Psychonomic Bulletin & Review (2019) 26:48–76
https://doi.org/10.3758/s13423-018-1508-8

THEORETICAL REVIEW

Model-guided search for optimal natural-science-category training exemplars: A work in progress

Robert M. Nosofsky[1] · Craig A. Sanders[1] · Xiaojin Zhu[2] · Mark A. McDaniel[3]

Published online: 9 July 2018

Abstract

Under the guidance of a formal exemplar model of categorization, we conduct comparisons of natural-science classification learning across four conditions in which the nature of the training examples is manipulated. The specific domain of inquiry is rock classification in the geologic sciences; the goal is to use the model to search for optimal training examples for teaching the rock categories. On the positive side, the model makes a number of successful predictions: Most notably, compared with conditions involving focused training on small sets of training examples, generalization to novel transfer items is significantly enhanced in a condition in which learners experience a broad swath of training examples from each category. Nevertheless, systematic departures from the model predictions are also observed. Further analyses lead us to the hypothesis that the high-dimensional feature-space representation derived for the rock stimuli (to which the exemplar model makes reference) systematically underestimates within-category similarities. We suggest that this limitation is likely to arise in numerous situations in which investigators attempt to build detailed feature-space representations for naturalistic categories. A low-parameter extended version of the model that adjusts for this limitation provides dramatically improved accounts of performance across the four conditions. We outline future steps for enhancing the current feature-space representation and continuing our goal of using formal psychological models to guide the search for effective methods of teaching science categories.

Keywords Models of category learning · Perceptual categorization and identification · Similarity

A key component of science education involves learning the fundamental categories of the target domain. For example, in botany, students learn wide varieties of plant types; in entomology, insect types; and, in the domain that is the focus of the present work, in the geologic sciences, students learn classifications of rocks.

In our recent work (Nosofsky, Sanders, Gerdom, Douglas, & McDaniel, 2017; Nosofsky, Sanders, & McDaniel, 2018a, b), we have initiated a long-range project that has the aim of using formal models of human category learning to help guide the search for effective ways of teaching such scientific classifications. The general research strategy is to simulate alternative classification-teaching methods using the formal models themselves. Empirical tests would then focus on those teaching methods that the model simulations predict would be most successful (for illustrative examples in highly simplified categorization domains, see, e.g., Khajah, Lindsey, & Mozer, 2014; Mathy & Feldman, 2016; Patil, Zhu, Kopec, & Love, 2014). Confirmation of the predictions could then lead to implementing the methods in the science classroom.

Although we believe that the idea is a good one in principle, it is highly ambitious for a number of reasons. Most important, although a variety of sophisticated formal models of human classification have been developed in the fields of psychological and cognitive science (for a comprehensive review, see Pothos & Wills, 2011), almost all rigorous tests have been in laboratory experiments involving artificial category structures and highly controlled stimuli that vary along just a few salient dimensions. By comparison, in real-world category domains such as rock types, the category structures

Electronic supplementary material The online version of this article (https://doi.org/10.3758/s13423-018-1508-8) contains supplementary material, which is available to authorized users.

✉ Robert M. Nosofsky
nosofsky@indiana.edu

[1] Psychological and Brain Sciences, Indiana University Bloomington, 1101 E. Tenth Street, Bloomington, IN 47405, USA

[2] University of Wisconsin-Madison, Madison, WI, USA

[3] Washington University in St. Louis, St. Louis, MO, USA

are highly complex, and the stimuli are composed of numerous dimensions that are difficult to describe and discern. Thus, the extent to which the formal models may scale up successfully to account for real-world category learning is highly uncertain.

To provide a foothold on this ambitious goal of applying formal category learning models to guide more effective instruction of science categories, in the present article, we report our initial attempts at implementing the research strategy. Using a formal exemplar-memory model of categorization (Nosofsky, 1986) to guide our investigation, the specific question that we explore is whether the model can be used to search for optimal sets of training examples for teaching the categories. Importantly, such research can be viewed as a two-way street: To the extent that laboratory tests of the model predictions are successful, it provides a firmer basis for implementing the teaching methods in real-world classroom settings. On the other hand, if the predictions fail, then important new diagnostic information is provided concerning limitations of the proposed model. Such information can be used to develop improved versions of the model or to suggest alternative models. The model-building and testing process can then be continued in iterative fashion.

To anticipate, we will see some of both of these outcomes—successes and failures—in the present work. On the positive side, the most notable result concerns a successful prediction involving the variable of category training-set size on learning and generalization. Across a reasonably broad range of its parameter settings, the exemplar model correctly predicts that (a) performance on old training items is significantly enhanced in conditions with small sets of training examples, but that (b) generalization to novel transfer items is significantly enhanced in a condition in which learners instead experience a large-size swath of training examples from each category. Importantly, although related research has previously pointed in this direction, we will argue that influential past studies have confounded manipulations related to category-training size with other factors, such as the total number of training trials devoted to each category.

On the negative side, our work will also reveal some departures from the exemplar model's predictions of which specific training examples will lead to optimal performance in a set of small category-training-size conditions. We will then take initial steps of revising the model (and its associated machinery) in light of these failures. Furthermore, we will provide a clear and promising direction for a more comprehensive form of revision. Importantly, we believe that the limitations that we identify are likely to hold generally for numerous others investigations in which researchers attempt to provide rigorous model-based accounts of category learning in naturalistic domains. Thus, we believe that the path that we carve out is likely to have broad, instructive value for the field.

In our next section, we provide a brief review of our recent efforts at modeling rock-classification learning and generalization. Building upon that work, we then outline the new theoretical and empirical efforts reported in this article in which we attempt to use the model to search for effective methods of teaching scientific classifications.

Review of the formal model and its initial tests

Sketch of the formal model

The model that is used to guide the present work is a well-known *exemplar* model termed the *generalized context model* (GCM; Medin & Schaffer, 1978; Nosofsky, 1984, 1986, 2011). According to the model, people represent categories by storing individual training exemplars of the categories in memory, and classify objects on the basis of their similarity to the stored exemplars. A number of more sophisticated models that elaborate upon the GCM in important ways have been developed in the field (e.g., Anderson, 1991; Love, Medin, & Gureckis, 2004; Nosofsky & Palmeri, 1997; Sanborn, Griffiths, & Navarro, 2010; Vanpaemel & Storms, 2008); however, because it is a fairly simple model that has already been applied successfully across diverse domains, it seemed reasonable to use the GCM as our starting point.

In a simple descriptive version of the model, the probability that item i is classified in Category J is found by summing the similarity of i to all exemplars of Category J and then dividing by the summed similarity of i to all exemplars of all categories:

$$P(J|i) = \frac{\sum_{j \in J} s_{ij}}{\sum_K \left(\sum_{k \in K} s_{ik} \right)}, \tag{1a}$$

where s_{ij} is the similarity of item i to exemplar j.

Typical applications of the GCM adopt a multidimensional scaling (MDS) approach (Kruskal & Wish, 1978; Shepard, 1980) to computing the similarity between each pair of exemplars. In MDS, objects are represented as points in a multidimensional space, and similarity is a decreasing function of distance in the space. In the present applications, we assume a simple Euclidean distance metric for computing the distance between each pair of exemplars i and j:

$$d_{ij} = \left[\sum \left| x_{im} - x_{jm} \right|^2 \right]^{1/2}, \tag{2a}$$

where x_{im} is the value of exemplar i on dimension m. (More elaborate versions of the model, discussed later in the article, extend the Equation 2a distance function with a set of "attention weight" parameters that systematically modify the structure of the space in which the exemplars are embedded.) Furthermore, following Shepard (1957, 1987), similarity is presumed to be an exponentially decreasing function of distance in the space:

$$s_{ij} = \exp\left(-c{\cdot}d_{ij}\right), \tag{3a}$$

where c is an overall sensitivity parameter. The sensitivity parameter describes the rate at which similarity declines with distance and provides a measure of overall discriminability in the space.

In their recent applications of the model to predicting rock classification, Nosofsky et al. (2018b) found that the fits of the model were significantly improved by adopting a probabilistic-storage assumption. Instead of assuming that all training exemplars were stored in memory, each training exemplar had some probability p_{store} of being stored (see Nosofsky et al., 2018b, for extended discussion).[1] We adopt this probabilistic-storage assumption again in the present applications.

Derivation of the MDS space The key to implementing the model is to derive the MDS space in which the exemplars are embedded and from which the similarities s_{ij} in Equations 1–3 are computed. In numerous past tests of the model, the derivation of the MDS space was straightforward, because the tests involved the use of highly controlled, simple stimuli varying along just a few salient dimensions (for reviews, see Ashby, 1992; Nosofsky, 1992; for examples involving use of high-dimensional MDS solutions for predicting semantic categorization, see, e.g., Storms, De Boeck, & Ruts, 2000; Verheyen, Ameel, & Storms, 2007; Voorspoels, Vanpaemel, & Storms, 2008). In a real-world category domain such as rocks, however, the derivation of the MDS space becomes a highly ambitious task. The stimuli that compose such categories vary along a very large number of dimensions, many of which are difficult to describe or discern.

Thus, as a prerequisite to testing the GCM in the rock-classification domain, Nosofsky, Sanders, Meagher, and Douglas (2018) engaged in extensive similarity-scaling studies of the rock stimuli.[2] In these studies, observers provided similarity judgments among pairs of items drawn from a set composed of 360 rock pictures (10 categories of each of the broad divisions of igneous, metamorphic, and sedimentary rocks, with 12 samples of each of the categories). MDS methods were then applied to fit the similarity-judgment data and thereby embed the stimuli in the space (for details, see Nosofsky et al. 2018). As noted by Nosofsky et al. (2018), because of the very large number of stimuli that needed to be scaled, the number of observations per individual cell of the similarity-judgment matrix was very small (there are more than 100,000 cells in a 360 × 360 similarity-judgment matrix). Thus, despite obtaining hundreds of judgments from each of more than 250 participants, the data were noisy at the level of individual cells. Nevertheless, at least at a global level, the results of the MDS analysis appeared to be remarkably straightforward and impressive. First, an eight-dimensional scaling solution provided an excellent fit (97.2% of the variance accounted for) to an aggregate form of the similarity-judgment data; the aggregate data measured the average similarity between all members of each pair of the 30 rock categories (e.g., the average similarity between all members of the categories *granite* and *diorite*). Second, the derived dimensions of the MDS solution had natural psychological interpretations. In particular, the coordinate values of the exemplars on the individual eight dimensions correlated highly with an independent group of subjects' direct ratings of the stimuli on the attributes of lightness/darkness of color, average grain size, roughness/smoothness, shininess, organization, chromaticity, hue, and certain shape-related components. Interactive displays of the derived eight-dimensional solution are provided in the online website (https://osf.io/w64fv/) associated with Nosofsky et al.'s (2018) study. Finally, and perhaps most important, when used in combination with the MDS solution, the GCM was able to achieve good first-order quantitative predictions of rock-classification learning and generalization in an independent set of categorization experiments involving the same stimuli (Nosofsky et al., 2018b). We provide a brief review of the modeling results from those initial rock category-learning experiments in the next section.

To anticipate, despite the promising initial results noted above, our present research will lead us to the conclusion that more work is needed to develop a still more precise and comprehensive scaling representation for the rock stimuli used in our experiments. Given the complexities of the rock stimulus domain, and the vast amount of data required to measure similarities among the large number of to-be-scaled stimuli, this outcome seems a reasonable one.

Review of recent applications of the GCM to rock category learning

In one recent experiment, Nosofsky et al. (2018b) had participants learn to classify the rock stimuli into the complete set of 10 igneous-rock categories in their rock-pictures collection. The goal was to use the GCM, in combination with the derived MDS solution for the rocks, to account for participants' category-learning and generalization performance. The key independent variable that was manipulated was the nature of

[1] Rather than referring to each individual-trial presentation of a rock instance, the p_{store} parameter refers to the probability that a single representation of each rock instance has been stored, along with its correct category label, by time of the test phase. Although standard applications of the GCM make provision for differential memory strengths of stored exemplars (Nosofsky, 1988, 1991), for simplicity in the present applications we assume that all individual instance-based representations have equal strength in memory.

[2] We should note that, in most of our studies, we have limited consideration to cases involving the classification of *pictures* of rocks. In principle, however, the scaling solution could be extended to include information pertaining to nonvisual properties such as hardness, density, and so forth.

the training exemplars. Across two conditions, participants first engaged in an instance-based training phase involving multiple presentations of three training exemplars of each of the 10 categories. In the *center* condition, the three training exemplars were those closest in distance to the centroid of each category distribution defined in the eight-dimensional scaling solution for the rocks (see Fig. 1, top panel, for a schematic two-dimensional illustration). In the *coverage* condition, the three training exemplars more completely covered the entire rock-category distribution; however, there was far less training on central exemplars than in the center condition (see Fig. 1, bottom panel). (For illustrative examples involving pictures of the actual rock categories, see Fig. 3 of Nosofsky et al., 2018b)

Following the training phase, participants engaged in a test phase that included presentations of the original training exemplars as well as novel rock samples from the categories. Nosofsky et al. (2018b) measured the mean probability—for each of the individual 10 rock categories—with which the participants correctly classified the center and coverage training exemplars across the conditions, as well as the probability with which participants generalized correctly to novel transfer items from each of the categories. Estimating only two free parameters—the overall sensitivity parameter and the probabilistic-storage parameter (described earlier in our introduction)—Nosofsky et al. found that the model provided an excellent account of this rich set of classification-probability data. Successful predictions from the model were also observed in a conceptual-replication experiment in which, rather than learning to classify only igneous rocks, participants learned to classify rock pictures into a mix of igneous, metamorphic, and sedimentary rock categories. In short, these initial tests provided extremely promising results in support of the idea that the GCM could be "scaled up" to account in a parsimonious fashion for classification performance in a complex, high-dimensional natural-category domain.

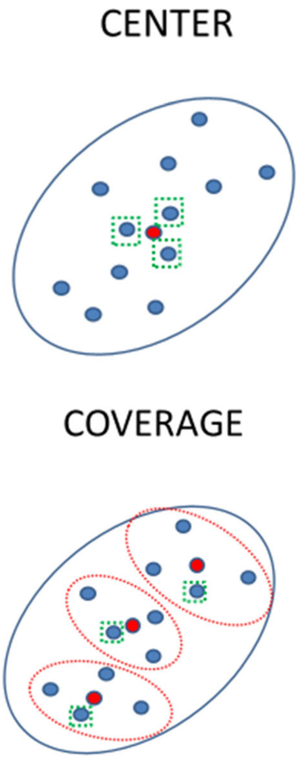

Fig. 1 Schematic illustration of the center and coverage conditions tested by Nosofsky et al. (2018b). Top panel: In the center condition, the three training examples (dotted squares) were the items closest to the centroid (red dot) of the entire category distribution. Bottom panel: In the coverage condition, the category distribution was divided into three nonoverlapping clusters that covered the entire distribution, and the training examples (dotted squares) were the items closest to the centroid of each individual cluster. (Color figure online)

Experiment

Search for optimal training sets

Given the success of the model in our recent rock-category-learning experiments, we were now prepared for our next step of using the model to search for enhanced techniques of teaching the rock classifications. Although a variety of such teaching issues might be pursued, the specific question that we addressed in the current research was, Which training exemplars might serve as optimal ones if the goal were to maximize observers' overall proportion of correct classifications at time of test? Importantly, in addressing this question, we presumed that—except for the choice of specific training exemplars—all other aspects of the methods from the Nosofsky et al. (2018b) experiments would remain basically the same. For example, in the current experiment, we again used a training phase followed by a test phase, with the same number of training and test trials used in the previous experiments. Participants again learned to classify rocks into 10 distinct igneous-rock categories, with 12 samples per category. In addition, in the conditions in which the model was used to generate strong, a priori predictions of performance (see below), there were roughly three training exemplars per category. (However, we also tested a larger training-set-size condition; as will be seen, the results from this condition turned out to provide some key results of major theoretical and practical interest.) More detailed aspects of the training and test procedures (e.g., stimulus presentation sequences, nature of feedback, stimulus and feedback durations) were also the same as in the earlier experiments. By holding fixed these methodological components across the experiments, it seemed reasonable to assume invariance of the best-fitting parameter values across the studies (cf. Wills & Pothos, 2012), thereby allowing for true, a priori quantitative predictions of performance in the newly tested conditions. As our model-based analyses will show, even if there were some variations in best-fitting parameter values (due, for example, to the fact that different populations of participants were tested), the pattern of predictions of the overall levels of performance across the new conditions would remain basically the same.

Before describing the specific conditions that were conducted, we should note a limitation in the measure that we used to define "optimal" test performance. In our view, ideally, the measure would pertain only to new, untrained objects from the categories. After all, in the real world, a learner will seldom experience the exact same rock sample twice; instead, the real interest is in learners' ability to generalize their knowledge to previously unseen examples of the categories. Thus, in future work, our goal would be to reserve some completely separate transfer set of rock stimuli for assessing true generalization performance.

Unfortunately, in our present work, we are limited to use of a fixed test set of stimuli for which we have obtained our MDS solution for the individual rocks. The test set includes all items(i.e., both old training items and new transfer items). Because the specific training items will differ across the experimental conditions (see below), so will the specific novel transfer items that are left over. Thus, across some conditions, there will be stimulus-specific differences in which items serve as training exemplars and which serve as transfer items. Accordingly, for the present study, we decided it made the most sense to define the objective function to be optimized as the overall proportion of correct classifications measured across all individual items in the test set, regardless of whether the items were training or transfer items.

The four training conditions

We implemented four different conditions in the present experiment. As noted above, in all conditions, participants learned to classify the rock pictures into the 10 igneous-rock categories from the larger stimulus set that was scaled in the Nosofsky et al. (2018) MDS study. In all conditions, the same set of 120 unique igneous-rock exemplars were presented during the test phase. The conditions differed only in terms of the specific training instances that were used.

The first condition was a replication of the *coverage* condition tested in Experiment 1 of Nosofsky et al. (2018b) and that we described above. Because a different population of participants was tested in the present experiment, we needed to repeat the condition in order for it to serve as a source of comparison with the new conditions. We decided to not also retest the *center* condition illustrated in our Fig. 1 because Nosofsky et al. (2018b) had already confirmed the GCM's prediction that overall performance in the coverage condition would be better than in the center condition in two separate experiments involving different sets of categories.

We term the second condition the *small-size* (*ss*)*-optimal* condition. Using a "greedy-search" computer algorithm (see Appendix A for details), we located the sets of training exemplars from each of the 10 categories that the GCM predicted would yield the maximum overall proportion correct during the test phase. As explained previously, overall proportion correct was computed across all 120 test items, regardless of whether an item was an old training instance or a novel transfer item. In conducting the computer search for the ss-optimal training exemplars, the predicted proportion-correct value was computed under the assumption that the best-fitting parameters estimated in Nosofsky et al.'s (2018b) closely related experiment would remain invariant. However, as we show below, the pattern of proportion-correct scores across conditions is predicted to remain basically the same across a large range of the parameter settings. Whereas there were precisely three training exemplars per category (i.e., 30 total training exemplars) in the coverage condition, we allowed more flexibility in assigning numbers of training instances in the ss-optimal condition. Specifically, besides holding fixed the total number of unique training exemplars at 30, the only other constraint introduced was that there be at least two training exemplars from each of the 10 categories. As it turned out, the algorithm tended to assign fewer training instances to easy categories and more training instances to harder ones. (We provide more details about these selections in Appendix A.) Besides distributing the number of training instances across categories in a (theoretically) more effective manner, our impression was that the training instances selected for the ss-optimal condition had the property of "covering" their respective category distributions, in much the same manner as the coverage condition. Thus, the distinctions between the coverage and ss-optimal conditions turned out to be subtle ones.

Intuitively, in asking whether performance in the ss-optimal condition will indeed be superior to performance in the coverage condition, we are setting a high bar for our investigation. The reason is that the coverage condition is already expected to be a relatively good condition for fostering high levels of performance in the categorization test phase. In particular, according to the exemplar model, because the training instances in that condition "cover" the complete category distribution, they should support very good generalization performance to novel transfer items (cf. Posner & Keele, 1968). Indeed, as noted earlier in this section, in the previous study reported by Nosofsky et al. (2018b), overall performance in the coverage condition was significantly better than in the center condition across two separate experiments, as predicted by the GCM.

Thus, in the present study, rather than relying solely on a comparison of performance across the coverage and ss-optimal conditions, we decided to introduce a third condition that we term the *random-3* condition. In this condition, for each individual participant, three random training instances were chosen from each of the 10 categories. As shown below, our computer simulations indicated that overall performance in the coverage condition should be better than in the random-3 condition. This result is important, because the random-3 condition can be considered to provide a measure of the "average" level of performance that might be expected to be observed under the present types of training conditions.

Finally, we tested a condition that we term the *random-9* condition. In this condition, for each individual participant, there were *nine* randomly chosen instances from each of the 10 categories (rather than three). Thus, there was a greater total number of training instances in this condition than in the other conditions; however, because we held fixed across conditions the total number of trials during the training phase (see the Method section for details), there were fewer presentations of each individual training instance in this condition. Because the nature of the training phase is changed substantially compared with the other conditions, we can no longer assume invariance of the GCM parameters across the random-9 condition and the other three conditions. However, based on preliminary computer simulations reported below, we nevertheless predicted that overall test performance would be considerably better in the random-9 condition than in the other three conditions.

To provide clearer documentation for the hypotheses outlined above, in Fig. 2 we report quantitative predictions from the GCM of overall proportion correct in the test phase for the four conditions. The predictions are computed across a range of parameter settings from the model. The top panel ("moderate" *c*) displays the predictions with the value of the sensitivity parameter (*c*) held fixed at the value estimated in the previous study reported by Nosofsky et al. (2018b); the middle panel ("low" *c*) displays the predictions with the value of *c* reduced by 25%; and the bottom panel ("high" *c*) displays the predictions with the value of *c* increased by 25%. Within each panel, we plot the quantitative predictions as the value of exemplar-storage parameter (p_{store}) varies from .05 to 1.00 in increments of .05. The best-fitting value in Nosofsky et al.'s experiment was p_{store} = .90, so values in this region are of the greatest interest. As can be seen from inspection of the figure, among the three conditions in which there are roughly three training exemplars per each of the 10 categories, the model does indeed predict that overall performance will be best in the ss-optimal condition, followed by the coverage condition, and finally the random-3 condition. Furthermore, this prediction is robust, as it holds for a wide range of values of the sensitivity parameter, and for values of p_{store} ranging from .50 to 1.00.

As discussed previously, it does not seem reasonable to assume parameter invariance for the random-9 condition; indeed, because there are far fewer presentations of each individual training instance in the random-9 condition than in the other three conditions, it should almost certainly be expected that the value of the exemplar-storage parameter (p_{store}) would be lower in the random-9 condition than in the other three. (We defer more detailed discussion of how the parameters in

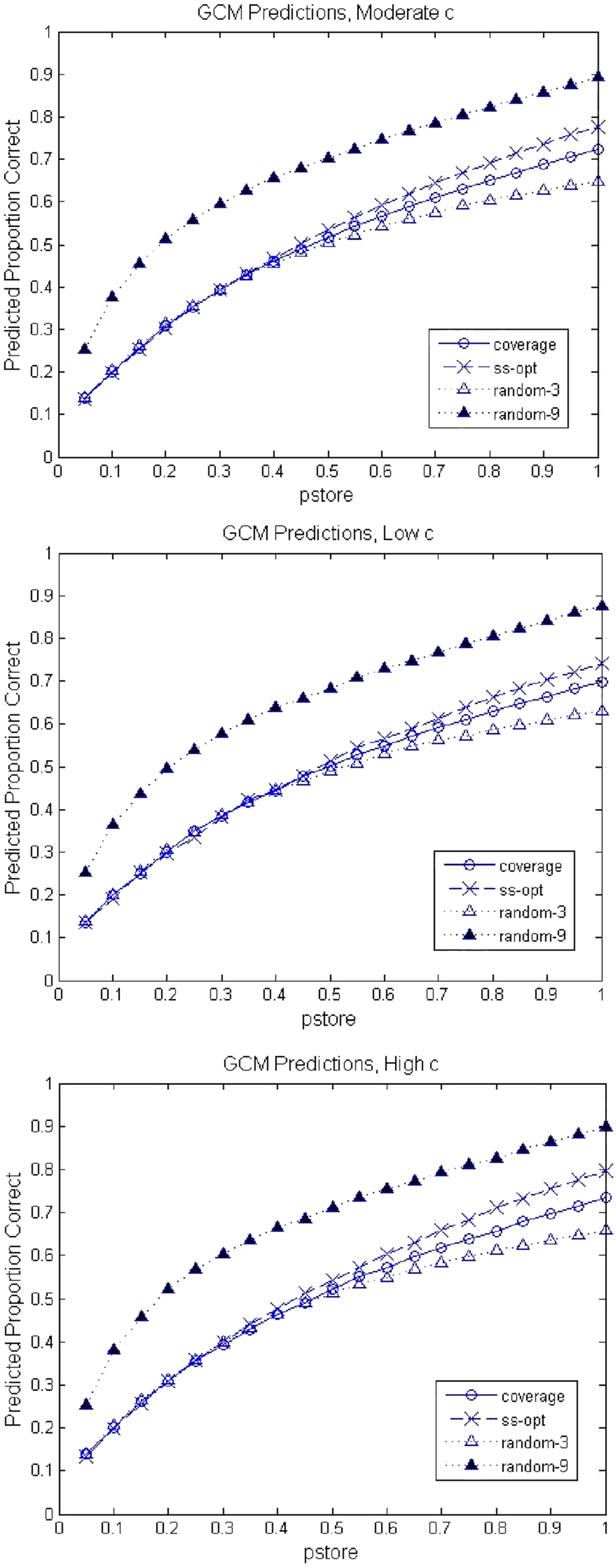

Fig. 2 Plots of GCM-predicted overall proportion correct in the test phase across the four conditions as a function of level of overall sensitivity (*c*) and exemplar-storage probability (p_{store}). ss-opt = small-set-size optimal

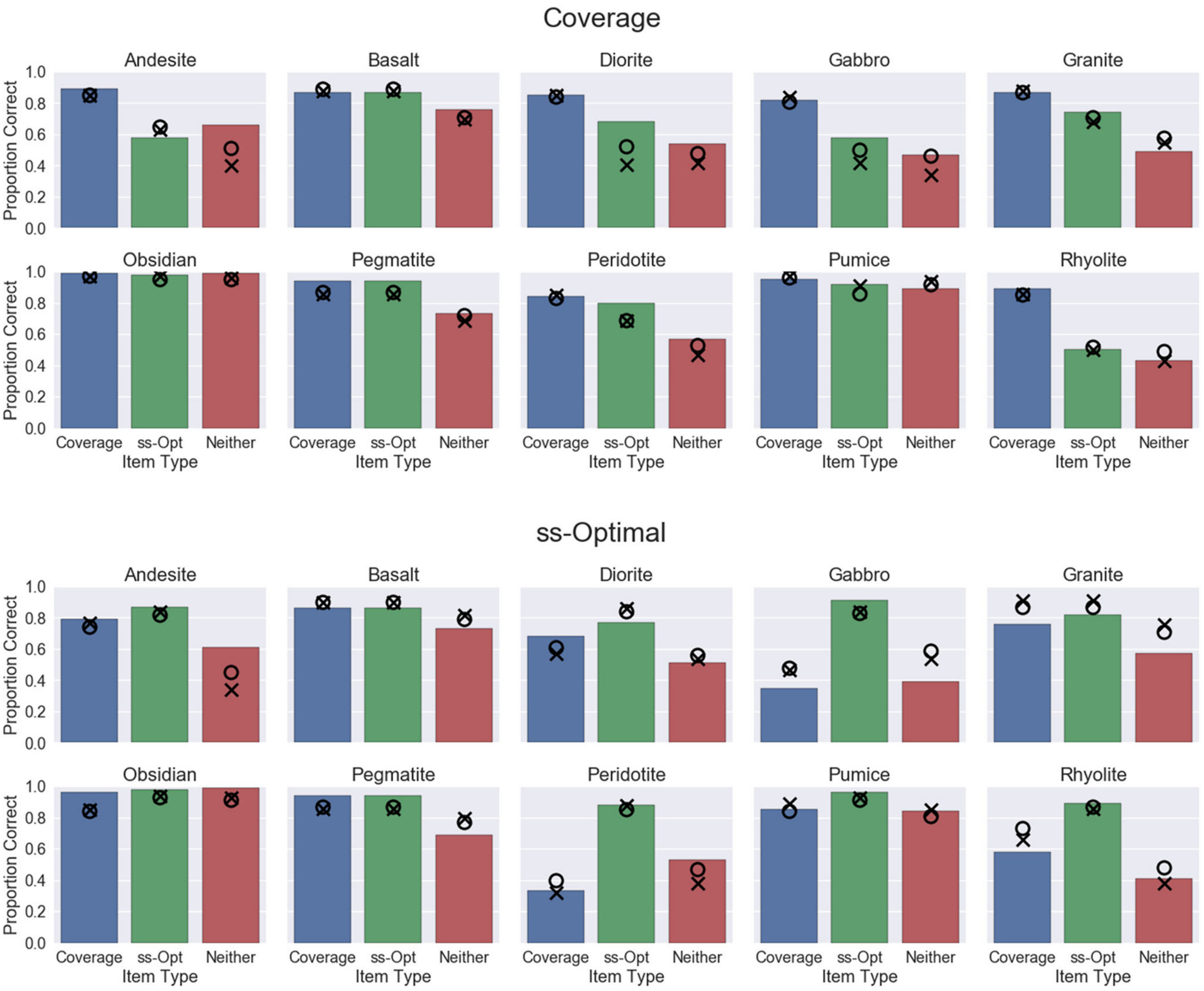

Fig. 3 Mean proportion correct for each of the item types in each of the 10 categories in the coverage and ss-optimal conditions. Colored bars = observed data (blue = coverage-training items, green = optimal-training items, red = neither items). Xs denote predictions from the baseline version of the GCM; open circles denote predictions from the baseline + c_w version of the GCM. (Color figure online)

the random-9 condition might compare with those in the other three conditions until later in our article.) Importantly, however, it can be seen from inspection of Fig. 2 that, even when the value of p_{store} is considerably lower in the random-9 condition than in the other three conditions, the model predicts that overall proportion correct will be greater in the random-9 condition than in the other three conditions.

This prediction involving the random-9 condition arises for two reasons. First, a much larger proportion of the test instances are old training instances in the random-9 condition than in the other three conditions; furthermore, as will be seen, the model predicts a robust training-instance advantage under the present conditions of testing. Second, even when the value of p_{store} is relatively low, our simulations indicated that the model predicts relatively good generalization performance to the new transfer stimuli. Intuitively, the reason is that the large number of training instances in the random-9 condition "cover" the complete category distributions, in much the same fashion as occurs in the coverage and ss-optimal conditions. We unpack the basis for these predictions in greater detail in the Theoretical Analysis section of our article.

Method

Participants The participants were 163 members of the Indiana University community. Roughly half were undergraduate students from introductory psychology courses who received credit toward a course requirement; the remaining half were paid for taking part in the experiment. The paid participants received $24 for a 2-hour experimental session, plus a possible $6 bonus for good performance (defined as 60% correct or better during the test phase of the experiment).

Participants were assigned randomly to the four conditions, with essentially the same proportion of for-credit and paid participants assigned to each condition. There was a total of 41, 44, 38, and 40 participants assigned to the coverage, ss-optimal, random-3, and random-9 conditions, respectively. 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 that form a subset of those used in the previous similarity-scaling studies reported by Nosofsky et al. (2018); the full set of rock images is available in the Rocks Library folder of the website associated with that article (https://osf.io/w64fv/). 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 inches wide and 1.7 inches tall. Participants sat approximately 20 inches from the computer screen, so each rock picture subtended a visual angle of approximately 6.0° × 4.9°. Further details regarding the manner in which the rock pictures were sampled and displayed are provided by Nosofsky et al. (2018).

The experiment was programmed in MATLAB and the Psychophysics Toolbox (Brainard, 1997). All participants were tested individually in private, sound-attenuated cubicles.

Procedure The statistical algorithm for choosing the training exemplars in the coverage condition has been described in detail by Nosofsky et al. (2018b). We described the procedure for choosing the training exemplars in the ss-optimal, random-3, and random-9 conditions in the introduction to this experiment. In Appendix A we provide a listing of the specific exemplars that served as training items in the coverage and ss-optimal conditions; recall that the specific training exemplars varied randomly across participants in the random-3 and random-9 conditions. The MDS coordinates of all 120 stimuli used in the four conditions are provided in the Rocks Library folder of the website https://osf.io/w64fv/. It turned out that, for most categories, one or occasionally two items served as both a coverage training exemplar and an ss-optimal training exemplar across those conditions; we refer to such items as “both” items. In addition, we refer to items that did not serve as either coverage or ss-optimal training exemplars in those conditions as “neither” items.

Table 1 Igneous-rock categories tested in the current experiment

Cat. #	Cat. name
1	Andesite
2	Basalt
3	Diorite
4	Gabbro
5	Granite
6	Obsidian
7	Pegmatite
8	Peridotite
9	Pumice
10	Rhyolite

In all conditions, the experiment was divided into a training phase and a test phase. In all conditions, the training phase consisted of six blocks of 60 trials each. Within each block of the coverage, ss-optimal and random-3 conditions, each of the 30 training exemplars was presented twice. Thus, each individual training instance was presented a total of 12 times in these conditions. By comparison, in the random-9 condition, each of the individual 90 training instances was presented a total of four times; the presentations of each instance were spread as evenly as possible across the six blocks of training (a total of two times in each of Blocks 1–3 and 4–6). The order of presentation of the training exemplars was randomized anew for each block and each participant, subject to the constraints just described.

On each trial of the training phase, a training exemplar was presented in the center of the screen, and the subject classified it into one of the ten subtype categories. The category response was indicated by pressing labeled number keys on the computer keyboard. A listing of the assignment of numbers to category names was displayed on the computer screen on each trial to facilitate this procedure. Following the response, corrective feedback was provided on the screen for 2 s (e.g., “Correct! Diorite”; or “Incorrect, Diorite”). At the end of each block, participants were informed of their overall proportion of correct responses. There was a 10-minute break between the training and test phases.

In the test phase, in all conditions, participants were tested on all 120 items from the igneous-rock set. 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 six blocks of 60 trials each in the test 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. (To hold fixed the number of feedback trials across all conditions, only one third of the test trials involving the 90 old exemplars in the random-9 condition were feedback trials.) 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 test block. The entire experiment took slightly less than 2 hours to complete.

Results

Prior to conducting the main analyses, we computed the overall proportion correct for each individual participant during the test phase and constructed histograms of the results. Visual inspection indicated extremely similar patterns of results for the for-credit and paid participants, so we combined those groups in all subsequent analyses. In addition, prior to conducting the main analyses, we deleted participants who appeared as severe outliers in the histogram plots. We deleted the data of four, three, three, and two outliers from the coverage, ss-optimal, random-3, and random-9 conditions, respectively.[3]

Although our central interest is in theoretical analyses that consider the ability of the formal model to account quantitatively for detailed aspects of the data, we get started by reporting elementary statistical analyses of some of the summary results.

In Table 2, we report overall proportion correct during the test phase, averaged across all items (old training and new transfer), in each of the four conditions. A more fine-grained breakdown is reported in Tables 3 and 4. Table 3 separates test items in the coverage and ss-optimal conditions according to whether they are “coverage-only” training items, “ss-optimal-only” training items, “both-conditions” training items, or “neither” items. Table 4 separates items across all four conditions according to whether they are old training items or new transfer items.

Inspection of Table 2 reveals that overall performance in the coverage condition was better than in the random-3 condition, as predicted by the GCM. In addition, overall performance in the random-9 condition was clearly better than in the remaining three conditions; although the GCM-based predictions are parameter dependent, this result too seems consonant with the model’s pattern of predictions (see Fig. 2). However, in contrast to the GCM’s predictions, overall performance in the ss-optimal condition did not exceed overall performance in the coverage condition; if anything, the results trended slightly in the opposite direction. To confirm these observations, we conducted a one-way ANOVA on the overall proportion-correct scores across the four conditions. The analysis revealed a significant effect of condition, $F(3, 147) = 18.09$, $MSE = .006$, $p < .001$. Planned comparisons revealed that overall performance in the random-9 condition was significantly better than in each of the other three conditions, average $t(147) = 5.75$, with $p < .001$ in all individual comparisons. In addition, as predicted, planned comparisons revealed that overall performance in the coverage condition was significantly better than in the random-3 condition, $t(147) = 2.19$, $p < .05$. However, in contrast to the GCM’s prior predictions, overall performance was not significantly different across the coverage and ss-optimal conditions, $t(147) = -1.06$, $p = .29$.

Table 2 Overall proportion correct during the test phase in each of the four conditions

Source	Coverage	ss-Optimal	Random-3	Random-9
Condition				
Observed data	.712	.693	.672	.797
Baseline model predictions	.663	.703	.609	.783
Baseline model + c_w predictions	.692	.712	.658	.792

Because the main goal of classification training is to foster generalization to new transfer items, the performance results for the new transfer items are perhaps of greater interest (see Table 4). A one-way ANOVA on new transfer-item performance again revealed a significant effect of condition, $F(3, 147) = 7.65$, $MSE = .007$, $p < .001$. Planned comparisons revealed that generalization performance in the random-9 condition was significantly better than the average across the coverage, ss-optimal, and random-3 conditions, $t(147) = 3.78$, $p < .001$. Note that individual-condition comparisons revealed significantly better generalization performance in the random-9 condition than in the ss-optimal and random-3 conditions, average $t(147) = 3.77$, with $p < .001$ in both cases; however, the improvement in generalization performance in the random-9 condition compared with the coverage condition did not reach significance, $t(147) = 1.67$, $p = .097$.

Because specific stimuli are chosen as old versus new items in the coverage and ss-optimal conditions, whereas old and new items are chosen randomly in the random-3 and random-

Table 3 Mean proportion correct by item type in the coverage and ss-optimal conditions

Condition	Cov.	ss-Opt.	Neith.	Both
Item type				
Coverage	.890	.607	.665	.892
ss-Optimal	.589	.877	.641	.892

Note. Cov = coverage-only training instance; ss-Opt = ss-optimal-only training instance; Neith = test stimuli that were not training instances in either condition; Both = test stimuli that were training instances in both conditions

[3] The extremely poor performance of these outlier participants was almost certainly due to factors such as lack of motivation, failure to understand instructions, and so forth. Such factors go beyond the scope of the present modeling efforts. In the test phase, the individual outliers had overall proportion correct .18, .24, .42 and .43 in the coverage condition; .13, .18 and .35 in the ss-optimal condition; .09, .12 and .28 in the random-3 condition; and .13 and .52 in the random-9 condition.

Table 4 Mean proportion correct by item type (old vs. new) in all four conditions

Condition	Old	New
Item type		
Coverage	.891	.653
ss-Optimal	.883	.630
Random-3	.905	.594
Random-9	.834	.685

9 conditions, the previous analyses confound general training conditions with specific stimuli. To remove this confound, we conducted focused analyses of performance on the old and new items across only the random-3 and random-9 conditions. A two-way mixed-model ANOVA using conditions (random-3 vs. random-9) as a between-subjects factor and item type (old vs. new) as a within-subjects factor revealed that old items were classified with significantly higher accuracy than were new ones, $F(1, 71) = 793.2$, $MSE = .002$, $p < .001$. In addition, there was a significant interaction between conditions and item type, $F(1, 71) = 99.07$, $MSE = .002$, $p < .001$. The interaction reflects that old items were classified with higher accuracy in the random-3 condition than in the random-9 condition, whereas the reverse occurred for new items.

The better performance on old items in the random-3 condition is not surprising, because participants received far more training on individual old items in the random-3 condition than in the random-9 condition. Of greater interest is that, despite the better performance on the old items in the random-3 condition, generalization performance on the new items was better in the random-9 condition. Generally speaking, if the goal is to promote generalization performance to new items, our results suggest that training with a broad sample of items from the category distribution is better than extensive training with a few items—even if old-item performance on the broad sample does not reach the same levels as is achieved with focused training on a few old items. Beyond the theoretical implications of the pattern of results, our finding is potentially highly significant for making recommendations regarding training strategies in natural-science category domains. We provide extensive discussion of this important result (and its relation to previous ones reported in the category-learning literature) in our General Discussion.

Finally, in Figs. 3 and 4, we provide a fine-grained breakdown of the data for each individual rock category in the four conditions. Figure 3 shows the results for the coverage and ss-optimal conditions. In these plots, the blue bars show performance on the coverage-training items, the green bars on the ss-optimal-training items, and the red bars on the neither items. (Because of small sample sizes at the individual rock-category level, we have reaggregated the results for the "both" items into the coverage items and ss-optimal items in these plots.) In addition, Fig. 4 shows the results for the random-3 and random-9 conditions, with the blue bars showing performance on the old training instances and the green bars on the novel transfer items.

As can be seen from inspection of Figs. 3 and 4, the general pattern of results reported in our summary-table analyses also tend to be seen within each of the individual rock categories. Specifically, across all four conditions, performance is consistently better on the training instances than on the novel transfer items. (Recall that in the coverage condition, the coverage items are training items and the ss-optimal items are novel transfer items; whereas the reverse holds in the ss-optimal condition.) In addition, there is a great deal of variability in overall performance levels across the different rock categories. Participants are extremely accurate, for example, in classifying members of the categories *obsidian* and *pumice*; moderately accurate in classifying *pegmatite* and *basalt*; and relatively inaccurate in classifying *gabbro* and *rhyolite*. (Performance levels for the remaining rock categories tend to be intermediate and vary somewhat across the different conditions.) A challenge for the formal model of classification learning is whether it can simultaneously characterize the patterns of training effects across the conditions as well as the varying performance levels observed across the different categories of rocks.

Theoretical analysis

Fitting the item-type accuracies across conditions and categories

We now turn to our central goal of considering the ability of the GCM to account for the observed classification data across the four conditions of testing. Although a target for future research is to account for performance at the level of individual participants (Lee & Pope, 2003; Nosofsky, 1986; Okada & Lee, 2016; Shen & Palmeri, 2016), at this early stage of the project in this complex domain we limit consideration to the major patterns seen in the averaged data.

We started the analysis by fitting a baseline version of the model to the proportion-correct data for the different item types shown in Figs. 3 and 4. Based on consideration of the summary results reported in the previous section, we know in advance that the baseline version of the model will have shortcomings, because its prediction that overall proportion correct in the ss-optimal condition would be greater than in the coverage condition was not confirmed. Nevertheless, by analyzing the baseline model's fit to the item-type data across the four conditions, we may gain deeper insights into where the model has its strengths and weaknesses.

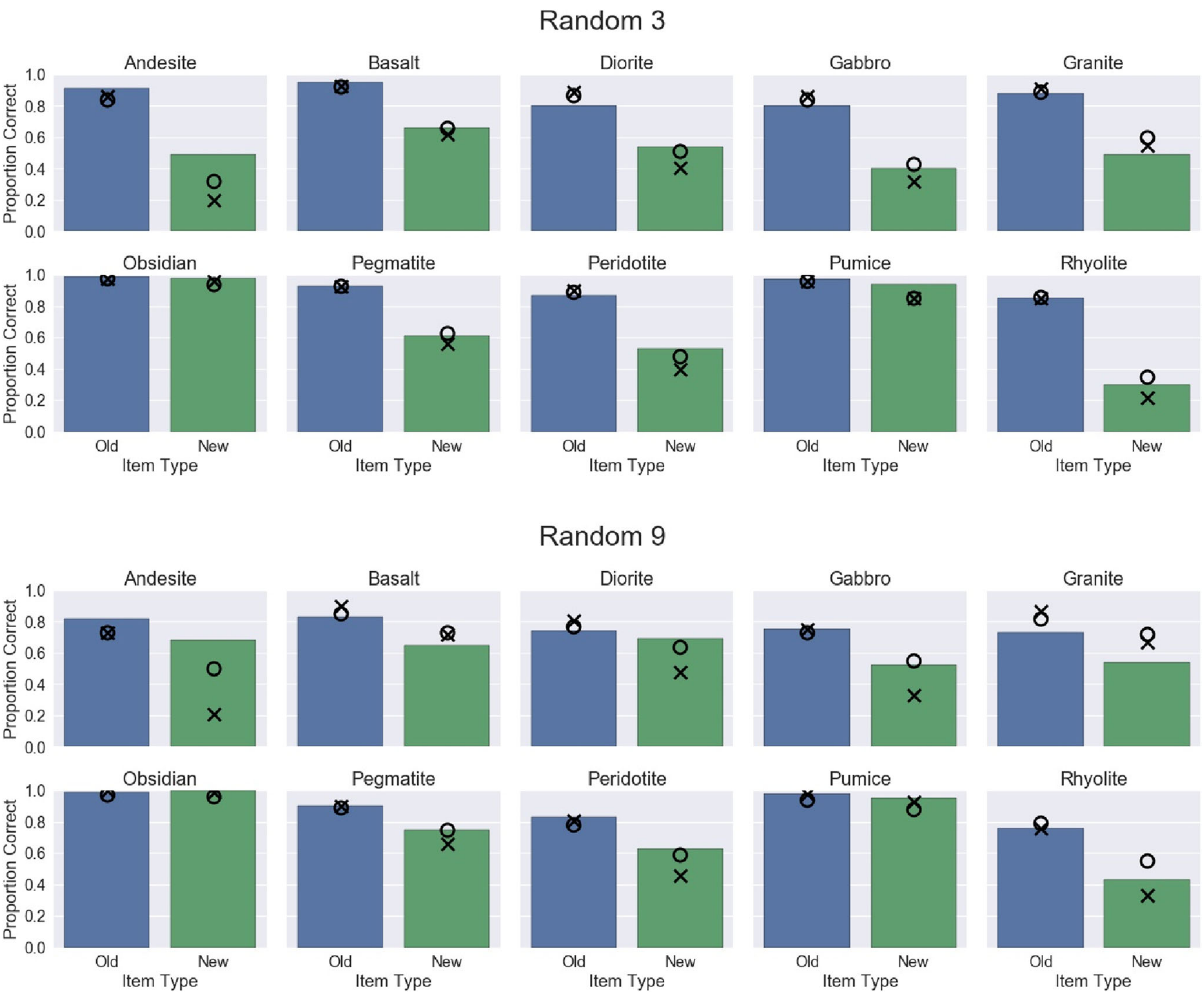

Fig. 4 Mean accuracy for each of the item types in each of the 10 categories in the random-3 and random-9 conditions. Colored bars = observed data (blue = old training items, green = new transfer items). Xs denote predictions from the baseline version of the GCM; open circles denote predictions from the baseline + c_w version of the GCM. (Color figure online)

We conducted computer searches to locate the values of the two free parameters in the model (c and p_{store}) that minimized the sum of squared deviations (SSD) between the predicted and observed item-type classification probabilities, combined across all conditions. Because of the probabilistic-storage assumption, fitting the present version of the model required the use of computer simulation. In particular, 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 1,000 individual simulation runs. Note that in the case of the coverage and ss-optimal conditions, the stored training instances for each simulation were always randomly generated from fixed parent sets (because all participants experienced the same training instances in these two conditions). However, in the case of the random-3 and random-9 conditions, each participant experienced a unique set of training instances. In conducting the simulations for these two conditions, the stored instances were randomly generated from the precise sets of training instances experienced by each individual participant.[4]

In conducting the fits, we held fixed the values of c and p_{store} across the coverage, ss-optimal and random-3 conditions; however, as we have discussed previously, because of the major changes in instance-training conditions, we estimated separate values of c and p_{store} for the random-9 condition.

[4] Extremely similar predictions were obtained, however, if the parent sets were chosen randomly anew for each individual simulation.

Finally, recall that in our Fig. 3 we aggregated the data for the "both" stimuli into the results of the coverage-training and ss-optimal-training items. So as not to count the "both" stimuli twice, in conducting the current fits, we separated out the "both" stimuli from the coverage-item and ss-optimal-item results, and required the model to fit each of the four item types separately (i.e., coverage, ss-optimal, neither, both). However, in displaying the model-fitting results, we again aggregate the predictions for the "both" stimuli back into the coverage and ss-optimal items (in order to aid in visual inspection).

The baseline model's item-type predictions for each individual category across the four conditions are indicated by the Xs in Figs. 3 and 4. The summary fits of the baseline model to the data from each of the individual conditions, as well as to the data combined across the four conditions, are reported in Table 5 (top row of each panel of table). As a further source of assessing the model's strengths and weaknesses, we show in Fig. 5 the observed and predicted data for the item types in each condition after averaging across all 10 categories.

Inspection of the figures and tables suggests that, at least as a first approximation, the baseline model does reasonably well at accounting for the complete set of data (with some important exceptions to be discussed below). For example, in all conditions, it predicts correctly that the training instances were classified with significantly higher accuracy than were the novel transfer items (see Figs. 3, 4, and 5). The old-item advantage arises because old test items provide a perfect match to their representations in memory, boosting their summed similarity to the correct target category. In addition, across all conditions, the model accounts well for the varying performance levels associated with the different categories (see Figs. 3 and 4): For example, the model predicts the extremely high classification accuracy for obsidian and pumice; the (in general) moderately high accuracy for basalt and pegmatite; and the low accuracy for gabbro and rhyolite. As documented in detail in the previous Nosofsky et al. (2018b) study, the basis for these successful individual-category predictions involves the structure of the rock categories and their positions in the multidimensional similarity space. In the obsidian and pumice categories, the individual-category members are tightly clustered; their compact structures yield high within-category similarity. Furthermore, for the present set of igneous rocks, obsidian and pumice lie in isolated regions of the similarity space, so there are few between-category confusions. By contrast, rhyolite is a highly dispersed category, with low within-category similarity (for an illustration involving pictures of the members of the rhyolite category, see Fig. 3 from Nosofsky et al., 2018b). Furthermore, both rhyolite and gabbro lie in much denser regions of the MDS space, so have high between-category similarity.

Another success for the baseline model, at least at a qualitative level, is that it accounts for the interaction between conditions and item type in the random-3 and random-9 conditions (see Fig. 5): Performance is better for the old items in the random-3 condition than in the random-9 condition, but is better for the new items in the random-9 condition than in the random-3 condition. The model accounts for this interaction because the estimate of the p_{store} parameter was higher in the random-3 condition than in the random-9 condition (we defer more detailed discussion of the best-fitting parameters until later in our article). This result seems sensible, because there were far more presentations of the individual old-training instances in the random-3 condition than in the random-9

Table 5 Fits of different versions of the GCM to the item-type data in Figs. 3 and 4

Model	# Free parms.	Coverage	ss-Optimal	Random-3	Random-9	Combined
Condition						
a Fit measured in terms of sum of squared deviations						
Baseline	4	0.376	0.387	0.161	0.408	1.333
Baseline + γ	6	0.392	0.372	0.150	0.381	1.294
Free parameters across all conditions	12	0.350	0.322	0.129	0.381	1.182
Mixed exemplar plus prototype	10	0.385	0.372	0.148	0.241	1.146
Baseline + c_w	6	0.201	0.339	0.072	0.121	0.732
Constrained Baseline + c_w	4	0.192	0.369	0.073	0.187	0.821
b Fit measured in terms of percentage variance accounted for						
Baseline	4	70.2	76.6	82.8	15.7	69.5
Baseline + γ	6	69.0	77.6	84.0	21.3	70.3
Free parameters across all conditions	12	72.3	80.6	86.3	21.3	72.9
Mixed exemplar plus prototype	10	69.5	77.5	84.2	50.2	73.8
Baseline + c_w	6	84.1	79.6	92.3	75.1	83.2
Constrained baseline + c_w	4	84.8	77.7	92.2	61.5	81.2

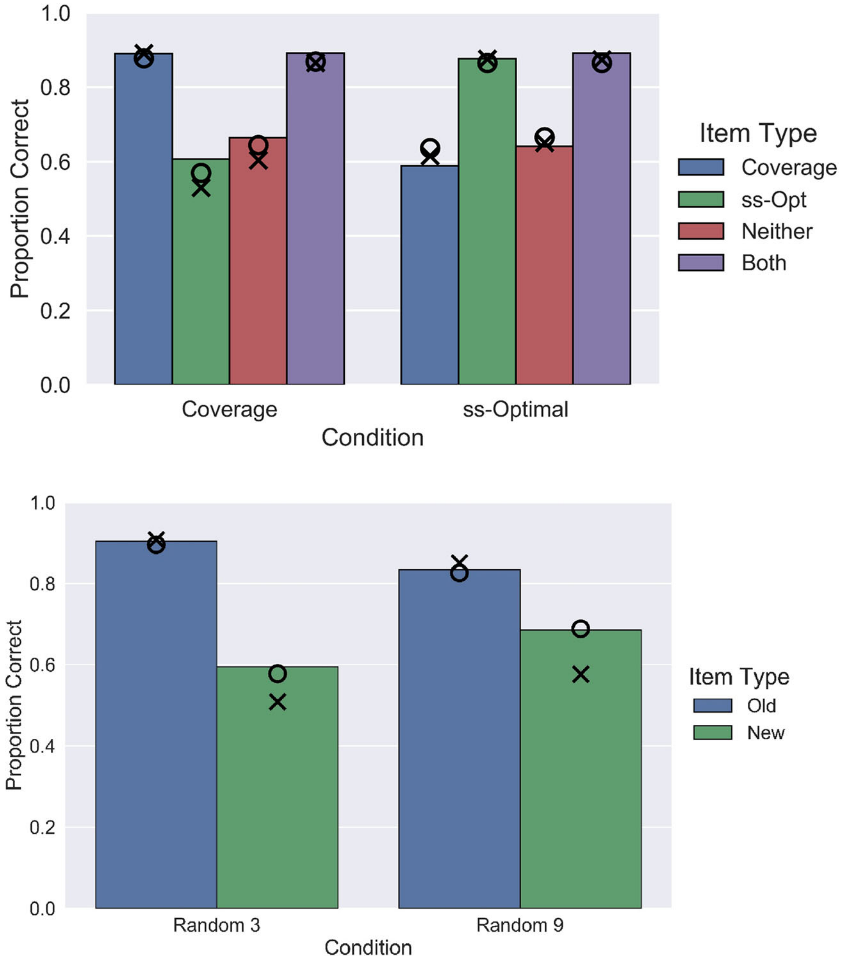

Fig. 5 Mean proportion correct for each of the item types, aggregated across all 10 categories, in all four conditions. Top panel: Coverage and ss-optimal conditions. Colored bars = observed data (blue = coverage-training items, green = ss-optimal-training items, red = neither items, purple = both items). Xs denote predictions from the baseline version of the GCM; open circles denote predictions from the baseline + c_w version of the GCM. Bottom panel: Random-3 and random-9 conditions. Colored bars = observed data (blue = old training items, green = new transfer items). Xs denote predictions from the baseline version of the GCM; open circles denote predictions from the baseline + c_w version of the GCM. (Color figure online)

condition. Despite the lower p_{store} value in the random-9 condition, generalization to new items is better in the random-9 condition than in the random-3 condition, because the random-9 training instances tend to provide better coverage of the complete category distributions than do the random-3 training instances.

As reported in the bottom panel of Table 5, averaged across the coverage, ss-optimal, and random-3 conditions, the model accounts for 76.5% of the variance in the item-type classification accuracies of the 10 categories. In our view, this result is respectable given that a very low-parameter model is being used to fit a large number of data entries in a complex, naturalistic domain. However, the model accounts for only 15.7% of the variance in the random-9 condition. There are three reasons for the poor fit to the random-9 data. First, unfortunately, the baseline-model fits to the random-9 condition include a single extreme misprediction involving new items from the andesite category (see Fig. 4, top-left panel of random-9 condition). If this single data point is removed from consideration, then the percentage of variance accounted for increases from 15.7% to 60.8%. Second, there is far less total variability in the data in the random-9 condition than in the other three conditions. The reason is that there is a much smaller difference in correct classification probabilities for old versus new items in that condition (see Table 4 and Fig. 5). Third, as can be seen from inspection of Figs. 4 and 5, the model systematically underpredicts the proportion of correct responses for new transfer items in this condition. (In fact, this same problem also tends to be seen in the coverage and random-3 conditions.) This result provides an initial clue about the locus of the shortcomings of the present application of the model. We expand considerably upon this issue in the next main section.

Finally, in the middle row of Table 2, we report the baseline model's predictions of overall proportion correct in each condition, averaged across all test items of all categories. As anticipated, although the model correctly predicts the ordering of performance for the random-9, coverage, and random-3 conditions, it is incorrect in its prediction that performance in the ss-optimal condition would be better than in the coverage condition. In addition, in struggling to account simultaneously for performance in the coverage, ss-optimal, and random-3 conditions (with parameters held fixed across those conditions), the model is slightly off quantitatively in its predictions of overall proportion correct for the coverage and random-3 conditions.

Before proceeding to the next major section of our theoretical analyses, we should note that we investigated various standard extensions of the baseline model in an attempt to improve its fits to the data. In one extension, for example, we added a response-scaling parameter to the model (Ashby & Maddox, 1993; McKinley & Nosofsky, 1995). In this version, each of the individual category-summed similarities in the Equation 1 decision rule is raised to the power γ. In the baseline model, with $\gamma = 1$, the observer "probability matches" to the relative summed-similarity of each category; whereas as γ grows larger than 1, the observer responds more deterministically with whichever category yields the largest summed similarity. As reported in Table 5, however, adding the response-scaling parameter to the baseline model left the fits virtually unchanged. In a further extension, rather than constraining the c, p_{store}, and γ parameters to be equal across the coverage, ss-optimal, and random-3 conditions, we allowed separate values of these parameters for each individual condition. As reported in Table 5, however, the improvements in fit yielded by this "free parameters across all conditions" model were relatively minor. Furthermore, as we will report in the next main section, adding category response-bias parameters to the Equation 1a decision rule as well as "attention-weight" parameters to the Equation 2a distance function (see Nosofsky, 1986, 1987, 2011) also resulted in only relatively minor improvements in fit to detailed forms of the classification data. In still another analysis, we extended the model by assuming that, with some probability, observers based classification decisions on the similarity of test items to category prototypes rather than to stored exemplars (see Appendix B for a description of the mixed exemplar-plus-prototype model). As reported in Table 5, despite the large number of free parameters granted to this mixed model, it still led to relatively minor improvements in overall fit. We now turn to an alternative set of theoretical analyses that did reveal a major locus of the model's shortcomings.

Fitting the data at the level of individual rocks

In our previous modeling analyses, as well as the major ones reported in Nosofsky et al. (2018b) study, the goal was to account for data aggregated at the level of "types" of rock instances broken down by their category membership and training versus transfer status. As argued by Nosofsky et al.'s (2018b), this goal seemed a reasonable starting one as we ventured into formal modeling in this complex natural-category domain. However, in view of the shortcomings of the model reported in our previous section, here we decided that greater insight might be achieved if we considered the model's predictions at the level of individual rock *tokens* rather than broad types of rocks. Specifically, focusing first on the coverage condition, we decided to fit the model to the complete classification-confusion matrix defined by the conditional probability with which participants classified each of the *individual* 120 rocks into each of the 10 categories during the test phase. Note that this classification-confusion matrix is composed of 1,200 data points (120 rows, one row per rock instance; by 10 columns, one per each category; however, because the conditional probabilities within each row sum to 1, there are only 1,080 data points that are truly free to vary). Also, as detailed below, because the entries in the confusion matrix vary considerably in magnitude, and therefore have widely varying error variance associated with them, we decided to use maximum-likelihood as a criterion of fit rather than minimization of a sum-of-squared-deviations criterion.

Given the very large number of data points and the complexity of the domain, it no longer seemed realistic to attempt fits with the extremely low-parameter baseline model. Instead, we decided to fit an elaborated version of the GCM with additional free parameters. First, we extended the Equation 1a decision rule with the response-scaling parameter (described above), a guessing parameter, and a set of category response-bias parameters. In this elaborated model, it is assumed that, on each trial, the observer guesses randomly among the 10 categories with probability g and bases his or her classification decisions on stored exemplar information with probability $1 - g$. (Although pure guessing likely occurs with very low probability under the present experimental conditions, the parameter may be needed to account for extremely small-magnitude but nonzero entries that are occasionally observed in the matrix.) Likewise, the category response-bias parameters are commonly used in fitting formal models to detailed stimulus-response confusion matrices (e.g., Luce, 1963); they reflect biases or preferences for responding with alternative categories independent of the specific test stimuli that are presented. Taken together, in this "full" GCM, the probability that item i is classified in Category J is given by

$$P(J|i) = (1-g)\frac{b_J\left(\sum_{j\in J} s_{ij}\right)^{\gamma}}{\sum_K b_K\left(\sum_{k\in K} s_{ik}\right)^{\gamma}} + \frac{g}{10}, \tag{1b}$$

where γ is the response-scaling parameter; g $(0 < g < 1)$ is the guessing probability; and b_J $(0 < b_J < 1,\ \sum b_J = 1)$ is the response bias associated with Category J.

Second, as is traditional in fitting the GCM to detailed classification-confusion data, the Equation 2a distance function was extended with a set of dimension-weight parameters:

$$d_{ij} = \left[\sum w_m \cdot |x_{im} - x_{jm}|^2\right]^{1/2}, \tag{2b}$$

where w_m ($0 < w_m < 1$, $\sum w_m = 1$) is the weight assigned to Dimension m. The weights describe the degree of "attention" that observers give to alternative dimensions in making their classification judgments. In cases in which some subset of dimensions is relevant for classification, with other dimensions being irrelevant, the attention weights often play a dramatic role in allowing the model to account for classification performance. In the present case, however, all dimensions tend to contribute useful information in allowing observers to classify the objects into the 10 different rock categories, so we expected the dimension-weight parameters to play a less dramatic role.

The present version of the model uses 20 free parameters for predicting the 1,200-cell classification-confusion matrix: the overall sensitivity parameter c and exemplar-storage parameter p_{store} (described previously); the response-scaling parameter γ and guessing-parameter g (Equation 1b); nine freely varying response-bias parameters b_J (Equation 1b); and seven freely varying dimension-weight parameters w_m (Equation 2b).

As noted above, we used a maximum-likelihood criterion in evaluating the model's fit to the classification-confusion matrix data. We used the Hooke and Jeeves (1961) computer algorithm to search for the best-fitting parameters; in an attempt to avoid local minima, we used 10 different random starting-parameter configurations in conducting the searches. Recall that because of the probabilistic-storage assumption, the predictions associated with any specific set of candidate parameters were derived across 1,000 simulations of the model. Finally, to evaluate the fit of different versions of the model with varying numbers of free parameters (see below), we used the BIC statistic, which penalizes a model for the number of free parameters it uses:

$$\mathrm{BIC} = -2\ln L + P\ln(N),$$

where L is the (maximum)-likelihood fit of the model, P is the number of free parameters, and N is the sample size upon which the fit is based. Smaller values of BIC indicate a better fit.

The key results of interest are shown in the top-left panel of Fig. 6, which plots, for the coverage condition of the present experiment, the 1,200 observed classification probabilities against the maximum-likelihood predicted ones. In this plot, all data points associated with *correct* classification probabilities are indicated by geometric forms, Xs, or asterisks; by contrast, all data points associated with *incorrect* classification probabilities are indicated by the small-size dots. The category membership associated with the correct-classification data points can be decoded through use of the figure legend and caption. So, for example, the open diamond toward the upper left of the scatter plot is showing a case in which a specific member of the *gabbro* category was correctly classified with probability .85, but in which the model predicted that it would be correctly classified into its category with probability only .25.

Given the extremely large number of data points and the complexity of the domain, we did not find it surprising to observe a good deal of scatter in the plot. What we did find to be highly instructive, however, was the systematic nature of the deviations from the perfect-prediction line: As can be seen, when there are large deviations for the correct-classification probabilities, they lie predominantly above the perfect-prediction line; whereas any large deviations for the incorrect-classification probabilities lie predominantly below the perfect-prediction line. In short, the model systematically underestimates the participants' ability to correctly classify many of the individual rocks.

A straightforward explanation for this tendency is that the current version of the MDS solution being used by the GCM tends to underestimate the degree of within-category similarity among the rock stimuli. (If within-category similarity were higher, then the summed similarity of test items to the members of their target categories would increase, boosting the correct-classification probabilities.)

In retrospect, this occurrence seems to have been almost inevitable, for multiple reasons. A first major reason is one that was anticipated by Nosofsky et al. (2018) in the context of their study that first attempted to derive the high-dimensional feature space for the rock stimuli. Although one major method for deriving the feature space involved the MDS analysis of pairwise similarity judgments of the rocks, Nosofsky et al. (2018, p. 531) noted that it was likely that various dimensions that might be critical in the context of a classification-learning task might not be highly salient in the context of a similarity-judgment task. For example, if the members of a given category of rocks shared a subtle but highly diagnostic feature that tended not to be present in the contrast categories, then such a feature would be likely to take on a great deal of salience in the context of the classification-learning task (Nosofsky, 1986). Yet, due to its subtle nature, it would have little influence on subjects' judgments in the generic pairwise similarity-judgment task, so it would be unlikely for that critical feature to be represented in the scaling solution derived from the similarity judgments. Thus, any model that made reference to the scaling solution to generate predictions in an independently conducted classification-learning task would then be at a severe disadvantage. As an approach to dealing with this issue, Nosofsky et al. (2018) collected direct dimension ratings along a large number of candidate dimensions for characterizing the rock stimuli and suggested the possibility that

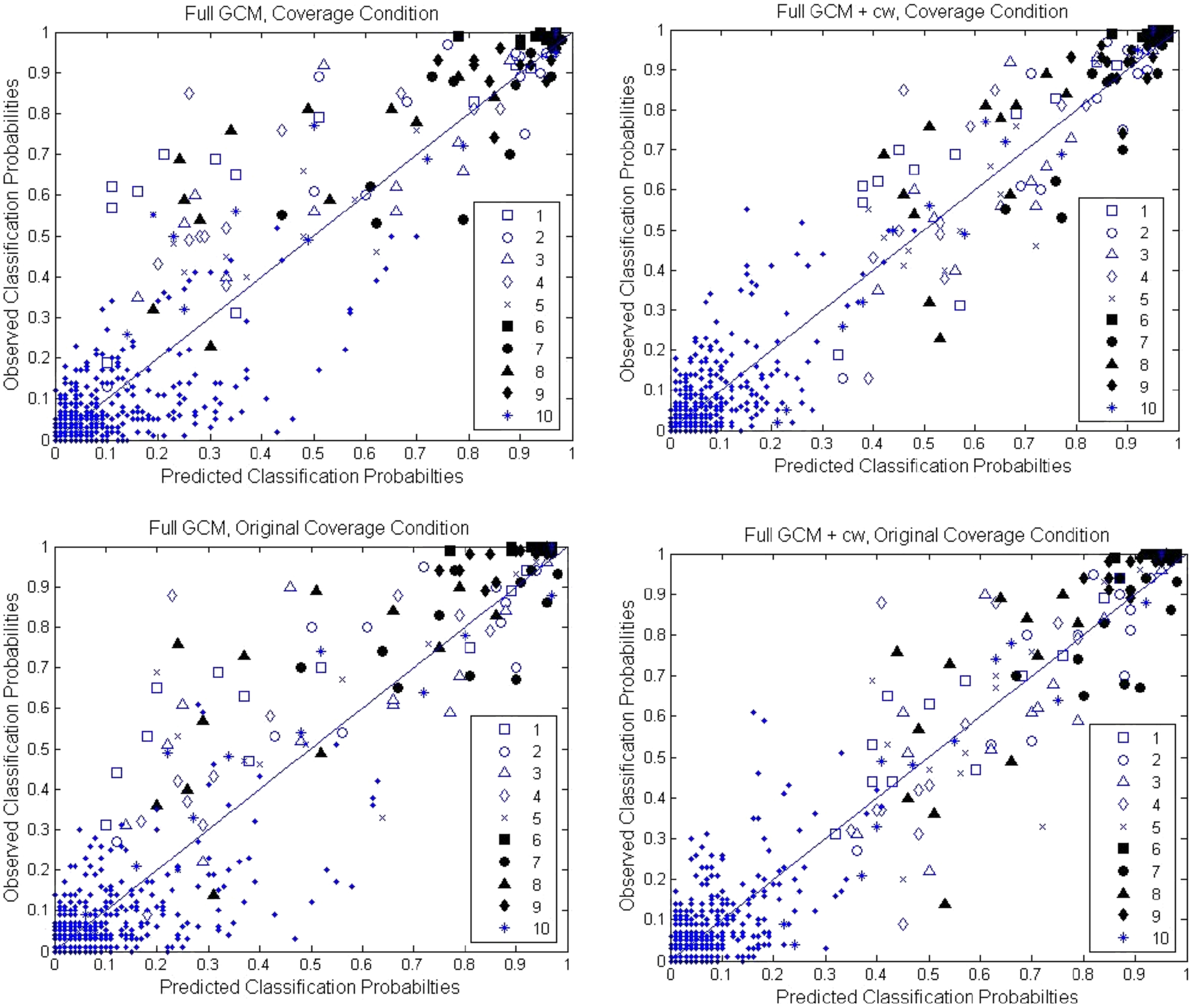

Fig. 6 Observed probability with which each individual rock instance was classified into each of the 10 categories plotted against the predicted probabilities from the GCM. Top left panel = coverage condition, full GCM; top right panel = coverage condition, full GCM + c_w; bottom left panel = original coverage condition, full GCM; bottom right panel = original coverage condition, full GCM + c_w. *Note.* 1 = andesite, 2 = basalt, 3 = diorite, 4 = gabbro, 5 = granite, 6 = obsidian, 7 = pegmatite, 8 = peridotite, 9 = pumice, 10 = rhyolite

the MDS solution could be extended by including some of these directly rated dimensions. However, it is impossible to anticipate all such potentially relevant dimensions in advance. Later in our article, we will provide what we believe is a clear-cut example of this problem in the context of the present classification-learning data.

A second major reason why the present scaling solution likely underestimates the degree of within-category similarity among the rock stimuli has to do with noise in the scaling solution.

Our argument is as follows: In the "true" psychological space in which the rocks are embedded, it is virtually certain that within-category similarities among the objects tend to be greater, on average, than between-category similarities. Now, suppose that one adds random noise to the locations of the objects in a configuration with this property. It should be intuitively clear to the reader that adding random noise will tend to decrease the relative degree of within-category similarity compared with between-category similarity. (We report simulation work in Appendix C to support this intuition.) Indeed, in the limit, as the amount of random noise that is added dominates the true locations, there would eventually be no difference in the relative degree of within-category versus between-category similarity among the objects.

Moreover, as we alerted the reader at the outset, the similarity-judgment data used for deriving the current high-dimensional scaling solution were in fact noisy at the individual-cell level: There was an average of only 1.82 similarity judgments per individual pair of rocks in the 360 × 360 matrix. Our claim has been that the MDS solution provides a reasonable first-order, global account of the dimensional structure of the objects and of the average similarity between members of the different pairs of categories. The current analysis, however, is the first to consider the precision of the predictions at the level of *individual* items.

In sum, the above considerations suggest the need to develop both a more *comprehensive* scaling solution for the rocks (that includes more dimensions of potential relevance to classification) as well as a scaling solution that is more *precise* in terms of the dimensions that are already represented. At the current juncture, this goal of deriving a comprehensive, high-precision, high-dimensional scaling solution for the hundreds of objects in our natural-category set must be viewed as an extremely long-range goal—we suggest routes to achieving that long-range goal in our General Discussion. In the meantime, however, we propose that important progress can be made with respect to the current modeling by applying a "repair" to the MDS computations. Our proposed repair is to extend the Equation 3a exponential-similarity computation by adding a within-category sensitivity parameter (c_w) to the model, viz.

$$s_{ij} = \begin{cases} \exp\left(-c \cdot d_{ij}\right), & \text{if } i \text{ and } j \text{ belong to different categories} \\ \exp\left(-c_w \cdot d_{ij}\right), & \text{if } i \text{ and } j \text{ belong to the same category} \end{cases} . \quad (3b)$$

Assuming that $c_w < c$, this computation will have the effect of "squeezing together" the locations of items belonging within the same category. As will be seen, it is still the case that the current MDS solution for the rocks is doing an enormous amount of work in enabling the predictions; the within-category sensitivity parameter is simply repairing the solution to enable even better predictions in this complex domain.

It is crucial to understand that the psychological claim here is not that it is the *observer* who applies a different sensitivity parameter to within-category versus between-category pairs. Such a claim would be tantamount to supposing a form of ESP that governs observers' classification judgments (the observers would need to have knowledge of the objective category membership of the individual objects in advance of making their classification judgments). Instead, the use of the within-category sensitivity parameter should be viewed as providing a mathematical repair to the independently-derived MDS solution to which the formal classification model makes reference.

The predictions from this extended GCM + c_w model for the coverage condition are shown next to the standard model's predictions in the upper-right panel of Fig. 6. It should be obvious from inspection that adding the single free parameter yields an enormous improvement in quantitative fit. Indeed, not only do the vast majority of data points come closer to the perfect-prediction line than for the standard model, but now the systematic deviations involving correct versus incorrect classification probabilities have all but disappeared. This visual impression is confirmed by the resulting fit indices for the two models, which are reported in the top panel of Table 6: Adding the single parameter improves the previous fit by more than 1,500 BIC points. To bring out the crucial role of the c_w parameter in another way, we also fitted a highly constrained five-parameter version of the GCM to the complete matrix of individual-item classification-confusion data. In this highly constrained version, we set all response-bias parameters equal to one another and assumed equal attention weights for all dimensions. Thus, the model made use of only the parameters c, p_{store}, g, γ, and c_w. Note that with only five free parameters, this constrained GCM + c_w model yields a better *absolute* fit (i.e., log-likelihood fit) than does the 20-parameter standard version of the model without c_w (see Table 6).[5]

To test the generality of these model-fitting results, we went back to the data from the original coverage condition conducted in Experiment 1 of Nosofsky et al. (2018b). (Recall that our present coverage condition was a replication of that original condition with a somewhat different population of participants. In the original study, we had fitted only the item-*type* data rather than the classification probabilities for the individual 120 rock instances.) The model predictions for the individual-item classification probabilities from the original coverage condition are shown in the bottom panels of Fig. 6—again, the left panel shows the predictions from the standard model, and the right panel the predictions from the GCM + c_w model. The detailed fit indices associated with the models are reported in the middle panel of Table 6. As can be seen from inspection, the correspondence in the pattern of model-fitting results across the present coverage condition and the originally tested one provides a remarkable replication.

To obtain yet additional evidence, we conducted the same sets of model-fitting analyses on the data from the ss-optimal condition from the present experiment. The results are shown in Fig. 7 and in the bottom panel of Table 6. Although not as dramatic as the results for the two coverage conditions, the same pattern of model-fitting results is observed for the ss-optimal condition. Thus, there is considerable support for the

[5] We also conducted explorations of another class of models for repairing the misestimated similarities from the MDS solution. In this alternative class, rather than adjusting the measured similarities based on whether objects belonged to the same or different categories, the similarities were adjusted on the basis of their computed distances in the MDS solution. The general idea is that because of noise in the derived scaling solution, there would be a tendency for items separated by small distances in the derived scaling solution to have even smaller "true" distances than MDS-derived distances. We provide a report of those alternative modeling explorations in a supplement to this article. In a nutshell, although these alternatives sometimes led to improvements in fit compared to the standard GCM, the improvements came nowhere close to the one achieved by the GCM + c_w model reported here. This result is perhaps not too surprising, because these alternative adjusted-distance models take no account of the problem that subtle features that become salient during classification learning may not have been represented in the initial similarity-judgment-derived MDS solution.

Table 6 Fits of different versions of the GCM to the complete individual-rock classification-confusion matrices observed in the coverage and ss-optimal conditions of the present experiment, and in the original coverage condition tested in Experiment 1 of Nosofsky, Sanders, and McDaniel (2018b)

Model	# Free parms.	Neg ln$L$	BIC	% Var
Coverage condition of present experiment				
Full GCM	20	13,863.3	27,916.5	87.1
Full GCM − c_w	**21**	**13,099.8**	**26,399.0**	**93.0**
Constrained GCM − c_w	5	13,493.0	27,033.5	92.1
Discrete-similarity GCM	14	15,588.4	31,309.8	85.2
Coverage condition tested in Nosofsky et al. (2018b)				
Full GCM	20	12,500.4	25,188.5	86.8
Full GCM − c_w	**21**	**11,871.4**	**23,939.8**	**92.3**
Constrained GCM − c_w	5	12,190.1	24,427.1	91.3
Discrete-similarity GCM	14	13,899.3	27,930.0	85.0
ss-optimal condition of present experiment				
Full GCM	20	15,505.6	31,203.2	88.5
Full GCM − c_w	**21**	**15,058.6**	**30,318.8**	**91.9**
Constrained GCM − c_w	5	15,515.9	31,079.8	89.5
Discrete-similarity GCM	14	17,842.3	35,819.2	83.7

Note. Parms. = parameters; Neg ln$L$ = negative ln-likelihood; BIC = Bayesian information criterion; % Var = percentage of variance accounted for. Boldface entries in each panel indicate results for the best-fitting model

hypothesis that the current MDS solution for the rock stimuli tends to underestimate the relative degree of within-to-between-category similarity, and that repairing the model through use of the within-category sensitivity parameter leads to significantly improved accounts of the rock-categorization data.

It is important to understand that the c_w parameter is acting to "repair" the MDS solution rather than "replacing" the solution. It is still the case that the MDS solution itself is providing a fundamental bedrock for enabling the successful predictions. To demonstrate this point, we fitted a "discrete-similarity" version of the GCM to the rock-classification data that made no use of the MDS solution. The discrete-similarity version made allowance for only three levels of interexemplar similarity: The self-similarity between an exemplar and its own representation in memory was set at unity; the similarity between distinct exemplars belonging to the same category ("within-category" similarity) was set at s_w; and the similarity between exemplars belonging to different categories ("between-category" similarity) was set at s_b. To give this discrete-similarity model maximum flexibility, we continued to estimate the parameters p_{store}, γ, g, and the set of response-bias parameters from the standard GCM. (Because the MDS solution was not used, no attention-weight parameters were estimated.) As shown in each of the panels of Table 6, despite using substantially more free parameters than the highly constrained five-parameter version of the GCM + c_w model, this discrete-similarity model fared dramatically worse at fitting the data in all cases (across the three conditions, the fit deteriorated by an average of 4,172 BIC points). In a nutshell, the continuous gradations in similarity between exemplars that are measured by the MDS solution are fundamental to allowing the GCM to predict the rock-classification data—they simply need to be adjusted to account for systematic underestimation of within-category similarities in the current solution.

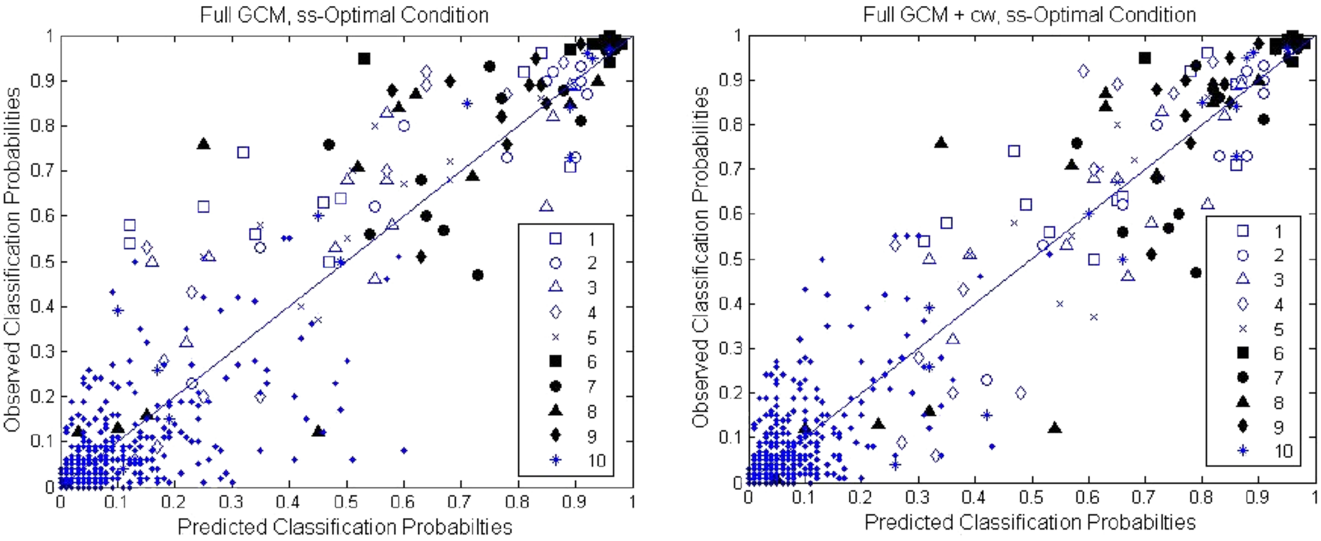

Fig. 7 Observed probability with which each individual rock instance was classified into each of the 10 categories plotted against the predicted probabilities from the GCM. Left panel = ss-optimal condition, full GCM; right panel = ss-optimal condition, full GCM + c_w. *Note.* 1 = andesite, 2 = basalt, 3 = diorite, 4 = gabbro, 5 = granite, 6 = obsidian, 7 = pegmatite, 8 = peridotite, 9 = pumice, 10 = rhyolite. (Color figure online)

Applying the baseline + c_w model to the item-type classification data

Having discovered the above-described limitation associated with the MDS solution, our next natural step is to circulate back to the original item-type data (see Figs. 3, 4, and 5) and apply the extended GCM + c_w model there as well. As in the original analyses, the goal is to characterize performance across all four conditions (coverage, ss-optimal, random-3, random-9) with a minimum of parameter estimation. Therefore, we return to applying the baseline version of the model (Equations 1a and 2a), except we extend it with the within-category sensitivity parameter c_w (Equation 3b). As in the original analyses, we hold the values of the free parameters (c, p_{store}, c_w) fixed across the coverage, ss-optimal and random-3 conditions, but estimate separate parameter values for the random-9 condition. We conducted computer searches for the values of the free parameters that minimized the SSD between the predicted and observed item-type probabilities computed across all 10 categories of the four conditions.

The predicted item-type classification probabilities from the above-described baseline + c_w version of the GCM are shown as open circles in Figs. 3 and 4. In addition, the predictions averaged across the 10 categories are shown as open circles in Fig. 5. The summary fits from the model are reported in Table 5. As can be seen in the table, there is dramatic overall improvement in the quantitative predictions compared to the fits yielded by the baseline model without the c_w parameter, with the total SSD reduced from 1.333 to 0.732. The dramatic improvements in fit arise in three of the four conditions (albeit with little change for the ss-optimal condition). Indeed, as can be seen from inspection of Table 5, this six-parameter baseline + c_w model yields an appreciably better SSD than did the 12-parameter version of the model without the c_w parameter.

Inspection of Fig. 5 reveals that the model pinpoints the item-type accuracies averaged across the 10 categories in all four conditions. Furthermore, inspection of Figs. 3 and 4 suggests very good fits to the individual-category results of all four conditions as well. We report in the bottom row of Table 2 the predicted overall proportion of correct classifications in each condition, averaged across all the test items—the initial analysis that motivated our investigation. Comparing the observed data in the top row of the table, one sees that although the model still incorrectly predicts a slight advantage for the ss-optimal condition compared with the coverage condition, the quantitative deviations are very small.

The best-fitting parameters from this baseline + c_w version of the GCM are reported in Table 7. As expected, in all conditions, the magnitude of the c_w parameter is less than that of the c parameter, reflecting the needed repair to the underestimated within-category similarities in the MDS solution. In addition, as expected, the estimate of the p_{store} parameter is lower in the random-9 condition than in the other conditions, reflecting the reduced number of presentations of individual training instances in the random-9 condition. We did not have strong prior hypotheses regarding how the magnitude of the c and c_w parameters might compare across the small training-set-size conditions and the random-9 condition; the current estimates suggest lower sensitivity in the random-9 condition. A reasonable explanation is that because individual instances were presented with much higher frequency in the small training-set-size conditions than in the random-9 condition, the memory representations associated with the training instances were more highly differentiated in the small-set-size conditions (Kılıç, Criss, Malmberg, & Shiffrin, 2017; Nosofsky, 1987; Shiffrin, Ratcliff, & Clark, 1990; Shiffrin & Steyvers, 1997). Nevertheless, we should note that the fit of the model is not very much worse if the values of c and c_w are constrained to be equal across the small-set-size conditions and the random-9 condition: As reported in Table 5, even this highly constrained version of the baseline + c_w model (which estimates only four free parameters) yields an appreciably better fit to the data than do any of the alternatives without the c_w parameter. In short, future research is needed before strong conclusions can be drawn regarding how overall sensitivity may vary with manipulations of the number and distribution of category training instances.

Table 7 Best-fitting parameters for the baseline + c_w version of the GCM fitted to the item-type data of Figs. 3 and 4

Condition	c	p_{store}	c_w
Parameters			
Coverage/ss-Optimal/Random-3	0.980	0.827	0.814
Random-9	0.777	0.662	0.498

General discussion

Summary

Our attempt to use the formal exemplar model to search for optimal training examples for teaching the present rock categories met with some successes and failures. On the positive side, the model predicted correctly: the observed ordering of overall performance across the random-9, coverage and random-3 conditions; the general patterns of item-type performance across all four conditions; and, for the most part, the difficulty levels of the 10 different categories across the four conditions. On the negative side, the model's prediction that overall proportion correct in the ss-optimal condition would exceed performance in the coverage condition was not confirmed. In addition, in more detailed quantitative tests, we discovered systematic shortcomings in which the model underpredicted correct classification probabilities associated with many of the individual rock instances in the coverage and ss-optimal conditions.

We conducted a variety of follow-up analyses to investigate the detailed basis for the failed predictions reviewed above. These analyses led us to the hypothesis that the current version of the MDS solution to which the exemplar model makes reference may tend to systematically underestimate within-category similarity relations among the rock exemplars. Indeed, we argued that, in retrospect, such an occurrence was perhaps inevitable when attempting to develop comprehensive, high-precision scaling solutions for objects in complex, high-dimensional naturalistic category domains. To remedy this problem, we developed an extended version of the model with an additional free parameter for repairing the underestimated within-category similarities. Application of the extended model led not only to vastly improved predictions of the classification probabilities associated with individual items in the coverage and ss-optimal conditions, but also to a very good quantitative account of the patterns of item-type performance across all four conditions.

Although the overarching project remains a work in progress, we believe that the achievements reached thus far are nevertheless highly significant and impressive. The upshot is that a model that requires the estimation of relatively few free parameters is providing very good overall quantitative accounts of performance in an exceedingly complex, naturalistic category domain. Thus, we continue to be encouraged that we are heading down a fruitful path. Furthermore, our present theoretical and empirical results point in clear-cut directions for taking next steps in the project; in addition, they already provide some strong recommendations for the strategies of choosing training examples that are the most likely to foster accurate generalization in natural-science classification. In the remainder of our General Discussion, we discuss in turn both of these points.

Improving the high-dimensional scaling solution

Having discovered a key shortcoming in the original model-based machinery, one approach is to repeat the procedure of a model-guided search for optimal training examples—except now using the repaired model—and then to conduct new empirical studies guided by the repaired model's modified predictions. At this juncture, however, a limitation of this approach is that our extension of the model with the within-category sensitivity parameter is intended only as a temporary repair and approximation. A more principled approach that would likely have more profound long-range benefits is to more directly address the problems that we hypothesize to be the underlying *causes* of the underestimated within-category similarities, namely, (1) the existence of rock characteristics that are not represented in the current MDS solution but that are nevertheless highly diagnostic for classification, and (2) noise in the locations of the represented objects in the high-dimensional feature space.

Missing diagnostic dimensions One approach to identifying diagnostic dimensions that are missing from the scaling solution is to search for systematic, category-specific mispredictions from the model and attempt to discern the basis for those mispredictions. To illustrate, consider the observed and predicted data from the complete classification-confusion matrix for the coverage condition that is shown in the top-left panel of Fig. 6. Among the data points that are most severely mispredicted are the four open squares that lie toward the upper-left corner of the scatter plot. These are all cases involving transfer items from the category *andesite*: The subjects in our experiment classified these transfer tokens with moderate to high accuracy, but the model severely underpredicted these accuracy levels. To gain insight, in the top panel of Fig. 8 we show the three training examples of andesite that appeared in the coverage condition, and in the bottom panel of Fig. 8 we show the four underpredicted transfer items. Note that all three training examples can be characterized as having a fine-grained homogeneous background, but with small pebbles or fragments glued into this background. These provide examples of rocks with what is known in geology as *porphyritic* texture (Tarbuck & Lutgens, 2014, p. 66). Although other rocks in the coverage-condition training set also had porphyritic texture, their occurrence in other categories was very sporadic and occasional, whereas porphyritic texture can be viewed as a unifying characteristic of the andesite training examples. The reader will further note that all of the underpredicted transfer items (in the bottom panel of Fig. 8) also had a clear-cut porphyritic texture. Although a dimension corresponding to "average grain size" figured prominently in the derived MDS solution for the rocks, this more subtle feature involving porphyritic texture was not explicitly represented. It seems highly plausible that this feature became salient in

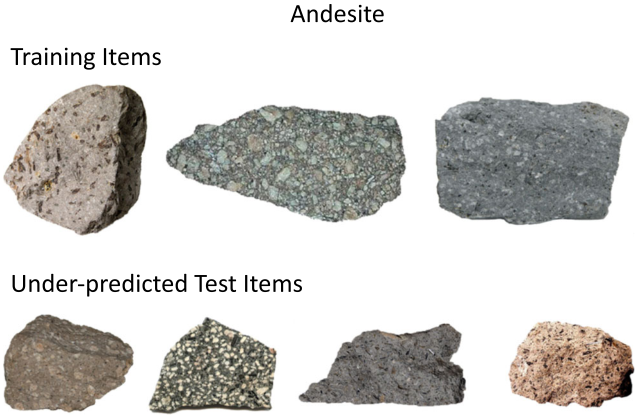

Fig. 8 Training items and underpredicted test items for andesite in the coverage condition. The scaling solution underestimates the similarity of these test items to the training items because it does not account for their shared porphyritic texture. (Color figure online)

the context of the classification-learning task and that many of the observers relied on it as a basis for classifying items into the andesite category. One approach to developing a more comprehensive feature-space representation for the rock stimuli would be to collect direct ratings of the extent to which each individual rock in the set has porphyritic texture (see Nosofsky et al., 2018, for extensive work on the collection of numerous types of direct dimension ratings for the present set of stimuli). The current MDS solution could then be extended simply by appending a new dimension defined by these direct porphyritic-texture ratings. A similar procedure to the one just illustrated could be attempted for other sets of systematically mis-predicted rock types in an attempt to build a more comprehensive feature-space representation for the objects in this complex naturalistic domain.

Noise reduction and automated scaling According to our current working hypotheses, a second problem that needs to be addressed is that, due to noise in the similarity-judgment scaling procedure, there is imprecision in the locations of the rock stimuli in the current MDS solution. An apparent example can be seen by again considering the top-left panel of Fig. 6. The data point represented by the open diamond toward the top-left of the scatterplot is a case in which subjects classified a transfer item from the category *gabbro* with high accuracy, but in which the model underpredicted this accuracy level. The training items for gabbro in the coverage condition are displayed in the top panel of Fig. 9; the underpredicted test item is displayed in the bottom panel. It seems apparent from inspection that the similarity of the test item to the first training item is extremely high; thus, from the perspective of the exemplar model, it is not surprising that subjects were so accurate in classifying the test item. Nevertheless, the points corresponding to these two rocks are located relatively far apart in the MDS solution.

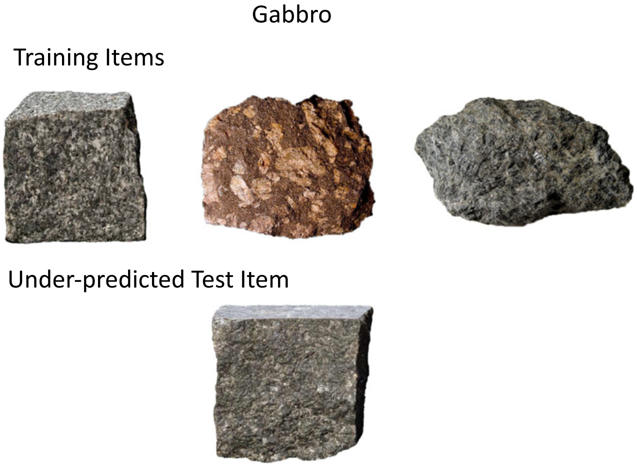

Fig. 9 Training items and the underpredicted test item for gabbro in the coverage condition. Despite their high similarity, the test item and the first training item are placed relatively far away in the MDS solution, likely due to noise in the similarity-based scaling procedure. (Color figure online)

Thus, a crucial direction for future research is to develop techniques that reduce noise in the positionings of the objects in the high-dimensional feature space. Unfortunately, this goal is a highly ambitious one and will likely require extensive new research and perhaps the application of major new scaling techniques before it can be accomplished. As we noted at the outset, the current MDS solution was derived by collecting similarity judgments among the extremely large number of pairs of distinct stimuli from a collection of 360 rock-picture tokens. Our goal, however, is not simply to model classification learning and generalization for the objects in this 360-member set, but to be able to extend the approach to predict classification learning and generalization for indefinitely large numbers of stimuli from this domain. Thus, the traditional approaches become intractable.

As one possible alternative, we have achieved some preliminary success in applying an integrated approach in which we combine MDS methods with the use of deep-learning convolutional neural networks (CNNs; e.g., Lecun, Bengio, & Hinton, 2015) to derive the high-dimensional feature space for these naturalistic stimuli (Sanders & Nosofsky, 2018). As is well known, CNNs have been used successfully to predict classification of natural images from large data bases. In a typical CNN architecture, elementary visual inputs are converted to higher-order features via connections to a series of hidden convolutional layers. These feed into a set of fully connected layers and a final output layer that generates the classification responses. A number of researchers have proposed and found support for the hypothesis that the patterns of activation developed in the layers of the CNN can serve as the feature space in which the naturalistic images are embedded (e.g., Battleday, Peterson, & Griffiths, 2017; Guest & Love, 2017; Lake, Zaremba, Fergus, & Gureckis, 2015; Peterson, Abbott, & Griffiths, 2016). Despite the preliminary successes described in these studies, the extent to which CNNs truly capture the details of human classification learning remains unknown. Thus, we have been exploring an approach that is complementary to the recent past applications. Rather than using CNNs to directly account for human classification judgments, we instead train them to predict the dimension values of individual exemplars derived from traditional MDS methods (Sanders & Nosofsky, 2018). Once the CNN is trained in this manner, new stimuli can be presented to the CNN and it can be used to automatically produce the coordinate values of the stimuli in the multidimensional psychological feature space. Thus, an unlimited number of stimuli from complex naturalistic domains can be scaled in this manner. The derived coordinate values can then be used in combination with formal models such as the GCM to predict classification learning and generalization (for related ideas, see Battleday et al., 2017).

Beyond scaling large sets of high-dimensional stimuli, such an approach may have other advantages compared to MDS methods that rely solely on similarity-judgment data. In particular, the CNN could potentially act to reduce the noise that is inherent in the similarity-judgment approach. For example, suppose that in the "true" psychological space, two stimuli reside very close to one another; however, due to noisy judgments, an observer may have provided a low similarity-judgment rating for the pair. In fitting the noisy data, a traditional MDS model might therefore position the two stimuli too far away from one another in the space. However, if used as input to the trained CNN, there would be an automatic correction, because the very similar inputs provided by the two stimuli would likely lead to similar outputs in the automated scaling solution. A great deal of future research is needed to test the viability of these ideas.

Alternative models of similarity and classification Finally, whereas the focus of our discussion thus far has been on repairing limitations of the current MDS solution for the rocks, we recognize of course that other limitations may reside in the formal models of similarity and classification themselves. The MDS similarity model is a purely spatial model that relies solely on measures of continuous distance. As an alternative, models based on matching and mismatching of discrete features, such as additive clustering (Shepard & Arabie, 1979) and tree models (Sattath & Tversky, 1977) might be applied. As discussed by (Nosofsky et al., 2018b, p. 347), for the present, highly complex stimulus domain, it seems most likely that hybrid models that combine spatial and discrete-feature components may be needed (e.g., Lee & Navarro, 2002; Nosofsky & Zaki, 2003; Verguts, Ameel, & Storms, 2004).

Likewise, the GCM provides only one candidate model of perceptual classification; numerous alternatives might fare considerably better. To take just one example, the SUSTAIN model (Love et al., 2004) is closely related to the GCM in terms of its computations; however, unlike the GCM, which stores each training item as a unique exemplar, SUSTAIN makes allowance for distinct exemplars to be combined in clusters represented by "subprototypes." This cluster-formation process results in category-compression effects, so perhaps could provide a process-level explanation for why the present exemplar-based modeling approach appears to have systematically underestimated within-category similarities among items. More generally, although the results are mixed, there is some evidence that category training itself can decrease within-category psychological distances, perhaps through forms of learned categorical perception (e.g., Goldstone, 1994; Gureckis & Goldstone, 2008; Viviani, Binda, & Borsato, 2007; but see Folstein, Gauthier, & Palmeri, 2012). Models that incorporate mechanisms for these potential forms of learned categorical perception might also provide significantly improved accounts of the present rock-classification data.

Number and variability of training instances

The central theme of our investigation was to generate true a priori predictions from the formal model of the consequences of using alternative sets of training examples on subsequent classification test performance. Although the prior predictions for the coverage, ss-optimal, and random-3 conditions were indeed parameter free, the same was not quite true of the random-9 condition (for reasons discussed extensively earlier in our article).

Nevertheless, in our view, the results from the random-9 condition are extremely informative in their own right and make a major new contribution to knowledge in this field. The general take-home message is that overall test performance in this condition—whether averaged over all individual items (old training and novel transfer) or considered for novel transfer items only—was better than in any of the conditions that used smaller numbers of training items. Although the generality of the finding clearly needs to be tested, the results point to the following tentative recommendation: If the goal is to foster generalization to novel transfer stimuli in a natural-science category domain, then it is better to train with a broad swath of training examples from each category rather than to focus training on a select few training examples.

Surprisingly, although related conclusions have emerged from previous work in the category-learning literature, we believe that our form of evidence has a unique slant. For example, a well-known and highly influential set of studies that bear on the issue are those reported by Homa and his colleagues (Homa, Cross, Cornell, Goldman, & Schwartz, 1973; Homa, Sterling, & Trepel, 1981; Homa & Vosburgh, 1976). Using artificial dot-pattern categories as materials, these investigators manipulated category size across conditions by varying the number of training instances that defined each category. They reported a robust pattern of results in which generalization to new transfer patterns improved with increases in category size. However, in their designs, each *individual* training instance was presented the same number of times. A consequence is that the larger-size categories received more total trials of training than did the smaller-size categories. By comparison, in our design, we held fixed across conditions the total number of training trials. In a real-world scenario involving the teaching of science categories, an educator may have fixed time or number of training trials available for accomplishing his or her goals, and simply devoting more total trials of category training may not be an option. Thus, our present results provide a major complement to those reported in the Homa studies.

Another classic study that bears on the issue is the one reported by Posner and Keele (1968). These researchers too used dot-pattern categories as materials but manipulated the *variability* of training instances across conditions. In a low-variability condition, each category was defined by training instances that were low distortions of a category prototype, whereas in a high-variability condition, each category was defined by training instances that were moderate distortions of a category prototype. This manipulation is related to our number-of-training-instances manipulation because, in general, as both the number of distinct training instances increases and as the variability of training instances increases, one expects better overall "coverage" of the complete category distribution. In line with the present results, Posner and Keele (1968) observed better generalization to new transfer items in the high-variability-training condition than in the low-variability one. However, in Posner and Keele's design, training did not end until observers met a criterion of correctly classifying all training instances from the study lists for two consecutive blocks of trials. Although Posner and Keele (1968, p. 356) did not report the detailed results from the training phase, they did note that, not surprisingly, observers in the high-variability condition made more errors in original learning than did observers in the low-variability condition. Almost certainly, therefore, observers in the high-variability condition received more total trials of training than did observers in the low-variability condition—a similar state of affairs as occurred in the category-size experiments reported by Homa and his colleagues.

A previous study that perhaps comes closest to our random-3 versus random-9 manipulation is one reported by Wahlheim, Finn, and Jacoby (2012). As in our investigation, these researchers trained participants to classify objects from natural categories—species of birds belonging to different bird-category families. Although the researchers' main interest was in investigating certain metacognitive aspects of their participants' category learning, their key experimental manipulation was similar to our own. In particular, for half the bird-category families, training involved a small-size/high-repetition condition: two species of each category repeated six times each (S_2R_6); whereas the remaining half of the bird-category families involved a large-size/low-repetition condition: six species of each category repeated two times each (S_6R_2). (Note that Wahlheim et al.'s S_2R_6 condition is analogous to our random-3 condition, and their S_6R_2 condition is analogous to our random-9 condition.) In line with our present results, Wahlheim et al. (2012) found that, at time of test, participants classified old training instances with higher accuracy in the small-size/high-repetition condition (S_2R_6), but generalized to members of novel species of each bird-category family with higher accuracy in the large-size/low-repetition condition (S_6R_2).

Although the take-home message from this aspect of the two studies is basically the same, we do note an important difference in our designs. In particular, in Wahlheim et al.'s case, the manipulations of repetition and category size were conducted within a mixed-categories design: Half the categories experienced by each participant were small-size/high-repetition and the other half were large-size/low-repetition. Thus, if participants had simply developed a general response bias for classifying objects into the large-size categories compared with the small-size ones, then that tendency alone could explain the observed pattern of generalization results. (The better performance on the old training instances in the high-repetition condition would be explained by the stronger memory traces developed for those items.) By comparison, we do not see an analogous response-bias explanation for our results, because, within each of our conditions, *all* categories were either small-size/high-repetition (random-3) or else large-size/low-repetition (random-9).[6]

Future research is needed to test the generality of our findings involving the apparent benefits conferred by our random-9 condition (compared with our small-size conditions). The needed research is both empirical and theoretical in nature. From an empirical standpoint, our present manipulations involved only a single contrast between the conditions (size-3 vs. size-9). Furthermore, other variables that might interact with the category-size manipulation—such as total number of training trials, number of to-be-learned categories, and difficulty level of the category distinctions—were held fixed. Future research might manipulate these variables in parametric fashion to test for the generality of the results.

From a theoretical standpoint, although the results from the random-9 condition appear to be in general accord with the predictions from the formal model, these predictions are not completely parameter free—a point that we have acknowledged in this article on several previous occasions. To achieve more precision in the predictions, deeper theories will be required that specify more detailed mechanisms that govern the settings of the currently estimated free parameters (i.e., p_{store} and c).[7] Once such theoretical advances are made, we might

[6] The response-bias possibility that we advance for the Wahlheim et al. (2012) study is not an idle concern. In particular, Cohen, Nosofsky, and Zaki (2001) reported a series of experiments that explicitly manipulated category variability. In their designs, one category had low variability and a second had high variability. At time of test, a critical test item was presented that had equal similarity to the nearest training exemplar of each category. In all experiments, participants tended to classify the critical item into the high-variability category with higher probability than was predicted by a baseline version of an exemplar-similarity model (the GCM). Cohen et al. (2001) argued for the role of a systematic response bias toward the high-variability category and made a case for a rational basis for that form of response bias.

[7] To take just one example, Kruschke's (1992) ALCOVE model builds upon the GCM by incorporating many of its components within the framework of an error-driven connectionist learning model. A potential advantage of this approach is that, rather than being specified as free parameters, the strength of associations between stored exemplars and categories are an emergent property of the learning characteristics of the network. It is a wide-open question what predictions ALCOVE would make for how the association strengths of the stored exemplars would be related across the random-9 condition and the small-set-size training conditions.

then also explore more intricate issues such as those suggested in the preceding paragraph. For example, answers to questions such as *which* exemplars form optimal training sets might vary in parameter-dependent ways with factors such as total number of training trials, category size, number of to-be-learned categories, and so forth. We believe that this kind of model sophistication would have tremendous applied value, as it could fruitfully limit plausible candidates for optimal training sets for a wide range of categories varying in their dimensional complexity and structure. Thus, rather than engaging in unconstrained empirical investigation of such issues, the model-guided search would make such research far more efficient and tractable.[8]

Other approaches to enhancing the teaching of the natural-science categories Finally, although the present investigation centered around the question of which training sets might be optimal, there are of course many other approaches to enhancing the teaching of natural-science categories, some of which may also profit from a formal model-guided approach. For instance, learning from optimal sets of training items can perhaps be further enhanced by specifying particularly effective presentation orders of those items (e.g., Mathy & Feldman, 2009, 2016; Pashler & Mozer, 2013; Spiering & Ashby, 2008). In a related vein, training items can be blocked by category or intermixed by category, and intermixing has been found to support better generalization to new items when training naturalistic categories such as chemical categories (Eglington & Kang, 2017) and artistic-style categories (Kang & Pashler, 2012; Kornell & Bjork, 2008). However, research with artificial categories has revealed that the relative advantage of one training order versus the other can vary depending on the structure of the to-be-learned categories (Carvalho & Goldstone, 2014). Accordingly, a formal modeling approach as developed in this article could be invaluable for anticipating how the effectiveness of blocked versus intermixed presentation might vary with the structure of the to-be-learned categories. Models and approaches that are closely related to GCM incorporate sensitivity to presentation order (e.g., Carvalho & Goldstone, 2017; Kruschke, 1992; Love et al., 2004) and in principle might be applied to this issue.

As well, recent approaches to dynamic adaptive training procedures, in which the frequency and spacing of particular instances are personalized to each learner (e.g., Lindsey, Shroyer, Pashler, & Mozer, 2014; Mettler & Kellman, 2010), offer potential promise for improving scientific-category learning. Some of these systems use the empirically based learning difficulty of particular target items along with the individual's past performance on particular target items to personalize the presentation of training items. For science-classification training, theoretical models that anticipate category difficulty, as developed herein, could potentially augment the effectiveness of dynamic adaptive training algorithms.

Yet another approach that is emerging in the basic cognitive research is specifying the types of explicit "coaching" that can enhance category learning. One type of coaching involves providing learners with explicit information about characteristic features or rules that are diagnostic for each category (e.g., Miyatsu, Gouravajhala, Nosofsky, & McDaniel, 2018; Pashler & Lovelett, 2017). For instance, Miyatsu et al. (2018) found that highlighting characteristic features of particular rock categories during training (by circling and describing specific features on the image of each training token) produced more accurate generalization to new instances than when learners were not provided such highlighting. Likewise, Eglington and Kang (2017) found that, in cross-experiment comparisons, explicit highlighting of diagnostic features improved learning and generalization in the domain of chemistry categories. Formal modeling could reveal how feature highlighting potentially changes learners' attention weighting across features, thereby modifying the structure of the psychological similarity space in which the category exemplars are embedded (Nosofsky, 1986). Such model-guided analysis might point to the particular forms of feature highlighting that would optimize the attention-weighting processes useful for generalizing to the members of alternative natural-science category structures.

Conclusions

Our model-guided search for optimal training instances for teaching the rock categories met with some successes and failures. The failures led us to major insights regarding the shortcomings of the present version of the model, and to a clear target path for improving upon it. Although our current account of the rock-categorization data is not completely parameter-free, it is still the case that a relatively low-parameter model is providing good fits to a rich set of data from a complex, naturalistic categorization domain. Furthermore, a tentative recommendation that emerges from the work is that, if the goal is to foster generalization to novel transfer stimuli in this domain, it is better to train with a broad swath of training examples from each category rather than to focus training on a select few training examples. Based on our results and consideration of findings exploring other dimensions of category training, we are confident that our long-range goal of using formal psychological models to help guide the search for effective methods of teaching science categories is a promising one.

[8] Indeed, this approach parallels successful high-throughput protein design efforts in biochemistry in which the structure/function of thousands of computer-designed protein candidates can be modeled, from which an optimal set can be identified for experimental synthesis and testing. By so doing, time-consuming and expensive experimental protein synthesis and characterization can be limited to a small set of fruitful candidates. Further, as is the intent of our investigations, the experimental results feed back to the computer models to further improve the accuracy of the modeling (Rocklin et al., 2017).

Acknowledgements This work was supported by National Science Foundation Grant 1534014 (EHR Core Research) to Robert Nosofsky, Bruce Douglas, and Mark McDaniel.

The authors thank Xuezhou Zhang for his contributions to computing the set of training instances in the ss-optimal condition of this experiment.

Appendix A

Description of the procedure for choosing the training examples in the optimal condition

As explained in the text, the objective function to be maximized was the GCM's prediction of overall proportion correct computed across all test items, holding fixed the values of the best-fitting parameters estimated in Experiment 1 of Nosofsky et al. (2018b). Due to combinatorial explosion, it was not feasible to conduct an exhaustive search of all possible sets of training examples to locate the theoretically optimal set. Instead, we relied on a heuristic greedy-search computer algorithm. On any given run of the algorithm, a starting training set was created by selecting at random a single item from each category. Then, on each iteration, the greedy-search algorithm added to the training set the single remaining exemplar (from the complete collection of 120 items) that yielded the highest predicted overall proportion correct. The iterations continued until the set-size limit of 30 was reached. This same procedure was conducted hundreds of thousands of times, and the set that maximized the objective function (subject to the constraint that at least two training examples from each category were included) was selected.

Following the selection of the theoretically optimal examples by the search algorithm, we decided to make two minor changes to the training set. In particular, we exchanged one token of granite for another, and one token of peridotite for another. (Using the numbering scheme developed in Nosofsky, Sanders, Meagher, and Douglas, 2018, we used token 10 of granite instead of token 12; and token 2 of peridotite instead of token 7.) We decided to make these exchanges based on our intuitive judgment that the new tokens provided better coverage of the to-be-learned category distributions than did the computer-chosen ones. Furthermore, our computer simulations indicated that the alternative training set yielded a prediction of overall proportion correct that was only .01 lower than the theoretically optimal one (for reasonably high settings of the p_{store} parameter). Despite these exchanges, for simplicity, we continue to refer to our second condition as the "ss-optimal" condition.

The specific training items used in the ss-optimal condition (as well as in the coverage condition) are listed in Table 8. The specific rock images and MDS coordinates associated with each of the listed training examples are available on the website https://osf.io/w64fv/. It is of interest to note that the greedy-search algorithm assigned fewer training examples to easier categories (obsidian, pegmatite, and pumice) and more training examples to difficult ones (andesite and rhyolite) in the ss-optimal condition.

Table 8 Training items and category membership values in the coverage and ss-optimal conditions

Training item #	Category	Training item #	Category
Coverage		ss-Optimal	
2	1	1	1
5	1	2	1
9	1	5	1
16	2	8	1
19	2	16	2
21	2	19	2
26	3	21	2
28	3	27	3
33	3	32	3
39	4	35	3
43	4	37	4
44	4	41	4
49	5	46	4
54	5	54	5
58	5	58	5
65	6	59	5
66	6	67	6
71	6	72	6
75	7	75	7
76	7	76	7
83	7	83	7
85	8	86	8
87	8	94	8
94	8	97	9
98	9	103	9
104	9	111	10
107	9	116	10
114	10	117	10
116	10	118	10
120	10	119	10

Note. 1 = andesite, 2 = basalt, 3 = diorite, 4 = gabbro, 5 = granite, 6 = obsidian, 7 = pegmatite, 8 = peridotite, 9 = pumice, 10 = rhyolite

Appendix B

Description of the mixed exemplar-plus-prototype model

According to prototype models (e.g., Nosofsky, 1986; Reed, 1972; Smith & Minda, 1998), 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. Nosofsky et al. (2018b) showed previously that a pure version of the prototype model provides very poor fits to the present forms of data; among the reasons is that it is unable to predict the robust advantage in classification accuracy observed for the old training examples compared with the new transfer stimuli in the test phase. Conceivably, however, observers may rely on some mixture of exemplar-plus-prototype information.

Let $P_{prot}(J|i)$ denote the prototype-model prediction of the probability that item i is classified in Category J, and let $P_{exm}(J|i)$ denote the corresponding probability for the exemplar model. In the mixed model, the overall probability that item i is classified in Category J is given by

$$P(J|i) \quad = \quad mix \times P_{exm}(J|i) + (1-mix) \times P_{prot}(J|i), \qquad (B1)$$

where mix ($0 \leq mix \leq 1$) is the probability that the observer relies on stored exemplars, and $1 - mix$ is the probability that the observer relies on the prototype.

In fitting the model, we held parameters fixed across the coverage, ss-optimal and random-3 conditions, but allowed completely separate free parameters for the random-9 condition. Within each set of conditions, we estimated the mixture parameter mix; the parameters c, $p_{store,}$ and γ for the exemplar model; and a separate sensitivity parameter c_{prot} for the prototype model, yielding a total of 10 free parameters. As reported in Table 5, this mixture model led to relatively minor improvements in fit compared to the pure, baseline version of the exemplar model.

Appendix C

Noisy MDS simulations

To confirm the intuitions regarding how adding random noise to the locations of objects in a "true" configuration would affect the patterns of within-to-between category similarity in the noise-modified configuration, we conducted some example simulations. To maintain comparability with the rock-stimuli modeling, we supposed that each category was composed of 12 objects, and that each object varied along eight independent dimensions. For simplicity, we restricted consideration to the case of two categories. (If one supposes that the two categories are neighboring categories in the high-dimensional space, with the other categories being distant from the two, then including the other categories would add needless complications to the present analysis.) We assumed that the categories both had multivariate normal structures. The population mean of Category A was set at the origin. For each simulation, the population mean of Category B was created by sampling randomly and independently from a normal distribution with mean zero and standard deviation *sdB* on each dimension. (As the magnitude of *sdB* increases, between-category similarity tends to decrease.) The 12 sample members of each category were then generated as follows. For each dimension, a deviation was randomly and independently sampled from a normal distribution with mean zero and standard deviation *sdW*. The deviation was added to the category population mean on that dimension to create the object's "true" location in the multidimensional space. (As the magnitude of *sdW* increases, within-category similarity tends to decrease.) In effect, this procedure created two clouds of points in the multidimensional space, with the clouds roughly centered about their respective population means. The magnitude of *sdB* controlled the distance between the means of the two clouds; the magnitude of *sdW* controlled the overall variability or expanse of each individual cloud.

Finally, we created "noise-distorted" configurations by defining a noise-parameter *sdN*. For each individual object and dimension, a deviation was randomly and independently sampled from a normal distribution with mean zero and standard deviation *sdN*, and the deviations were added to the objects' true locations on each dimension.

For any given collection of settings of *sdB*, *sdW*, and *sdN*, we conducted the simulations described above 10,000 times. For each simulation, we computed the average within-category similarity between all distinct pairs of points, and the average between-category similarity between all pairs of points. The similarity between any individual pair of points was given by $s = \exp(-\ d)$, where d is the Euclidean distance between the pair of points in the simulated space. We then computed the average value of the computed within-category similarity and the average value of the computed between-category similarity across all 10,000 simulations, as well as the ratio of those two measures.

The results for some example values of *sdB* and *sdW*, with the value of *sdN* varied parametrically in each case, is reported in Tables 9, 10, and 11. In each example, *sdB* is held fixed at 1, while *sdW* takes on the values .05, .1, or .2. (Again, larger values of *sdW* create higher-variability clouds of points, so that within-category similarity in the true configuration tends to decrease.) Regardless of the settings of *sdB* and *sdW*, it can be seen that as greater amounts of noise are added to the locations of the individual points (i.e., as, the magnitude of *sdN* increases), there is a dramatic lowering of average within-category similarity, a relatively smaller lowering of averaged between-category similarity, and a steady decrease in the ratio of averaged within-to-between category similarity. For very large values of *sdN*, both measures drop towards zero, and the ratio begins to approach one.

Table 9 Simulated average within-category and between-category similarities as a function of *sdN* in the case in which *sdB* = 1 and *sdW* = .05

sdN	Ave. within sim.	Ave. between sim.	Ratio
0	.412	.040	10.224
.1	.326	.039	8.267
.5	.080	.024	3.324
1.0	.016	.008	1.918
2.0	.001	.001	1.333

Table 10 Simulated average within-category and between-category similarities as a function of *sdN* in the case in which *sdB* = 1 and *sdW* = .1

sdN	Ave. within sim.	Ave. between sim.	Ratio
0	.341	.040	8.572
.1	.292	.039	7.547
.5	.078	.024	3.279
1.0	.016	.008	1.931
2.0	.001	.001	1.331

Table 11 Simulated average within-category and between-category similarities as a function of *sdN* in the case in which *sdB* = 1 and *sdW* = .2

sdN	Ave. within sim.	Ave. between sim.	Ratio
0	.235	.037	6.367
.1	.215	.036	5.980
.5	.071	.023	3.124
1.0	.015	.008	1.916
2.0	.001	.001	1.339

Note. Ave. within sim. = average within-category similarity; Ave. between sim. = average between-category similarity; *sdN* = magnitude of the location-noise parameter

References

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

Ashby, F. G. (Ed.). (1992). Multidimensional models of perception and cognition. Hillsdale: Erlbaum.

Ashby, F. G., & Maddox, W. T. (1993). Relations between prototype, exemplar, and decision bound models of categorization. *Journal of Mathematical Psychology*, *37*(3), 372–400.

Battleday, R. M., Peterson, J. C., & Griffiths, T. L. (2017). Modeling human categorization of natural images using deep feature representations. Retrieved from https://arxiv.org/abs/1711.04855

Brainard, D. H. (1997). The Psychophysics Toolbox. *Spatial Vision*, *10*, 433–436.

Carvalho, P. F., & Goldstone, R. L. (2014). Putting category learning in order: Category structure and temporal arrangement affect the benefit of interleaved over blocked study. *Memory & Cognition*, *42*(3), 481–495.

Carvalho, P. F., & Goldstone, R. L. (2017). The sequence of study changes what information is attended to, encoded, and remembered during category learning. *Journal of Experimental Psychology: Learning, Memory, and Cognition*, *43*, 1699–1719.

Cohen, A. L., Nosofsky, R. M., & Zaki, S. R. (2001). Category variability, exemplar similarity, and perceptual classification. *Memory & Cognition*, *29*(8), 1165–1175.

Eglington, G., & Kang, S. H. K. (2017). Interleaved presentation benefits science category learning. *Journal of Applied Research in Memory and Cognition*, *6*(4), 475–485.

Folstein, J. R., Gauthier, I., & Palmeri, T. J. (2012). How category learning affects object discrimination: Not all morphspaces stretch alike. *Journal of Experimental Psychology: Learning, Memory, and Cognition, 38*, 807–820.

Goldstone, R. L. (1994). Influences of categorization on perceptual discrimination. *Journal of Experimental Psychology: General, 123*, 178–200.

Guest, O., & Love, B. C. (2017). What the success of brain imaging implies about the neural code. *eLife*, *6,* e21397.

Gureckis, T. M., & Goldstone, R. L. (2008). The effect of the internal structure of categories on perception. Paper presented at the Proceedings of the 30th Annual Conference of the Cognitive Science Society, Austin.

Homa, D., Cross, J., Cornell, D., Goldman, D., & Schwartz, S. (1973). Prototype abstraction and classification of new instances as a function of number of instances defining the prototype. *Journal of Experimental Psychology*, *101*(1), 116.

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

Homa, D., & Vosburgh, R. (1976). Category breadth and the abstraction of prototypical information. *Journal of Experimental Psychology: Human Learning and Memory*, *2*(3), 322.

Hooke, R., & Jeeves, T. A. (1961). "Direct search" solution of numerical and statistical problems. *Journal of the ACM (JACM)*, *8*(2), 212–229.

Kang, S. H., & Pashler, H. (2012). Learning painting styles: Spacing is advantageous when it promotes discriminative contrast. *Applied Cognitive Psychology*, *26*(1), 97–103.

Khajah, M. M., Lindsey, R. V., & Mozer, M. C. (2014). Maximizing students' retention via spaced review: Practical guidance from computational models of memory. *Topics in Cognitive Science*, *6*(1), 157–169.

Kılıç, A., Criss, A. H., Malmberg, K. J., & Shiffrin, R. M. (2017). Models that allow us to perceive the world more accurately also allow us to remember past events more accurately via differentiation. *Cognitive Psychology*, *92*, 65–86.

Kornell, N., & Bjork, R. A. (2008). Learning concepts and categories: Is spacing the "enemy of induction"?. *Psychological Science*, *19*(6), 585–592.

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

Kruskal, J. B., & Wish, M. (1978). Multidimensional scaling (Vol. 11). New York: SAGE.

Lake, B. M., Zaremba, W., Fergus, R., & Gureckis, T. M. (2015). Deep neural networks predict category typicality ratings for images. In R. Dale, C. Jennings, P. Maglio, T. Matlock, D. Noelle, A. Warlaumont, & J. Yoshimi (Eds.), Proceedings of the 37th Annual Conference of the Cognitive Science Society. Austin: Cognitive Science Society.

LeCun, Y., Bengio, Y., & Hinton, G. (2015). Deep learning. *Nature*, *521*(7553), 436–444.

Lee, M. D., & Navarro, D. J. (2002). Extending the ALCOVE model of category learning to featural stimulus domains. *Psychonomic Bulletin & Review, 9*(1), 43–58.

Lee, M. D., & Pope, K. J. (2003). Avoiding the dangers of averaging across participants when using multidimensional scaling. *Journal of Mathematical Psychology, 47*(1), 32–46.

Lindsey, R. V., Shreyer, J. D., Pashler, H., & Mozer, M. C. (2014). Improving students' long-term retention through personalized review. *Psychological Science, 25,* 639–647.

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

Luce, R. D. (1963). Detection and recognition. In D. Luce (Ed.), Handbook of mathematical psychology (pp. 1-103). New York: Wiley.

Mathy, F., & Feldman, J. (2009). A rule-based presentation order facilitates category learning. *Psychonomic Bulletin & Review*, *16*(6), 1050-1057.

Mathy, F., & Feldman, J. (2016). The influence of presentation order on category transfer. *Experimental Psychology*, 63(1), 59-69.

McKinley, S. C., & Nosofsky, R. M. (1995). Investigations of exemplar and decision bound models in large, ill-defined category structures. *Journal of Experimental Psychology: Human Perception and Performance*, *21*(1), 128.

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

Mettler, E., & Kellman, P. (2010). Adaptive sequencing in perceptual learning. *Journal of Vision*, *10*(7), 1098.

Miyatsu, T., Gouravajhala, R., Nosofsky, R.M., & McDaniel, M.A. (2018). Feature highlighting enhances learning of complex natural-science categories. *Journal of Experimental Psychology: Learning, Memory, and Cognition*. https://doi.org/10.1037/xlm0000538

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

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

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

Nosofsky, R. M. (1988). Similarity, frequency, and category representations. *Journal of Experimental Psychology: Learning, Memory, and Cognition*, *14*(1), 54–65.

Nosofsky, R. M. (1991). Tests of an exemplar model for relating perceptual classification and recognition memory. *Journal of Experimental Psychology: Human Perception and Performance*, *17*(1), 3.

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

Nosofsky, R. M. (2011). The generalized context model: An exemplar model of classification. In E. M. Pothos & A. J. Wills (Eds.), Formal approaches in categorization (pp. 18-39). New York: Cambridge University Press.

Nosofsky, R. M., & Palmeri, T. J. (1997). An exemplar-based random walk model of speeded classification. *Psychological Review*, *104*(2), 266.

Nosofsky, R. M., Sanders, C. A., Gerdom, A., Douglas, B. J., & McDaniel, M. A. (2017). On learning natural-science categories that violate the family-resemblance principle. *Psychological Science*, *28*(1), 104-114.

Nosofsky, R. M., Sanders, C. A., & McDaniel, M. A. (2018a). A formal psychological model of classification applied to natural-science category learning. *Current Directions in Psychological Science*, *27*, 129–135.

Nosofsky, R. M., Sanders, C. A., & McDaniel, M. A. (2018b). Tests of an exemplar-memory model of classification learning in a high-dimensional natural-science category domain. *Journal of Experimental Psychology: General*, *147*, 328–353.

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

Nosofsky, R. M., & Zaki, S. R. (2003). A hybrid-similarity exemplar model for predicting distinctiveness effects in perceptual old-new recognition. *Journal of Experimental Psychology: Learning, Memory, and Cognition, 29*(6), 1194.

Okada, K., & Lee, M. D. (2016). A Bayesian approach to modeling group and individual differences in multidimensional scaling. *Journal of Mathematical Psychology, 70*, 35-44.

Pashler, H., & Lovelett, J. (2017). Does coaching promote perceptual category learning?. Talk given at the 58th Annual Meeting of the Psychonomic Society, Vancouver.

Pashler, H., & Mozer, M. C. (2013). When does fading enhance perceptual category learning?. *Journal of Experimental Psychology: Learning, Memory, and Cognition*, *39*(4), 1162.

Patil, K., Zhu, X., Kopec, L., & Love, B. (2014). Optimal teaching for limited-capacity human learners. *Advances in Neural Information Processing Systems* (*NIPS*). Retrieved from https://papers.nips.cc/paper/5541-optimal-teaching-for-limited-capacity-human-learners.pdf

Peterson, J. C., Abbott, J. T., & Griffiths, T. L. (2016). Adapting deep network features to capture psychological representations. Retrieved from https://arxiv.org/abs/1608.02164

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

Pothos, E. M., & Wills, A. J. (Eds.). (2011). Formal approaches in categorization. New York: Cambridge University Press.

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

Rocklin, G. J., Chidyausku, T. M., Goreshnik, I., Ford, A., Lemak, A., … Baker, D. (2017). Global analysis of protein folding using massively parallel design, synthesis, and testing, *Science*, *357*, 168-175.

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

Sanders, C. A., & Nosofsky, R. M. (2018). Using deep learning representations of complex natural stimuli as input to psychological models of classification. Proceedings of the 2018 Conference of the Cognitive Science Society, Madison.

Sattath, S., & Tversky, A. (1977). Additive similarity trees. *Psychometrika, 42*(3), 319-345.

Shen, J., & Palmeri, T. J. (2016). Modelling individual difference in visual categorization. *Visual Cognition*, *24*(3), 260–283.

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

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

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

Shepard, R. N., & Arabie, P. (1979). Additive clustering: Representation of similarities as combinations of discrete overlapping properties. *Psychological Review, 86*(2), 87.

Shiffrin, R. M., Ratcliff, R., & Clark, S. E. (1990). List-strength effect: II. Theoretical mechanisms. *Journal of Experimental Psychology: Learning, Memory, and Cognition*, *16*(2), 179.

Shiffrin, R. M., & Steyvers, M. (1997). A model for recognition memory: REM—Retrieving effectively from memory. *Psychonomic Bulletin & Review*, *4*(2), 145–166.

Smith, J. D., & Minda, J. P. (1998). Prototypes in the mist: The early epochs of category learning. *Journal of Experimental Psychology: Learning, Memory, and Cognition*, *24*(6), 1411.

Spiering, B. J., & Ashby, F. G. (2008). Initial training with difficult items facilitates information integration, but not rule-based category learning. *Psychological Science*, *19*(11), 1169-1177.

Storms, G., De Boeck, P., & Ruts, W. (2000). Prototype and exemplar-based information in natural language categories. *Journal of Memory and Language*, *42*(1), 51-73.

Tarbuck, E. J., & Lutgens, F. K. (2014). *Earth science* (14th ed.). Boston, MA: Pearson.

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

Verguts, T., Ameel, E., & Storms, G. (2004). Measures of similarity in models of categorization. *Memory & Cognition, 32*(3), 379–389.

Verheyen, S., Ameel, E., & Storms, G. (2007). Determining the dimensionality in spatial representations of semantic concepts. *Behavior Research Methods, 39*(3), 427 438.

Viviani, P., Binda, P., & Borsato, T. (2007). Categorical perception of newly learned faces. *Visual Cognition, 15*, 420-467.

Voorspoels, W., Vanpaemel, W., & Storms, G. (2008). Exemplars and prototypes in natural language concepts: A typicality-based evaluation. *Psychonomic Bulletin & Review*, *15*(3), 630-637.

Wahlheim, C. N., Finn, B., & Jacoby, L. L. (2012). Metacognitive judgments of repetition and variability effects in natural concept learning: Evidence for variability neglect. *Memory & Cognition*, *40*(5), 703-716.

Wills, A. J., & Pothos, E. M. (2012). On the adequacy of current empirical evaluations of formal models of categorization. *Psychological Bulletin*, *138*(1), 102.