# Contrasting Exemplar and Prototype Models in a Natural-Science Category Domain

**Robert M. Nosofsky (nosofsky@indiana.edu)**
Psychological and Brain Sciences, 1101 E. Tenth Street
Bloomington, IN 47405 USA

**Brian Meagher (bmeagher@iu.edu)**
Psychological and Brain Sciences, 1101 E. Tenth Street
Bloomington, IN 47405 USA

**Parhesh Kumar (pakuma@iu.edu)**
Psychological and Brain Sciences, 1101 E. Tenth Street
Bloomington, IN 47405 USA

**Abstract**

A classic issue in the cognitive-science of human category learning has involved the contrast between exemplar and prototype models. However, experimental tests to distinguish the models have relied almost solely on use of artificial categories composed of simplified stimuli. Here we contrast the predictions from the models in a real-world natural-science category domain – geologic rock types. Previous work in this domain used a set of complementary methods, including multidimensional scaling and direct dimension ratings, to derive a high-dimensional feature space in which the rock stimuli are embedded. The present work compares the category-learning predictions of exemplar and prototype models that make reference to this derived feature space. The experiments include conditions that should be favorable to prototype abstraction, including use of large-size categories, delayed transfer testing, and real-world natural category structures. Nevertheless, the results of the qualitative and quantitative model comparisons point toward the exemplar model as providing a better account of the observed results. Limitations and directions of future work are discussed.

**Keywords:** categorization; exemplar models; prototype models; high-dimensional similarity spaces

## Introduction

A classic issue in cognitive science concerns the manner in which people represent categories in memory and make decisions about category membership. Two of the major models of human categorization are exemplar and prototype models. According to exemplar models, people represent categories by storing numerous individual examples of the categories in memory and classify objects on the basis of their similarity to the stored examples (Medin & Schaffer, 1978; Nosofsky, 1986). By contrast, according to prototype models (Reed, 1972; Minda & Smith, 2001), people form summary representations, usually formalized as the central tendencies of the category distributions, and classify objects on the basis of their similarity to the prototypes.

There is an enormous past literature that has contrasted the predictions from exemplar and prototype models. However, to achieve needed controls, virtually all the research has contrasted the models in experiments using highly simplified perceptual stimuli and artificially designed category structures. The key idea in the present work was to contrast the models in a domain involving real-world natural-science categories involving complex, high-dimensional stimuli.

Our example target domain is rock classification in the geologic sciences. Recent research suggests that the same principles govern the structure of rock categories as govern the structure of numerous other categories in the natural world (Nosofsky, Sanders, et al., 2018a,b). For example, rock categories exhibit graded structures, with prototypical instances at their centers, but numerous less typical instances as well (cf. Rosch & Mervis, 1975). Also, as is the case with other natural categories, the boundaries separating different rock categories from one another are often fuzzy, and the category distributions can sometimes even overlap (cf. McCloskey & Glucksberg, 1978). Finally, we think rock classification is an intriguing domain to study because, despite people's general familiarity with rocks, a relatively small percentage of people arrives to the lab with very much prior knowledge of their detailed category structures. Thus, one can maintain careful control of people's learning histories in laboratory settings.

In recent work, Nosofsky and colleagues demonstrated success in using an exemplar model to account for classification learning and generalization in the rocks domain (Nosofsky et al., 2018a, 2019). However, most of the experiments involved cases in which there were only three training exemplars per category and in which transfer tests took place immediately after initial training. In the present work we conducted a rock-category learning experiment that we thought would be quite challenging to the predictions from exemplar models. First, participants learned to classify rock images into 10 igneous-rock categories defined by 9 training examples each (90 total). Past research has characterized size-9 categories as large in size and conducive to prototype abstraction (Homa et al., 1981). Second, our experiment included a delayed testing phase, with a one-week interval between initial training and subsequent test. Classic work argues that any exemplar-based category knowledge tends to be short-lived, but that the representation of the prototype is durable over time (Homa et al., 1981; Posner & Keele, 1970). Third, as we have already emphasized, the rocks domain is a real-world natural-category domain involving complex high-dimensional stimuli. If natural

categories are organized around prototypes, as theorized in classic work, then testing the models in this natural domain might also confer advantages to prototype models.

Our model comparisons will involve both qualitative and quantitative contrasts between the competing models. Initial qualitative contrasts will examine overall performance patterns during a test phase that includes old training examples from the categories and novel transfer items of varying degrees of similarity to the original training examples. Included among the novel transfer items are a set of photoshopped rock images that we explicitly constructed to be highly similar to specific training items from each category, yet clearly discriminable from the parent rocks from which they were created. We will refer to these photoshopped transfer stimuli as the *high-similarity neighbors* (*HSN*s): An example is shown in Figure 1. The intuition is that if the training examples of the categories are really stored in memory, then when tested with a HSN, people may be reminded of the specific parent training rock from which it was created, thereby enhancing their classification performance (Ross et al., 1990). In addition to assessing overall qualitative performance patterns across different item types of the categories, we also conduct detailed quantitative fitting of the models to classification-probability data observed at the individual-item level. As explained below, the models are fitted to the data by making reference to a high-dimensional feature-space representation for the rock stimuli that has been derived using a variety of complementary methods. The goal is to evaluate the ability of the competing models to account for extremely rich sets of classification data: namely, the probability with which participants classify each individual rock image into each of the 10 candidate categories during both immediate and delayed test phases.


Figure 1. Example HSN transfer item and its parent rock.

## Method

### Participants

There were 67 participants from the Indiana University community. The participants all had normal or corrected-to-normal vision and all reported having normal color vision. All reported that they had little or no previous experience in rock classification. Each participant received $40 in compensation, $15 for an initial session involving a training phase and an immediate-test phase, and $25 for completing a one-week-delayed test phase. Nine participants did not return for the one-week-delayed test phase; we do not include these participants' data in the analyses.

### Stimuli and Apparatus

The stimuli were 120 rock images from the igneous-rock image set described in previous articles (Nosofsky et al., 2018a,b). In addition, we created 30 photoshopped rock images (the high-similarity neighbors; HSNs), with each HSN highly similar to a specific "parent rock" image used during the training phase. The HSNs were either slightly lighter or darker than the corresponding parent image and were rotated a variable number of degrees from the original orientation of the parent rock. In addition, the edge of the parent rock was manually cropped to change the parent rock's shape outline. As a manipulation check, we conducted similarity-scaling work involving the HSNs (a detailed report goes beyond the scope of this article) to confirm that they were judged to have a relatively high degree of similarity to their parent rocks but were clearly discriminable from them.

The experiment was run on PCs and each participant was tested individually in a private, sound-attenuated booth.

### Procedure

Each of the 10 rock categories comprised 15 members: 9 randomly selected training items, 3 standard transfer items, and 3 HSN transfer items. The experiment started with a training phase: each of the 90 total training examples was shown once per block in a random order across 6 blocks for a total of 540 training trials. On each trial of the training phase, a rock image was presented on the screen and the participant attempted to classify it into one of the 10 categories. Immediate feedback was provided on each trial informing the participant of the correct category response.

Following training there was an immediate test phase. For each category, in addition to the 9 old training examples, participants were presented with the 3 novel standard transfer items and the 3 HSN transfer items. The test phase was organized into 6 blocks of 75 trials each, with each item presented 3 times in a balanced random order across blocks. To keep participants engaged in the task, corrective feedback was provided on one-third of the trials in which old training examples were presented; no feedback was provided on trials in which new transfer items were presented. Following an approximately one-week delay, participants were tested in a delayed test phase. The procedure for the delayed test phase was the same in all respects as the immediate one, except no corrective feedback was provided on any trials.

## Results

In Figure 2 we plot mean proportion correct in both test phases, averaged across all 10 categories, for the three main item types: the old training examples, the standard transfer items, and the HSN transfer items. As is clear from inspection, in both test phases, performance was best on the old training examples, intermediate on the HSN transfer items, and worst on the standard transfer items. The finding that performance was better on the old training examples than

on the novel transfer items is consistent with the view that classification decision making was based, at least in part, on memories for the individual training examples themselves. The pattern of results challenges the predictions from pure prototype models, because each of the item types has roughly equal distance to the category prototypes.

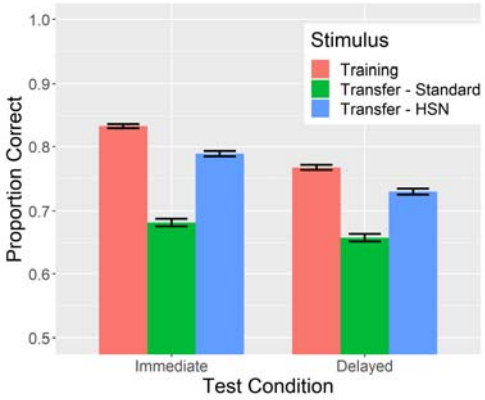


Figure 2: Mean proportions of correct responses for the item types in the immediate and delayed conditions. Error bars are within-subject standard errors.

A more nuanced version of the prototype model that has been proposed in the literature is a mixed *prototype-plus-rote-memory* (PRM) model (e.g., Medin & Schaffer, 1978; Minda & Smith, 2001). According to this model, the prototype plays the dominant role in the category representation. However, the model makes allowance for the possibility that learners also form all-or-none rote memories for some of the examples. If an old example is tested, and the rote memory for that example has been formed, then people can use the rote memory to correctly classify the item; otherwise they rely on the prototype. Clearly, the PRM model blurs the distinction between exemplar and prototype models. In addition, unlike the pure prototype model, the PRM model can predict a performance advantage for the old training examples. However, as formalized in the literature, the PRM model fails to predict any difference in performance between the standard transfer items and the HSN transfer items -- both tend to be the same distance from their category prototypes. By contrast, the exemplar model predicts an advantage for the HSN transfer items, because it explicitly assumes that people can make use of their specific exemplar-based memories to generalize to novel items. As is seen in Figure 2, in both the immediate and delayed tests, overall performance was higher for the HSN transfer items than for the standard transfer items – again strongly suggesting the operation of exemplar-based generalization processes.

Figure 3 shows the results broken down by the individual 10 categories in the immediate test (the results from the delayed test were extremely similar). Overall, most of the individual categories show the same pattern of results as for the grand-averaged data, so the effects are reasonably general. However, there are a couple of exceptions (basalt and diorite); we consider these in more detail below.

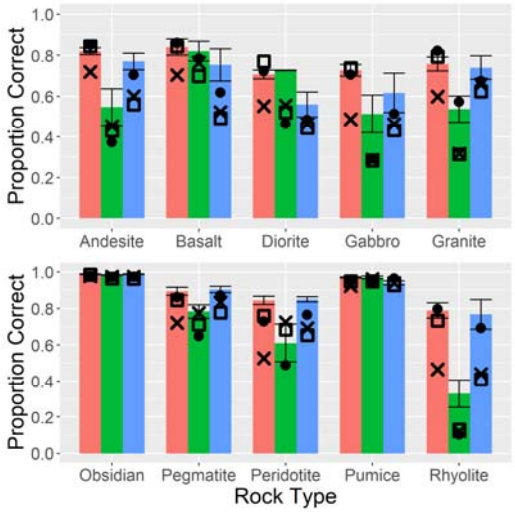


Figure 3: Immediate Test. Observed (red=training, green=standard transfer, blue=HSN) and predicted proportion correct by individual rock category. Dots= exemplar predictions, x's=prototype predictions, squares= PRM predictions.

## Formal Modeling Analyses

Although the qualitative comparisons described above provide initial evidence pointing in the direction of exemplar-based generalization processes, the central goal of the work involves quantitative model-fitting comparisons between the competing models. As representatives from the classes of exemplar and prototype models we use Nosofsky's (1986) *generalized context model* (GCM) and its prototype analogue (Nosofsky, 1987, p. 102). The steps in the model-fitting are as follows: First, one derives a high-dimensional feature-space representation in which the rock stimuli are embedded; second, one computes similarity relations between test items and the individual rock exemplars and category prototypes by making reference to this feature space; and third, one substitutes the computed similarities into well-known choice rules that predict the probability with which each individual test item is classified into each of the alternative categories.

### Feature-Space Representation

Using a variety of methods, Nosofsky et al. (2018b, 2019; Sanders & Nosofsky, 2018, 2020) have devoted extensive work to building a high-dimensional feature-space representation for a set of 360 rock-image samples, including the 120 igneous-rock samples tested in the present experiment. In some studies, they conducted multidimensional-scaling (MDS) analyses (Lee, 2001; Shepard, 1980) of matrices of pairwise similarity-judgment data for the rock images. Based on a combination of overall fit and interpretability of the resulting dimensions, the researchers initially settled on an 8-dimensional scaling solution, in which the dimensions could be easily interpreted in terms of: lightness vs. darkness, average grain size, smoothness vs. roughness, shininess, organization, chromaticity, red-green hue, and shape-related components

(for interactive displays of the MDS solution, see https://osf.io/w64fv/). Although these dimensions provided an excellent account of the similarity-judgment data, subsequent research made clear that, in the context of independent classification-learning experiments, observers also made use of more subtle emergent dimensions that were highly diagnostic of category membership (Nosofsky et al., 2019; Sanders & Nosofsky, 2020). Nosofsky et al. (2019) supplemented the original MDS space with these "missing" dimensions by having independent groups of participants provide direct dimension ratings for them. The supplementary dimensions included: porphyritic texture, presence of holes, pegmatitic texture, greenness of hue, and conchoidal fracture (see Nosofsky et al., 2019, for details). In addition, for purposes of the present study, we also collected additional similarity-judgment data (analyzed through MDS) and ratings on the supplementary dimensions in order to embed the 30 HSN items in the same feature-space representation as the original 120 igneous rock images. A detailed description of these MDS analyses for the HSN stimuli goes beyond the scope of this brief report.

### Similarity Computation

As in past work, we assume that the original eight MDS dimensions and the supplementary ones lie in a common Euclidean space. The distance between test item $i$ and exemplar $j$ is given by

$$d_{ij} = [\ \sum|x_{im} - x_{jm}|^2 + \sum w_m|x'_{im} - x'_{jm}|^2\ ]^{1/2}, \quad (1)$$

where $x_{im}$ is the value of item $i$ on dimension $m$ in the MDS space; $x'_{im}$ is the value of item $i$ on supplementary-dimension $m$; and $w_m$ is the weight given to supplementary-dimension $m$. Here, we have assumed for simplicity that the observers give equal attention-weight to the original 8 MDS dimensions; for scaling convenience, these weights are set at one. However, weights on each of the supplementary dimensions need to be estimated because the manner in which the psychological scales of the directly rated supplementary dimensions relate to one another and to the initial MDS dimensions is unknown – see Nosofsky et al,. 2019 for extensive discussion.

Following Shepard (1987), the similarity between item $i$ and exemplar $j$ is an exponential decay function of their distance,

$$s_{ij} = \exp(-c \cdot d_{ij}), \quad (2)$$

where $c$ is an overall sensitivity parameter that measures the rate at which similarity declines with distance in the space.

The values of the prototypes on each dimension are computed by averaging across the dimension values associated with each of the individual training examples of each category. The similarities of any given test item to the prototypes are then computed by using equations analogous to Equations 1 and 2 above.

### Classification Decision Rules

According to the GCM, the probability with which item $i$ is classified in Cat $J$ is found by summing the similarity of $i$ to all the training examples of Cat $J$ and dividing by the summed similarity of $i$ to all training examples of all categories:

$$P(J|i) = \frac{(\sum_{j\in J} s_{ij})^{\gamma}}{\sum_K (\sum_{k\in K} s_{ik})^{\gamma}} \quad (3)$$

where $\gamma$ is a response-scaling parameter. As $\gamma$ grows larger in magnitude, observers respond more deterministically with the category that yields the largest summed similarity.

According to the prototype-plus-rote-memory (PRM) model, if test-item $i$ is a member of Cat $J$, then the probability that it is correctly classified into Cat $J$ is given by

$$P(J|i) = p_{mem} + (1-p_{mem}) \cdot [\ (s_{iJ})^{\gamma} / \sum (s_{iK})^{\gamma}], \quad (4a)$$

where $s_{iJ}$ is the similarity of item $i$ to the prototype of Cat $J$ (computed using equations analogous to Eqs. 1 and 2 above); and $p_{mem}$ is the probability that the observer uses a rote memory for test-item $i$ to correctly classify it into Cat $J$. If test-item $i$ is not a member of Cat $J$, then the probability that the observer classifies it into Cat $J$ is given by

$$P(J|i) = (1-p_{mem}) \cdot [\ (s_{iJ})^{\gamma} / \sum (s_{iK})^{\gamma}]. \quad (4b)$$

The pure-prototype model is a special case of (4a) and (4b) in which $p_{mem} = 0$.

The present version of the GCM makes use of 7 free parameters: the sensitivity parameter $c$, response-scaling parameter $\gamma$, and 5 weights ($w_m$) associated with each of the 5 supplementary dimensions. As explained in previous articles (e.g., Nosofsky & Zaki., 2002, p. 926), the parameters $c$ and $\gamma$ are mathematically non-identifiable within the prototype model (one can estimate only their product), so we arbitrarily set $\gamma$=1 in the prototype models. Thus, the PRM model uses 7 free parameters ($c$, $p_{mem}$, 5 $w_m$'s); and the pure-prototype model uses 6 free parameters (with $p_{mem}$=0). The model fits are evaluated by using the Bayesian Information Criterion, BIC = -2ln($L$) + $P$ln($N$), where $L$ is the maximum-likelihood of the data given the model, $P$ is the number of free parameters in the model, and $N$ is the number of observations in the data set. The model that minimizes BIC is viewed as providing the best fit. Using multiple random starting configurations, we used the Hooke and Jeeves (1961) search algorithm to locate the best-fitting parameters.

## Model-Fitting Results

In Figure 4 we provide scatterplots of the maximum-likelihood fits from the GCM and pure-prototype model to the data from the immediate test phase. (The scatterplots for the delayed test phase showed extremely similar patterns, but space limitations prevent us from displaying them in this brief report.) Recall that in each test phase, observers classified 150 rock items into 10 different categories. Each point in each scatterplot indicates the probability that a particular item

was classified into a particular category; thus, there are 1500 points in each plot. For each item, the probability of correct classification is indicated using a geometric symbol (decoded in the figure caption). In addition, for each item, the probabilities of the remaining nine incorrect classifications are indicated with small dots. The y-axis shows the observed probabilities, whereas the x-axis shows the predicted ones.

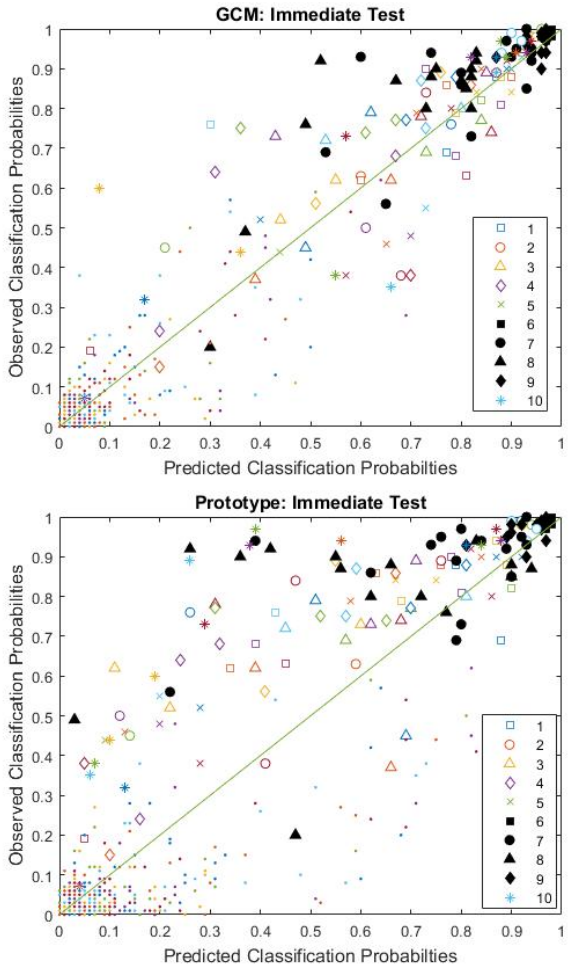


Fig 4. Item-level predictions from the models. 1=andesite, 2=basalt, 3=diorite, 4=gabbro, 5-granite, 6=obsidian, 7=pegmatite, 8=peridotite, 9=pumice, 10=rhyolite.

Comparing the results of the GCM (top panel) to those of the prototype model (bottom panel), it is clear that the GCM provides a far superior fit to the data. The conclusion drawn from visual comparisons of the Figure 4 scatter plots is confirmed by the BIC fits of the models to the data, which are reported in Table 1: In both the immediate and delayed test phases, the BIC is dramatically smaller for the GCM than for the prototype model. To gain deeper insights into these results, in Figure 3 we show the averaged predictions from the models for each of the individual item types in each category – the predictions from the GCM are shown as black dots and those of the prototype model as x's. In general, the GCM predicts well the results for the old training items and for the HSN transfer items; however, it tends to under-predict the accuracies associated with the standard transfer items. We consider likely reasons for the latter result below. By contrast, the prototype model under-predicts the accuracies associated with all three item types in most of the categories. Thus, the model-fitting results converge strongly with the qualitative patterns described earlier in suggesting a strong role of exemplar-based classification processes.

| Model | Condition | BIC |
|---|---|---|
| GCM | Immediate | 39,952 |
| Prototype | Immediate | 45,304 |
| PRM | Immediate | 41,090 |
| | | |
| GCM | Delayed | 46,921 |
| Prototype | Delayed | 49,461 |
| PRM | Delayed | 47,350 |

Table 1: BIC Fits of Models

It was more difficult to tell apart the quantitative predictions of the exemplar model and the PRM model (full-prediction scatterplot not shown in this brief report). This is not surprising, because the rote memory model's assumption that numerous individual exemplars are stored blends together the assumptions from the two classes of models. Nevertheless, the BIC fit achieved by the exemplar model is still far better than that achieved by the PRM model in both the immediate and delayed conditions (see Table 1). More importantly, focused qualitative comparisons continue to favor the predictions from the exemplar model. The predictions from the PRM model for the individual item types in each category are depicted as open squares in Figure 3. Although this version of the prototype model predicts correctly the accuracies for the old training examples (because it allows rote memory for those items), it still under-predicts the accuracies for the HSN items because it makes no allowance for generalization from the stored examples.

Earlier in our article, we noted that for a couple of categories (basalt and diorite), participants did not classify the HSN transfer items with higher accuracy than the standard transfer items. Interestingly, as can be seen in Figure 3, the models provide a partial account of this finding. Recall that the specific stimuli that served as HSN-transfer items versus standard-transfer items were selected randomly for each category. Apparently, by chance, the particular stimuli chosen to serve as HSN-transfer items in the basalt and diorite categories were located in difficult-to-classify regions of the rock-feature space. For example, they may have been located near the boundaries that separated basalt and diorite from contrasting categories in the space. The more important result is that, considered across all categories, the HSN-transfer stimuli were classified with higher accuracy than the standard ones, providing clear evidence for exemplar-generalization processes.

Earlier in our article, we also acknowledged a tendency for the exemplar model to underestimate accuracies associated with the standard transfer stimuli. One possibility is that the result may be pointing in the direction of a high-parameter mixed model that allows generalization to both exemplar *and*

prototype representations. Although a detailed report goes beyond the scope of this article, we formalized such a model and fitted it to the data: In brief, the high-parameter mixed model yielded slightly improved BIC fits to the data, but still systematically under-estimated the accuracies associated with the standard transfer stimuli.

We believe that a more likely reason for the under-estimation is that more work remains to develop a fully comprehensive feature space for the rock stimuli. As documented by Nosofsky et al. (2019), participants in these experiments appear to be highly adept at discovering features that are useful for purposes of classification and that tie categories together (cf. Austerweil & Griffiths, 2013; Schyns et al., 1998). Although Nosofsky et al. made progress in identifying such supplementary dimensions for the present domain, more work along these lines is undoubtedly needed.

## Discussion

### Summary

This study is among the first to have conducted rigorous comparisons between predictions of exemplar and prototype models in a high-dimensional natural-category domain in which the history of training examples experienced by the learners was placed under careful experimental control. In addition, the design had attributes that theorists have argued should be conducive to prototype abstraction, including reasonably large category sizes, large numbers of distinct training examples, delayed testing, and use of naturally-occurring category structures. Nevertheless, the patterns of qualitative results and quantitative modeling comparisons provided compelling evidence for a role of exemplar-based classification and generalization processes.

### Relations to Other Recent Work

A line of closely related recent work is the interesting and ambitious study reported by Battleday, Peterson, and Griffiths (2019). These researchers collected classification choice data for 10,000 images from 10 real-world categories (Krizhevsky & Hinton, 2009). In apparent contrast to the results reported in the present study, these researchers found that prototype models yielded quantitative fits to the data that were as good as or better than the exemplar model. However, there are numerous differences between the studies that make comparisons of the results difficult. First, rather than using MDS-based approaches, Battleday et al. obtained feature representations for the images by extracting activations from the final pooling layer of deep-learning networks that had been trained to classify the images. It is an open question whether these highly derived pooling-layer activations should be regarded as fundamental "building-block" features composing objects. Second, the good-fitting prototype models used complex Mahalanobis-distance functions, with ten times as many free parameters as the distance function used for the good-fitting exemplar model. Third, because the researchers tested categories for which the observers had a lifetime of prior knowledge, the training examples that guided the development of the category representations were unspecified. To fit the data, the researchers therefore randomly sampled training items from the complete category distributions across different runs of the models. By contrast, in the present study, the learning histories of the participants were placed under strict experimental control and the specific training exemplars were known. Any one of these major differences in approach across our studies would likely have an enormous impact on the results and conclusions.

### Limitations and Future Research Directions

We made efforts in this research to test conditions that previous work suggested should be challenging to exemplar models. Nevertheless, it may be that more extreme manipulations are needed to provide compelling evidence for a role of prototype abstraction. Thus, future research might test even larger-size categories or greater delays between initial study and subsequent test than were used here.

In addition, the present research considered models at only two endpoints of a continuum between specific exemplar storage and prototype abstraction. Important models have been developed that lie intermediate along this continuum, such as models that allow for formation of multiple prototypes or clusters (Anderson, 1991; Love, Medin, & Gureckis, 2004; Sanborn, Griffiths, & Navarro, 2010; Vanpaemel & Storms, 2008). The extent to which the rock categories examined here may be well characterized in terms of multiple distinct clusters remains unknown. Thus, future research should examine whether the multiple-prototype models might yield improved accounts of the present data.

Finally, within the framework of the exemplar model, more work is needed to flesh out the detailed cognitive processes involved in learning, forgetting, and selectively attending to diagnostic information. In the present work, for simplicity, we assumed that the complete set of presented exemplars was stored, and with each exemplar stored at full strength. More general versions of the model allow for probabilistic exemplar storage and with individual examples stored with differential memory strengths. Likewise, any forgetting that occurred across immediate and delayed testing was modeled primarily in terms of decreases in the overall memory sensitivity parameter ($c$ in Equation 2). However, a fuller account of forgetting would make allowance for processes such as probabilistic loss of stored examples from memory, decreases in memory strength, and loss of memory for the category labels attached to the stored examples. Finally, a core assumption of exemplar models has always been that observers selectively attend to those dimensions of items that are most diagnostic for purposes of categorization. However, in cases in which observers are attempting to discriminate simultaneously among members of 10 categories, as in the present experiment, the relevant directions in the space will vary dramatically depending on which particular category contrasts the observer is considering at any point in time. Future work is needed to flesh out the kinds of dynamic changes in selective attention that likely operate during the time course of categorization decision making.

## Acknowledgements

This work was supported by NSF grants EHR-1534014 and DUE-1937361.

## References

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

Austerweil, J. L., & Griffiths, T. L. (2013). A nonparametric Bayesian framework for constructing flexible feature representations. *Psychological review*, *120*(4), 817.

Battleday, R., Peterson, J. C., & Griffiths, T. L. (2019). Capturing human categorization of natural images at scale by combining deep networks and cognitive models. arXiv preprint arXiv: 1904.12690.

Homa, D., Sterling, S., & Trepel, L. (1981). Limitations of exemplar-based generalization and the abstraction of categorical information. *JEP: Human Learning and Memory*, *7*(6), 418.

Hooke, R., & Jeeves, T. A. (1961). ``Direct Search"Solution of Numerical and Statistical Problems. *Journal of the ACM (JACM)*, *8*(2), 212-229.

Krizhevsky, A., & Hinton, G. (2009). Learning multiple layers of features from tiny images (Technical report, University of Toronto, 2009).

Lee, M. D. (2001). Determining the dimensionality of multidimensional scaling representations for cognitive modeling. *Journal of Mathematical Psychology*, *45*(1), 149-166.

Love, B. C., Medin, D. L., & Gureckis, T. M. (2004). SUSTAIN: a network model of category learning. *Psychological Review*, *111*(2), 309.

McCloskey, M. E., & Glucksberg, S. (1978). Natural categories: Well defined or fuzzy sets?. *Memory & Cognition*, *6*(4), 462-472.

Medin, D. L., & Schaffer, M. M. (1978). Context theory of classification learning. *Psychological review*, *85*(3), 207.

Minda, J. P., & Smith, J. D. (2001). Prototypes in category learning: the effects of category size, category structure, and stimulus complexity. *JEP: Learning, Memory, and Cognition*, *27*(3), 775.

Nosofsky, R. M. (1986). Attention, similarity, and the identification–categorization relationship. *JEP: General*, *115*(1), 39.

Nosofsky, R. M. (1987). Attention and learning processes in the identification and categorization of integral stimuli. *JEP: Learning, Memory, and Cognition*, *13*(1), 87.

Nosofsky, R. M., Sanders, C. A., & McDaniel, M. A. (2018a). Tests of an exemplar-memory model of classification learning in a high-dimensional natural-science category domain. *JEP: General*, *147*(3), 328.

Nosofsky, R. M., Sanders, C. A., Meagher, B. J., & Douglas, B. J. (2018b). Toward the development of a feature-space representation for a complex natural category domain. *Behavior Research Methods*, *50*(2), 530-556.

Nosofsky, R. M., Sanders, C. A., Meagher, B. J., & Douglas, B. J. (2019). Search for the missing dimensions: Building a feature-space representation for a natural-science category domain. *Computational Brain & Behavior*, 1-21.

Nosofsky, R. M., & Zaki, S. R. (2002). Exemplar and prototype models revisited: Response strategies, selective attention, and stimulus generalization. *JEP: Learning, Memory, and Cognition*, *28*(5), 924.

Posner, M. I., & Keele, S. W. (1970). Retention of abstract ideas. *Journal of Experimental psychology*, *83*(2p1), 304.

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

Rosch, E., & Mervis, C. B. (1975). Family resemblances: Studies in the internal structure of categories. *Cognitive psychology*, *7*(4), 573-605.

Ross, B. H., Perkins, S. J., & Tenpenny, P. L. (1990). Reminding-based category learning. *Cognitive Psychology*, *22*(4), 460-492.

Sanborn, A. N., Griffiths, T. L., & Navarro, D. J. (2010). Rational approximations to rational models: alternative algorithms for category learning. *Psychological review*, *117*(4), 1144.

Sanders, C., & Nosofsky, R. M. (2018). Using Deep-Learning Representations of Complex Natural Stimuli as Input to Psychological Models of Classification. *Proceedings of the 2018 Conference of the Cognitive Science Society*, Madison, WI.

Sanders, C. A., & Nosofsky, R. M. (2020). Training Deep Networks to Construct a Psychological Feature Space for a Natural-Object Category Domain. *Computational Brain & Behavior*, 1-23.

Schyns, P. G., Goldstone, R. L., & Thibaut, J. P. (1998). The development of features in object concepts. *Behavioral and brain Sciences*, *21*(1), 1-17.

Shepard, R. N. (1980). Multidimensional scaling, tree-fitting, and clustering. *Science*, *210*(4468), 390-398.

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

Vanpaemel, W., & Storms, G. (2008). In search of abstraction: The varying abstraction model of categorization. *Psychonomic bulletin & review*, *15*(4), 732-749.