Tests of an Exemplar-Memory Model of Classification Learning in a High-Dimensional Natural-Science Category Domain

Robert M. Nosofsky and Craig A. Sanders
Indiana University Bloomington

Mark A. McDaniel
Washington University in St. Louis

Experiments were conducted in which novice participants learned to classify pictures of rocks into real-world, scientifically defined categories. The experiments manipulated the distribution of training instances during an initial study phase, and then tested for correct classification and generalization performance during a transfer phase. The similarity structure of the to-be-learned categories was also manipulated across the experiments. A low-parameter version of an exemplar-memory model, used in combination with a high-dimensional feature-space representation for the rock stimuli, provided good overall accounts of the categorization data. The successful accounts included (a) predicting how performance on individual item types within the categories varied with the distributions of training examples, (b) predicting the overall levels of classification accuracy across the different rock categories, and (c) predicting the patterns of between-category confusions that arose when classification errors were made. The work represents a promising initial step in scaling up the application of formal models of perceptual classification learning to complex natural-category domains. We discuss further steps for making use of the model and its associated feature-space representation to search for effective techniques of teaching categories in the science classroom.

Keywords: category learning, computational modeling, exemplar model, multidimensional scaling, similarity

A fundamental component of science education is learning the key categories of the target domain. For example, in botany, students learn to classify wide varieties of plants; entomologists are expert in the classification of insects; and, in the target domain that is the focus of the present work, students in the geologic sciences are trained in the identification and classification of rocks.

For readers who believe that learning to classify rocks is easy and straightforward, consider an example. The top row of Figure 1 displays pictures of three tokens of granite, a type of rock with which many people are familiar. The bottom row displays pictures of three tokens of a closely related rock known as diorite. Learning the subtle features that can be used to discriminate between these highly similar categories of rocks would clearly be a challenging task.

There are a wide variety of different techniques that might be used to teach real-world scientific categories such as rock types. A common aspect of almost all such techniques involves presentation of instances of the to-be-learned categories; however, there are wide-open issues about the best ways for instance-based training to proceed. Indeed, many of these issues have been researched in classic and recent studies of category learning conducted in the laboratory. Among the numerous questions that arise are: (a) which training instances should be used? (e.g., Posner & Keele, 1968); (b) in what order should the training instances be presented? (e.g., Elio & Anderson, 1984; Pashler & Mozer, 2013; Mathy & Feldman, 2009); (c) how does the use of explicit rule-based instruction influence the instance-based codings that are stored in memory (e.g., Allen & Brooks, 1991; Medin & Smith, 1981)? A major challenge is that conducting empirical studies to systematically navigate through the vast set of combinations of teaching possibilities would be an extraordinarily time-consuming process (see Lindsey, Mozer, Huggins, & Pashler [2013] for one proposed approach to navigating through the empirical search space).

A key idea that we advance in the present research is that the search may be made more efficient by application of successful formal models of perceptual category learning. The idea would be to investigate different teaching possibilities by simulating the techniques on the models themselves. One could then focus empirical investigations on those techniques that the models predict...
would be most successful (e.g., Khajah, Lindsey, & Mozer, 2014; Mathy & Feldman, 2016; Patil, Zhu, Kopec, & Love, 2014). Although this idea seems like a good one in principle, there is a major practical stumbling block in its application. In particular, although the field has advanced a number of highly sophisticated models of perceptual classification learning (for a comprehensive review, see Pothos & Wills, 2011), essentially all tests of such models have been in simplified domains involving highly controlled, low-dimensional stimuli and artificial category structures. There is little if any knowledge of the extent to which the formal models may scale up successfully to account for category learning in the real world.

The central goal of the present research is to begin such an investigation by testing a well-known exemplar-memory model (Nosofsky, 1986, 2011) on its ability to account for the learning of real-world rock categories. Our starting focus is on the exemplar model because it has received extensive support in the basic category learning literature across a wide range of artificial category-learning studies and a range of dependent measures, including learning profiles, generalization patterns, and response latencies (e.g., Medin & Schaffer, 1978; Nosofsky, 1986, 1987, 1988; Nosofsky & Palmeri, 1997; Shin & Nosofsky, 1992). Moreover, it serves as a foundation for a number of other highly significant models of category learning (e.g., Anderson, 1991; Kruschke, 1992; Love, Medin, & Gureckis, 2004; Pothos & Bailey, 2009; Vanpaemel & Storms, 2008). In addition, researchers have successfully expanded its application to account for interesting aspects of the classification of natural language concepts (e.g., Smits, Storms, et al., 2002; Storms, De Boeck, & Ruts, 2000; Voorstepols, Vanpaemel, & Storms, 2008). Finally, in recent work, Nosofsky, Sanders, Gerdon, Douglas, and McDaniel (2017) showed that the exemplar model accounted successfully for certain qualitative effects involving the learning of rock classifications at subordinate versus superordinate levels of categorization. Hence, the exemplar model seemed like a promising starting candidate for the present investigation.

Conducting quantitative tests of the exemplar model, however, requires the specification of a “feature space” in which the to-be-classified objects are embedded (Ashby, 1992; Nosofsky, 1992). In particular, according to the formal model, people represent categories by storing individual exemplars in memory, and classify objects on the basis of their similarity to the stored exemplars (see Theoretical Analysis sections for details). The examples are represented as points in a multidimensional feature space; similarity between examples is a decreasing function of their distance in the space. Importantly, the model further assumes that similarity is not an invariant relation but a context-dependent one: Similarity relations among exemplars change systematically because observers attend to alternative sets of dimensions that compose the exemplars in different categorization contexts (Medin & Schaffer, 1978; Medin & Smith, 1981; Nosofsky, 1984, 1986). Thus, to use the model to predict categorization, one needs to specify the underlying multidimensional feature-space representation for the objects.

In numerous past tests of the exemplar model, the derivation of the feature space has been straightforward, because the studies involved the classification of simple stimuli varying along a small number of salient dimensions: Examples includes shapes varying in size and angle; colors varying in brightness and saturation; or schematic faces varying along manipulated dimensions such as eye separation, mouth height, and so forth (for a review, see Nosofsky, 1992). In such cases, the derived multidimensional feature space is a fairly direct reflection of the highly controlled, experimentally manipulated dimensions that compose the objects.

In a real-world natural-category domain such as rocks, however, the derivation of a feature-space representation becomes an enormously ambitious task: Natural stimuli such as rocks vary along a very large number of complex dimensions that may be difficult to describe or discern (for similar issues in domains involving semantic concepts, see De Deyne et al., 2008; Verheyen, Ameel, & Storms, 2007). In addition, because we seek some generality in our tests of the model in this domain, we use multiple tokens of the numerous rock categories that our participants must learn. Thus, a large number of stimuli must be positioned in the high-dimensional feature space to which the model makes reference. Therefore, as a prerequisite for conducting the category-learning experiments reported in this article, Nosofsky, Sanders, Meagher, and Douglas (2017) first engaged in an extensive project to derive a feature-space representation for the large set of rock stimuli that we used in the present experiments. We start by providing a brief review of that previous work before moving to a report of two new category-learning experiments. We should note that although our focus is on classification learning of pictures of rocks, there is no reason in principle why the feature space cannot be extended to include information about nonvisual properties of the rocks such as heft, hardness and so forth.

**Derivation of the Feature-Space Representation**

The stimuli used in the present experiments are drawn from a set of 360 rock pictures. Specifically, using college-level geology texts as a guide (Marshak, 2013; Tarbuck & Lutgens, 2015), we selected 10 common subtypes of each of the broad categories of igneous, metamorphic and sedimentary rocks. The subtypes are listed in Table 1. Using web searches and guidance from an expert in geoscience education, we obtained pictures of 12 tokens of each
of the 30 subtypes to produce the stimulus set (see Nosofsky, Sanders, Meagher, et al., 2017, for details). 1

Nosofsky, Sanders, Meagher, et al. (2017) used two complementary, mutually informative approaches to deriving a feature-space representation for the rock stimuli. The first approach relied on similarity-scaling methods. In this approach, participants provided similarity judgments among the pairs of rock tokens. Multidimensional-scaling (MDS) procedures were then used to provide similarity judgments among the pairs of rock tokens. Multidimensional-scaling (MDS) procedures were then used to model the pairwise similarity judgments (Kruskal & Wish, 1978; Shepard, 1980). In brief, in MDS, objects are represented as points in a psychological space, and similarity between objects is a decreasing function of distance in the space. Thus, the MDS procedure positioned each of the rock stimuli in an M-dimensional feature space so as to optimize the fit to the similarity-judgment data.

It is interesting to note that because there were 360 rock tokens, this approach produced an enormous pairwise similarity-judgment matrix (there are 129,600 cells in a 360 × 360 matrix). Thus, despite the fact that Nosofsky, Sanders, Meagher, et al. (2017) recruited a very large number of participants for the study, the number of data entries per individual cell of the matrix was small, so the similarity-judgment data were very noisy at the individual-cell level. Nevertheless, there is a great deal of redundancy and mutually constraining data in such a matrix, because each individual row of the matrix provides information concerning the similarity of a single rock token to each of the other 360 tokens.

As it turned out, despite the small number of data entries per individual cell of the matrix, the MDS procedure produced highly regular and systematic results (for details, see Nosofsky, Sanders, Meagher, et al., 2017). First, the rock stimuli arranged themselves along highly interpretable psychological dimensions. Interactive displays of an 8-dimensional MDS solution for the stimuli are provided in the online website https://osf.io/w64fv/ associated with the Nosofsky, Sanders, Meagher, et al. (2017) study. The derived dimensions of the scaling solution for the rocks could be reasonably well interpreted in terms of (a) lightness/darkness of color, (b) average grain size, (c) roughness/smoothness of texture, (d) dullness/shininess, (e) degree of organization, (f) chromaticity or saturation, and (g) a circular dimension of hue. In addition, there were hints that shape-related attributes formed part of the MDS solution as well. Second, although the data were noisy at the level of individual cells, the MDS solution provided accurate quantitative accounts of aggregated forms of the similarity-judgment data. In particular, Nosofsky et al. considered a collapsed 30 × 30 matrix in which each cell i–j corresponded to the average similarity between all tokens from subtype i and subtype j (e.g., the average similarity between tokens of granite and tokens of diorite). The 8-dimensional MDS solution accounted for 97.2% of the variance in the similarity judgments of this collapsed matrix.

As a second approach to deriving a feature-space representation, Nosofsky, Sanders, Meagher, et al. (2017) collected direct dimension ratings for the stimuli along a battery of 18 rating scales (see Table 2; for a similar approach in research on medical decision making, see Swets et al., 1991). For example, in one condition, participants provided ratings of the darkness/lightness of each of the 360 rocks on a 1–9 scale. The authors developed the set of rated dimensions based on characterizations of rocks provided in college-level geology textbooks (e.g., Marshak, 2013; Tarbuck & Lutgens, 2015), as well as on preliminary similarity-scaling work with these stimuli that had already been conducted (Nosofsky, Sanders, Meagher, et al., 2017). In this approach, the feature-space representation for the rocks was simply the averaged rating for each of the rocks along each of the rated dimensions.

The MDS and direct-ratings approaches each have advantages and disadvantages. In the case of MDS, certain dimensions that may be crucial for making fine-grained distinctions between different categories of rocks may be ignored in the context of a generic similarity-judgment task, especially because the judgments were made by novice observers with little or no prior expertise in rock classification. Thus, such dimensions would not appear in the derived feature-space representation, and models of human category learning that rely on such a representation would be severely handicapped. In addition, as noted previously, natural objects such as rocks are composed of a very large number of complex dimensions. Similarity-scaling techniques may be limited in their power to reliably extract all dimensions, even if observers do make use of them in judging similarities. Thus, by collecting the direct dimension-ratings data, one could potentially supplement the MDS solution with information that was not extracted from the similarity-scaling technique.

Conversely, the direct-ratings approach has its own limitations. First, the response function that is involved in the translation of psychological scale values onto the direct ratings is unknown. Second, the manner in which values along separate dimensions interact needs to be specified. Third, not all dimensions that enter into participants’ perceptions of the stimuli may be easily accessible. Indeed, the reason why similarity-scaling techniques are so valuable is to overcome these kinds of shortcomings.

Because neither approach appeared to have a clear a priori advantage in general, in the present research we considered each approach to provide an alternative candidate for a feature-space representation for the rocks. (Ultimately, as we suggest in our General Discussion, the optimal feature-space representation for the rocks may be one that combines elements of the MDS and direct-dimension-ratings approaches.) Thus, we fitted

---

1 We thank Bruce Douglas, director of the Indiana University Judson Mead Geologic Field Station and Senior Lecturer in Geological Sciences at Indiana University, for his expert guidance.
Following alternative training conditions, we investigated participants’ category learning of the complete collection of 120 images of igneous-rock tokens in our rocks stimulus set – 10 subtypes of 12 tokens each (see Table 1 for a listing of the subtypes). Across two conditions, we manipulated the structure of the training exemplars that participants experienced during an initial learning phase. We then tested participants’ ability to classify the old training examples and to generalize to new items in a transfer phase. We refer to the two training conditions as the center condition and the coverage condition.

In both conditions, there were three training exemplars per subtype, for a total of 30 distinct training exemplars. As explained in more detail below, in the center condition participants were trained on rocks that occupied central locations of the category distributions corresponding to each rock subtype. By contrast, in the coverage condition, participants were trained on rocks that more fully covered the complete category distribution corresponding to each subtype; however, coverage-trained participants received far less experience with central rocks from the distributions than did participants in the center condition (i.e., for most categories, coverage-trained participants had experience with one central rock, whereas center-trained participants always experienced three central rocks per category).

Our choice of these alternative training conditions was motivated intuitively around the idea that major alternative theories of categorization place differential emphasis on the type of information that is retained in people’s category representations. According to prototype models (e.g., Posner & Keele, 1968; Reed, 1972; Smith & Minda, 1998), people abstract the “central tendency” of category distributions: To the extent that abstracting the central tendency is fundamental to category learning, then performance might be enhanced by the type of training received in the center condition. By contrast, according to exemplar models (Medin & Schaffer, 1978; Nosofsky, 1986) and closely related “clustering” models of categorization (e.g., Anderson, 1991; Love, Medin, & Gureckis, 2004; Sanborn, Griffiths, & Navarro, 2010; Vanpaemel & Storms, 2008), the focus is on retention of more details of the individual training exemplars themselves: Because the central tendency of the distribution does not necessarily have a privileged representational status, training conditions that provide more complete coverage of the full distribution might be preferable.

To help motivate some fundamental predictions from the model, we now outline the specific sets of steps for selecting training examples in the center and coverage conditions. The steps are illustrated schematically in Figure 2 using two-dimensional depictions; it should be remembered, however, that the computations described below are actually occurring in a high-dimensional space. In both conditions, we used the 8-dimensional MSD solution for the rock stimuli from Nosofsky, Sanders, Meagher, et al. (2017) as a generating vehicle. In the center condition, we first computed the centroid of each complete subtype distribution (red circle in top panel of Figure 2). We then selected as training exemplars of the subtype the three members of that subtype that were closest in Euclidean distance to the centroid (three dashed blue squares in top panel of Figure 2). In the coverage condition, we conducted a K-means clustering analysis for each subtype distribution, specifying three clusters per subtype. This analysis created three nonoverlapping, spatially compact clusters that were mutually exhaustive of the subtype distribution (three dashed elliptical contours in bottom panel of Figure 2). We then computed the centroid of each individual cluster (red circles in each cluster of the bottom panel of Figure 2). Finally, we selected as training exemplars the single example of each individual cluster that was closest in distance to that cluster’s centroid (three dashed blue squares in bottom panel of Figure 2).

To provide the reader with a sense of the different types of training exemplars selected for the two training conditions, we illustrate the complete set of rock pictures from the subtype rhyolite – see Figure 3. We should note that according to the similarity-scaling and dimension-ratings analyses conducted by Nosofsky, Sanders, Meagher, et al. (2017), rhyolite was among the most dispersed of the 30 rock subtypes. Thus, it serves as a useful example to bring out the contrast between the center and coverage conditions. We acknowledge that for many other subtypes the contrast between the center and coverage training examples is more subtle (e.g., in some cases, the particular items selected for the center and coverage conditions, though not identical, might be fairly similar to one another). We provide illustrations of the complete set of center and coverage examples used for all categories tested in this article in the online website https://osf.io/s3v9r/.

---

Table 2
Listing of Dimensions in the Direct Dimension-Ratings Experiment of Nosofsky, Sanders, Meagher, and Douglas (2017)

<table>
<thead>
<tr>
<th>Dimension</th>
</tr>
</thead>
<tbody>
<tr>
<td>Continuous dimensions</td>
</tr>
<tr>
<td>1. Lightness of color</td>
</tr>
<tr>
<td>2. Average grain size</td>
</tr>
<tr>
<td>3. Roughness</td>
</tr>
<tr>
<td>4. Shininess</td>
</tr>
<tr>
<td>5. Organization</td>
</tr>
<tr>
<td>6. Variability of color</td>
</tr>
<tr>
<td>Present–absent dimensions</td>
</tr>
<tr>
<td>7. Visible grain</td>
</tr>
<tr>
<td>8. Fragments</td>
</tr>
<tr>
<td>9. Stripes or bands</td>
</tr>
<tr>
<td>10. Holes</td>
</tr>
<tr>
<td>11. Physical layers</td>
</tr>
<tr>
<td>12. Salient special feature</td>
</tr>
<tr>
<td>Conditional continuous dimensions</td>
</tr>
<tr>
<td>13. Variability of size of grain</td>
</tr>
<tr>
<td>14. Angular/rounded fragments</td>
</tr>
<tr>
<td>15. Straight/curved stripes</td>
</tr>
<tr>
<td>Derived ratings from color matching</td>
</tr>
<tr>
<td>16. Brightness (Munsell value)</td>
</tr>
<tr>
<td>17. Saturation (Munsell chroma)</td>
</tr>
<tr>
<td>18/19. Hue (Munsell hue coded as circular coordinates)</td>
</tr>
</tbody>
</table>
The top panel of Figure 3 illustrates the center training examples. Although it is difficult to imagine the centroid of a complex 8-dimensional space, these training examples are roughly average among the complete set of rhyolite category members along dimensions such as lightness/darkness, average grain size, smoothness/roughness, and organization; and also lie toward the center of the two-dimensional color circle. The bottom panel of Figure 3 illustrates the coverage training examples for this subtype. It seems intuitively obvious that one training example covers the “banded” rhyolites; a second training example covers the members of the rhyolite category that contain small fragments; and the third training example covers the members of the category with a more homogeneous, fine-grained texture.

We provide a detailed formal statement of the exemplar-memory model in a subsequent Theoretical Analysis section of this article. Leaving aside the quantitative details for that section, here we highlight several main qualitative predictions stemming from the model. The predictions for transfer performance are illustrated schematically in Figure 4. The left panel displays predicted-proportion correct for the different item types in the center training condition, and the right panel displays predicted-proportion correct for the different item types in the coverage training condition. The black bars correspond to the center items, the white bars to the coverage items, and the gray bars to items that are neither center nor coverage items (“neither” items). A first (fairly obvious) prediction is a cross-over interaction between center and coverage items and training condition. Because participants are trained on the center items in the center condition, but are not trained on the coverage items, the exemplar model tends to predict better performance on the center items in the center condition. Analogously, because participants are trained on the coverage items in the coverage condition, but are not trained on the center items, the model tends to predict better performance on the coverage items than on the center items in the coverage condition. A second, more subtle, prediction is that the magnitude of the performance difference between the center and coverage items should be greater in the center condition than in the coverage condition. The reason is that, in the center condition, the coverage items receive relatively little support from nearby training items, so there should be a big dip in performance for the coverage items (Figure 2, top panel). By contrast, in the coverage condition, all the center items tend to receive at least some support from a nearby training item, because one of the three coverage clusters tends to be centrally located in the category distribution (Figure 2, bottom panel). Thus, the ex-

![Figure 2. Schematic illustration of the algorithms for choosing training examples in the center condition (top panel) and the coverage condition (bottom panel).](image)

![Figure 3. Illustration of the center training examples (top panel) and the coverage training examples (bottom panel) for the subtype rhyolite. Stimuli enclosed by red ovals are the center training examples; stimuli enclosed by blue rectangles are the coverage training examples.](image)
The exemplar model predicts a smaller dip in performance for the untrained center items (compared to the trained coverage items) in this condition.

Next consider the neither items, which are untrained in both conditions. Both intuition and basic findings with artificial categories (Posner & Keele, 1968) suggest that neither items would be categorized more accurately after coverage training than after center training. The reasoning is that by design the coverage training provides more information than does center training regarding the range of variability of the instances within each category; this experience with variable training instances (in the coverage condition) might be expected to support better transfer for new items. In contrast to this expectation, the formal model predicted roughly equal performance on the “neither” items across the two conditions. In addition to testing this counterintuitive prediction, we develop and test additional fine-grained quantitative predictions from the exemplar model in the Theoretical Analysis section—such as the extent to which alternative rock subtypes are confused with one another based on their positions in the multidimensional similarity space. In addition, we competitively test predictions emanating from the two alternative feature-space representations (similarity scaling vs. direct dimension ratings) that we described earlier in the introduction. Finally, we contrast the exemplar-model predictions with those from an alternative prototype model of classification.

Method

The study received approval from the Indiana University Institutional Review Board.

Participants. The participants were 66 members of the Indiana University community who were paid for taking part in the experiment. The participants received $24 for a 2-hr experimental session, plus a possible $6 bonus for good performance (defined as 60% correct or better during the transfer phase of the experiment). Data from 6 participants were lost as a result of computer crashes or an individual deciding to withdraw from the experiment. Participants were assigned randomly to the center and coverage conditions, with 33 participants completing the center condition and 27 participants completing the coverage condition. All participants had normal or corrected-to-normal vision and claimed to have normal color vision. All participants reported little or no past experience involving rock classification.

Materials. The stimuli were 120 pictures of rocks. These stimuli form a subset of those used in the previous similarity-scaling studies reported by Nosofsky, Sanders, Meagher, et al. (2017), and the full set of rock images is available in the “Rocks Library” folder of the website associated with that article (https://osf.io/w64fv/). We obtained the pictures from web searches, and used photo-shopping procedures to remove background objects and idiosyncratic markings such as text labels. In the present experiment, there were 10 subtypes from the broad category of igneous rocks (see Table 1), and 12 tokens of each of the 10 subtypes.

The stimuli were presented on a 23-inch LCD computer screen. The stimuli were displayed on a white background. Each rock picture was approximately 2.1 in. wide and 1.7 in. tall. Participants sat approximately 20 in. from the computer screen, so each rock picture subtended a visual angle of approximately 6.0° × 4.9°. Although we cannot precisely estimate the camera-distance of each individual rock picture, we selected all pictures in consultation with an expert in geology education such that they made clearly visible the salient characteristic features of the rock. The experiments were programmed in MATLAB and the Psychophysics Toolbox (Brainard, 1997). All participants were tested individually in private, sound-attenuated cubicles.

Procedure. The algorithms for choosing the training exemplars in the center and coverage conditions were described in our introduction to Experiment 1.2 We reiterate here that in the center condition the center items were the training exemplars and the coverage items were untrained; in the coverage condition the reverse held; and in both conditions the “neither” items were untrained. It turned out that, for most categories, a particular item served as both a center and a coverage training example across the conditions; we separate out the analysis of such items in our Results and Formal Modeling sections.

In each condition (center and coverage), the category-learning experiment was divided into a training phase and a transfer phase. The training phase consisted of 6 blocks of 60 trials each. Within each block, each of the 30 training exemplars was presented twice. The order of presentation of the training exemplars was randomized anew for each block and each participant. On each trial, a training exemplar was presented in the center of the screen and the subject classified it into one of the 10 subtype categories. The subject indicated the category response by pressing labeled number keys on the computer keyboard. A listing of the assignment of numbers to category names was displayed on the screen for 2 sec (e.g., “Correct! Diorite”; or “Incorrect, Diorite”). At the end of each

2 Regarding the K-means clustering step in the coverage condition, the solution that is produced by the algorithm will often vary with the starting configuration of clusters that is specified. Thus, for each subtype, we ran the algorithm ‘12 choose 3’ times, with each run using a different combination of three tokens from the subtype as starting-configuration cluster “anchors.” The starting configuration of clusters was then defined by assigning each of the 12 tokens to the cluster defined by the closest anchor. We chose as the K-means clustering solution the final configuration from all the starting configurations that yielded the smallest total within-cluster distance.
block, participants were informed of their overall proportion of correct responses. There was a 10-min break between the training and transfer phases.

In the transfer phase, participants were tested on both the 30 old training examples and the 90 new transfer stimuli. Participants were informed that in addition to classifying the old training exemplars, they would be tested on new stimuli from each of the rock categories. There were 6 blocks of 60 trials each in the transfer phase. Each of the 120 stimuli was presented once during blocks 1–2, once during blocks 3–4, and once during blocks 5–6. The order of presentation of the stimuli was randomized anew for each pair of blocks and each participant. To keep participants engaged in the task, we continued to provide corrective feedback on the trials in which the old training exemplars were presented; however, no corrective feedback was presented on trials involving the new transfer stimuli. Participants were informed that on such trials the computer would simply display the word “Okay” to indicate that the response was recorded. Participants were informed of their overall proportion correct at the end of each individual transfer block.

Results

Training performance. Because the results of major theoretical interest involve the patterns of performance in the transfer phase, we report the detailed analyses of the training-phase data in Appendix A. In brief, as would be expected, performance improved during the six blocks of training, with accuracy on the center items generally being higher than on the coverage items; however, by the final block, accuracy was very high for both types of training items.

Transfer performance. The main summary results from the transfer phase are displayed in Figure 5. The left panel displays overall proportion correct on the main types of stimuli in the center condition, whereas the right panel reports the results from the coverage condition. The display of the item types in this figure is a bit more fine-grained than we previewed in our introduction. In particular, for most of the subtypes, a particular item served as both a center training example and a coverage training example across the two conditions. We refer to such items as “both” stimuli. In Figure 5, the “both” stimuli are not included in the computations of overall proportion correct for the center stimuli (black bars) and the coverage stimuli (white bars); instead, they are separated out as distinct data points (cross-hatched bars). Thus, the computed proportions of correct responses for the center and coverage items are with respect to nonoverlapping sets of stimuli.

Inspection of Figure 5 reveals that the central qualitative predictions derived from the exemplar model were all confirmed. First, performance was better on the center items than on the coverage items in the center condition, but the reverse pattern was observed in the coverage condition. This observation is confirmed by statistical test. A 2 × 2 mixed-model ANOVA using condition (center vs. coverage) as a between-groups factor and item type (center vs. coverage) as a within-groups factor yielded a significant interaction, \( F(1, 58) = 663.2, \text{MSE} = 0.004, p < .001, \eta_p^2 = .920 \).

Tests of simple effects revealed that the center stimuli were classified more accurately than the coverage stimuli in the center condition, \( F(1, 58) = 680.3, p < .001 \); but the reverse pattern was observed in the coverage condition, \( F(1, 58) = 123.9, p < .001 \). Most important, as predicted, the magnitude of the absolute performance differences between the center and coverage items was greater in the center condition than in the coverage condition, \( t(58) = 9.24, p < .001 \).

Second, as is evident in Figure 5, performance on the neither items was comparable across the two training conditions. This nearly equivalent performance on the “neither” items was also predicted by the exemplar-memory model (deriving the result requires a detailed quantitative application of the model, presented in the next section). A statistical comparison confirmed that there was no difference between the conditions on neither-item performance, \( t(58) = -0.133, p = .90 \).

Finally, it is interesting to note that performance on the “both” stimuli was significantly better in the center condition than in the coverage condition (see Figure 5), \( t(58) = 4.46, p < .001 \). The exemplar model predicts this pattern of results because the “both” items have a greater number of close training-example neighbors in the center condition than in the coverage condition. That is, within each subtype category, although the “both” item is a training example in both conditions, there are two other close neighbors of that item that also receive training in the center condition (see Figure 2, top panel). By contrast, in the coverage condition, the two other training examples tend to lie far away from the “both” item (see Figure 2, bottom panel). The greater redundancy of training in the center condition thereby boosts performance on the “both” items in the center condition.

Figure 6 provides a fine-grained breakdown of the data for each individual rock category in the center condition (top panel) and the coverage condition (bottom panel). (Because of the much smaller sample sizes at the individual subtype level, to remove noise from the data, we have reggregated the results from the “both” stimuli into the center and coverage stimuli in these plots.) Inspection of the two panels reveals two key results. First, the patterns of performance observed for the center, coverage, and “neither” stimuli in the aggregate plots of Figure 5 tend to also be seen for each of the individual subtypes: Performance is better for the center stimuli than the coverage stimuli in the center condition, but the reverse occurs...
in the coverage condition. Moreover, the dip in performance for the untrained stimuli is generally greater in the center condition than in the coverage condition. Thus, these key patterns of results are robust. Second, there is a great deal of variability in overall performance levels across the different subtypes. Participants are extremely accurate, for example, in classifying members of the subtypes obsidian and pumice; moderately accurate in classifying members of subtypes such as pegmatite and basalt; and relatively inaccurate in classifying members of the subtypes diorite, gabbro, granite, and rhyolite. To confirm this latter observation, we computed overall proportion correct scores for each of the 10 subtypes. One-way ANOVAs revealed that the proportion-correct scores across the subtypes were significantly different from one another in both the center condition \( F(5.34, 170.93) = 96.8, \, MSE = 1.86, \, p < .001, \, \eta^2_p = .751 \) and the coverage condition \( F(5.36, 139.25) = 47.9, \, MSE = 1.33, \, p < .001, \, \eta^2_p = .648 \). A challenge for the formal model of classification learning is whether it can simultaneously characterize the patterns of training effects across the center and coverage conditions as well as the varying performance levels observed across the different subtypes of rocks.

**Theoretical analysis.**

*The formal model.* The version of the exemplar model tested in this article is known as the *generalized context* model (GCM; Nosofsky, 1986, 2011). We reviewed the conceptual underpinnings of the GCM in the introduction. In this section we provide a concise formal statement of a simplified, descriptive version of the GCM.

---

3 Because of violations of sphericity, we applied the Greenhouse-Geisser correction in these tests.
model and investigate its ability to account quantitatively for the rock classification-learning data.

According to the GCM, the probability that stimulus $i$ is classified into Category $J$ ($C_J$) is found by summing the similarity of $i$ to all training exemplars of $C_J$ and then dividing by the summed similarity of $i$ to all training exemplars of all categories:

$$P(C_J|i) = \frac{\sum_{j \in C_J} s_{ij}}{\sum_{k} \sum_{j \in C_k} s_{ij}}$$

(1)

where $s_{ij}$ denotes the similarity of test item $i$ to exemplar $j$. The parameter $\gamma$ is a response-scaling parameter that describes the degree of determinism in participants’ response strategies: As $\gamma$ grows larger, participants respond with higher probability with the category that yields the largest summed similarity (Ashby & Maddox, 1993; McKinley & Nosofsky, 1995; Nosofsky & Zak, 2002; for a process-model interpretation, see Nosofsky & Palmeri, 1997). For simplicity in this article, however, we set $\gamma = 1$.

The exemplars are represented as points in a multidimensional similarity space. The distance between exemplars $i$ and $j$ ($d_{ij}$) is given by the (weighted) Minkowski power model:

$$d_{ij} = \left[ \sum w_m |x_m - s_{jm}|^r \right]^\frac{1}{r},$$

(2)

where $x_m$ is the value of exemplar $i$ on dimension $m$; $w_m$ is the “attention weight” given to dimension $m$; and $r$ defines the distance metric of the space. In the main version of the GCM that we fit to the data, the $x_m$ coordinate parameters are given by the 8-dimensional MDS solution for the rock stimuli that was derived in Nosofsky, Sanders, Meagher, et al.’s (2017) similarity-scaling study. (Recall, however, that another purpose of this research is to also test the direct dimension-ratings approach to deriving the coordinate parameters.) In addition, we assume that participants give equal attention weight to all dimensions, so the $w_m$ weight parameters are all set at unity. Finally, the value of the power exponent is set at $r = 2$, yielding the familiar Euclidean metric for computing distances among the stimuli.

Following Shepard (1987), the similarity between exemplars $i$ and $j$ is presumed to be an exponential decay function of their distance in the psychological space:

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

(3)

where $c$ is a sensitivity parameter that determines the rate at which similarity declines with distance. An important conceptual aspect of the exponential decay function is that it formalizes the idea that exemplars that are in close proximity to the test item in the multidimensional space will have the major influence on how it is classified. Exemplars of moderate or far distance from the test item have a relatively small influence (see also Medin & Schaffer, 1978). Finally, the $s_{ij}$ values computed in Equation 3 are substituted into Equation 1 to generate the classification predictions made by the model.

The version of the GCM sketched above makes the strong assumption that all training exemplars are stored deterministically in memory. In preliminary model-fitting analyses, we found evidence that the fits could be improved by making allowance for the idea that the training exemplars are stored probabilistically in memory. Specifically, in our modeling, we assumed that each individual training exemplar was stored with probability $p_{store}$, and we generated predictions from the model by making use of computer simulation. For each run of the simulation, the set of stored training exemplars was randomly generated in accord with the probabilistic-storage assumption. Given that set of stored training exemplars, the system of Equations 1–3 was used to generate the classification predictions from the model for that run of the simulation. The overall predictions from the model were then obtained by averaging across the predictions from 1000 individual simulation runs.4

The formal model sketched above makes use of only two free parameters: the sensitivity parameter $c$ from Equation 3, and the probabilistic-storage parameter $p_{store}$. As will be seen, the goal is to use this low-parameter model to fit large numbers of freely varying data points in this complex, natural-science category-learning paradigm.

**Fitting the item-type data from the center and coverage conditions.** In the transfer phase of the present paradigm, participants classified each of 120 rock pictures into 10 different categories across 2 conditions of testing. An ambitious target goal might be to test models on their ability to characterize the data at this extremely fine-grained level of detail. However, this target goal seems unrealistic at this stage. First, given the very complex nature of this rock-stimulus domain, our feature-space representations likely provide only first-order approximations rather than complete accounts of the detailed similarity relations among the stimuli. Second, if the data are broken down into extremely fine-grained levels of detail, then noise in the data will limit one’s ability to diagnose places where the models are performing well versus poorly. Therefore, in the present work we set more tractable starting goals by considering the ability of the models to account for various aggregated forms of the data. As will be seen, different forms of data aggregation will provide alternative perspectives on the nature of participants’ rock-classification performance and on the strengths and weaknesses of the candidate models. Also, although a crucial target goal for the project is to eventually characterize performance at the level of individual participants (e.g., Carroll & Wish, 1974; Lee & Pope, 2003; Nosofsky, 1986; Okada & Lee, 2016), in these initial research stages we focus mainly on the key trends observed in the averaged data.

The first form of aggregated data is the one that we already provided in our Figure 5 and Figure 6 analyses, namely the patterns of performance on the center, coverage, “neither,” and “both” items for each of the subtypes across the center and coverage conditions. We conducted computer searches to locate the values of the two free parameters in the model ($c$ and $p_{store}$) that

---

4 For example, if $p_{store} = .90$, then, on average, across different runs of the simulation, .90 of the training exemplars would be stored in memory. However, because the storage process is stochastic, the exact proportion of stored exemplars, and the particular exemplars that were stored, would vary across the different runs of the simulation. So, for example, in Simulation 1, all training exemplars might be stored except, say, training exemplars 3 and 17. To compute the classification probabilities for each test-item $i$, one would sum the similarity of item $i$ to all training exemplars except for 3 and 17 and use Equation 1 to generate the predicted classification probabilities for that simulation run. Next, in Simulation 2, all training exemplars might be stored except for exemplars 5, 21, and 25. Now, the summed similarities in Equation 1 would be computed across all training exemplars except 5, 21, and 25 to generate the predicted classification probabilities for all the test items. The final classification predictions from the model would be the average of the predicted classification probabilities across all 1000 simulation runs.
minimized the sum of squared deviations between these predicted and observed item-type classification probabilities. The best-fitting predictions from the model are shown as open circles in the bar-graphs of Figures 5 and 6. Inspection of the summary data that are averaged across subtypes (see Figure 5) reveals that the model captures all of the major qualitative trends that we described previously and does a good job of quantitatively fitting the data. In addition, inspection of the top and bottom panels of Figure 6 reveals that, across both the center and coverage conditions, the model does a reasonably good job of predicting the overall performance levels for each of the individual subtypes (while simultaneously capturing the patterns of item-type performances within each subtype). So, for example, it naturally predicts the extremely accurate classification performance that was observed for obsidian and pumice; the intermediate performance levels for basalt and pegmatite; and the relatively poor performance for diorite, gabbro, granite, and rhyolite.

The summary fits for the exemplar model are reported in the top panel of Table 3. The 2-parameter model accounts for 87.0% of the variance in the classification probabilities associated with the 4 different item types across the 10 subtypes of the 2 learning conditions (roughly 80 data points). Given the complexity of this natural-science domain, these initial fits seem quite good. According to the values of the best-fitting parameters, the classification learners in this experiment stored individual training examples with probability $p_{store} = .84$ (with the value of the sensitivity parameter equal to $c = 1.79$). Our examination of the individual-subject data revealed two participants in the coverage condition with much lower overall accuracy than the remaining participants. When the model is fitted to the data with these two outlier participants removed, the fit to the averaged data is essentially the same as before, and the value of the exemplar-storage parameter rises to $p_{store} = .90$ (with $c = 1.82$).

As a source of comparison, we fitted a prototype model of classification to the item-type data (e.g., Reed, 1972; Smith & Minda, 1998). According to the prototype model, people represent each of the subtype categories in terms of the central tendency of the training exemplars of each category. Except for the form of the category representation, the prototype model is the same as the exemplar model, using the same system of Equations 1–3 as discussed previously. (Rather than summing similarities to stored exemplars as in Equation 1, the evidence for Category J is given simply by the similarity of test item i to the category-j prototype—see Nosofsky, 1986, for a detailed formal statement.) The best-fitting predictions from the prototype model, averaged across the subtypes, are shown in Figure 7. As can be seen, for the center condition, the prototype model does a fine job of predicting the data. This result is not surprising: If the training examples are packed around the central tendency of the category distribution, then there is little functional difference between computing the similarity of a test item to those central training examples versus the central tendency itself. The key result is that the prototype model fails dramatically to account for the pattern of performance in the coverage condition. As can be seen, whereas participants were 91% accurate in classifying the coverage stimuli in the coverage condition, the best-fitting prototype model predicts 69% accuracy in this condition. The magnitude of this mis-prediction is very large given that the summary results are computed across all coverage items of all subtypes. It can be viewed as a qualitative shortcoming of that alternative model.

To understand the reason for this shortcoming, consider again the schematic illustration of the coverage condition in the bottom panel of Figure 2. Although the three training exemplars are spread out across the category distribution, the central tendency computed across those three exemplars is still fairly close to the central tendency of the entire distribution. The two coverage-training examples near the edges of the distribution are located far from this central tendency. Because they are dissimilar to this prototype representation, they receive little support from it relative to the prototypes of competing categories, so the prototype model predicts relatively low accuracy on these far-away coverage items. Given this qualitative shortcoming of the prototype model, it is not surprising that its measured quantitative fit (only 21.3% of the variance accounted for) is far worse than that of the exemplar model (see Table 3).

Finally, we emphasize that the difference in the comparative fits of the exemplar and prototype models (in the coverage condition) is not some artifact of fitting the models to data averaged across the participants. In particular, in a follow-up analysis we fitted each model to each individual participant’s classification data. Because the probabilistic-storage process is stochastic, the particular subset of exemplars stored for any given participant cannot be determined in advance using the present model-fitting methods. Therefore, for simplicity, we set $p_{store} = 1$ in the present analyses and assumed that all exemplars were stored. Thus, we fitted the models to each participant’s data by estimating, separately for each

<table>
<thead>
<tr>
<th>Model</th>
<th>SSD</th>
<th>% Variance</th>
</tr>
</thead>
<tbody>
<tr>
<td>Experiment 1</td>
<td></td>
<td></td>
</tr>
<tr>
<td>Exemplar, MDS</td>
<td>.568</td>
<td>87.0</td>
</tr>
<tr>
<td>Prototype, MDS</td>
<td>3.422</td>
<td>21.3</td>
</tr>
<tr>
<td>Exemplar, Dim. ratings</td>
<td>1.236</td>
<td>71.2</td>
</tr>
<tr>
<td>Experiment 2</td>
<td></td>
<td></td>
</tr>
<tr>
<td>Exemplar, MDS</td>
<td>.761</td>
<td>80.0</td>
</tr>
<tr>
<td>Prototype, MDS</td>
<td>2.229</td>
<td>40.9</td>
</tr>
<tr>
<td>Exemplar, Dim. ratings</td>
<td>1.981</td>
<td>48.1</td>
</tr>
</tbody>
</table>

Note. SSD = sum of squared deviations; % Var. = percentage of variance accounted for; MDS = rock coordinates given by the multidimensional scaling solution derived from similarity-judgment data; Dim. ratings = rock coordinates given by the mean ratings obtained in the direct dimension-ratings experiment.

In the model-fitting analyses, we separated out the “both” stimuli from the center and coverage stimuli in computing the model fit; however, the prediction dots for the individual subtypes collapse the “both” results back into the center-item and coverage-item results in the same manner as done for the observed data. To provide the prototype model with additional flexibility that was analogous to the probabilistic-storage assumption in the exemplar model, we also made allowance for a guessing process in the prototype model. With probability $g$ the observer guessed randomly among the 10 categories (because, e.g., of the absence of a prototype with sufficient similarity to the test item), and with probability $1 - g$ classification decision making was governed by the prototype process.
participant, a single value of the sensitivity parameter \( c \) that minimized the SSD between predicted and observed classification probabilities for that participant.

As reported in the top panel of Table 4, in the coverage condition, the SSD yielded by the exemplar model (mean SSD = 1.84) was far smaller in magnitude than the SSD yielded by the prototype model (mean SSD = 4.58), \( t(26) = -8.42, p < .001 \). In addition, the exemplar model yielded a better fit than did the prototype model to the data of 25 of the 27 participants from the coverage condition. As would be expected, the difference in fits between the exemplar and prototype models was much smaller in the center condition (see Table 4, top panel) – the reason being that there is little functional difference in computing similarity to a central tendency versus three points that are clustered around the central tendency. However, even in the center condition, the SSD yielded by the exemplar model (mean SSD = 1.23) was significantly smaller than the one yielded by the prototype model (mean SSD = 1.40), \( t(32) = -4.39, p < .001 \); with the exemplar model yielding a better fit than the prototype model to the data of 25 of the 33 participants.

**Fitting the patterns of between-category confusions.** The analyses from the previous section focused on the proportion of correct responses observed for the different item types within each rock category. Another perspective on the data is obtained by considering which categories of rocks are confused with one another when errors are made. To examine this question, we computed collapsed category-confusion matrices for both the center and coverage conditions. In these matrices, cell \( i-j \) gives the probability with which members of category \( i \) were classified into category \( j \). These collapsed categorization-confusion matrices are reported in Appendix B (Tables B1 and B2).

The exemplar model predicts that the structure of the confusion matrices should reflect the configuration of rock categories in the high-dimensional similarity space. There should be high probabilities of confusion between rock categories that are positioned close to one another in the space, and low probabilities of confusion between rock categories that are positioned far apart. Furthermore, classification performance should be accurate for rock categories located in isolated portions of the space, whereas classification performance should be poor for rock categories located in dense regions.

To test these predictions, we fitted the exemplar model to the category confusion-matrix data. Note that the present goal involves fitting the model to fine-grained data involving numerous low-frequency cells (because numerous categories of rocks are rarely confused with one another). Therefore, because the error variance associated with cells of different frequencies is not homogeneous, we decided to use maximum-likelihood as a criterion of fit rather than the sum-of-squared-deviations measure used in our previous analyses. Furthermore, to evaluate the fits of models with differing numbers of free parameters, we used the Bayesian Information Criterion (BIC; Schwarz, 1978). According to the BIC, the fit of a model is given by

\[
\text{BIC} = -2\ln L + P \ln(N),
\]

where \( L \) is the (maximum) likelihood of the data given the model, \( P \) is the number of free parameters used by the model, and \( N \) is the sample size. The term \( P \ln(N) \) in Equation 4 penalizes a model for the number of free parameters that it uses. The model that minimizes the BIC is considered to provide the most parsimonious account of the data.

Because the goal was to fit the fine-grained confusion data, in addition to fitting the 2-parameter version of the GCM described previously, we decided to fit some extended versions of the model. First, classic approaches to the analysis of confusion-matrix data make allowance for the role of response bias in influencing the structure of the matrices (e.g., Luce, 1963; Shepard, 1957; Smith, 1980; Townsend, 1971). As discussed by Nosofsky (2011), in the biased GCM, the probability that stimulus \( i \) is classified in category \( J \) is given by

\[
P(C_j|i) = b_j \left( \sum_{i 
eq j} s_{ij} \right)^{\gamma} \sum_k \left[ b_k \left( \sum_{i \neq k} s_{ik} \right)^{\gamma} \right],
\]

where \( b_j (0 \leq b_j \leq 1, \sum b_k = 1) \) is the response-bias associated with category \( J \). Among other things, the bias parameters allow the model to account for asymmetric patterns of confusions among categories. For example, if the bias for responding granite is high and the bias for responding diorite is low, then members of diorite will be classified as granite more than the reverse. The second extension was to allow the attention weights defined in Equation 2 to vary as free parameters. Past research has demonstrated that in cases in which some dimensions are far more relevant than others for the purpose of classification, the attention-weight parameters play a dramatic role in allowing the GCM (related to the parameter) to account for category-learning behavior (e.g., Krusche, 1992; Nosofsky, 1984, 1986, 1987). In the present case, in which observers are learning simultaneous discriminations among 10 dif-

---

7 The likelihood function assumed that the distribution of responses into categories was multinomial in form and that the response distributions associated with each category were independent. Thus, because observers were classifying objects into 10 categories across 2 experimental conditions, the overall log-likelihood fit was given by

\[
\ln(L) = \sum_{i} \left[ \ln(N_{Ci}) - \sum_{j \neq i} \ln(f_{i,j}) + \sum_{j \neq i} f_{i,j} \cdot \ln(p_{ij}) \right]
\]

where \( N_{Ci} \) is the overall frequency with which members of category \( i \) were presented in condition \( C \) (\( C = 1 \), center; \( C = 2 \), coverage); \( f_{i,j} \) is the observed frequency with which members of category \( i \) were classified in category \( j \) in condition \( C \); and \( p_{ij} \) is the predicted probability with which members of category \( i \) were classified into category \( j \) in condition \( C \).
ficient categories, all dimensions tend to play a significant role, so extreme forms of selective-attention learning are probably not taking place. Nevertheless, it seems reasonable to hypothesize that observers may tend to focus somewhat more on some dimensions than on others in learning the present rock categories. Based on these lines of reasoning, we therefore fitted four versions of the model to the data: (1) the baseline model that allows only $c$ and $p_{store}$ to vary; (2) the baseline model with the addition of the bias parameters; (3) the baseline model with the addition of the attention weights; and (4) a “full” version of the model with the addition of both the bias parameters and the attention weights.

The fits of the four versions of the model to the category-confusion data are reported in Table 5 (top panel). According to the BIC, the best-fitting version of the model is the full version that allows both the bias parameters and attention weights to vary. Nevertheless, before examining the predictions from the full model, we start by considering the predictions from the 2-parameter baseline version in order to assess what might be viewed as prior predictions from the model.

In the top panel of Figure 8 we plot the observed confusion probabilities among the rock categories against the predicted probabilities from the baseline model. The solid dots represent cases in which the rocks were correctly classified into their respective categories, whereas the crosses represent the between-category confusion probabilities. To remove noise from the presentation in this figure, we have averaged across the two conditions of testing (center and coverage) in computing both the observed and predicted confusion probabilities. In addition, in the case of the between-category error probabilities (the crosses), we have averaged across each pair of cells $i-j$ and $j-i$.

Inspection of the figure suggests that, using only two free parameters, the baseline model is already doing a fair job of accounting for the detailed structure in the categorization-confusion matrices. The correlation between the observed and predicted correct-response probabilities across the categories is $r = .958$, and the correlation for the between-category error probabilities is $r = .815$.

To provide some intuition for the basis for these predictions, in Figure 9 we provide a plot of a condensed three-dimensional scaling solution for the rock stimuli. To produce the figure, we started with a 3-dimensional MDS solution for the 360 individual rock tokens that was derived in Nosofsky, Sanders, Meagher, et al.’s (2017) similarity-scaling study (described in the introduction to this article). We then computed the centroid of the 12 tokens that defined each rock category in this three-dimensional space and plotted these centroids in Figure 9. (Because participants learned to categorize only the igneous rocks in this experiment, we have restricted the plot to only the igneous-rock categories.) The radius of each sphere in Figure 9 is proportional to the variance of the locations of the rock tokens that define each category, averaged across the three dimensions. So, for example, the small-size sphere that is associated with obsidian indicates that the tokens that defined this category all occupied similar locations of the MDS solution; whereas the large-size sphere associated with rhyolite indicates that the tokens of this category occupied highly variable locations of the MDS solution.

We emphasize that the Figure 9 plot provides only a very rough approximation to the category-similarity structure and should be interpreted with a good deal of caution. First, Nosofsky, Sanders, Meagher, et al.’s (2017) formal analyses indicated clearly that high-dimensional solutions were necessary to provide an adequate account of the similarity relations among the rock stimuli. Second, whereas the spheres in Figures 9 might suggest that the individual categories were compact, the individual tokens that composed the

Table 4

<table>
<thead>
<tr>
<th>Model</th>
<th>Coverage</th>
<th>Center</th>
</tr>
</thead>
<tbody>
<tr>
<td><strong>Experiment 1</strong></td>
<td></td>
<td></td>
</tr>
<tr>
<td>Exemplar</td>
<td>1.84</td>
<td>1.23</td>
</tr>
<tr>
<td>Prototype</td>
<td>4.58</td>
<td>1.40</td>
</tr>
<tr>
<td><strong>Experiment 2</strong></td>
<td></td>
<td></td>
</tr>
<tr>
<td>Exemplar</td>
<td>1.59</td>
<td>1.29</td>
</tr>
<tr>
<td>Prototype</td>
<td>2.91</td>
<td>1.32</td>
</tr>
</tbody>
</table>

Note. SSD = sum of squared deviations.

Table 5

<table>
<thead>
<tr>
<th>Model</th>
<th>BIC</th>
<th>Correlation: Correct responses</th>
<th>Correlation: Between-category errors</th>
</tr>
</thead>
<tbody>
<tr>
<td><strong>Experiment 1</strong></td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>Baseline, MDS</td>
<td>5,178.5</td>
<td>.958</td>
<td>.815</td>
</tr>
<tr>
<td>Baseline + Bias parameters</td>
<td>4,556.1</td>
<td>.951</td>
<td>.839</td>
</tr>
<tr>
<td>Baseline + Attention weights</td>
<td>4,862.9</td>
<td>.963</td>
<td>.833</td>
</tr>
<tr>
<td>Full</td>
<td>4,083.5</td>
<td>.951</td>
<td>.872</td>
</tr>
<tr>
<td>Baseline, Dim. ratings</td>
<td>5,243.3</td>
<td>.826</td>
<td>.859</td>
</tr>
<tr>
<td><strong>Experiment 2</strong></td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>Baseline, MDS</td>
<td>4,418.8</td>
<td>.959</td>
<td>.920</td>
</tr>
<tr>
<td>Baseline + Bias parameters</td>
<td>3,730.2</td>
<td>.964</td>
<td>.957</td>
</tr>
<tr>
<td>Baseline + Attention weights</td>
<td>4,131.6</td>
<td>.975</td>
<td>.928</td>
</tr>
<tr>
<td>Full</td>
<td>3,349.0</td>
<td>.988</td>
<td>.976</td>
</tr>
<tr>
<td>Baseline, Dim. ratings</td>
<td>6,247.1</td>
<td>.833</td>
<td>.818</td>
</tr>
</tbody>
</table>
categories could in fact occupy highly variable locations, thereby producing category distributions that might overlap in the multidimensional space. We developed the Figure 9 plot only for the purpose of providing rough intuitive guidance from a low-dimensional scaling solution that can be visualized.

As can be seen from the condensed three-dimensional scaling solution, the rock categories obsidian and pumice are relatively compact and lie in isolated regions of the similarity space. Thus, the model predicts correctly that the members of these rock categories are classified with very high accuracy (see confusion matrices in Appendix B). By contrast, the tokens of the rock category rhyolite are highly dispersed; in addition, rhyolite lies in a dense region of the similarity space, with a number of neighboring categories. Thus, the model predicts correctly that the members of the rhyolite category are classified with low accuracy (see confusion matrices in Appendix B).

With regard to the between-category errors, category pairs such as diorite-granite and andesite-rhyolite are located relatively close together in the similarity space, so the model predicts correctly their relatively high confusion probabilities. Among the limitations of the baseline model are that it underpredicts the confusion probabilities among pairs such as peridotite-pegmatite (predicted: .040, observed: .098). We can only speculate about the basis for these mis-predictions at the present juncture, and we provide a more systematic discussion of the issue in our General Discussion. Here, we note, for example, that peridotite and pegmatite do indeed differ significantly along a large number of dimensions in the MDS solution. However, it is notable that both categories include members that are green in color (whereas none of the other rock categories had members with this feature). Possibly, this match in color was the main basis for the higher-than-predicted confusion errors observed between these categories.

In the bottom panel of Figure 8 we provide a plot of the observed confusion probabilities against the predicted confusion probabilities from the full version of the model. Naturally, there is improvement in the fine-grained predictions, with the model now yielding a correlation of $r = .872$ between the observed and predicted between-category error probabilities (the correlation for the correct responses remains high at $r = .951$). Nevertheless, some of the original limitations from the baseline model are still present (there is still a tendency, albeit reduced, for the model to underpredict the proportion of peridotite-pegmatite confusions), and we did not observe easy-to-interpret differences in response bias across the categories or attention-weighting across the stimulus dimensions. We consider these results in more detail after presenting some related analyses in our Experiment 2.

Using the direct dimension-ratings as a feature-space representation. Recall that in addition to deriving MDS solutions for the rock stimuli by fitting similarity-judgment data, Nosofsky, Sanders, Meagher, et al. (2017) also obtained direct dimension-ratings data for the rocks along a battery of 18 dimensions (see Table 2). Thus, as described in our introduction, rather than using the MDS solution as a feature-space representation for the rocks, an alternative idea is to assume that the feature-space representation is given by the averaged ratings of the rocks along

\[ \begin{align*}
\text{Figure 8.} & \text{ Experiment 1: Scatter plots of observed against predicted between-category confusion errors. Top panel: predictions from baseline model, bottom panel: predictions from full model. See the online article for the color version of this figure.}
\end{align*} \]

\[ \begin{align*}
\text{Figure 9.} & \text{ Condensed three-dimensional scaling solution for the Experiment 1 rock stimuli derived from the similarity-scaling study of Nosofsky, Sanders, Meagher, et al. (2017).}
\end{align*} \]
each of the 18 directly rated dimensions. However, to use this approach, one needs to specify the weight given to each rated dimension in computing overall distances in the space. Note that in using the MDS approach, the salience of each dimension was already calibrated with respect to an independent set of similarity-judgment data. Thus, dimensions that played a greater role in influencing the similarity judgments had greater variance in the coordinate-parameter locations of the stimuli than did dimensions that played a smaller role. To achieve an analogous form of calibration, Nosofsky et al. estimated the attention weight associated with each directly rated dimension that provided a best fit to the similarity-judgment data (for a closely related approach in a domain involving semantic concepts, see Zeigenfuse & Lee, 2010). In the present analyses, we held those previously estimated weights fixed in using the exemplar model to predict the present independent set of categorization data. Except for the use of this alternative weighted feature-space representation, the application of the exemplar model is essentially the same as described in the previous sections of this article.

The fits from this dimension-ratings version of the GCM are shown along with those of the MDS version in the top panels of Tables 3 and 5. As can be seen, in application to both sets of categorization data, the dimension-ratings version performs more poorly than does the MDS version, suggesting some limitations in this alternative approach. We consider possible reasons for these limitations after presenting the results from our closely related Experiment 2.

Discussion

The results from Experiment 1 provide some preliminary support for the proposal that the GCM, when used in combination with a high-dimensional MDS representation for the to-be-classified objects, can be applied to account successfully for category learning in a complex, natural-science domain. Although there is room for improvement, a low-parameter version of the model yielded reasonably accurate quantitative accounts of a variety of forms of the rock-picture categorization data. In Experiment 2 we sought some generality for the results by testing participants in a category-learning study involving a new set of rock categories.

Experiment 2

In our previous experiment, we limited consideration to classification learning of a set of 10 subtypes from only the broad category of igneous rocks. Conceivably, some aspect of the dimensional structure of igneous rocks could be conducive to enabling successful applications of the exemplar model. For example, as explained in college-level geology texts (e.g., Tarbuck & Lutgens, 2015, Chapter 3), the variations among numerous igneous rocks can be well characterized in terms of two major dimensions: the lightness/darkness of color of the rocks, and the average grain size of the rocks. (The lightness/darkness of color turns out to be related to the proportion of silica content in the igneous rocks’ chemical composition; the average grain size is related to the rate at which the rock solidified, which is influenced by whether the rock cooled while deep underground, near the surface, or in the air.) Perhaps the exemplar model performed well in Experiment 1 only because the MDS solution could capture these two major sources of variation among this particular subset of rock types. Therefore, to provide further tests of the model, in Experiment 2 we tested participants in another category-learning task in which we sampled subtypes from all three broad categories of rocks: igneous, metamorphic, and sedimentary. The specific subtypes are listed in Table 6.

Beyond the changed set of categories, the main experimental manipulation was again the center/coverage training-example manipulation from Experiment 1. Again, we predicted a cross-over interaction, in which performance would be best on the center items in the center condition, but best on the coverage items in the coverage condition. Furthermore, for the same reasons as explained previously, we predicted that the magnitude of the performance difference between the center and the coverage items would be greater in the center condition than in the coverage condition. As in Experiment 1, we also predicted that overall correct classification performance across the different subtypes, as well as the pattern of between-category confusion errors, would be systematically related to the configuration of the rock categories in the multidimensional similarity space.

A final prediction involved an across-experiment comparison of performance between two specific subtypes of rocks. Recall that in Experiment 1, participants classified obsidian with extremely high accuracy, but classified diorite with relatively low accuracy. Our explanation was that obsidian was located in an isolated region of the similarity space with no competing categories, whereas diorite was located in a dense region with many competitors. An alternative possibility is that the performance levels for these different categories may instead have reflected some accidental stimulus-specific properties associated with the rocks. For example, perhaps the different perceptual features that compose obsidian are easy to integrate and remember as a unit, whereas the features that compose diorite are more difficult to remember. To test between these alternatives, we chose the remaining rock subtypes in Experiment 2 so as to systematically alter the similarity structure of the category space. Specifically, we deleted most of the subtypes from Experiment 1 that were close neighbors of diorite (namely, andesite, gabbro, granite and rhyolite), and added a new subtype to Experiment 2 that was a close neighbor of obsidian (namely, anthracite). In Figure 10 we illustrate a condensed three-dimensional scaling solution for the Experiment-2 rock categories (analogous to the Figure-9 plot that we described previously). Although this low-dimensional representation should be treated with caution, the figure illustrates that diorite now lies in a relatively isolated portion of the space, whereas obsidian lies in a denser region than in Experiment 1. Thus, we predict a cross-over interaction, in which diorite is classified with higher accuracy in Experiment 2 than in Experiment 1, but in which obsidian is classified with lower accuracy in Experiment 2 than in Experiment 1.

8 Following Nosofsky, Sanders, Meagher, et al. (2017), we used a hybrid distance metric that merged city-block and Euclidean distance components in order to accommodate the mix of separable and integral dimensions that composed the rock stimuli—see Nosofsky, Sanders, Meagher, et al. (2017) for details.
metamorphic rocks, diamonds

Table 6
Rock Subtypes Tested in Experiment 2

|------------|-----------|------------|-------------|-----------|---------------|-----------|-------------|------------|----------------|---------------|

Method

Participants. The participants were 65 members of the Indiana University community. Data from 2 participants were lost as a result of computer crashes. There were 32 randomly assigned participants who completed the center condition and 31 randomly assigned participants who completed the coverage condition. All other aspects of the experiment involving the participants were the same as in Experiment 1.

Materials and procedure. All aspects of the materials and procedure were the same as in Experiment 1 except for the new set of rock categories that participants learned to classify (see Table 6).

Results

Training performance. The detailed analyses of training performance are reported in Appendix A; the pattern of results was the same as in Experiment 1.

Transfer performance. As shown in Figure 11, the pattern of summary results from Experiment 2 replicated in most respects the pattern observed in Experiment 1. The predicted cross-over interaction between training condition (center/coverage) and item type (center/coverage) was again confirmed, $F(1, 61) = 672.3, MSE = .002, p < .001, \eta^2_p = .917$; and the tests of simple effects in each training condition were again significant [center condition: $F(1, 61) = 608.7, p < .001$; coverage condition: $F(1, 61) = 146.3, p < .001$]. In addition, as predicted by the exemplar model, the magnitude of the absolute performance difference between the center and coverage items was greater in the center condition than in the coverage condition, $t(61) = 8.69, p < .001$.

Also as was observed in Experiment 1, classification of the “both” stimuli was more accurate in the center condition than in the coverage condition, $t(61) = 2.65, p = .01$. Interestingly, unlike Experiment 1, there was a slight performance advantage for the “neither” stimuli in the coverage condition ($M = .673$) compared with the center condition ($M = .636$), $t(61) = -2.65, p = .01$, a result that we discuss below.

The proportion-correct scores for the item types within each individual subtype are displayed in the top and bottom panels of Figure 12. As was the case in Experiment 1, the patterns of item-type results that were observed in the aggregate plots of Figure 11 also tend to be seen within each of the individual subtypes. At the same time, there was considerable variation in overall performance levels across the different subtypes in both the center condition [$F(9, 279) = 96.4, MSE = 1.174, p < .001, \eta^2_p = .757$] and the coverage condition [$F(6.43, 192.97) = 109.9, MSE = 1.738, p < .001, \eta^2_p = .786$] (see footnote 3).

In Figure 13 we provide a focused comparison of overall performance on diorite and obsidian across Experiments 1 and 2. As predicted by the exemplar model, there was a cross-over interaction between item type and experiment, $F(1, 121) = 323.0, MSE = .012, p < .001, \eta^2_p = .727$: Diorite was classified more accurately in Experiment 2 than in Experiment 1, $F(1, 121) = 208.8, MSE = .017, p < .001, \eta^2_p = .633$; whereas the reverse was observed for obsidian, $F(1, 121) = 114.3, MSE = .007, p < .001, \eta^2_p = .486$.

Theoretical analysis.

Fitting the item-type data from the center and coverage conditions. We fitted the exemplar model to the item-type classification data by using the same procedures as in Experiment 1.
The summary fits are reported in the bottom panel of Table 3 (best fitting parameters $c = 2.03$, $p_{store} = .88$), and the best-fitting predictions from the model are displayed as open circles in Figures 11 and 12.\(^9\) Again, the model provides excellent quantitative fits to the item-type data observed at the aggregate level (see Figure 11). It also does a reasonably good job of capturing the variability in performance observed at the subtype level in both the center and coverage conditions (see Figure 12). (Its main limitation is that it, in the center condition, it somewhat overpredicts performance on basalt and somewhat underpredicts performance on rock gypsum.) Finally, as shown in Figure 13, the model captures in reasonably good quantitative detail the predicted cross-over interaction in our focused comparison between diorite and obsidian across Experiments 1 and 2.

Although the overall patterns of performance at the aggregate level were extremely similar across Experiments 1 and 2 (compare

\(^9\) We should point out that, at least with respect to fitting the present forms of aggregate data, the best-fitting parameters from the exemplar model are nearly identical to the ones that were estimated in Experiment 1 (after deleting the data from the two outlier participants from that experiment). Indeed, if one holds fixed the Experiment 1 parameter values, there is essentially no change in the fit of the model to the present data (see Wills & Pothos, 2012, for extended arguments pertaining to the importance of parameter invariance in model evaluation). We are probably at too early a stage in the present research to always expect this form of parameter invariance, as there may be good psychological reasons why overall sensitivity and exemplar-storage probability might vary depending on the particular exemplars and category structures that need to be learned. In our General Discussion, we discuss a variety of factors that might influence the values of these parameters and that would need to be incorporated in a more fully specified version of the exemplar-memory model.
The best-fitting predictions from the prototype model for the Experiment-2 item-type data are shown in aggregate form in Figure 14. The pattern of results is the same as in Experiment 1: Although the model does a good job of capturing the results from the center condition, it again severely underpredicts the accuracies for the coverage items in the coverage condition. Thus, it again falls far short of the exemplar model in its summary-fit statistics (see Table 3, bottom panel).

Using the same method of analysis as in Experiment 1, we also fitted the individual-subject data. These fits are shown in the bottom panel of Table 4, and the patterns converge with the group fits. For the coverage condition, the SSD yielded by the exemplar model (mean SSD = 1.59) was far smaller than that of the prototype model (mean SSD = 2.91), t(30) = −8.09, p < .001. In addition, in the coverage condition, the exemplar model yielded better fits than did the prototype model to the data of 29 of the 31 participants. By comparison, and as one might expect, there was relatively little difference in the fits of the models to the center-condition data (exemplar: mean SSD = 1.29; prototype: mean SSD = 1.32; t(31) = −0.60, p = .28; with the exemplar model yielding better fits to the data of 17 of the 32 participants).

Fitting the patterns of between-category confusions. The collapsed category-confusion matrices from Experiment 2 are reported in Appendix B. The summary fits of the four versions of the exemplar model to these confusion data are reported in the bottom panel of Table 5. As found in Experiment 1, according to the BIC, adding both the bias parameters and the attention weights led to improvements in fit. The scatterplot of observed against predicted confusion probabilities from the baseline model (without bias and attention-weight parameters) is shown in the top panel of Figure 15, whereas the scatterplot from the full model is shown in the bottom panel. Again, the two-parameter model already provides a reasonably good account of the matrix of confusion probabilities. It accounts extremely accurately for the correct classification probabilities associated with the rock categories (correlation between observed and predicted proportions r = .959), and also predicts well most of the between-category confusion errors (r = .920). The model’s main limitation is that it underpredicts the confusion errors between basalt and pumice (predicted confusion probability = .021, observed confusion probability = .087). Basalt and pumice differ dramatically on numerous dimensions in the similarity space. Interestingly, however, a few specific tokens of basalt in our stimulus set had holes, which is a major characteristic feature of pumice. The matches between basalt and pumice on this relatively rare feature are the likely cause of the higher-than-predicted confusion errors between these subtypes. As shown in the bottom panel of Figure 15, making allowance for the bias and attention-weight parameters yields an extremely accurate fit of the exemplar model to the complete matrix of category-confusion data: For the full model, the correlation between the observed and predicted correct response probabilities was r = .988, and the correlation for the between-category confusion errors was r = .976.

Using the direct dimension-ratings as a feature-space representation. The fits of the direct dimension-ratings version of the exemplar model to the item-type data are reported in the bottom panel of Table 3, and the fits to the detailed matrices of category-confusion data are reported in the bottom panel of Table 5. As can be seen, the direct dimension-ratings version again performs poorly compared to the MDS version of the exemplar model. Thus, there is now a good deal of converging evidence to suggest limitations in the direct dimension-ratings approach to deriving a feature-space representation for the rock stimuli. We discuss likely reasons for these limitations in our General Discussion.

One possibility is that in the coverage condition the “neither” items may indeed have relatively high similarity to a trained coverage item in their target category but also have similarity to trained coverage items at the edge of competing categories. By contrast, in the center condition, the “neither” items may not share high similarity to the trained center items of either the target or competing categories. On balance then, the ratio of within-to-between category similarity may be roughly the same for the “neither” items across the center and coverage conditions, thereby explaining the roughly equivalent levels of classification that were observed for these items across the conditions.
The fields of cognitive psychology and cognitive science have seen tremendous advances in the development of formal models of human classification learning. However, most tests of those models have been in highly simplified domains involving stimuli with relatively few dimensions and in which participants learn to classify objects into artificial categories. The purpose of the present work was to begin an investigation of the extent to which one such model might be applicable to predicting performance in a real-world category domain involving complex stimuli varying along numerous dimensions. In particular, we tested Nosofsky’s (1986, 2011) exemplar-based GCM on its ability to predict category learning and generalization in a domain involving pictures of a variety of subtypes of igneous, metamorphic, and sedimentary rocks.

As explained in our introduction, however, tests of the exemplar model require the specification of a feature-space representation in which the to-be-classified stimuli are embedded. In work leading up to the present project, Nosofsky, Sanders, Meagher, et al. (2017) conducted extensive similarity-scaling and dimension-ratings studies in an effort to derive such feature spaces for the present stimuli. The utility of those derived feature-space representations for predicting performance in independent tasks was not tested in that previous work: Instead, such tests were an associated goal of the present research. Thus, the present work should be viewed as providing joint tests of the utility of the exemplar model and the derived feature-space representations in this natural-category domain.

Given the complexity of the world of igneous, metamorphic, and sedimentary rock categories, we do not expect to achieve the same levels of quantitative precision as in past tests of the exemplar model in simpler domains. Indeed, we envision the present approach as involving a long-range project in which continued modifications and improvements to both the model and the feature-space representation will take place in light of the experimental results. Here, we have initiated the first steps in this long-range project. These initial steps are nevertheless theoretically significant in light of persistent questions about whether exemplar models can be fruitfully extended to real-world category learning (Murphy, 2002, 2016; Smith, Zakrzewski, Johnson, & Valleau, 2016).

In our view, the results of these initial tests are quite promising. Across two experiments, a low-parameter version of the exemplar model achieved good overall predictions of a variety of forms of classification performance. These achievements included predicting participants’ performance on different item types within the different rock categories and how this performance varied with the types of training examples that learners experienced; predicting the overall levels of classification accuracy across the different rock categories; and predicting the patterns of between-category confusions that arose when classification errors were made. Moreover, the good predictions were achieved in two different learning contexts: one involving the complete set of igneous-rock categories in our rocks-stimulus set, and the second involving a mix of igneous, metamorphic, and sedimentary rock categories.

**General Discussion**

The fields of cognitive psychology and cognitive science have seen tremendous advances in the development of formal models of human classification learning. However, most tests of those models have been in highly simplified domains involving stimuli with relatively few dimensions and in which participants learn to classify objects into artificial categories. The purpose of the present work was to begin an investigation of the extent to which one such model might be applicable to predicting performance in a real-world category domain involving complex stimuli varying along numerous dimensions. In particular, we tested Nosofsky’s (1986, 2011) exemplar-based GCM on its ability to predict category learning and generalization in a domain involving pictures of a variety of subtypes of igneous, metamorphic, and sedimentary rocks.

As explained in our introduction, however, tests of the exemplar model require the specification of a feature-space representation in which the to-be-classified stimuli are embedded. In work leading up to the present project, Nosofsky, Sanders, Meagher, et al. (2017) conducted extensive similarity-scaling and dimension-ratings studies in an effort to derive such feature spaces for the present stimuli. The utility of those derived feature-space representations for predicting performance in independent tasks was not tested in that previous work: Instead, such tests were an associated goal of the present research. Thus, the present work should be viewed as providing joint tests of the utility of the exemplar model and the derived feature-space representations in this natural-category domain.

Given the complexity of the world of igneous, metamorphic, and sedimentary rock categories, we do not expect to achieve the same levels of quantitative precision as in past tests of the exemplar model in simpler domains. Indeed, we envision the present approach as involving a long-range project in which continued modifications and improvements to both the model and the feature-space representation will take place in light of the experimental results. Here, we have initiated the first steps in this long-range project. These initial steps are nevertheless theoretically significant in light of persistent questions about whether exemplar models can be fruitfully extended to real-world category learning (Murphy, 2002, 2016; Smith, Zakrzewski, Johnson, & Valleau, 2016).

In our view, the results of these initial tests are quite promising. Across two experiments, a low-parameter version of the exemplar model achieved good overall predictions of a variety of forms of classification performance. These achievements included predicting participants’ performance on different item types within the different rock categories and how this performance varied with the types of training examples that learners experienced; predicting the overall levels of classification accuracy across the different rock categories; and predicting the patterns of between-category confusions that arose when classification errors were made. Moreover, the good predictions were achieved in two different learning contexts: one involving the complete set of igneous-rock categories in our rocks-stimulus set, and the second involving a mix of igneous, metamorphic, and sedimentary rock categories.

**General Discussion**

The fields of cognitive psychology and cognitive science have seen tremendous advances in the development of formal models of human classification learning. However, most tests of those models have been in highly simplified domains involving stimuli with relatively few dimensions and in which participants learn to classify objects into artificial categories. The purpose of the present work was to begin an investigation of the extent to which one such model might be applicable to predicting performance in a real-world category domain involving complex stimuli varying along numerous dimensions. In particular, we tested Nosofsky’s (1986, 2011) exemplar-based GCM on its ability to predict category learning and generalization in a domain involving pictures of a variety of subtypes of igneous, metamorphic, and sedimentary rocks.

As explained in our introduction, however, tests of the exemplar model require the specification of a feature-space representation in which the to-be-classified stimuli are embedded. In work leading up to the present project, Nosofsky, Sanders, Meagher, et al. (2017) conducted extensive similarity-scaling and dimension-ratings studies in an effort to derive such feature spaces for the present stimuli. The utility of those derived feature-space representations for predicting performance in independent tasks was not tested in that previous work: Instead, such tests were an associated goal of the present research. Thus, the present work should be viewed as providing joint tests of the utility of the exemplar model and the derived feature-space representations in this natural-category domain.

Given the complexity of the world of igneous, metamorphic, and sedimentary rock categories, we do not expect to achieve the same levels of quantitative precision as in past tests of the exemplar model in simpler domains. Indeed, we envision the present approach as involving a long-range project in which continued modifications and improvements to both the model and the feature-space representation will take place in light of the experimental results. Here, we have initiated the first steps in this long-range project. These initial steps are nevertheless theoretically significant in light of persistent questions about whether exemplar models can be fruitfully extended to real-world category learning (Murphy, 2002, 2016; Smith, Zakrzewski, Johnson, & Valleau, 2016).

In our view, the results of these initial tests are quite promising. Across two experiments, a low-parameter version of the exemplar model achieved good overall predictions of a variety of forms of classification performance. These achievements included predicting participants’ performance on different item types within the different rock categories and how this performance varied with the types of training examples that learners experienced; predicting the overall levels of classification accuracy across the different rock categories; and predicting the patterns of between-category confusions that arose when classification errors were made. Moreover, the good predictions were achieved in two different learning contexts: one involving the complete set of igneous-rock categories in our rocks-stimulus set, and the second involving a mix of igneous, metamorphic, and sedimentary rock categories.
By comparison, an alternative prototype model of classification failed to account for the patterns of category learning and generalization, showing insensitivity to the particular examples on which observers were trained (see also Smits et al. [2002] and Storms et al. [2000] for related findings in experiments that investigated people’s classification performance involving natural-language concepts). Future research is needed to test whether prototype models might perform more successfully in the rock-category domain under alternative conditions of training. For example, in the present experiments, there was only a 10-min delay between initial training and subsequent testing in the transfer phase; we continued to provide feedback on the original training examples during the transfer phase; and there was only a small number of training examples per category. Perhaps prototype-based information is retained in people’s category representations only after long delays between study and test (Posner & Keele, 1968) and when category training-size is large (Homa, Sterling, & Trepel, 1981; but see McKinley & Nosofsky, 1995). Nevertheless, because the structure of many of the rock categories does not seem to conform to a simple, compact “family-resemblance” structure (see Nosofsky, Sanders, Gerdon, et al., 2017, Nosofsky, Sanders, Meagher, et al., 2017), we suspect that the severe limitations of prototype models observed in this study will continue to be observed in variants of the present real-world category-learning paradigm.

Although the main version of the exemplar applied in the present research made use of only two free parameters, one of those parameters—the probabilistic-storage \( p_{\text{store}} \) parameter—has not generally been used in fitting data from artificial category-learning experiments. Instead, it has typically been assumed that all exemplars are stored with probability one. The key finding that motivated our use of \( p_{\text{store}} \) in the present work is that the standard exemplar model (with \( p_{\text{store}} = 1 \)) tended to predict accuracy differences between trained and untrained items that were slightly too large. By assuming that exemplar storage is probabilistic, the predicted accuracy differences between the trained and untrained items were reduced. There are a wide variety of reasons why probabilistic-exemplar storage may have played a larger role in the present experiments compared to previous ones. Besides the high-dimensional nature of the stimulus domain, the participants in our experiments were simultaneously trying to learn 10 categorizations. By comparison, in typical artificial-category learning experiments, the objects are composed of far fewer dimensions and only two or three category distinctions need to be learned. The probabilistic-storage process may be reflecting the higher information-processing demands of the present tasks. In any case, in our view, the idea that exemplar storage is a probabilistic process is a reasonable one, and such an assumption might lead to improved exemplar-based accounts of performance in a variety of other category-learning paradigms as well.

Limitations and Future Research Directions

Despite the preliminary successes associated with the exemplar model, there were also some clear limitations as well as various key issues that were not studied in this initial research. In the following sections, we discuss some of these limitations as well as directions for future research.

Alternative scaling approaches and processing assumptions.

Part of our investigation was aimed at comparing two candidate feature-space representations for the rock stimuli, which were derived in the earlier Nosofsky, Sanders, Meagher, et al. (2017) study that we reviewed in our introduction. Among the two candidates, the MDS solution derived from the similarity-judgment data yielded better predictions of the categorization data than did the representation based on the averaged direct dimension ratings. We emphasize that, heading into this comparison between the two feature-space candidates, we believed that the direct dimension ratings might fare better than the MDS approach at predicting the category-learning data, and found the present outcome to be highly instructive. It is true that the direct dimension-ratings model yielded worse fits to the original similarity-judgment data collected by Nosofsky, Sanders, Meagher, et al. (2017) than did the eight-dimensional MDS solution. However, in that previous situation, fitting the eight-dimensional MDS model to the similarity-judgment data involved the estimation of a very large number of free parameters (8 coordinate parameters for each of 360 stimuli) compared with the number of free parameters used by the direct dimension-ratings model. Note that in transitioning to the context of the present category-learning experiments, the very large set of parameters from the MDS model was held fixed and the parameters were no longer free to vary. When high-parameter models are fitted to one data source, and their parameters are then held fixed, the models may often not generalize well when used to predict independent data sources. In the present case, however, the high-parameter MDS model did generalize well: Apparently, the high-dimensional MDS solution captures more of the fine-grained similarity structure among the rock stimuli that is relevant for classification than do the direct dimension ratings.

There are numerous potential reasons for the limitations associated with the use of the direct dimension ratings as a feature-space representation. First, certain characteristics of the rocks that play a significant role in influencing psychological similarity and classification may not be easily accessible for verbal description. Such characteristics may have been left out of the dimension-ratings battery. Second, in the current modeling, it was assumed that the psychological coordinates of the stimuli in the feature space were linearly related to the mean ratings of the rocks on each dimension. Instead, the relation between psychological and rated scale values may be highly nonlinear. Third, in the modeling, we used simple additive rules for combining distances along individual rated dimensions to compute multidimensional distances (for details, see Nosofsky, Sanders, Meagher, et al., 2017); however, the correct combination rule may be far more complex. Future research is needed to systematically address each of these potential factors.

Although the MDS solution derived from the similarity judgments fared much better at predicting the categorization data, there were some limitations associated with this approach as well. One issue is that dimensions that are important for the purpose of making fine-grained discriminations between different rock categories may receive relatively little emphasis in the context of a generic similarity-judgment task. For example, both breccia and conglomerate (two types of sedimentary rocks) are composed of fragments that are cemented together in unorganized fashion; however, the fragments composing breccia are sharp and angular whereas the fragments composing conglomerate are rounded and
smoother. Observers may not consider this fine-grained detail involving the shape of the fragments in making their similarity judgments. Conceivably, in future work, the information obtained in the direct dimensions-rating study might be used to supplement the psychological dimensions that were derived from the similarity-judgment data in order to develop a more complete representation. Still another avenue to explore in future work involves the use of deep-learning network models that might automate the derivation of the high-dimensional feature representations of the stimuli from this natural category domain (e.g., Guest & Love, 2017; Lake, Zaremba, Fergus, & Gureckis, 2015; Peterson, Abbott, & Griffiths, 2016).

Another issue is that more sophisticated processing models may be needed for characterizing similarity relations among complex objects such as rocks. In the MDS approach, similarity is assumed to be some decreasing function of continuous distance in a space. Alternative models of similarity, such as Tversky’s (1977) influential feature-contrast model (FCM), are based on matching and mismatching of discrete features. An important aspect of the FCM is that it allows for “boosts” in similarity between objects in cases in which they have a greater number of common matching features. It is difficult to capture such boosts using approaches based solely on continuous distance. We suspect that the role of common features becomes particularly important in cases in which the matching features are relatively rare. For example, as noted earlier in our article, the present modeling approach underestimated the extent of between-class confusions between pumice and basalt. A major characteristic feature of pumice is the presence of holes; some tokens of basalt also contain holes. The match on this relatively rare feature is a likely reason for the higher-than-predicted confusions between these two categories of rocks. One approach to addressing this issue might be to formulate hybrid models that combine contributions of continuous distance and matching discrete features in determining the overall similarity between objects (e.g., Lee & Navarro, 2002; Nosofsky & Zaki, 2003; Verguts, Ameel, & Storms, 2004).

Regardless of the manner of deriving the feature-space representation and the detailed processing assumptions associated with the model, future work will need to expand upon the current versions to account for expert forms of classification (e.g., Biederman & Shiffrar, 1987; Gauthier, Skudlarski, Gore, & Anderson, 2000; Palmeri, Wong, & Gauthier, 2004; Shen, Mack, & Palmeri, 2014; Tanaka, Curran, & Sheinberg, 2005). Geologic experts have undoubtedly learned to attend to subtle characteristics of the rock stimuli that enable improved categorization performance compared with novice learners. An exciting and highly fruitful direction for future research is to conduct systematic comparisons between the feature-space representations for the rock stimuli derived from the judgments of novices and experts.

On the “curse of dimensionality.” A limitation of human mental capabilities is an inability to visualize and intuit the structure of high-dimensional spaces. However, various mathematical and computer-simulation analyses suggest that sets of problematic phenomena arise in high-dimensional spaces that do not occur in low-dimensional spaces. For example, although the details must depend on the generating process, one theme is that as dimensionality increases, Euclidean distance computations do not behave as one might expect, with all items tending to become equally distant from one another (e.g., Aggarwal, Hinneburg, & Keim, 2001; Beyer, Goldstein, Ramakrishnan, & Shaft, 1999). A closely related theme is that with increased dimensionality, all points tend to occupy “corners” of the space, with none occupying the “center.”

In view of these often discussed aspects of the “curse of dimensionality,” the relative success of the present modeling approach seems all the more interesting and impressive. The bottom line is that the approach seems to have “worked” (at least under the present conditions of testing). First, as documented in the similarity-scaling study reported by Nosofsky, Sanders, Meagher, et al. (2017), an eight-dimensional Euclidean scaling solution provided good accounts of the similarity-judgment data collected for the present rock stimuli, with the derived dimensions being highly interpretable. Second, the derived solution was used successfully in combination with the proposed exemplar model to provide good overall accounts of the present category-learning and generalization data. And the finding that the prototype and exemplar models yielded similar predictions of performance in the “center” conditions of the present experiments suggests that the center training examples may in fact have occupied relatively central locations of the individual rock-category structures.

Given the challenges in understanding the nature of high-dimensional spaces, at this juncture we can only speculate about why the “curse” did not appear to impede the exemplar model’s success under the present testing conditions. One possibility is that the curse manifests itself only at very high dimensionalities, and the present Euclidean-distance approach remains well behaved at eight dimensions. Another possibility is that the processes that give rise to the generation of the structure of natural categories place constraints on the configuration of objects in high-dimensional spaces, and that the statistical models that have been used in analyzing the curse of dimensionality do not obey these constraints. For example, perhaps any given individual rock category exhibits significant variation along only a few dimensions, reducing the category’s effective dimensionality. These questions are extremely complex and challenging ones, and it will require extensive future research to resolve them.

Prior knowledge and contextual influences. For purposes of simplicity, we assumed in our modeling that all learning of the rock categories took place within the context of the experiments themselves. Although our participants reported little or no prior experience in rock classification, it seems likely that many had at least some prior knowledge of some of the categories. Such prior knowledge would need to be incorporated in the formal models to capture the category-learning data in even greater detail (e.g., Rehder & Murphy, 2003). One approach to representing such prior knowledge within the framework of the exemplar model is to load the model with prior examples and assume that the observer compares test items to both training examples from the experiment and to the prior examples as well (Heit, 1994). A related approach involves the use of exemplar-based connectionist network models in which prior knowledge influences the setting of the starting association weights between exemplar and category nodes (Choi, McDaniel, & Busemeyer, 1993). To adopt such approaches, one would need to conduct careful measurement of the prior knowledge that observers bring with them to the experiment. This direction is an important one for future research.

Likewise, in the real world, rock classification often does not operate in context-free situations but rather takes place in particular environments that provide clues about the identity of the rocks. For example, in a major process that generates sedimentary rocks, weathering and erosion break rocks into small pieces that...
are deposited in shallow seas, lakes, or rivers. Over time, layers of the deposited rocks are compacted and cemented together. Thus, layered rocks found in regions that were once river beds or shallow seas are likely to be sedimentary rocks. An effective classifier would incorporate such prior knowledge to guide his or her categorization decisions. One approach to formalizing such a process within the framework of the exemplar model is to develop theories of the response-bias parameters that are part of the model (see Equation 5). In the scenario described above, for example, response biases for classifying items as members of the categories of sedimentary rocks should increase, and response biases for classifying items as members of other rock categories would decrease.

Using the model to optimize teaching. With continued progress in the development of the model and the feature-space representation, our hope is that the present approach can usefully be translated to real-world classroom and field settings to improve the teaching of scientific classifications.

For example, one question to be decided in teaching scientific categories is which training examples should be used. In the present work, we limited consideration to cases in which there were three training examples per category and the examples were chosen using either a “center” algorithm or a “coverage” algorithm. Although not the focus of our article, we can consider the exemplar model’s predictions of overall proportion correct, averaged across all 120 test items, during the transfer phase of our experiments. It turns out that, given the parameter settings reported in our article, the exemplar model predicts slightly higher overall proportion correct in both experiments under the coverage-training condition than under the center-training condition (Experiment 1: coverage = .688, center = .652; Experiment 2: coverage = .750, center = .730). The observed data were in general accord with these predictions (Experiment 1: coverage = .723, center = .686; Experiment 2: coverage = .754, center = .719), a source of evidence that provides some further confirmation of the modeling approach.

The more important point, however, is that the “center” and “coverage” cases are but two examples from an essentially infinite collection of different possibilities. An extremely wide variety of combinations of such possibilities might be explored in empirical studies. Our proposal is that applications of the model itself could provide theoretical guidance to organize such research and suggest which possibilities are the most promising ones. The idea would be to use the model itself to predict the learning and generalization outcomes that would be yielded by the different training sets (for recent work, see Carvalho & Goldstone, 2017; Kruschke, 1992). Closely related models have been proposed and tested that presume that training examples are grouped into multiple “clusters” and where the specific clusters that are formed depend strongly on the precise sequencing of the examples (e.g., Anderson, 1991; Love et al., 2004; Sanborn et al., 2010). In the same manner as illustrated in this article, such models could be tested on their ability to account for effects of sequencing in the present natural.

The coverage advantage was statistically significant in both Experiment 1, \( t(56) = -2.34, p < .05 \); and Experiment 2, \( t(61) = -2.68, p < .01 \). The Experiment 1 computations are with the two outlier participants removed from the analysis.

It is of interest to compare our center-coverage manipulation to a related one reported in a recent study of Hornsby and Love (2014). These researchers conducted a category-learning study in which novice participants learned to classify mammograms as normal or tumorous. One group was trained exclusively on items that, according to a nornal study, were easy and unambiguous to classify. A second group was trained on a mixture of easy, medium, and hard items. On the surface, Hornsby and Love’s “easy” group might seem roughly comparable to the present “center” condition, whereas the mixture group seems roughly comparable to the present “coverage” condition. Furthermore, because Hornsby and Love (2014) observed significantly better overall classification transfer performance for the “easy” group than the “mixture” group, their pattern of results would appear to be the opposite of the one we have just reported here. However, there are numerous differences between the two studies that make such comparisons difficult to interpret. Two key differences include the following. First, the Hornsby-Love (Hornsby & Love, 2014) paradigm involved a single category contrast: normal versus tumorous. By comparison, in the present study, participants were learning an enormous number of simultaneous, embedded category contrasts, with the directions of contrast in the multidimensional space varying across each pair of categories. The types of training examples that lead to better transfer performance are likely very different across such conditions. Second, detailed findings reported by Hornsby and Love (2014, pp. 74–75) suggest that the mammograms may have involved probabilistically overlapping category structures, with the “hard” training items located in regions that were more typical of the contrast category than the target category. As explained by Gigüere and Love (2013; see also Nosofsky & Stanton, 2005), including such training items adds noise to the exemplar-based decision process. By comparison, in the present study, although some of the “coverage” training items were located toward the edges of the category distributions, it may seldom have been the case that they overlapped into the contrast-category regions. In a nutshell, a great deal of future research would be needed to develop more rigorous comparisons across these two studies.
science category domain. Successful versions of such models could then be simulated to search for optimal training-example sequences. In an even more ambitious approach, sequence-sensitive models could be used in an online manner to assess the current nature of a learner’s category representation based on his or her recent history of categorization choices while performing the task. The model could then be used to suggest which specific training examples would lead to the greatest enhancements in the individual learner’s future performance (e.g., Markant & Gureckis, 2014; Mettler & Kellman, 2010).

Conclusion

In sum, in the present research we have taken promising initial steps in testing a formal model of human classification learning in a complex, high-dimensional, natural-science category domain. We also identified certain limitations of the model and its associated feature-space representation that will stimulate future research to improve upon the current version. With continued development and testing, our hope is that the model can eventually be put to use for suggesting effective techniques for the teaching of scientific classifications in the classroom and the field. Such work would contribute to the goal of developing important bridges between the domains of formal modeling in cognitive psychology and theory-driven approaches to improving education.

Context of the Research

Throughout their careers, Nosofsky and McDaniel have been interested in the nature of category learning and representation. A dominant theme in Nosofsky’s work involves the development and testing of formal computational models, whereas McDaniel has placed emphasis on the translation of principles of cognitive psychology to education and the learning sciences. They are currently engaged in a collaborative project funded through the Education Core Research Program of the National Science Foundation. The theme of the project is to use formal models of human classification to help guide the search for effective strategies of teaching science categories in the classroom. The example target domain is rock classification in the geologic sciences. The general research strategy is to simulate alternative teaching methods using successful formal models, and to focus empirical testing efforts on those methods that the models predict are most promising. Before implementing the strategy, however, it is essential to verify that the formal models that have been successful in accounting for performance in artificial category-learning experiments conducted in the laboratory will scale up successfully to account for the learning of complex, high-dimensional real-world categories. The work reported in this article represents the initial steps in achieving this goal. Modeling the learning and representation of natural-science categories presents some interesting and significant challenges. Thus, in addition to the goal of translation to education, the project is revealing key theoretical issues that will be of fundamental interest to the fields of cognitive psychology and cognitive science.

References


( Appendices follow )
Appendix A

Analysis of the Training-Phase Data

In Figure A1 we plot the results from the training phase of Experiments 1 (top panel) and 2 (bottom panel). The figure plots mean proportion correct as a function of blocks (1-6) and condition (center and coverage). As can be seen, in both experiments, proportion correct increases dramatically across training blocks, with accuracy in the center condition being higher than accuracy in the coverage condition. However, by the final blocks, accuracy for both the center and coverage conditions is very high. We conducted 2 (condition) × 6 (blocks) mixed-model analyses of variance of the proportion-correct scores in each experiment. In Experiment 1, there was a significant effect of condition \(F(1,58) = 32.5, \text{MSE} = 1.58, p < .001, \eta^2_p = .359\); a significant effect of blocks \(F(2.42, 140.6) = 620.7, \text{MSE} = 4.52, p < .001, \eta^2_p = .915\); and a significant interaction between the two factors \(F(2.42, 140.6) = 17.9, \text{MSE} = .130, p < .001, \eta^2_p = .235\). Likewise, in Experiment 2, there was a significant effect of condition \(F(1, 61) = 9.60, \text{MSE} = .285, p < .01, \eta^2_p = .136\); a significant effect of blocks \(F(2.58, 157.35) = 606.9, \text{MSE} = 4.54, p < .001, \eta^2_p = .909\); and a significant interaction \(F(2.58, 157.35) = 7.22, \text{MSE} = .054, p < .001, \eta^2_p = .106\).

![Figure A1](image.png)

*Figure A1.* Mean proportion correct as a function of conditions and blocks in the training phase of Experiments 1 (top panel) and 2 (bottom panel).
# Appendix B

## Collapsed Category-Confusion Matrices

### Table B1

**Collapsed Category-Confusion Matrices in Experiment 1**

<table>
<thead>
<tr>
<th>Category</th>
<th>1</th>
<th>2</th>
<th>3</th>
<th>4</th>
<th>5</th>
<th>6</th>
<th>7</th>
<th>8</th>
<th>9</th>
<th>10</th>
</tr>
</thead>
<tbody>
<tr>
<td><strong>Center condition</strong></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>1</td>
<td>.623</td>
<td>.035</td>
<td>.097</td>
<td>.049</td>
<td>.018</td>
<td>.003</td>
<td>.040</td>
<td>.015</td>
<td>.004</td>
<td>.118</td>
</tr>
<tr>
<td>2</td>
<td>.119</td>
<td>.632</td>
<td>.004</td>
<td>.098</td>
<td>.001</td>
<td>.005</td>
<td>.001</td>
<td>.004</td>
<td>.114</td>
<td>.021</td>
</tr>
<tr>
<td>3</td>
<td>.019</td>
<td>.007</td>
<td>.655</td>
<td>.020</td>
<td>.200</td>
<td>.003</td>
<td>.063</td>
<td>.029</td>
<td>.000</td>
<td>.003</td>
</tr>
<tr>
<td>4</td>
<td>.051</td>
<td>.086</td>
<td>.122</td>
<td>.598</td>
<td>.025</td>
<td>.005</td>
<td>.028</td>
<td>.023</td>
<td>.000</td>
<td>.062</td>
</tr>
<tr>
<td>5</td>
<td>.002</td>
<td>.001</td>
<td>.142</td>
<td>.007</td>
<td>.506</td>
<td>.007</td>
<td>.068</td>
<td>.019</td>
<td>.000</td>
<td>.249</td>
</tr>
<tr>
<td>6</td>
<td>.000</td>
<td>.002</td>
<td>.000</td>
<td>.000</td>
<td>.000</td>
<td>.992</td>
<td>.005</td>
<td>.000</td>
<td>.000</td>
<td>.001</td>
</tr>
<tr>
<td>7</td>
<td>.004</td>
<td>.003</td>
<td>.014</td>
<td>.005</td>
<td>.071</td>
<td>.012</td>
<td>.823</td>
<td>.044</td>
<td>.001</td>
<td>.024</td>
</tr>
<tr>
<td>8</td>
<td>.058</td>
<td>.069</td>
<td>.036</td>
<td>.084</td>
<td>.012</td>
<td>.001</td>
<td>.114</td>
<td>.584</td>
<td>.018</td>
<td>.024</td>
</tr>
<tr>
<td>9</td>
<td>.006</td>
<td>.000</td>
<td>.003</td>
<td>.002</td>
<td>.000</td>
<td>.001</td>
<td>.003</td>
<td>.004</td>
<td>.974</td>
<td>.008</td>
</tr>
<tr>
<td>10</td>
<td>.130</td>
<td>.129</td>
<td>.059</td>
<td>.045</td>
<td>.040</td>
<td>.038</td>
<td>.011</td>
<td>.024</td>
<td>.055</td>
<td>.470</td>
</tr>
<tr>
<td><strong>Coverage condition</strong></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>1</td>
<td>.621</td>
<td>.061</td>
<td>.071</td>
<td>.107</td>
<td>.029</td>
<td>.000</td>
<td>.019</td>
<td>.029</td>
<td>.005</td>
<td>.059</td>
</tr>
<tr>
<td>2</td>
<td>.082</td>
<td>.755</td>
<td>.011</td>
<td>.061</td>
<td>.001</td>
<td>.004</td>
<td>.002</td>
<td>.020</td>
<td>.051</td>
<td>.012</td>
</tr>
<tr>
<td>3</td>
<td>.038</td>
<td>.024</td>
<td>.613</td>
<td>.072</td>
<td>.188</td>
<td>.001</td>
<td>.026</td>
<td>.026</td>
<td>.001</td>
<td>.011</td>
</tr>
<tr>
<td>4</td>
<td>.087</td>
<td>.146</td>
<td>.053</td>
<td>.522</td>
<td>.028</td>
<td>.011</td>
<td>.043</td>
<td>.087</td>
<td>.003</td>
<td>.019</td>
</tr>
<tr>
<td>5</td>
<td>.008</td>
<td>.005</td>
<td>.179</td>
<td>.044</td>
<td>.601</td>
<td>.004</td>
<td>.088</td>
<td>.034</td>
<td>.002</td>
<td>.034</td>
</tr>
<tr>
<td>6</td>
<td>.000</td>
<td>.000</td>
<td>.000</td>
<td>.001</td>
<td>.003</td>
<td>.994</td>
<td>.002</td>
<td>.000</td>
<td>.000</td>
<td>.000</td>
</tr>
<tr>
<td>7</td>
<td>.006</td>
<td>.003</td>
<td>.019</td>
<td>.012</td>
<td>.028</td>
<td>.006</td>
<td>.819</td>
<td>.099</td>
<td>.004</td>
<td>.004</td>
</tr>
<tr>
<td>8</td>
<td>.047</td>
<td>.021</td>
<td>.047</td>
<td>.055</td>
<td>.021</td>
<td>.001</td>
<td>.135</td>
<td>.638</td>
<td>.012</td>
<td>.024</td>
</tr>
<tr>
<td>9</td>
<td>.004</td>
<td>.016</td>
<td>.000</td>
<td>.003</td>
<td>.000</td>
<td>.000</td>
<td>.000</td>
<td>.004</td>
<td>.962</td>
<td>.010</td>
</tr>
<tr>
<td>10</td>
<td>.167</td>
<td>.078</td>
<td>.065</td>
<td>.053</td>
<td>.038</td>
<td>.009</td>
<td>.010</td>
<td>.040</td>
<td>.211</td>
<td>.519</td>
</tr>
</tbody>
</table>

**Note.** The entry in cell $i$–$j$ gives the conditional probability with which members of category $i$ were classified into category $j$. 1 = andesite; 2 = basalt; 3 = diorite; 4 = gabbro; 5 = granite; 6 = obsidian; 7 = peridotite; 8 = pumice; 9 = rhyolite.

### Table B2

**Collapsed Category-Confusion Matrices in Experiment 2**

<table>
<thead>
<tr>
<th>Category</th>
<th>1</th>
<th>2</th>
<th>3</th>
<th>4</th>
<th>5</th>
<th>6</th>
<th>7</th>
<th>8</th>
<th>9</th>
<th>10</th>
</tr>
</thead>
<tbody>
<tr>
<td><strong>Center condition</strong></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>1</td>
<td>.705</td>
<td>.022</td>
<td>.005</td>
<td>.236</td>
<td>.010</td>
<td>.000</td>
<td>.008</td>
<td>.010</td>
<td>.003</td>
<td>.000</td>
</tr>
<tr>
<td>2</td>
<td>.003</td>
<td>.963</td>
<td>.002</td>
<td>.001</td>
<td>.008</td>
<td>.016</td>
<td>.000</td>
<td>.001</td>
<td>.008</td>
<td>.000</td>
</tr>
<tr>
<td>3</td>
<td>.000</td>
<td>.007</td>
<td>.826</td>
<td>.003</td>
<td>.157</td>
<td>.003</td>
<td>.000</td>
<td>.002</td>
<td>.003</td>
<td>.000</td>
</tr>
<tr>
<td>4</td>
<td>.000</td>
<td>.014</td>
<td>.951</td>
<td>.003</td>
<td>.002</td>
<td>.000</td>
<td>.000</td>
<td>.000</td>
<td>.000</td>
<td>.000</td>
</tr>
<tr>
<td>5</td>
<td>.004</td>
<td>.001</td>
<td>.122</td>
<td>.004</td>
<td>.857</td>
<td>.004</td>
<td>.000</td>
<td>.000</td>
<td>.000</td>
<td>.000</td>
</tr>
<tr>
<td>6</td>
<td>.003</td>
<td>.078</td>
<td>.001</td>
<td>.003</td>
<td>.002</td>
<td>.579</td>
<td>.006</td>
<td>.020</td>
<td>.282</td>
<td>.027</td>
</tr>
<tr>
<td>7</td>
<td>.117</td>
<td>.011</td>
<td>.000</td>
<td>.088</td>
<td>.005</td>
<td>.115</td>
<td>.359</td>
<td>.211</td>
<td>.048</td>
<td>.047</td>
</tr>
<tr>
<td>8</td>
<td>.059</td>
<td>.000</td>
<td>.022</td>
<td>.003</td>
<td>.042</td>
<td>.024</td>
<td>.245</td>
<td>.529</td>
<td>.041</td>
<td>.036</td>
</tr>
<tr>
<td>9</td>
<td>.010</td>
<td>.004</td>
<td>.001</td>
<td>.004</td>
<td>.030</td>
<td>.197</td>
<td>.013</td>
<td>.026</td>
<td>.713</td>
<td>.002</td>
</tr>
<tr>
<td>10</td>
<td>.015</td>
<td>.012</td>
<td>.004</td>
<td>.049</td>
<td>.000</td>
<td>.001</td>
<td>.144</td>
<td>.050</td>
<td>.016</td>
<td>.707</td>
</tr>
<tr>
<td><strong>Coverage condition</strong></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>1</td>
<td>.837</td>
<td>.008</td>
<td>.004</td>
<td>.088</td>
<td>.019</td>
<td>.002</td>
<td>.022</td>
<td>.021</td>
<td>.000</td>
<td>.001</td>
</tr>
<tr>
<td>2</td>
<td>.003</td>
<td>.989</td>
<td>.005</td>
<td>.002</td>
<td>.000</td>
<td>.010</td>
<td>.000</td>
<td>.000</td>
<td>.000</td>
<td>.000</td>
</tr>
<tr>
<td>3</td>
<td>.000</td>
<td>.004</td>
<td>.841</td>
<td>.005</td>
<td>.144</td>
<td>.001</td>
<td>.002</td>
<td>.001</td>
<td>.000</td>
<td>.000</td>
</tr>
<tr>
<td>4</td>
<td>.025</td>
<td>.004</td>
<td>.004</td>
<td>.923</td>
<td>.003</td>
<td>.003</td>
<td>.011</td>
<td>.012</td>
<td>.001</td>
<td>.015</td>
</tr>
<tr>
<td>5</td>
<td>.003</td>
<td>.001</td>
<td>.134</td>
<td>.005</td>
<td>.849</td>
<td>.001</td>
<td>.000</td>
<td>.000</td>
<td>.000</td>
<td>.000</td>
</tr>
<tr>
<td>6</td>
<td>.005</td>
<td>.035</td>
<td>.001</td>
<td>.012</td>
<td>.004</td>
<td>.818</td>
<td>.022</td>
<td>.022</td>
<td>.077</td>
<td>.004</td>
</tr>
<tr>
<td>7</td>
<td>.137</td>
<td>.008</td>
<td>.006</td>
<td>.065</td>
<td>.008</td>
<td>.396</td>
<td>.197</td>
<td>.015</td>
<td>.071</td>
<td>.017</td>
</tr>
<tr>
<td>8</td>
<td>.091</td>
<td>.001</td>
<td>.034</td>
<td>.012</td>
<td>.026</td>
<td>.039</td>
<td>.282</td>
<td>.422</td>
<td>.035</td>
<td>.058</td>
</tr>
<tr>
<td>9</td>
<td>.007</td>
<td>.013</td>
<td>.000</td>
<td>.005</td>
<td>.015</td>
<td>.265</td>
<td>.018</td>
<td>.019</td>
<td>.653</td>
<td>.004</td>
</tr>
<tr>
<td>10</td>
<td>.013</td>
<td>.003</td>
<td>.002</td>
<td>.005</td>
<td>.000</td>
<td>.017</td>
<td>.090</td>
<td>.039</td>
<td>.006</td>
<td>.825</td>
</tr>
</tbody>
</table>

**Note.** The entry in cell $i$–$j$ gives the conditional probability with which members of category $i$ were classified into category $j$. 1 = basalt; 2 = diorite; 3 = obsidian; 4 = pumice; 5 = anthracite; 6 = marble; 7 = dolomite; 8 = micrite; 9 = rock gypsum; 10 = sandstone.