Psychon Bull Rev
DOI 10.3758/s13423-016-1208-1

BRIEF REPORT

Similar to the category, but not the exemplars: A study of generalization

Nolan Conaway[1,2] · **Kenneth J. Kurtz**[1]

Abstract Reference point approaches have dominated the study of categorization for decades by explaining classification learning in terms of similarity to stored exemplars or averages of exemplars. The most successful reference point models are firmly grounded in the associative learning tradition—treating categorization as a stimulus generalization process based on inverse exponential distance in psychological space augmented by a dimensional selective attention mechanism. We present experiments that pose a significant challenge to popular reference point accounts which explain categorization in terms of stimulus generalization from exemplars, prototypes, or adaptive clusters. DIVA, a similarity-based alternative to the reference point framework, provides a successful account of the human data. These findings suggest that a successful psychology of categorization may need to look beyond stimulus generalization and toward a view of category learning as the induction of a richer model of the data.

Keywords Categorization · Concepts · Classification learning · Generalization · Formal models · Stimulus generalization theory · Neural network models

Introduction

A central goal of categorization research is to understand how people learn to correctly classify a set of items based on experience. The most complete and widely accepted accounts of classification learning conform to a ‘reference point’ framework—according to which learners make classification decisions by evaluating the similarity of a target to stored locations in the dimensional space of the objects in the domain (for overviews, see Murphy, 2002; Pothos and Wills, 2011). There has been extensive debate over the particulars within the reference point framework (e.g., Homa, 1984; Nosofsky, 1992a; Smith and Minda, 2000; reviewed in Murphy, 2002). For example, the prototype view (Homa et al., 1979; Minda & Smith, 2001; Posner & Keele, 1968; Reed, 1972; Rosch & Mervis, 1975) posits that categories are represented by the central tendency (prototype) of the category members experienced by the learner. Exemplar theory (Brooks, 1978; Kruschke, 1992; Medin & Schaffer, 1978; Nosofsky, 1984; 1986) states that category representations consist of a collection of observations stored in memory. There is also an intermediate position that posits reference points summarizing clusters of examples (Love et al., 2004; Vanpaemel & Storms, 2008); see also Anderson (1991) and Rosseel (2002). Our present question is about a core design principle held in common across all of these: that learners categorize based on similarity to stored reference points.

Generally speaking, the term “reference point model” can refer to any member of broad class of formal approaches in which category representations consist of stored points in a multidimensional space and inputs can be mapped into that space for the purposes of determining geometric

✉ Nolan Conaway
nconaway@wisc.edu

[1] Department of Psychology, Binghamton University, Binghamton, NY, USA

[2] Present address: University of Wisconsin-Madison, 1202 West Johnson Street, Madison, WI 53706, USA

Published online: 15 December 2016

distances. Here, we use the term to describe canonical versions of the reference point framework in which classification is explained in terms of two underlying explanatory principles. The first follows from Shepard's (1957, 1987) influential work in stimulus generalization: leading models (ALCOVE, Kruschke (1992); SUSTAIN, Love et al. (2004); GCM, Nosofsky (1986)) assume that similarity to reference points is computed according to an inverse exponential function of distance in a task-independent psychological space. By consequence, the close proximity of a target to a stored reference point produces strong evidence in favor of the category associated with that reference point. The impact of similarity to a reference point drops off sharply with distance (the sensitivity or specificity of the region of activation around a reference point is determined by a free parameter). Exemplar models lack any form of abstraction in the selection of reference points and therefore manifest stimulus generalization theory directly; prototype and cluster models deviate only in the presence of an abstraction process allowing reference points to localize as the average (centroid) of multiple observations. The second explanatory principle is the use of dimensional selective attention such that the spatial distance between a target and reference points is computed under the potential stretching or shrinking of each dimension. The dimensions are assumed to exist within a 'psychological space' space of the inputs, typically estimated using Multidimensional Scaling (MDS) analyses conducted on pairwise similarity data (i.e., confusion matrices; see Shepard, 1957, 1987). Beyond attentional stretching and shrinking of dimensions, the psychological space is assumed to be task-independent: stimulus representation is not altered in the course of category learning, and the same representation is used in many different cognitive processes (e.g., classification, recognition, inference; see Nosofsky, 1992b).

Exemplar models, based on stimulus generalization theory supplemented with selective attention, have achieved unparalleled levels of success in fitting human performance in traditional artificial classification learning. To be specific, either the GCM or ALCOVE (an implementation of the exemplar view as a network model with error-driven learning of attentional and associative weights) have provided compelling accounts of how people learn: ill-defined categories like the 5-4 problem (Medin and Schaffer, 1978; Nosofsky, 1984), elemental classifications (Shepard et al., 1961; Nosofsky et al., 1994a), the relation between categorization and memory (Nosofsky, 1986), the inverse base rate effect (Kruschke, 1992, 2001, 2003), and attentional filtration versus condensation of stimulus dimensions (Gottwald & Garner, 1972; Kruschke, 1993). The available evidence has led to fairly broad acceptance of the idea that classification learning is well explained in terms of stimulus generalization plus selective attention. Much of the ongoing debate in classification learning research has focused on the possible need for hybrid models or separate systems that form explicit verbal rules (Anderson & Betz, 2001; Ashby et al., 1998; Denton et al., 2008; Erickson & Kruschke, 1998; 2002; Nosofsky et al., 1994b).

Our approach: testing the explanatory power of categorization as stimulus generalization

We present experiments challenging the idea that people make classification responses using stimulus generalization from stored reference points within a task-independent psychological space of the stimuli. The design of the behavioral studies follows from a priori predictions made by two formal models: Nosofsky's (1984, 1986) Generalized Context Model (GCM), and the *DIV*ergent *A*utoencoder model (DIVA; Kurtz, 2007, 2015). The GCM is the canonical exemplar-based model for predicting the overall ease of learning a classification problem and classification performance after learning. This is accomplished by computing the attention-weighted similarity of targets to all stored exemplars. Classification behavior is modeled based on the relative summed similarity of the stimulus to members of each category—where the category with the greatest summed similarity will garner the greatest classification probability.

We contrast the GCM's predictions with an approach that does not conform to stimulus generalization theory. DIVA (Kurtz, 2007, 2015) offers a theoretical alternative to the reference point framework by representing each category as a generative model (Ng & Jordan, 2002) of the statistical regularities among its members. The DIVA model (see Fig. 1) is instantiated fully within the connectionist tradition—at the start of a training run, DIVA is initialized with a collection of input units encoding feature values of the objects within the domain, a hidden recoding layer shared among the categories, and divergent output channels corresponding to each of the known classes. Rather than learning to associate reference points with class labels, the weights in the DIVA network are trained auto-associatively—to learn

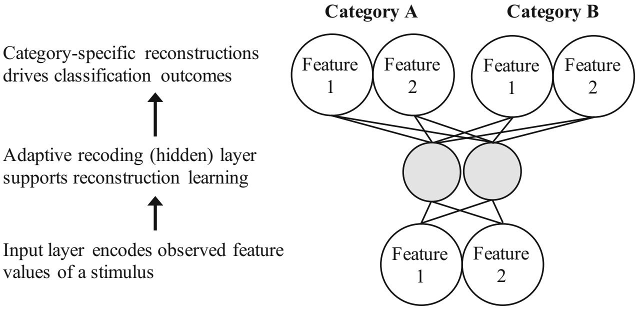
Category A
Category B
Category-specific reconstructions drives classification outcomes
Adaptive recoding (hidden) layer supports reconstruction learning
Input layer encodes observed feature values of a stimulus
Feature 1
Feature 2
Feature 1
Feature 2
Feature 1
Feature 2

Fig. 1 Depiction of the DIVA network

to predict features within each category—using the standard backpropagation technique (Rumelhart et al., 1986): when an object is observed in concurrence with its correct category label, the network's weights are updated to minimize reconstructive error along the correct category channel. Classification is then based on the relative amount of reconstructive error across category channels—DIVA is most likely to classify an item along the channel with the least error. In sum, the model learns to successfully reconstruct the members of each category along their category channel and produces distorted versions of the input when trying to reconstruct the input along the wrong channel. The training examples for each category and novel examples that are sufficiently like them will be well reconstructed along the appropriate channel (likely to be seen as a category member), while anything else will be poorly reconstructed (likely to be rejected as a category member).

DIVA and the GCM are comparable in that classification under both models is driven by similarity. Importantly, however, they differ in their use of similarity to classify objects. The GCM classifies based on attention-weighted distance to exemplar reference points and explicitly relies on Shepard's law. In contrast, DIVA classifies examples based on relative reconstructive success (how well an example conforms to the type of input that the model has learned to recode and decode with minimal distortion along each category channel). Importantly, reconstructive success can deviate from the predictions of attention-weighted distance, and thus the contrast between DIVA and the GCM highlights the central role of stimulus generalization in category learning. We can therefore evaluate the role of stimulus generalization in category learning by comparing the predictions made by the two models.

A priori model simulations

We began with a comparison of the generalization performance of the two models after training on a variation of the well-known exclusive-OR (XOR) category structure. XOR categories are commonly studied in psychology and machine learning research. To give an illustrative case: one category might consist of white squares (00) and black circles (11), while the contrast category consists of black squares (10) and white circles (01). The usual set-up for XOR classification learning tasks uses binary stimulus dimensions—meaning that there can be no test of generalization. We instead used a two-dimensional, continuous adaptation of the XOR structure that maintains the logical structure of the categories while allowing a test of generalization. We found that the two models were broadly consistent in their predictions for learning and generalization.

However, a much more interesting outcome arose when we tested a variation of XOR with the training set altered: one category remained intact, but the other was reduced by half such that one of the four traditional quadrants was left untrained (see Fig. 2). This alteration makes a considerable impact by changing the non-linearly separable XOR problem into one that can be learned in terms of a single diagonal boundary that separates the categories. Given training on this partial-XOR structure, we found that DIVA tended to produce two qualitatively distinct generalization behaviors across different training runs with the same parameterization: (1) extending the Complete (two-quadrant) category to the untrained area or (2) extrapolating the Reduced (one-quadrant) category to the untrained area. Note that the second pattern parallels the standard XOR structure.

The second prediction is highly notable because the critical test items in the untrained quadrant are more proximal to the exemplars of the Complete category. Specifically, the central exemplar in the untrained quadrant is, on average, 1.67 city blocks away from members of the Complete category and 3 city blocks away from members of the Reduced category (consistent results arise using a Euclidean metric; our simulations use the city block metric due to the separable dimensions in the stimuli we study, see Garner (1974)). To be clear, the second predicted pattern of generalization amounts to a dissociation between proximity to exemplars and categorization: the target is classified as a member of the Reduced category even though it is closer to known members of the Complete category.

To estimate the frequency of Reduced category generalization in DIVA, we conducted a 'grid-search' to generate predictions for partial-XOR generalization over a wide range of settings for DIVA's four parameters: number of hidden nodes, learning rate, initial weight range, and a focusing parameter, β (Conaway and Kurtz, 2014; Kurtz, 2015). At each point in the search, DIVA was initialized 2000 times (random initial weights and random presentation sequence) and trained for 12 blocks (12 observations of each of the training exemplars). The exemplars were represented within a ±1 space, DIVA's output units were linear, and its hidden units were logistic. For each point in the grid-search, we calculated the proportion of initializations that produced the pattern of interest operationalized as a clear majority

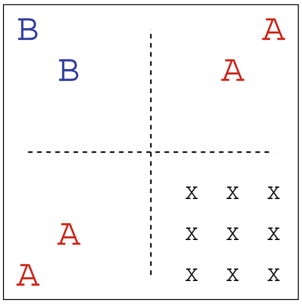

Fig. 2 Partial XOR categories. Critical transfer items are marked with *X*

(six or more out of nine) of items in the critical generalization region being classified as a member of the Reduced category. Results are plotted as a density distribution across all points in the grid search (Fig. 3). As can be seen, there are many parameterizations of DIVA that produce the key behavior of extrapolating the Reduced category.

To provide a description of what enables this behavior in DIVA, we first examined the parameterizations which most commonly produced large rates of extrapolation (≥ 65 % of initializations). These parameterizations had in common a smaller recoding space ($2-5$ hidden units), a strong learning rate ($0.9 - 1.0$), and a small to moderate weight range (typically less than ± 1.5). DIVA's β parameter did not appear to affect the rate of extrapolation. In-depth examination of fully trained DIVA networks revealed the model's basis for this prediction. In every observed instance of extrapolation, all of DIVA's hidden units encoded the value of just one of the two stimulus dimensions (e.g., size). The model learns strongly positive weights connecting the hidden units to the Complete category channel, indicating it has learned a key within-category feature correlation: within the Complete category, exemplars are either small and black or large and white. At test, this channel's reconstructions follow the correlation directly—novel exemplars are reconstructed along a diagonal line capturing its training items, and interpolating all other items between the two clusters of Complete category exemplars.

The network, however, learns a more varied basis to reconstruct the Reduced category items from this hidden recoding. Whereas the weights connected to the feature encoded in the hidden representation (i.e., size) are strongly positive, the weights connected to the opposite feature (i.e., color) are strongly negative, indicating that DIVA has learned the opposite correlation within the Reduced category. At test, the Reduced category channel therefore reconstructs novel items along the negative diagonal, capturing the Reduced category exemplars and extending into the untrained area. In effect, the model learns feature correlations capturing each category's exemplars, and it learns that the direction of the correlation is opposite between the categories. See Fig. 4 for a depiction of this solution.

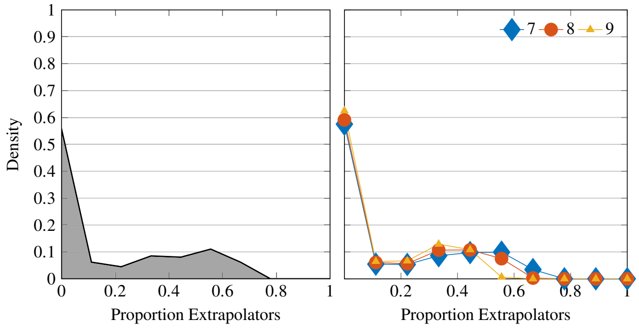

Fig. 3 DIVA 'grid-search' results. *Left* Any initialization which produced at least 6/9 Reduced category responses to items in the untrained quadrant is counted as an extrapolator. *Right* Results visualized with alternative response thresholds (7-9 Reduced category responses)

By contrast, true to its foundation in stimulus generalization, the GCM's performance was typified by generalization based on the more proximal exemplars of the Complete category (although neutral generalization can be achieved with a more extreme value for the sensitivity parameter). Importantly, the GCM's predictions are not specific to exemplar representations: stimulus generalization from learned category prototypes or clusters of examples does not change the underlying similarity dynamics and therefore produces the same behavior. As such, the GCM's performance effectively stands in for the whole explanatory framework based on stimulus generalization from reference points.

Experiment 1

The a priori model predictions clearly motivate a behavioral study to test human generalization performance after learning the partial-XOR categories. While behavioral experiments using XOR categories are ubiquitous in the category learning literature, the partial-XOR structure has only been lightly studied. Bourne (1982) and Nosofsky (1991) employed a partial-XOR structure using a binary, four-dimensional stimulus set, but their analyses focus primarily on typicality measures and neither report addresses how individual learners generalize to the untrained quadrant.

We were interested in determining whether human learners would extrapolate the Reduced category to the critical region despite the items in that quadrant being closer to the training examples in the two quadrants of the Complete category. In doing so, we test the core premise of categorization as stimulus generalization. If DIVA makes a

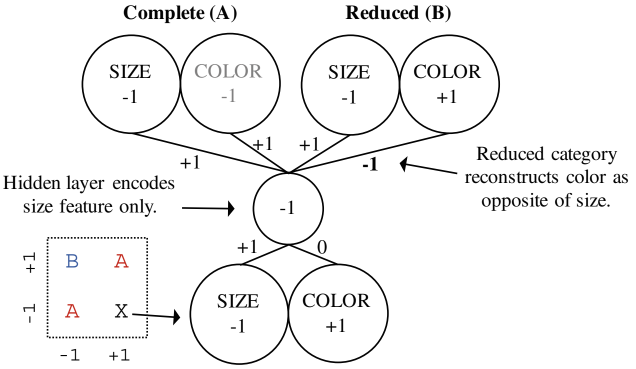
Complete (A)
Reduced (B)
SIZE -1
COLOR -1
SIZE -1
COLOR +1
+1
+1
+1
-1
Reduced category reconstructs color as opposite of size.
Hidden layer encodes size feature only.
-1
+1
0
B
A
A
X
+1
-1
-1
+1
SIZE -1
COLOR +1

Fig. 4 Depiction of DIVA's solution supporting extrapolation of the Reduced (B) category to critical items *X*. The above network also correctly classifies the training items. The values depicted are purely illustrative—hidden activations are logistic and bias units are used in our DIVA simulations. Other solutions are learned to support Proximity-based generalization

psychologically valid prediction, then we should expect a mix of two profiles of generalization performance; if the reference point framework is correct, then generalization should be proximity-based or neutral, but never driven by extrapolation of the Reduced category.

Participants and materials

Thirty undergraduates from Binghamton University participated toward partial fulfillment of a course requirement. Stimuli were squares varying in shading and size (see Fig. 5 for samples). Exemplars were automatically generated at seven positions on each dimension (7 shading x 7 size = 49 examples). In an independent scaling study with these materials, we found the two dimensions to be nearly equal in perceptual salience. The assignment between perceptual and conceptual dimensions was randomized across participants.

Procedure

Participants completed 96 training trials (12 randomized blocks of eight items). In order to equate for category frequency, the two Reduced category exemplars were presented twice within each block. This way of handling the unbalanced classification was considered the best option, although it does create variation in item presentation frequency (see Nosofsky, 1988). Little is known about how item frequency impacts generalization. After training, participants completed 49 generalization trials consisting of items sampled at seven positions on each dimension. All training examples were included (intermixed) in the generalization phase.

Participants were informed that there would be test trials prior to beginning the experiment. The instructions were neutral in that they did not encourage learners to engage in hypothesis testing to discover a rule (see Kurtz et al., 2013). On each trial, a single stimulus was presented on a computer screen and learners were prompted to make a classification decision by clicking one of two buttons labeled ‘Alpha’ and ‘Beta’. During the training phase, learners were given corrective feedback on their selection. Feedback was not provided during the generalization phase. We note that as part of this experiment we also collected data in a full-XOR condition (a description of these data can be found in Conaway and Kurtz, 2015).

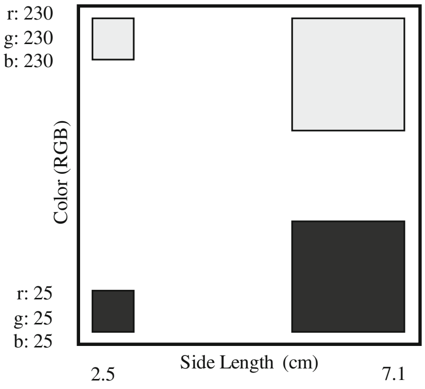

Fig. 5 Sample stimuli (not drawn to scale)

Results

Learners had little trouble acquiring the categories: classification performance was highly accurate by the end of the training phase (typified by perfect accuracy in the final training block). To assess generalization performance, we calculated the proportion of Reduced category responses made by each learner for the nine novel test items in the untrained quadrant. A histogram of the data (see Fig. 6) shows two qualitatively distinct groups of learners. A majority of the learners were proximity-based generalizers of the Complete category, but a substantial group of learners systematically generalized from the (more distant) Reduced category. Specifically, 9/30 learners were extrapolators who produced six or more Reduced category responses to the test items. The average generalization gradient for each subgroup is shown in Fig. 7. The proximity ($n = 20$) and extrapolation ($n = 9$) subgroups did not differ in their training accuracy, $t(27) = 0.012$, $p > 0.6$. See Table 1 for descriptive statistics for the performance of these groups on the training exemplars.

Discussion

Our behavioral experiment was broadly supportive of DIVA’s predictions. After training on a partial-XOR classification, a subset of learners extrapolated the Reduced category to novel items that are more proximal to members

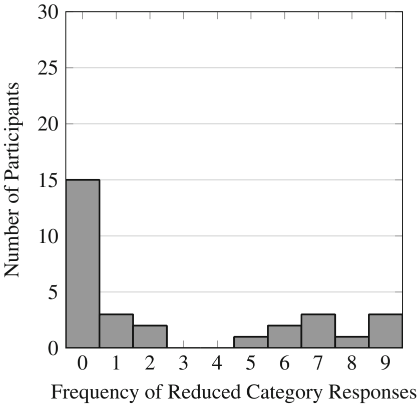

Fig. 6 Histogram of Reduced category responses to items in the untrained area

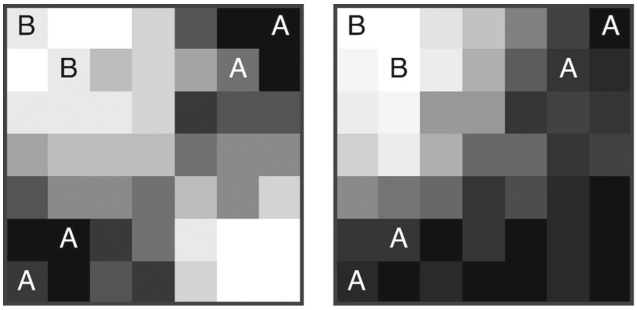

Fig. 7 Generalization gradients for the extrapolation (*left*, $n = 9$) and proximity (*right*, $n = 20$) subgroups (one participant who made 5 Reduced category responses was excluded)

of the Complete category. This pattern of generalization, while not modal, occurred frequently and therefore demands psychological explanation. The challenge for models is to account for the distribution of two commonly occurring profiles. While the GCM can to handle the majority result (proximity-based generalization), its commitment to stimulus generalization theory undercuts the model's ability to predict the second major profile. By contrast, DIVA could successfully predict the occurrence (and relative rate) of both proximity and extrapolation-based generalization. The observed proportion of human learners who were extrapolators fits well with the density distribution shown in Fig. 3. Since DIVA is free from a commitment to similarity as distance to reference points, it is able to predict extrapolation of the Reduced category to novel exemplars in the missing quadrant.

The presence of the extrapolator subgroup would appear to conflict with the core claims made by reference point theories, and it would appear to support DIVA's account that people acquire knowledge of the within-class feature relationships. However, it is important to more carefully consider how Reduced category extrapolation could be explained within the reference point framework. For example, because items in the untrained area are dissimilar to the known members of the Reduced and Complete categories, reference point models could predict that some learners are guessing at test. Is it possible that the learners in our extrapolation subgroup simply classified the novel items at random? Some doubt can be shed on this account based on the design of our experiment: because we tested participants on a total of nine exemplars within the critical region, learners using a guessing strategy are most likely to produce 4–5 Reduced category responses to these exemplars, and larger differentials are less probable. We, however, observed only one participant who produced 4–5 Reduced category responses (see Fig. 6). Nonetheless, lacking relevant behavioral data, we cannot entirely dismiss guessing as an account of our results.

Table 1 Subgroup accuracies on the training exemplars presented during the training and generalization phases (standard deviation in *parenthesis*)

	Training	Generalization
Proximity	0.93(0.06)	0.96(0.09)
Extrapolation	0.92(0.11)	0.89(0.12)

Alternatively, it is also possible that learners encoded as an additional feature whether the stimulus is along the positive (Complete category) or the negative (Reduced category) diagonal. This additional feature would render the novel items in the untrained quadrant more proximal to members of the Reduced category, producing extrapolation. Problematically, this account would violate a core claim made by stimulus generalization as applied to category learning: that the stimulus encoding is independent of the classification learning task (Nosofsky, 1992b). But, as with the guessing account, we cannot rule out feature learning without further experiments.

In Experiments 2A and 2B, we report behavioral studies that explicitly address these two accounts (*Guessing* and *Feature Learning*). In each experiment, we also provide a replication of the core phenomenon: Reduced category extrapolation to novel items that are more proximal to members of the Complete category.

Experiment 2A

The guessing account of Experiment 1 proposes that, because the critical items are fairly distant from known members of both categories, many participants classified the critical items at random, and some learners by chance produced six or more Reduced category responses. We can address this account using the modification of the partial-XOR classification depicted in Fig. 8. In this variation, the

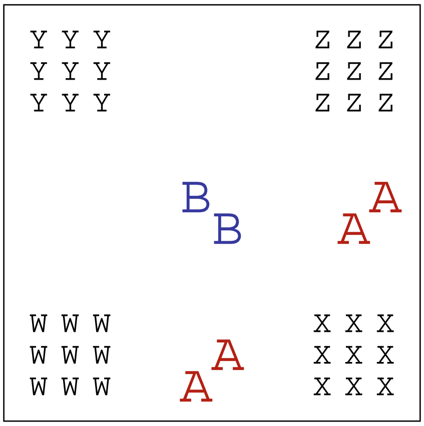

Fig. 8 Partial XOR categories tested in Experiment 2A. Critical transfer items are marked with *W*, *X*, *Y*, and *Z*

categories (A and B) lie in the lower-right quadrant, yielding four untrained areas (W, X, Y, and Z). Under this set-up, critical items X possess the same status as in Experiment 1: learners can either generalize the more proximal Complete category, or extrapolate the Reduced category. W, Y, and Z are, however, more distant from the training examples compared to X. Thus, if extrapolation simply occurs due to guessing, then extrapolators should respond at random to all the critical items. If, in contrast, the extrapolation we observed in Experiment 1 was deliberate, then extrapolators should also systematically classify items in the other critical regions.

Participants and materials

Thirty undergraduates from Binghamton University participated toward partial fulfillment of a course requirement. Stimuli were identical to Experiment 1 with the exception that they were generated at 12 (rather than seven), points along each dimension.

Procedure

The training phase was identical to Experiment 1. After training, participants completed 42 generalization trials consisting of nine items sampled from each of the four corners of the stimulus space (W, X, Y, Z; see Fig. 8), as well as the six unique training items. All other aspects of the procedure are identical to Experiment 1.

Results

As in Experiment 1, learners easily mastered the partial-XOR classification: accuracy was near-perfect in the final training block ($M = 0.99$, $SE = 0.01$). To assess generalization performance, we calculated the proportion of Reduced category responses made by each learner to the nine critical items in each quadrant of the stimulus space (W, X, Y, Z). Histograms of these data can be found in Fig. 9. As depicted in the lower-right panel of Fig. 9, we successfully replicated the core result of Experiment 1: a subset of learners (8/30) systematically generalized the Reduced category to critical items X, and a second group (21/30) generalized the proximal Complete category to these items. One participant produced 5 Reduced category responses and did not fall into either group. As before, the proximity and extrapolation subgroups did not differ in their training accuracy, $t(27) = 0.58$, $p > 0.5$.

Figure 10 depicts responses from the extrapolation and proximity subgroups to the critical items in each quadrant. If the extrapolators had simply classified critical items X at random, then they should not have systematically classified the critical items in areas W, Y, and Z. Extrapolator responses in these areas were beyond chance (4.5/9 Reduced category responses): items from W were classified into the Complete category, $t(7) = 3.93$, $p = 0.006$, items from Y were classified into the Reduced category, $t(7) = 4.48$, $p < 0.003$, and items from Z were classified into the Complete category, $t(7) = 3.32$, $p = 0.013$. The subgroups did not differ in their response to areas W, Y, and Z ($ps > 0.3$), but they did differ in their responses to X, $t(27) = 22.3$, $p < 0.001$.

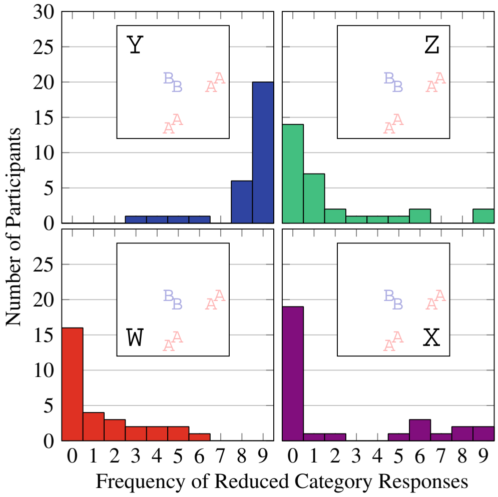

Fig. 9 Frequency of Reduced category responses to the critical areas tested in Experiment 2A

Summary

Experiment 2A provides strong evidence against a “guessing” account of Experiment 1. We observed that individuals who extrapolated the Reduced category also systematically classified critical items in other regions of the stimulus space. Thus, these learners are not likely to have assumed a random response strategy simply because the critical items are distant from the training examples. Instead, our

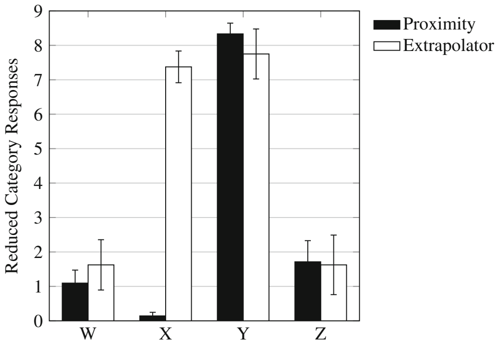

Fig. 10 Frequency of Reduced category responses to all critical areas tested in Experiment 2A. *Error bars* reflect 1 standard error

extrapolators appear to have deliberately classified critical items into the Reduced category.

Experiment 2B

Rather than learning within-class feature relationships (as proposed by DIVA), it is also possible that our extrapolators classified on the basis of distance to reference points within a higher dimensional space. Specifically, it is possible that, through classification training on the partial-XOR categories, some learners encoded a new feature essentially representing whether the stimulus lies along the positive (Complete category) or the negative (Reduced category) diagonal. At test, this feature would render the critical items more proximal to members of the Reduced category, thus producing the extrapolation profile.

It is worth nothing that feature learning processes are beyond the scope of stimulus generalization as applied to categorization. Not only is the task-independence of features a core assumption in reference point theories (Nosofsky, 1992b), but the very construct of similarity only possesses explanatory power if constraints are placed on the underlying feature representation (see Goldstone, 1994; Goodman, 1972; Medin et al., 1993; Murphy and Medin, 1985). Without proper constraints on the features, reference point models are capable of explaining theoretically any result.

To address the feature learning account of Reduced category extrapolation, we conducted a second replication and extension of Experiment 1. The key difference is that, after generalization, participants are asked to rate the similarity between pairs of examples. If Reduced category extrapolation is due to a feature learning process, then each learner's similarity ratings should correspond to their generalization performance: individuals who extrapolate should view the critical items as more similar to the Reduced category than the Complete category. The alternative pattern of results (extrapolators view the Complete category as more similar to the critical items) would stand against the very foundation of the reference point view: that perceived similarity is the basis for classification.

Participants and materials

Thirty-one undergraduates from Binghamton University participated toward partial fulfillment of a course requirement. Stimuli were identical to Experiment 1.

Procedure

The training phase was identical to Experiment 1. However, to reduce the overall number of trials, the generalization phase consisted only of targeted critical trials (rather than the entire gradient). After training, participants completed 15 generalization trials consisting of the nine critical items plus the six unique training items (see Fig. 2).

Before training and after generalization, participants completed a pairwise similarity phase. On each trial, two stimuli were presented side-by-side, and participants were asked to rate the similarity of the pair on a 1 (not at all similar) to 9 (highly similar) scale. Each similarity rating phase consisted of *(A)* trials in which a critical item (X, in Fig. 2) was paired with a training exemplar, and *(B)* trials in which training items from opposite quadrants were paired with one another. The two trial types were randomly intermixed. The data from the second type (12 trials total), as well as the data collected prior to training, was collected for a separate project—here we will focus our discussion on the ratings between critical items and training exemplars after generalization. To reduce the number of trials (there are 48 total pairs of critical and training items), we tested only a subset of the possible pairs. We constructed the similarity rating phase using four critical items (the four items from the lower-right corner of the space), and four training items (the outermost examples from the Complete category), and both examples from the Reduced category, producing a total of 16 trials.

Results

As in Experiment 1 and 2A, the partial-XOR classification was not difficult for learners to acquire: accuracy was near-perfect in the final training block ($M = 0.98$, $SE = 0.02$). We also observed a notably larger number of extrapolators (defined as individuals who produced 6+ Reduced category responses to items in the critical area): whereas $14/31$ learners extrapolated the Reduced category to the untrained area, $15/31$ learners generalized the proximal Complete category. One learner produced five Reduced category responses and did not fall into either group. The proximity and extrapolation subgroups did not differ in their training accuracy, $t(28) = 1.66$, $p = 0.11$. Figure 11 depicts the histogram of the frequency of Reduced category extrapolation.

The key question for this experiment concerns whether individuals in the extrapolation subgroup view the items in the untrained area as more similar to members of the Reduced category. If these individuals extrapolated as a result of having learned a new feature (encoding the valance of the diagonal), then they should more highly rate the similarity of Reduced-Critical pairs compared to Complete-Critical pairs. Similarity ratings for each subject are depicted in Fig. 12. As can be seen in Fig. 12, only two (of 14) extrapolators produced greater similarity ratings for Reduced-Critical pairs (and only one of these individuals produced considerably greater ratings). Although a

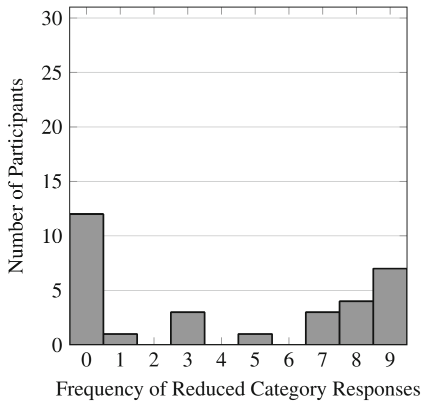

Fig. 11 Experiment 2B histogram of Reduced category responses to items in the untrained area

traditional T-Test reveals no difference between Extrapolator responses to Reduced-Critical and Complete-Critical pairs, $t(13) = 1.20$, $p = 0.25$, the outlying data point renders a nonparametric test more appropriate. A Wilcoxon signed-rank test supports a statistical difference between Extrapolator responses to Reduced-Critical and Complete-Critical pairs, $Z = 90$, $p = 0.012$. The extrapolation and proximity subgroups did not differ in their ratings of Reduced-Critical pairs, $t(28) = 0.75$, $p = 0.46$, but the proximity subgroup produced marginally greater similarity ratings for Complete-Critical pairs, $t(28) = 1.94$, $p = 0.06$.

Summary

Experiment 2B provides strong evidence against the "feature learning" account of Experiment 1: the majority of individuals in the extrapolation subgroup did not view the critical items as more similar to the Reduced category. By consequence, these learners are not likely to have extrapolated the Reduced category as a result of a feature learning process by which the Reduced category became more similar to the untrained area. These results in fact contradict the very core of the reference point account of category learning: many learners viewed critical items as a good match to the Reduced category, but more similar to members of the Complete category.

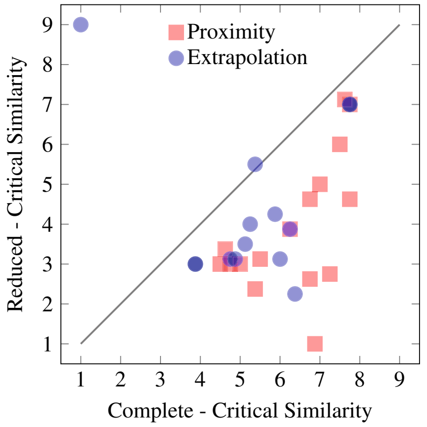

Fig. 12 Experiment 2B averaged similarity ratings between critical items and members of the Reduced and Complete categories. Each participant is plotted as a separate *point*

General discussion

According to the dominant reference point view, category knowledge is represented in terms of stored reference points that are associated with category labels. The computation of similarity to reference points is based on distance in a task-independent psychological space supplemented with an optimized set of selective attention weights. The idea that classification decisions are a function of distance (or similarity) to stored reference points accords with Shepard's (1957, 1987) universal law of generalization. Indeed, many reference point models explicitly adopt Shepard's mathematical formulation of stimulus generalization (e.g., Kruschke, 1992; Nosofsky, 1984; see also Love et al., 2004).

We report behavioral results that challenge the idea that human category learning can be reduced to stimulus generalization. Learners received classification training on a variant of the XOR category structure with two continuous dimensions and one of the four quadrants absent from the training set (partial-XOR). In one experiment (and two replications), we found that 25–50 % of learners generalized by extrapolating the Reduced (one-quadrant) category to exemplars in the missing quadrant—even though the test items were more proximal to members of the Complete (two-quadrant) category. Follow up experiments establish that these learners did not classify items in the missing quadrant by chance, and that they did not view members of the Reduced category as more similar to the items in the missing quadrant.

These results raise a challenge to accounts of categorization based on stimulus generalization from stored points of reference within a task-independent psychological space. To be clear, the argument we put forward is not that categorization cannot be explained on the basis of some form of similarity–not only is DIVA a similarity-based model, but the broader concept of similarity is extremely flexible and there are several ways of modifying reference point models to enable apparent extrapolation against distance to exemplars. For example, it is still possible that learners may have modified their category representation upon exposure to stimuli in the untrained quadrant during the test phase (i.e., Zaki and Nosofsky, 2007). This account would suggest that, initially, learners randomly responded to items in the untrained quadrant. After committing to a response, learners may have recruited additional reference points associated with the category that they responded with. Subsequent exposures to members of the untrained quadrant would then follow the classification given to the first item. Although it is difficult to directly test this account with the available data, a close inspection revealed that participant generalization often deviated from early responses–indeed, some extrapolators even classified the first item as a member of the Complete category. Zaki and Nosofsky (2007) point to the potential impact of this phenomenon when learners are provided a weak training regimen and then asked to complete a lengthy test phase. The present circumstances are not a good fit: learners were performing nearly perfectly by the end of a training phase of substantial length before a considerably shorter test phase.

Other formulations of the reference point approach aside from exemplar-based models are equally committed to similarity as distance to reference points. That is, the central tendency of each of the two obvious clusters of the Complete category (or that of more finely divided clusters) offers no advantage in terms of proximity to the test items in the untrained quadrant. Similarly, a pure prototype account of the Complete category would reduce the bimodal distribution to the central tendency at the origin—which is closer to the untrained quadrant than the members of the Reduced category.

Our results are therefore problematic for similarity-based accounts of categorization, as constrained by the core assumptions of applying stimulus generalization to category learning. We argue that the present classification learning data are not well explained by attention-weighted distance to stored points of reference, assuming that (1) reference points are represented in a task-independent psychological space of the stimuli, and (2) the reference points correspond to exemplars or averages of exemplars. Instead, investigators must look beyond categorization as similarity to reference points in studying the processes used to learn and represent category knowledge.

Looking outside of the reference point framework, another line of interpretation of our results arises from a rule-based perspective on category learning. Noting that extrapolators generalize in accord with the logical structure of full-XOR, is it possible that these learners acquired the categories according to a verbalizable logical rule? If so, a sophisticated rule-based models of category learning (i.e., RULEX; Nosofsky and Palmeri, 1998; Nosofsky et al., 1994b) might provide an account of the generalization phenomena observed. We have not conducted simulations to address this possibility, though it should be noted that RULEX follows the fundamental design principle that simpler rules are learned before more complex rules. A simple rule can be formed to pick out the Reduced category (e.g., dark AND small). A rule based on the Complete category (e.g., bright AND small OR dark AND large) would be more complex, although a shorter version of the rule is also viable: (e.g., bright OR large). Both of the simpler logical forms predict proximity-based generalization (the Reduced category-based rule excludes items that are bright and large; the simple Complete category-based rule includes items that are bright and large), whereas only the complex form predicts extrapolation.

Remarkably, the present phenomenon—that two distinct profiles of generalization can arise after training on the partial-XOR structure—is problematic to nearly all varieties of psychological theory. However, this is not a mystery without a clue. The successful simulation results with DIVA show that a formal model of category learning can predict the behavioral finding. What about DIVA underlies its success? One important property of DIVA is that the model does not entail an a priori commitment to representational constructs (exemplars, prototypes, rules). Instead, DIVA uses error-driven learning to induce a model of the distributional character of each category. Examination of learned DIVA networks reveals that the model is highly sensitive to the within-category feature correlations describing each of the partial-XOR classes. Extrapolation of the Reduced category follows directly from the knowledge that members of the Reduced and Complete categories follow opposite correlations—exemplars within the critical region are therefore more consistent with DIVA's knowledge of the Reduced category.

DIVA's success can more broadly be attributed to its ability to more flexibly construct internal representations—a property that is lacking in traditional reference point approaches. However, DIVA's learning is in no sense underconstrained. Because DIVA's learning objective is autoassociative (i.e., feature reconstruction), each category's representation can be considered as a reconstruction space into which any stimulus can be projected. The location and shape of the space reflects the model's expectations about the category, and the training items act as the primary constraint upon these properties: DIVA will find a local minimum solution to the optimization problem of reducing reconstructive error for known category members. The remainder of space gets carved up in different ways depending on the nature of the solution. In this manner, DIVA divides the entire input space into regions that are better handled by each category channel (we note the contrast here with Kruschke's (1992) critique of back-propagation as it operates in a traditional multi-layer perceptron architecture). Our work with DIVA on partial-XOR has shown that the reconstruction space associated with the Complete category is fairly stable, while the shape of the space associated with the more localized Reduced category varies considerably depending on the random initial weights and presentation order of the training exemplars. Extrapolation is achieved under a particular orientation of the Reduced category reconstruction space—to the extent that the space overlaps with the untrained quadrant, generalization by extrapolation ensues.

Conclusions

It is largely unheard of for exemplar models to fail to account for human performance on the type of straightforward artificial classification learning tasks that are their ‘home turf’. Here we have addressed the link between categorization and stimulus generalization—is it a matter of distance from reference points or is it a process more like that represented by DIVA based on building generative models of the categories? The present findings provide evidence of a weakness in the stimulus generalization account of human category learning: it is possible to be a good match to the category, but a poor match to its exemplars.

Acknowledgments A previous version of this work appeared in the Proceedings of the Thirty-Seventh Annual Conference of the Cognitive Science Society (Conaway & Kurtz, 2015). We thank Sarah Laszlo, Robert Nosofsky, Greg Murphy and other colleagues for helpful comments on this work. We additionally thank Sarah Laszlo for access to her server. Finally, we thank the members of the Learning and Representation in Cognition (LaRC) laboratory and the Binghamton University Modeling Meeting.

References

Anderson, J. R. (1991). The adaptive nature of human categorization. *Psychological Review*, *98*(3), 409.

Anderson, J. R., & Betz, J. (2001). A hybrid model of categorization. *Psychonomic Bulletin and Review*, *8*(4), 629–647.

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.

Bourne, L. E. (1982). Typicality effects in logically defined categories. *Memory and Cognition*, *10*(1), 3–9.

Brooks, L. (1978). Nonanalytic concept formation and memory for instances. In Rosch, E., & Lloyd, B. B. (Eds.) *Cognition and Categorization*. Hillsdale, N. J.

Conaway, N. B., & Kurtz, K. J. (2014). Now you know it, now you don’t: Asking the right question about category knowledge. In Bello, P., Guarini, M., McShane, M., & Scassellati, B. (Eds.) *Proceedings of the 36th Annual Conference of the Cognitive Science Society* (pp. 2062–2067). Austin, TX: Cognitive Science Society.

Conaway, N. B., & Kurtz, K. J. (2015). A dissociation between categorization and similarity to exemplars. In Noelle, D. C., & Dale, R. (Eds.) *Proceedings of the 37th Annual Conference of the Cognitive Science Society* (pp. 435–440.) Austin, TX: Cognitive Science Society.

Denton, S. E., Kruschke, J. K., & Erickson, M. A. (2008). Rule-based extrapolation: A continuing challenge for exemplar models. *Psychonomic Bulletin and Review*, *15*(4), 780–786.

Erickson, M. A., & Kruschke, J. K. (1998). Rules and exemplars in category learning. *Journal of Experimental Psychology: General*, *127*(2), 107.

Erickson, M. A., & Kruschke, J. K. (2002). Rule-based extrapolation in perceptual categorization. *Psychonomic Bulletin and Review*, *9*(1), 160–168.

Garner, W. R. (1974). *The processing of information and structure*. Erlbaum: Potomac, MD.

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

Goodman, N. (1972). Seven strictures on similarity. In Goodman, N. (Ed.) *Problems and Projects* (pp. 437–447). Indianapolis: Bobbs-Merrill.

Gottwald, R. L., & Garner, W. R. (1972). Effects of focusing strategy on speeded classification with grouping, filtering and condensation tasks. *Perception and Psychophysics*, *11*, 179–182.

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

Homa, D., Rhoads, D., & Chambliss, D. (1979). Evolution of conceptual structure. *Journal of Experimental Psychology: Human Learning and Memory*, *5*(1), 11.

Kruschke, J. K. (1992). ALCOVE: An exemplar-based connectionist model of category learning. *Psychological Review*, *99*(1), 22–44.

Kruschke, J. K. (1993). Human category learning: Implications for backpropagation models. *Connection Science*, *5*(1), 3–36.

Kruschke, J. K. (2001). The inverse base rate effect is not explained by eliminative inference. *Journal of Experimental Psychology: Learning Memory, and Cognition*, *27*, 1385–1400.

Kruschke, J. K. (2003). Attentional theory is a viable explanation of the inverse base rate effect: A reply to Winman, Wennerholm, and Juslin (2003). *Journal of Experimental Psychology: Learning Memory, and Cognition*, *29*, 1396–1400.

Kurtz, K. J. (2007). The divergent autoencoder (DIVA) model of category learning. *Psychonomic Bulletin and Review*, *14*, 560–576.

Kurtz, K. J. (2015). Human category learning: Toward a broader explanatory account. *Psychology of Learning and Motivation*, *63*, 77–114.

Kurtz, K. J., Levering, K. R., Stanton, R. D., Romero, J., & Morris, S. N. (2013). Human learning of elemental category structures: Revising the classic result of Shepard, Hovland, and Jenkins (1961). *Journal of Experimental Psychology: Learning Memory, and Cognition*, *39*(2), 552–572.

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

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

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

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, and Cognition*, *27*(3), 775.

Murphy, G. L. (2002). *The big book of concepts*. Cambridge: MIT Press.

Murphy, G. L., & Medin, D. L. (1985). The role of theories in conceptual coherence. *Psychological Review*, *92*(3), 289.

Ng, A. Y., & Jordan, M. (2002). On discriminative vs generative classifiers: A comparison of logistic regression and naive bayes. *Advances in Neural Information Processing Systems*, *14*, 841–848.

Nosofsky, R. M. (1984). Choice, similarity, and the context theory of classification. *Journal of Experimental Psychology: Learning Memory, and Cognition*, *10*(1), 104–114.

Nosofsky, R. M. (1986). Attention, similarity, and the identification-categorization relationship. *Journal of Experimental Psychology: General*, *115*(1), 39–57.

Nosofsky, R. M. (1988). Similarity, frequency, and category representations. *Journal of Experimental Psychology: Learning Memory, and Cognition*, *14*(1), 54.

Nosofsky, R. M. (1991). Typicality in logically defined categories: Exemplar-similarity versus rule instantiation. *Memory and Cognition*, *19*(2), 131–150.

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

Nosofsky, R. M. (1992b). Similarity scaling and cognitive process models. *Annual Review of Psychology*, *43*(1), 25–53.

Nosofsky, R. M., & Palmeri, T. J. (1998). A rule-plus-exception model for classifying objects in continuous-dimension spaces. *Psychonomic Bulletin and Review*, *5*(3), 345–369.

Nosofsky, R. M., Gluck, M. A., Palmeri, T. J., & McKinley, S. C. (1994a). Comparing models of rule-based classification learning: a replication and extension of Shepard, Hovland, and Jenkins (1961). *Memory and Cognition*, *22*(3), 352–369.

Nosofsky, R. M., Palmeri, T. J., & McKinley, S. C. (1994b). Rule-plus-exception model of classification learning. *Psychological Review*, *101*(1), 53–79.

Posner, M. I., & Keele, S. W. (1968). On the genesis of abstract ideas. *Journal of Experimental Psychology*, *77*, 353–363.

Pothos, E. M., & Wills, A. J. (2011). *Formal approaches in categorization.* Cambridge University Press.

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

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

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

Rumelhart, D. E., Hinton, G. E., & Williams, R. J. (1986). Learning representations by back-propagating errors. *Nature*, *323*, 533–536.

Shepard, R. N. (1957). Stimulus and response generalization: A stochastic model relating generalization to distance in psychological space. *Psychometrika*, *22*(4), 325–345.

Shepard, R. N. (1987). Toward a universal law of generalization for psychological science. *Science*, *237*(4820), 1317–1323.

Shepard, R. N., Hovland, C. I., & Jenkins, H. M. (1961). Learning and memorization of classifications. *Psychological Monographs: General and Applied*, *75*(13), 1.

Smith, J. D., & Minda, J. P. (2000). Thirty categorization results in search of a model. *Journal of Experimental Psychology: Learning Memory, and Cognition*, *26*(1), 3.

Vanpaemel, W., & Storms, G. (2008). In search of abstraction: the varying abstraction model of categorization. *Psychonomic Bulletin and Review*, *15*, 732–749.

Zaki, S. R., & Nosofsky, R. M. (2007). A high-distortion enhancement effect in the prototype-learning paradigm: Dramatic effects of category learning during test. *Memory and Cognition*, *35*(8), 2088–2096.