepard et al.'s (1961) six types, Nosofsky and Palmeri (1996) found a shift in the order of acquisition using integral-dimension, as opposed to separable-dimension, stimuli (see also Shepard & Chang, 1963). The key finding was that Type II was the second most difficult category structure to learn, as opposed to being the second easiest. Nosofsky and Palmeri (1996) offered an elegant explanation based on the idea that dimensional selective attention is much harder to apply to integral-dimension stimuli. Without selective attention, ALCOVE captures the reversal by showing increased difficulty in the acquisition of Type II (see also Kruschke, 1992). For DIVA, there is no obvious reason for the network to run any differently under these two cases. A speculative account of the phenomenon is that integral-dimension stimuli may be encoded in an impoverished fashion as the learner attempts to extract the underlying integral dimensions or construct ad hoc features. Accordingly, Type II becomes harder to learn, because it is impossible to make even partial progress in learning the XOR structure without having two fully realized features (a single feature is highly predictive for Types I, III, IV, and V).

Returning now to the larger picture, DIVA is the first model to produce a successful qualitative fit to the benchmark data of Shepard et al. (1961) without the use of quasilocalist encoding and a mechanism of dimensional selective attention. The Shepard et al. results have been considered to arise largely on the basis of the number of relevant features (i.e., the number of features requiring attention): Type I requires only one feature to specify the classification basis; Type II requires two features; and the remaining types require all three features (e.g., Kruschke, 1992). DIVA offers an alternative account in which learning by backpropagation, specifically in a divergent autoencoder architecture, accords with human performance. DIVA shows the correct ordering among the six types as a result of its learning process of extracting principle components that minimize the loss of information across the memberships of the two categories.

According to Kruschke (1992, 1993), standard backpropagation fails to capture key patterns of human learn-

ing because of the flexibility inherent in the algorithm. Backpropagation learning operates by performing weight adjustments to establish a multidimensional recoding space in which the members of each category are situated on opposite sides of a hyperplane (i.e., linearly separable). This hyperplane can pass through the recoding space at any orientation. By contrast, ALCOVE can only stretch or shrink the original representational space along the dimensional axes. Kruschke (1992, 1993) argued that this limitation is a psychologically valid constraint. The use of backpropagation within the DIVA architecture produces a different outcome. The output nodes of DIVA code for features (in accord with PCA) rather than for categories (in accord with multiple regression). This alters the learning dynamics and removes the division of the recoding space into category-based regions. In DIVA, learning is directed toward collectively approximating a set of functions, each of which characterizes the membership of a single category, as opposed to approximating a single function directed toward finding a way of distinguishing the categories.

A set of hidden node activations was recorded after training a $3-2-3 \times 2$ DIVA network on each of the six types (see Table 2; note that for ease of viewing, the stimulus items are displayed using 0/1 values, although ± 1 values were used in the simulations). Each cell in the table shows the activations of the two hidden nodes. These particular solutions do not always occur, but they are representative. The solutions for Type I and Type II both use the four corners of the representational space: (0 0), (0 1), (1 0), and (1 1). Specifically, each corner of the representational space codes for two items, one from each category. The property of maximizing interitem separation can clearly be seen; note, however, that separation is only maximized within each category. To offer an analogy, this is like being in a crowded room and trying to keep your distance from the members of one group while being indifferent to your proximity to members of another group. In the Type I solution, each hidden node codes for one of the two irrelevant features. It is somewhat counterintuitive that the recodings are devoted to information that is irrelevant to the category structure. The reason for this is that there is no within-category variability for the critical feature (it is always on for one category and off for the other). The DIVA network captures this regularity most simply by adjusting the strength of the bias assigned to the first output node in each channel, and therefore does not have to devote the input-to-hidden and hidden-to-output weights to this part of the problem. In the Type II solution, the first hidden node codes for the first stimulus feature and the second hidden node codes for the irrelevant feature. The second relevant feature in the XOR function does not need to be represented in the recoding because within each channel it is perfectly predicted by the first feature. The more complex category structures lead to various compromises that optimize separation in the recoding space for the members of each category under the constraints imposed by the decreased levels of category coherence.

Table 2
Sample Hidden Node Activations for DIVA Trained on Shepard et al. (1961)

	Category Structure					
Item	Type I	Type II	Type III	Type IV	Type V	Type VI
1 1 0	(1 0) a	(1 0) a	(.9 0) a	(.9 0) a	(1 0) a	(1 0) a
1 1 1	(1 1) a	(1 .9) a	(.8 1) a	(.5 .5) a	(1 .2) b	(1 .6) b
1 0 1	(0 1) a	(1 1) b	(1 1) b	(1 1) a	(.3 1) a	(.8 1) a
1 0 0	(0 0) a	(1 0) b	(1 0) a	(1 .1) b	(1 .8) a	(.3 .2) b
0 1 0	(1 0) b	(0 0) b	(0 0) b	(0 0) b	(.2 0) b	(0 0) b
0 1 1	(1 1) b	(0 1) b	(0 1) a	(0 .9) a	(0 0) a	(.3 .6) a
0 0 1	(0 1) b	(0 1) a	(.1 1) b	(1 1) b	(0 1) b	(0 1) b
0 0 0	(0 0) b	(0 0) a	(.2 0) b	(.5 .5) b	(0 .4) b	(0 .2) a

Note—a and b denote the assignment of items to the two categories within each type.

As stated above, a core design principle of DIVA is learning along coordinated channels—that is, collectively solving a multiplicity of functions. This is also a good way to characterize the processing that takes place within each channel. The reconstructive process taking place at each output node is actually computing a traditional MLP classification function in which the number of classes is the number of different values that the particular feature takes on in the training set. For example, in Type I learning (e.g., a unidimensional rule on the first feature), the items (**1** 0 1), (**1** 1 1), (**1** 1 0), and (**1** 0 0) are processed along the same category channel. Reducing the error on each output node requires learning to correctly predict the value of that feature for each of the four patterns. Since there are two possible values for each feature, the reconstruction task specifically for the third output node is a binary classification problem of distinguishing the patterns (1 0 **1**) and (1 1 **1**) from the patterns (1 0 **0**) and (1 1 **0**). This characterization in terms of *dimensional classifications* extends to the processing that takes place in parallel at each of the output nodes.

This type of analysis helps to clarify why some learning tasks are harder than others for DIVA. For the Type I problem, each channel includes two two-way dimensional classifications like the one just described, and the third output node is a simple one-way dimensional classification (all of the patterns assigned to the channel share the same value for this feature). Type II learning consists of three two-way dimensional classifications, but since two dimensions are perfectly correlated in the XOR function, two of the dimensional classifications in each channel can be jointly solved. Unlike reference point models that benefit in Type II from ignoring the irrelevant dimension, DIVA benefits from the fact that the two relevant dimensions perfectly predict one another within each category.

The remaining Types III–VI are more difficult because they require learning three distinct dimensional classifications on each channel. The high level of difficulty for Type VI can be understood in terms of another property of dimensional classification. In the Type VI problem, one category channel is assigned the patterns (0 1 1), (1 0 1), (1 1 0), and (0 0 0). It can be seen that each of the three two-way dimensional classifications is balanced—that is, half of the patterns have one value and half of the patterns have the other. Types III–V are easier to solve because at least one of the dimensional classifications is an "odd-

ball" classification, in which three of the four patterns assigned to each channel share a common value whereas the remaining "oddball" has the alternate value. This directly parallels another way of looking at these three category structures in terms of the number of dimensions along which a unidimensional rule-plus-exception (RULE+) classification can be formulated (Nosofsky et al., 1994). The presence of at least one oddball dimensional classification markedly reduces the difficulty of finding a region of weight space that simultaneously supports all of the dimensional classifications.

The number of oddball dimensional classifications varies among Types III, IV, and V, even though they are learned at similar rates: Type III has two, Type IV has three, and Type V has only one. Accordingly, Type V is unique among the six types in that it is fairly difficult to learn and the solution is highly constrained. Part of the challenge faced by DIVA is reaching a set of recodings that support reconstruction on both category channels. When the system can move freely toward the optimal solution (i.e., under a strict range on the initial random weights), it tends to solve Type V in a fairly straightforward fashion. Alternatively, when the initial weights bias one or both channels in the direction of a suboptimal solution, it is more difficult to reach a coordinated recoding scheme. This effect is predicted only for a category structure with a solution that is high in both difficulty and constraint.

Linear Separability

There is an ambiguity that arises in interpreting the theoretical implications of Shepard et al. (1961), because two factors are in play at the same time. Type II is a nonlinearly separable (NLS) function over two dimensions, and Type IV is a linearly separable (LS) function over three dimensions. Medin and Schwanenflugel (1981, Experiment 4) used subsets of the training items from Type III and Type IV of Shepard et al. (1961) to examine the role of linear separability with the number of relevant dimensions held constant. Medin and Schwanenflugel (1981) found no evidence of a learning advantage for the particular linearly separable categories that they tested. Their results remain an important benchmark for model comparison and have exerted a major influence on theory development.

Most importantly, a strong prediction of prototype-based accounts is greater ease of learning for LS categories. Simply put, unless each member of a category is closer to the central tendency of its own category than to that of the contrasting category, successful classification based on similarity to prototypes is an unpromising proposition. By contrast, exemplar accounts with sensitivity to individual instances have no such difficulty. Along similar lines, simple neural network models (without a hidden layer) lack the ability to acquire NLS classifications. Standard backpropagation networks are capable of learning NLS categories, but they still make a strong and incorrect prediction: that linear boundaries should be much easier to acquire. In sum, oversensitivity to linear boundaries (Kruschke, 1992) has plagued accounts of human learning outside the reference point framework.

The immediate question, then, is whether the success of DIVA in simulating the Type II advantage will extend to a test of the LS-versus-NLS category structures used by Medin and Schwanenflugel (1981). A $3 - 2 - 3 \times 2$ DIVA network was tested using the default values (learning rate of 1.0 and random initialization range of ± 0.5). Each training run consisted of 25 passes through the training set. Mean levels of classification accuracy at the end of training were 92% for NLS and 87% for LS. The overall rates of learning were very similar, with a possible slight advantage for the NLS category structure. Therefore, a well-known problem in simulating human learning using backpropagation (LS category structures being learned too easily and NLS category structures being learned with too much difficulty) is overcome. The DIVA account reveals how a learning system based on abstracting statistical regularities, as opposed to computing weighted similarity to item-specific reference points, can handle NLS category structures as human learners do. By building coordinated statistical models of the categories, rather than by explicitly searching for a discrimination function, DIVA shows the appropriate insensitivity to linear separability.

Modeling the 5–4 Problem: A Case of Weak Category Structure

The success of the exemplar view of categorization rests in no small part on extensive behavioral and computational tests of the 5–4 problem (Medin & Schaffer, 1978; Minda & Smith, 2002; Nosofsky, 2000; Nosofsky, Kruschke, & McKinley, 1992; Nosofsky, et al., 1994; Smith & Minda, 2000). The 5–4 problem (see Table 3) consists of nine training items and a set of transfer items that are based on four binary-valued features. The classification problem is designed to be linearly separable, even though the training set includes three weak category members with only two out of four category-consistent feature values. ALCOVE and RULEX fit the data well in terms of goodness of fit to item-by-item human classification accuracy and capturing two qualitative patterns that specifically challenge abstraction-based accounts.

Table 3
Fitting DIVA to the 5–4 Problem

	Probability of Classification Response A at Test	
Stimulus	DIVA	Medin & Schaffer (1978)
A1 (1 1 1 0)	.72	.78
A2 (1 0 1 0)	.94	.88
A3 (1 0 1 1)	.92	.81
A4 (1 1 0 1)	.75	.88
A5 (0 1 1 1)	.85	.81
B1 (1 1 0 0)	.26	.16
B2 (0 1 1 0)	.24	.16
B3 (0 0 0 1)	.08	.12
B4 (0 0 0 0)	.07	.03
T1 (1 0 0 1)	.52	.59
T2 (1 0 0 0)	.45	.31
T3 (1 1 1 1)	.80	.94
T4 (0 0 1 0)	.44	.34
T5 (0 1 0 1)	.56	.50
T6 (0 0 1 1)	.54	.62
T7 (0 1 0 0)	.25	.16

The first of these qualitative patterns is that learners are more accurate in classifying Stimulus A2 (1 0 1 0), which has two features in common with the A prototype (1 1 1 1), than they are in classifying Stimulus A1 (1 1 1 0), which has three prototypical features. (For ease of viewing, here and in Table 3, values of 0 and 1 are used to describe the stimulus items.) Whereas the prototype view clearly makes the opposite prediction, exemplar-based models effectively capture this result (Medin & Schaffer, 1978; Nosofsky et al., 1992). The second pattern emerges in transfer performance on the Category A prototype (1 1 1 1), which is not included in the training set. The prototype (Stimulus T3) is the most accurately classified among the transfer items, but it is not classified with greater accuracy than some of the nonprototypical training items, A2 (1 0 1 0) and A3 (1 0 1 1), which are more distant from the central tendency of the category. One caveat with regard to these results is that they are not uniformly observed across variations in stimulus materials and method (Medin & Schaffer, 1978; Smith & Minda, 2000; but see Nosofsky, 2000).

A $4-2-4 \times 2$ DIVA network was applied to the 5–4 problem, using the default parameter settings of learning rate of 1.0 and random initialization range of ± 0.5. The binary features were encoded using ± 1 values. Average performance on all items (training and transfer examples) was computed across 50 network initializations. Each network was tested after 16 blocks of training. Sensitivity testing showed qualitative consistency across parameter settings. DIVA successfully captured the two notable patterns found in human learning. The model showed a greater mean probability of correct responding for the less prototypical Stimulus A2 (.94) than for Stimulus A1 (.72). The transfer test for the prototype of Category A yielded a mean probability of correct responding (T3 = .80), which was the highest accuracy rate among the transfer items but lower than the two training items (A2 = .94, A3 = .92). DIVA showed a traditional prototype enhancement effect (see, e.g., Posner & Keele, 1968), which is one of the best established findings in the category learning literature. Like exemplar-based accounts, DIVA displays prototype sensitivity without an explicit mechanism of prototype formation.

In simulating the set of phenomena surrounding the 5–4 problem, DIVA captures the advantages of the prototype view (enhanced classification accuracy on an untrained prototype) and the advantages of the exemplar view (accuracy levels on individual items that belie their similarity to the underlying prototype; good qualitative fits for both old and new items). DIVA achieves these results by building coordinated statistical models of the categories, not by storing reference points in the form of exemplars or prototypes.

The present research has placed emphasis on the qualitative pattern of results from the 5–4 structure. The detailed quantitative fit of the parameter-free DIVA model to the aggregate data of Medin and Schaffer (1978, Experiment 2) is not as good (sum of squared deviations = .128; see Table 3) as that achieved by a four-parameter exemplar model (Medin & Schaffer, 1978; Nosofsky et al., 1992). In future work, DIVA will be tested for its ability to generate quantitative fits as close to the human data as those achieved using reference point models. The validity of aggregated learning data has recently come under some question as a result of individual subject analyses showing that aggregate performance is often inconsistent (even qualitatively) with individual profiles (Nosofsky et al., 1994; see also Ashby, Maddox, & Lee, 1994; Lee & Webb, 2005). Therefore, it will also be a priority to assess DIVA's ability to fit individual subject data. DIVA produces variable performance across training runs primarily as a result of the random initial weights. This may correspond with variations in individual performance based on dispositional or situational initial conditions that bias the learner toward a particular solution path. Other ways to capture individual variation include the learning rate and parameters that scale or modify the choice rule.

To summarize, DIVA produced a good qualitative fit to a benchmark data set that was designed specifically to advance the case for exemplar-based accounts. DIVA succeeds by constructing category representations that summarize statistical tendencies while also being tailored to the individual items. In order to reduce the reconstructive error on weakly structured categories, DIVA discovers a set of weights that preserve both general and specific information about the content of training set. DIVA is the first model that shows exemplar sensitivity and prototype sensitivity without employing the reference point formulation of dimensionally weighted similarity to quasilocal representations. Just as DIVA successfully matched the level of difficulty shown by humans in learning different types of category structures, the model also effectively captures the relative difficulty of individual items.

DISCUSSION

DIVA is the first model outside the reference point framework that successfully accommodates major psychological benchmarks. The relative ordering of the six types of classification learning in Shepard et al.'s (1961) influential research has been widely thought to reflect the operation of a system with the capacity to encode specific exemplars and to compute similarity only on the basis of the relevant dimensions (e.g., Kruschke, 1992). In the original work, Shepard and colleagues raised the possibility of either an abstraction-based or a selective attention-based solution. On the basis of the present findings, the relative ordering of the six types can be understood in terms of the ease with which a set of internal representations can be constructed using the same set of weights to simultaneously support recoding/decoding of the members of both categories. In further simulation experiments, it was found that DIVA is not overly sensitive to linear boundaries. DIVA also successfully accounted for human performance on Medin and Schaffer's (1978) 5–4 problem—another foundational data set that helped to consolidate widespread acceptance of the reference point framework and its core design principles of dimensional selective attention and item-specific representation. Throughout the simulation experiments, a number of

novel predictions and alternate interpretations of behavioral data were generated, making clear the value of the model above and beyond its data fits. Finally, although it is beyond the scope of the present report, DIVA has been shown to effectively capture important phenomena in the domain of inference learning (see Love et al., 2004; Markman & Ross, 2003; Yamauchi, Love, & Markman, 2002; Yamauchi & Markman, 1998) and to avoid the problem of catastrophic forgetting in the traditional demonstration cases showing that standard backpropagation networks lose their initial learning under sequential processing conditions (Kruschke, 1992; McCloskey & Cohen, 1989; Ratcliff, 1990; see also French, 1999, and Kruschke, 1993).

Further Directions for Cognitive Simulation

The data fits to this point have been qualitative in nature and oriented toward aggregate data. Researchers are increasingly taking into account the profiles of individual learners (e.g., Minda & Smith, 2002; Nosofsky & Johansen, 2000; Nosofsky et al., 1994; Rehder & Hoffman, 2005a, 2005b), so a goal for continued testing of DIVA is to establish whether the model can quantitatively fit individual learning curves. The model parameters (e.g., learning rate and weight initialization range) will be evaluated as a way to account for individual differences in learning. Some immediate speculations are that the range of the initial random weights corresponds to the likelihood of switching away from an initial approach to a classification problem, and that the learning rate corresponds to the amount of impact each trial has on the construction of an overall solution. Another source of power for generating quantitative fits to aggregate or individual data with DIVA is the use of a parameter-tuned version of the choice rule for generating response probabilities from the output node activations.

Research is under way to apply the DIVA model to problems based on continuous-valued inputs. An important target phenomenon is human learning performance on filtration-versus-condensation tasks (Kruschke, 1993; see also Garner, 1974 and Posner, 1964). In a *filtration* task, only one of two available perceptual dimensions is relevant to classification. (The task is so named because the irrelevant dimension can be filtered out.) In a *condensation* task, both of the dimensions are relevant and the two dimensions must be condensed—that is, considered in combination—in order to achieve classification success. Kruschke (1993) found that human learners showed a consistent filtration advantage from start to finish of training, and that ALCOVE—but not standard backpropagation—effectively captured the filtration advantage. This is a challenging test case for DIVA, because the phenomenon appears to result from a shift of attention away from the irrelevant feature. DIVA does not tend to ignore information, since its core task is to optimally reconstruct the input.

Rehder and Hoffman (2005a) used an eye-tracking methodology to study category learning and argued that observed patterns of visual attention are best explained in terms of a hybrid account of human category learning with a gradual probabilistic component and an all-or-none rule-based component. Their evidence runs counter to models like ALCOVE, because the eye movements consistent with dimensional selective attention begin at about the time at which proficient performance is reached, rather than emerging gradually on the basis of trial-by-trial error-corrective learning. DIVA correctly predicts broad attention to all available features under all category structures during the period in which learning is actually taking place. This is because the model uses the full set of available information as inputs and reconstructive targets, so information does not tend to be ignored while the coordinated statistical models underlying category understanding are being formed. The question is, what would explain a learner's beginning to ignore irrelevant dimensions only after having mastered the classification task? One speculation is that, once learning has succeeded, people are likely to look for procedural shortcuts, such as disregarding information that need not be considered. Along these lines, a DIVA network trained on a unidimensional rule (i.e., Type I from Shepard et al., 1961) would perform well on a partial pattern consisting only of the relevant dimension, but would perform poorly on a partial pattern that lacked the relevant dimension. Therefore, it is not difficult to imagine learners exploiting this efficiency, once they were confident they had mastered the categories. Additionally, the threshold for asserting successful learning might be an individual difference variable.

Another area of increased emphasis in categorization research is the underlying brain basis for cognitive performance. Although DIVA is brain style in the connectionist tradition, the account is not constrained by specific neuroscientific findings. Separate systems views are widespread in the cognitive neuroscience literature due to evidence of dissociations between implicit/procedural and explicit/declarative modes of category learning (e.g., Ashby, Alfonso-Reese, Turken, & Waldron, 1998; Ashby & Maddox, 2005; Smith, Patalano, & Jonides, 1998), and between classification and memory tasks (Knowlton & Squire, 1993; Reed, Squire, Patalano, Smith, & Jonides, 1999). Nosofsky and colleagues (e.g., Nosofsky & Johansen, 2000; Nosofsky & Zaki, 1998; Zaki & Nosofsky, 2001) offer a range of evidence that exemplar-based accounts can account for a broad range of learning phenomena without the need for multiple systems. Can DIVA succeed in a similar manner? In addition, categorization researchers have sought to explain classification and old–new recognition tasks within a common modeling framework (e.g., Nosofsky, 1988, 1991). With DIVA, it should be possible to model recognition performance in terms of the best reconstruction across category channels or as a function of reconstructive success across all category channels. Further predictions are that training DIVA to a high criterion will lead to high levels of recognition for category exceptions (Palmeri & Nosofsky, 1995; Sakamoto & Love, 2004) and that DIVA will show false recognition of novel yet highly typical items.

Directions for Model Development

One of the aims in this work has been to minimize the role of parameter fitting and modeler intervention in the

simulation process. A remaining goal is to develop a precise and a priori procedure for direct translation from a description of a psychological task to a complete specification of the network and its settings. Toward this goal, unless the consistent appropriateness of two hidden nodes turns out to reflect underlying processing constraints in human category learning, DIVA will be generalized to dynamically prune (e.g., Castellano, Fanelli, & Pelillo, 1997; Karnin, 1990) or recruit (Fahlman & Lebiere, 1990; Love et al., 2004) hidden nodes. This may offer an account of how DIVA can handle a problem like the Type VI category structure, which is not amenable to the default learning mode. In addition, mechanisms will be sought to directly determine the learning rate and range of random weight initialization based on properties of the task and/or the learner.

The focus of this article has been on supervised category learning, but DIVA is naturally extensible to unsupervised learning. Instead of relying on a target signal, the channel selected for weight update is the one that produces the best reconstruction. On early trials, the random initial weights cause one of the channels to win. This channel is updated to become better able to successfully reconstruct the current stimulus item. When the task conditions specify the number of classes, the number of category channels is fixed accordingly; when there are no constraints on the number of categories, an additional channel is recruited when the established channel(s) fail(s) to produce a satisfactory reconstruction. The new channel shares the set of input-to-hidden weights common to all channels, and the remaining weights are randomly initialized. The new channel is trained on its seed example and thereafter competes normally to produce the best reconstruction. The addition of channels is not presently a core design principle of DIVA, but it provides a straightforward means by which the system can autonomously originate or expand a classification scheme. Incidental unsupervised tasks (Love, 2002; Wattenmaker, 1991), in which the learner is unaware of being in a categorization task, are modeled using a standard autoencoder without divergent channels.

DIVA and Natural Concepts

The DIVA account is consistent with the probabilistic view of categorization and the graded structure of natural categories. At the same time, certain aspects of the theory view (Murphy & Medin, 1985) are also realized, even though important elements of the theory view having to do with causal cores and the role of background knowledge are not yet integrated. Assigning category membership is not a matter of finding the best match between an input and stored reference points, as in the similarity-based probabilistic models criticized in the theory view. In DIVA, a categorization judgment takes the form of an assessment of the extent to which a stimulus is well accounted for by the statistical model underlying a particular category. More specifically, whether or not a stimulus is understood to be a dog depends on the result of attempting to recode and decode the stimulus according to a statistical model of *dog* properties experienced under particular task conditions. The statistical models underlying categories are not strictly a product of the structure of the environment, but are mediated by the specific tasks that constitute the interaction between the learner and the data. Along these lines, DIVA learns to recode the initial form of the stimulus features in a top-down manner, using the representational space of the hidden nodes. DIVA constructs psychological representations in concert with learning to categorize, as opposed to assuming their availability (and their fixedness) from an external and independent source. Once again, despite these elements of consistency with the theory view, the role of explicit structure and explanatory relationships within and between natural categories is not yet realized.

Another promising aspect of DIVA is its potential extension to a broader range of higher cognitive functions. For example, the psychological similarity of *horse* to *cow* can be understood as a matter of how effectively a representative instance or summary of the *horse* category can be reconstructed by a statistical model representing the *cow* category. The question of their similarity becomes the question: How good a *cow* is a *horse*? As such, reconstructive error is a goodness-of-fit measure that can be converted to a similarity judgment. To further illustrate, comparing *cat* or *wolf* to the category *dog* would produce little distortion, but a target like *bookshelf* or *sea slug* would elicit a highly distorted output. The interpretation of *wolf-as-dog* is coherent; that of *bookshelf-as-dog* is not. This account is also consistent with the influential constraint (Tversky, 1977) that similarity judgments can be asymmetric; that is, *wolf-as-dog* yields a different similarity outcome than does *dog-as-wolf*. The typicality of category members (Rosch & Mervis, 1975) can also be determined by testing an exemplar relative to its category. A dog is a typical member of the *animal* category (or a particular dog is typical of the *dog* category) to the extent that its reconstructive error is low. Category-based induction tasks (Osherson, Smith, Wilkie, Lopez, & Shafir, 1990; Rips, 1975) can potentially also be modeled in terms of reconstruction-based interpretations of similarity, typicality, and coverage.

An open question is whether a person categorizing a stimulus actually experiences the stimulus in terms of the resulting reconstruction. Consider the case of assimilating a wolf to the *dog* category. The resulting reconstruction would distort features toward the *dog* category. The belief that one is seeing a dog, not a wolf, might make the teeth seem not quite so fang-like and the eyes not quite so predatory. This can account for construal and feature interpretation processes (Kurtz & Dietrich, 2007; Wisniewski & Medin, 1994) through which the understanding of a stimulus is infused with the semantics of the category guiding the interpretation. The feature reconstructions produced by DIVA can also potentially account for schema-like memory distortions.

Conclusion

The theoretical view of categories as task-coordinated statistical models appears to have much to offer. The approach is grounded in powerful and formally understood statistical and computational procedures. DIVA repre-

sents a major conceptual departure from the small set of models that have achieved comparable success in accounting for human category learning. Specifically, DIVA does not employ nodes that code specifically for exemplars or sets of exemplars (i.e., quasilocal reference points) and does not categorize according to the similarity match between inputs and reference points. Rather than using dimensional selective attention to diagnostically stretch or shrink dimensions, DIVA performs a task-driven recoding or transformation of inputs into a distributed representational space. As opposed to computing the error that drives learning as the deviation between a category guess and the correct category, DIVA uses the deviation between the reconstructed and original forms of the input. DIVA captures signature phenomena traditionally associated with the use of rules, prototypes, and exemplars while explicitly implementing none of them. The mechanism of divergent autoencoding is highly general, and it is hoped that the approach may prove broadly applicable to pattern recognition problems in a variety of psychological and applied domains.

AUTHOR NOTE

This work was partially supported by a B/START award (MH68412-01) from the National Institute of Mental Health to K.J.K. I thank the many colleagues with whom I have had helpful discussions of these matters, Robert Nosofsky and two anonymous reviewers for providing invaluable feedback, and, most of all, D.E.R. Correspondence concerning this article should be addressed to K. J. Kurtz, Department of Psychology, P.O. Box 6000, Binghamton University, Binghamton, NY 13902 (e-mail: kkurtz@binghamton.edu).

REFERENCES

AHN, W., KIM, N., LASSALINE, M., & DENNIS, M. (2000). Causal status as a determinant of feature centrality. *Cognitive Psychology*, **41**, 361-416.

ANDERSON, J. A., SILVERSTEIN, J. W., RITZ, S. A., & JONES, R. S. (1977). Distinctive features, categorical perception and probability learning: Some applications of a neural model. *Psychological Review*, **84**, 413-451.

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

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

ASHBY, F. G., & MADDOX, W. (2005). Human category learning. *Annual Review of Psychology*, **56**, 149-178.

ASHBY, F. G., MADDOX, W., & BOHIL, C. (2002). Observational versus feedback training in rule-based and information-integration category learning. *Memory & Cognition*, **30**, 666-677.

ASHBY, F. G., MADDOX, W. T., & LEE, W. W. (1994). On the dangers of averaging across subjects when using multidimensional scaling on the similarity-choice model. *Psychological Science*, **5**, 144-151.

BALDI, P., & HORNIK, K. (1989). Neural networks and principal components analysis: Learning from examples without local minima. *Neural Networks*, **2**, 53-58.

BECKER, S., & PLUMBLEY, M. (1996). Unsupervised neural network learning procedures for feature extraction and classification. *Applied Intelligence*, **6**, 185-203.

BEN-DAVID, S., & SCHULLER, R. (2003). Exploiting task relatedness for multiple task learning. In *Proceedings of Computational Learning Theory (COLT)* (pp. 567-580).

BOULARD, H., & KAMP, Y. (1988). Auto-association by multilayer perceptrons and singular value decomposition. *Biological Cybernetics*, **59**, 291-294.

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

CARUANA, R. (1995). Learning many related tasks at the same time with backpropagation. In G. Tessauro, D. S. Touretzky, & T. K. Leen (Eds.), *Advances in neural information processing systems* (Vol. 7, pp. 657-664). San Mateo, CA: Morgan Kaufmann.

CASTELLANO, G., FANELLI, A. M., & PELILLO, M. (1997). An iterative pruning algorithm for feedforward neural networks. *IEEE Transactions on Neural Networks*, **8**, 519-531.

CHEN, S., & SUN, T. (2005). Class-information-incorporated principle component analysis. *Neurocomputing*, **69**, 216-223.

COTTRELL, G. W., MUNRO, P., & ZIPSER, D. (1988). Image compression by backpropagation: An example of extensional programming. In N. E. Sharkey (Ed.), *Advances in cognitive science* (Vol. 3). Norwood, NJ: Ablex.

DEMERS, D., & COTTRELL, G. (1993). Nonlinear dimensionality reduction. In S. J. Hanson, J. Cowan, & L. Giles (Eds.), *Advances in neural information processing systems* (Vol. 5, pp. 580-587). San Mateo, CA: Morgan Kaufmann.

DUDA, R., & HART, P. (1973). *Pattern classification and scene analysis*. New York: Wiley.

EDELMAN, S., & INTRATOR, N. (1998). Learning as extraction of low-dimensional representations. In R. Goldstone, P. Schyns, & D. Medin (Eds.), *Mechanisms of perceptual learning* (pp. 353-376). San Diego: Academic Press.

ELLIS, R., & HUMPHREYS, G. L. (EDS.). (1999). *Connectionist psychology: A text with readings*. Hove, U.K.: Psychology Press.

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

FAHLMAN, S. E., & LEBIERE, C. (1990). The cascade-correlation learning architecture. In D. S. Touretzky (Ed.), *Advances in neural information processing systems* (Vol. 1, pp. 524-532). San Mateo, CA: Morgan Kaufmann.

FRENCH, R. (1999). Catastrophic forgetting in connectionist networks. *Trends in Cognitive Science*, **3**, 128-135.

FRIED, L., & HOLYOAK, K. (1984). Induction of category distributions: A framework for classification learning. *Journal of Experimental Psychology: Learning, Memory, & Cognition*, **10**, 234-257.

GARNER, W. R. (1974). *The processing of information and structure*. Hillsdale, NJ: Erlbaum.

GLUCK, M. A., & BOWER, G. (1988). From conditioning to category learning: An adaptive network model. *Journal of Experimental Psychology: General*, **119**, 105-109.

GLUCK, M. A., & MYERS, C. E. (1993). Hippocampal mediation of stimulus representation: A computational theory. *Hippocampus*, **3**, 491-516.

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

GOLDSTONE, R. L., & KERSTEN, A. (2003). Concepts and categories. In A. F. Healy & R. W. Proctor (Eds.), *Comprehensive handbook of psychology: Experimental psychology* (Vol. 4, pp. 599-621). New York: Wiley.

GUENTHER, F. (1995). Speech sound acquisition, coarticulation, and rate effects in a neural network model of speech production. *Psychological Review*, **102**, 594-621.

HAMPTON, J. (1979). Polymorphous concepts in semantic memory. *Journal of Verbal Learning & Verbal Behavior*, **18**, 441-461.

HARNAD, S., HANSON, S., & LUBIN, J. (1995). Learned categorical perception in neural nets: Implications for symbol grounding. In V. Honavar & L. Uhr (Eds.), *Symbol processors and connectionist network models in artificial intelligence and cognitive modeling: Steps toward principled integration* (pp. 191-206). San Diego: Academic Press.

INTRATOR, N., & EDELMAN, S. (1997). Learning low-dimensional representations via the usage of multiple-class labels. *Network*, **8**, 259-281.

JAPKOWICZ, N. (2001). Supervised versus unsupervised binary-learning by feedforward neural networks. *Machine Learning*, **42**, 97-122.

JAPKOWICZ, N., HANSON, S. J., & GLUCK, M. A. (2000). Nonlinear

autoassociation is not equivalent to PCA. *Neural Computation*, **12**, 531-545.
JAPKOWICZ, N., MYERS, C., & GLUCK, M. (1995). A novelty detection approach to classification. In *Proceedings of the 14th International Joint Conference on Artificial Intelligence* (Vol. 1, pp. 518-523). Montreal.
JOLIFFE, I. T. (1986). *Principal component analysis*. New York: Springer.
KARNIN, E. D. (1990). A simple procedure for pruning back-propagation trained neural networks. *IEEE Transactions on Neural Networks*, **1**, 239-242.
KNAPP, A. G., & ANDERSON, J. A. (1984). Theory of categorization based on distributed memory storage. *Journal of Experimental Psychology: Learning, Memory, & Cognition*, **10**, 616-637.
KNOWLTON, B. J., & SQUIRE, L. R. (1993). The learning of categories: Parallel brain systems for item memory and category knowledge. *Science*, **262**, 1747-1749.
KOHONEN, T. (1977). *Associative memories*. Berlin: Springer.
KOLEN, J. F., & POLLACK, J. B. (1990). Back-propagation is sensitive to initial conditions. *Complex Systems*, **4**, 269-280.
KRAMER, M. A. (1991). Nonlinear principal components analysis using autoassociative neural networks. *American Institute of Chemical Engineers Journal*, **37**, 233-243.
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. (2005). Category learning. In K. Lamberts & R. Goldstone (Eds.), *The handbook of cognition* (pp. 183-201). London: Sage.
KURTZ, K. J. (2005). The Divergent Autoencoder (DIVA) account of human category learning. In B. Bara, L. Barsalou, & M. Bucciarelli (Eds.), *Proceedings of the 27th Annual Conference of the Cognitive Science Society* (pp. 1214-1219). Mahwah, NJ: Erlbaum.
KURTZ, K. J., & BECK, J. (2007). *On the locus of learning in supervised classification: A componential analysis*. Manuscript submitted for publication.
KURTZ, K. J., & DIETRICH, E. (2007). *Construing categories*. Manuscript submitted for publication.
KURTZ, K. J., MARTIN, M., & WALKER-HODKIN, A. (2007). *On the roles of abstraction and attention in human category learning: Revisiting a classic result*. *Manuscript in preparation*.
KURTZ, K. J., & SMITH, G. (2007). *The ORACL account of the internal structure of concepts*. Manuscript in preparation.
LEE, M. D., & WEBB, M. R. (2005). Modeling individual differences in cognition. *Psychonomic Bulletin & Review*, **12**, 605-621.
LOVE, B. C. (2002). Comparing supervised and unsupervised category learning. *Psychonomic Bulletin & Review*, **9**, 829-835.
LOVE, B. C., MEDIN, D. L., & GURECKIS, T. M. (2004). SUSTAIN: A network model of category learning. *Psychological Review*, **111**, 309-332.
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.
MARESCHAL, D., & FRENCH, R. M. (2000). Mechanisms of categorization in infancy. *Infancy*, **1**, 59-76.
MARESCHAL, D., QUINN, P. C., & FRENCH, R. M. (2002). Asymmetric interference in 3- to 4-month-olds' sequential category learning. *Cognitive Science*, **26**, 377-389.
MARKMAN, A. B., & ROSS, B. H. (2003). Category use and category learning. *Psychological Bulletin*, **129**, 592-613.
MARR, D. (1982). *Vision*. San Francisco: Freeman.
MATSUKA, T. (2004). Generalized exploratory models of human category learning. *International Journal of Computational Intelligence*, **1**, 8-15.
MCCLELLAND, J. L., & RUMELHART, D. E. (1985). Distributed memory and the representation of general and specific information. *Journal of Experimental Psychology: General*, **114**, 159-188.
MCCLELLAND, J. L., & RUMELHART, D. E. (1986). A distributed model of memory. In D. E. Rumelhart & J. L. McClelland (Eds.), *Parallel distributed processing: Explorations in the microstructure of cognition. Vol 2: Applications* (pp. 170-215). Cambridge, MA: MIT Press.
MCCLOSKEY, M., & COHEN, N. J. (1989). Catastrophic interference in connectionist networks: The sequential learning problem. In G. H. Bower (Ed.), *The psychology of learning and motivation* (Vol. 24, pp. 109-165). New York: Academic Press.
MEDIN, D. L. (1989). Concepts and conceptual structure. *American Psychologist*, **44**, 1469-1481.
MEDIN, D. L., GOLDSTONE, R. L., & GENTNER, D. (1993). Respects for similarity. *Psychological Review*, **100**, 254-278.
MEDIN, D. L., & ROSS, B. H. (1989). The specific character of abstract thought: Categorization, problem solving, and induction. In R. J. Sternberg (Ed.), *Advances in the psychology of human intelligence* (Vol. 5, pp. 189-223). Hillsdale, NJ: Erlbaum.
MEDIN, D. L., & SCHAFFER, M. M. (1978). Context theory of classification learning. *Psychological Review*, **85**, 207-238.
MEDIN, D. L., & SCHWANENFLUGEL, P. J. (1981). Linear separability in classification learning. *Journal of Experimental Psychology: Human Learning & Memory*, **7**, 355-368.
MINDA, J. P., & SMITH, J. D. (2001). Prototypes in category learning: The effects of category size, category structure, and stimulus complexity. *Journal of Experimental Psychology: Learning, Memory, & Cognition*, **27**, 775-799.
MINDA, J., & SMITH, J. (2002). Comparing prototype-based and exemplar-based accounts of category learning and attentional allocation. *Journal of Experimental Psychology: Learning, Memory, & Cognition*, **28**, 275-292.
MURPHY, G. L. (2002). *The big book of concepts*. Cambridge, MA: MIT Press.
MURPHY, G. L. (ED.) (2005). The study of concepts inside and outside the laboratory: Medin versus Medin. In W. K. Ahn, R. L. Goldstone, B. C. Love, A. B. Markman, & P. W. Wolff (Eds.), *Categorization inside and outside the laboratory* (pp. 179-195). Washington, DC: American Psychological Association.
MURPHY, G. L., & MEDIN, D. L. (1985). The role of theories in conceptual coherence. *Psychological Review*, **92**, 289-316.
NOSOFSKY, R. M. (1986). Attention, similarity and the identification-categorization relationship. *Journal of Experimental Psychology: General*, **115**, 39-57.
NOSOFSKY, R. M. (1988). Similarity, frequency, and category representations. *Journal of Experimental Psychology: Learning, Memory, & Cognition*, **14**, 54-65.
NOSOFSKY, R. M. (1991). Tests of an exemplar model for relating classification and recognition memory. *Journal of Experimental Psychology: Human Perception & Performance*, **17**, 3-27.
NOSOFSKY, R. M. (2000). Exemplar representation without generalization? Comment on Smith and Minda's (2000) "Thirty categorization results in search of a model." *Journal of Experimental Psychology: Learning, Memory, & Cognition*, **26**, 1735-1743.
NOSOFSKY, R. M., GLUCK, M. A., PALMERI, T. J., MCKINLEY, S. C., & GLAUTHIER, P. (1994). Comparing models of rule-based classification learning: A replication and extension of Shepard, Hovland, and Jenkins (1961). *Memory & Cognition*, **22**, 352-369.
NOSOFSKY, R. M., & JOHANSEN, M. K. (2000). Exemplar-based accounts of "multiple-system" phenomena in perceptual categorization. *Psychonomic Bulletin & Review*, **7**, 375-402.
NOSOFSKY, R. M., KRUSCHKE, J., & MCKINLEY, S. (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. (1998). A rule-plus-exception model for classifying objects in continuous-dimension spaces. *Psychonomic Bulletin & Review*, **5**, 345-369.
NOSOFSKY, R. M., PALMERI, T. J., & MCKINLEY, S. (1994). Rule-plus-exception model of classification learning. *Psychological Review*, **101**, 55-79.
NOSOFSKY, R. M., & ZAKI, S. R. (1998). Dissociations between categorization and recognition in amnesic individuals: An exemplar-based interpretation. *Psychological Science*, **9**, 247-255.
OJA, E. (1983). *Subspace methods of pattern recognition*. New York: Wiley.
OJA, E. (1989). Neural networks, principal components, and subspaces. *International Journal of Neural Systems*, **1**, 61-68.

O'REILLY, R. (1998). Six principles for biologically based computational models of cortical cognition. *Trends in Cognitive Science*, **2**, 455-462.
OSHERSON, D. N., SMITH, E. E., WILKIE, O., LOPEZ, A., & SHAFIR, E. (1990). Category-based induction. *Psychological Review*, **97**, 185-200.
PALMERI, T. J. (1999). Learning categories at different hierarchical levels: A comparison of category learning models. *Psychonomic Bulletin & Review*, **6**, 495-503.
PALMERI, T. J., & NOELLE, D. C. (2002). Concept learning. In M. Arbib (Ed.), *Handbook of brain theory and neural networks* (pp. 234-237). Cambridge, MA: MIT Press.
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.
POGGIO, T., & GIROSI, F. (1990). Regularization algorithms for learning that are equivalent to multilayer networks. *Science*, **247**, 978-982.
POSNER, M. I. (1964). Information reduction in the analysis of sequential tasks. *Psychology Review*, **71**, 491-504.
POSNER, M. I., & KEELE, S. (1968). On the genesis of abstract ideas. *Journal of Experimental Psychology*, **77**, 353-363.
RATCLIFF, R. (1990). Connectionist models of recognition memory: Constraints imposed by learning and forgetting functions. *Psychological Review*, **97**, 285-308.
REED, S. (1972). Pattern recognition and categorization. *Cognitive Psychology*, **3**, 382-407.
REED, J. M., SQUIRE, L. R., PATALANO, A. L., SMITH, E. E., & JONIDES, J. (1999). Learning about categories that are defined by object-like stimuli despite impaired declarative memory. *Behavioral Neuroscience*, **113**, 411-419.
REHDER, B. (2003). A causal-model theory of conceptual representation and categorization. *Journal of Experimental Psychology: Learning, Memory, & Cognition*, **29**, 1141-1159.
REHDER, B., & HOFFMAN, A. B. (2005a). Eyetracking and selective attention in category learning. *Cognitive Psychology*, **51**, 1-41.
REHDER, B., & HOFFMAN, A. B. (2005b). Thirty-something categorization results explained: Selective attention, eyetracking, and model of category learning. *Journal of Experimental Psychology: Learning, Memory, & Cognition*, **31**, 811-829.
RIPS, L. J. (1975). Inductive judgments about natural categories. *Journal of Verbal Learning & Verbal Behavior*, **14**, 665-681.
ROBERTS, S., & PASCHLER, H. (2000). How persuasive is a good fit? A comment on theory testing. *Psychological Review*, **107**, 358-367.
RODGERS, J., & ROWE, D. C. (2002). Theory development should begin (but not end) with good empirical fits: A comment on Roberts and Pashler. *Psychological Review*, **109**, 599-604.
ROGERS, T., & MCCLELLAND, J. L. (2004). *Semantic cognition: A parallel distributed processing approach*. Cambridge, MA: MIT Press.
ROSCH, E., & MERVIS, C. (1975). Family resemblances: Studies in the internal structure of categories. *Cognitive Psychology*, **7**, 573-605.
RUMELHART, D. E. (1989). Toward a microstructural account of human reasoning. In S. Vosniadou & A. Ortony (Eds.), *Similarity and analogical reasoning* (pp. 298-312). New York: Cambridge University Press.
RUMELHART, D. E. (1990). Brain style computation: Learning and generalization. In S. F. Zornetzer, J. L. Davis, & C. Lau (Eds.), *An introduction to neural and electronic networks* (pp. 405-420). San Diego: Academic Press.
RUMELHART, D. E., DURBIN, R., GOLDEN, R., & CHAUVIN, Y. (1995). Backpropagation: The basic theory. In Y. Chauvin & D. E. Rumelhart (Eds.), *Mathematical perspectives on neural networks* (pp. 533-566). Mahwah, NJ: Erlbaum.
RUMELHART, D. E., HINTON, G. E., & WILLIAMS, R. J. (1986). Learning internal representations by error propagation. In D. E. Rumelhart, J. L. McClelland, & the PDP Research Group (Eds.), *Parallel distributed processing: Explorations in the microstructure of cognition. Vol 1. Foundations* (pp. 318-362). Cambridge, MA: MIT Press, Bradford Books.
RUMELHART, D. E., MCCLELLAND, J. L., & THE PDP RESEARCH GROUP (1986). *Parallel distributed processing: Explorations in the microstructure of cognition. Vol. 1: Foundations*. Cambridge, MA: MIT Press, Bradford Press.
SAEGUSA, R., SAKANO, H., & HASHIMOTO, S. (2004). A nonlinear principal component analysis on image data. In *Proceedings of the 14th IEEE International Workshop on Machine Learning for Signal Processing*.
SAKAMOTO, Y., & LOVE, B. (2004). Schematic influences on category learning and recognition memory. *Journal of Experimental Psychology: General*, **133**, 534-553.
SCHYNS, P., GOLDSTONE, R., & THIBAUT, J. (1998). The development of features in object concepts. *Behavioral & Brain Sciences*, **21**, 1-54.
SHANKS, D. (1991). Categorization by a connectionist network. *Journal of Experimental Psychology: Learning, Memory, & Cognition*, **17**, 433-443.
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. L., & JENKINS, H. M. (1961). Learning and memorization of classifications. *Psychological Monographs*, **75**, 42.
SMITH, E. E., & MEDIN, D. L. (1981). *Categories and concepts*. Cambridge, MA: Harvard University Press.
SMITH, E. E., PATALANO, A. L., & JONIDES, J. (1998). Alternative strategies of categorization. *Cognition*, **65**, 167-196.
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.
WATTENMAKER, W. D. (1991). Learning modes, feature correlations, and memory-based categorization. *Journal of Experimental Psychology: Learning, Memory, & Cognition*, **17**, 908-923.
WISNIEWSKI, E. J., & MEDIN, D. L. (1994). On the interaction of theory and data in concept learning. *Cognitive Science*, **18**, 221-281.
XIE, X., & SEUNG, H. (2003). Equivalence of backpropagation and contrastive Hebbian learning in a layered network. *Neural Computation*, **15**, 441-454.
YAMAUCHI, T., & MARKMAN, A. (1998). Category learning by inference and categorization. *Journal of Memory & Language*, **39**, 124-148.
YAMAUCHI, T., LOVE, B., & MARKMAN, A. (2002). Learning nonlinearly separable categories by inference and classification. *Journal of Experimental Psychology: Learning, Memory, & Cognition*, **28**, 585-593.
ZAKI, S. R., & NOSOFSKY, R. M. (2001). A single-system interpretation of dissociations between recognition and categorization in a task involving object-like stimuli. *Cognitive, Affective, & Behavioral Neuroscience*, **1**, 344-359.
ZHANG, G. (2000). Neural networks for classification: A survey. *IEEE Transactions on System, Man, & Cybernetics*, **30**, 451-462.

NOTE

1. By coincidence, there is a neural network model of speech production (Guenther, 1995) also called DIVA. The two models are entirely independent of one another.

(Manuscript received September 8, 2006;
revision accepted for publication December 20, 2006.)