Tests of a Hybrid-Similarity Exemplar Model of Context-Dependent Memorability in a High-Dimensional Real-World Category Domain

Robert M. Nosofsky[1] and Adam F. Osth[2]

1. Indiana University Bloomington
2. Melbourne School of Psychological Sciences,
   The University of Melbourne

Robert Nosofsky
Department of Psychological and Brain Sciences
1101. E. Tenth Street
Indiana University
Bloomington, IN 47405
nosofsky@indiana.edu

Running Head: Exemplar Model of Context-Dependent Memorability

Word Count: 15,630

The following materials associated with this article are available on the OSF website https://osf.io/a958r/: the complete sets of individual-subject data from Experiments 1 and 2; the complete set of rock-image stimuli used in the experiments along with their category assignments; the multidimensional scaling solution for the rock images; the set of distinctive-feature ratings for the rock images; the set of goodness-of-category-example ratings obtained for the rock images; and a file that indexes the specific rock images that are new to this study or that had already been used in previous studies.

Abstract

We conduct tests of a hybrid-similarity exemplar model on its ability to account for the context-dependent memorability of items embedded in high-dimensional category spaces. According to the model, recognition judgments are based on the summed similarity of test items to studied exemplars. The model allows for the idea that “self-similarity” among objects differs due to matching on highly salient distinctive features. Participants viewed a study list of rock images belonging to geologically defined categories where the number of studied items from each category was manipulated. Following study, the participants’ old-new recognition memory performance was tested. We also manipulated across experiments the nature of the encoding task used during the study phase: Experiment 1 used a category-description matching task, whereas Experiment 2 used more neutral encoding instructions. Hit rates were markedly lower in Experiment 2 than in Experiment 1 and participants relied less on the presence of distinctive features for recognizing old items in the second experiment. With a minimum of parameter estimation, the hybrid-similarity model provided good accounts of a wide variety of fundamental benchmark phenomena across the two experiments. These included changing levels of memorability due to contextual effects of category size, within- and between-category similarity, and the presence of distinctive features. However, the hybrid model and a variety of extensions of the model fell short in accounting for the variability in hit rates within the class of old target items themselves. We discuss future directions for potentially improving upon the current predictions from the model.

Keywords: memorability, old-new recognition, computational modeling, categorization, similarity

Public Significance Statements

This study suggests that people's ability to remember visual images of real-world objects is influenced by the nature of other objects that are studied in the same learning context. For example, people exhibit strong false memories for novel objects that belong to large-size categories experienced during a learning episode.

People tend to exhibit strong true memories for studied objects that possess distinctive features that make them stand out from other objects in a learning context. The prevalence of this tendency varies with the encoding task and varies across different individuals.

The study proposes and tests a computational model that formalizes the idea that items have different degrees of "self-match" to their representations in memory. The model provides a good account of numerous benchmark findings involving the context-dependent nature of memorability but falls short in capturing the variation in true memories associated with individual studied items.

Modern research indicates that images of individual real-world objects vary considerably in their memorability: Some objects appear to be relatively easy to remember, whereas memories of other objects are more error-prone. This differential memorability is often studied in tasks of old-new recognition, in which observers are presented with lists of items to study, followed by test lists composed of old and new items. Certain old items are consistently recognized with higher accuracy than are others; and people also tend to false-alarm to certain new items more than others (Bainbridge, 2019; Kramer, Hebart, et al., 2023).

The high level of consistency between participants in which items are well recognized or poorly recognized has led many researchers to conclude that memorability is primarily dependent on the stimulus itself. This principle has carried forward into modeling investigations, in which researchers have reported successful attempts to build deep-learning networks that can predict which images will and will not be memorable (Bylinskii, Goetschalckx, et al., 2022; Dubey, Peterson, et al., 2015; Khosla, Raju, et al., 2015; Needel & Bainbridge, 2022). Although these deep-learning networks have provided impressive predictions of individual-item memorability, a limitation is that they don't provide insights into the psychological reasons why certain images are more memorable than others.

A more fundamental limitation is that these deep learning models do not take account of the enormous role that context can play in influencing memorability: as we review below, an extensive cognitive-psychology literature has shown that memory is highly dependent on the context of the study list in which the items are embedded -- a stimulus can be memorable or non-memorable depending on the other stimuli that accompany it on the list.

In the present work, rather than using deep-learning networks from the machine-learning tradition, we advance and test a cognitive-modeling approach to understanding individual item memorability. Specifically, we test predictions of memorability from the perspective of summed-similarity exemplar models of recognition (Gillund & Shiffrin, 1984; Hintzman, 1988; Kahana & Sekuler, 2002; Nosofsky, 1988, 1991; Osth & Dennis, 2024). Such models presume that items on a study list are stored as individual exemplars in memory. Presentation of a test probe gives rise to a global activation of memory via a summing of similarity of the probe to the stored study exemplars. The greater the degree of global activation, the higher is the probability that the observer judges the test probe to be old. Crucially, in such models, the extent to which a stimulus is memorable or not depends on the context of the learning episode. The reason, as explained below, is that the summed-similarity signal itself is highly context dependent, because a different memory set results in a different summed-similarity signal. In two experiments, we tested old-new recognition performance in a high-dimensional, real-world category domain, and manipulated a variety of factors related to the context of the learning episodes. We then evaluated the ability of different versions of summed-similarity exemplar models to capture the context-dependent nature of individual item memorability.

Study-List Context, Memory, and Exemplar Models

An extensive cognitive-psychology literature has shown that memory is highly dependent on the context of the study list in which the items are embedded. Although there are many empirical demonstrations of this property, an example that is very relevant to the current work is the demonstration of category size effects, where increasing the number of items from the same category on the study list increases false alarm rates to novel members of the same category (e.g., Robinson & Roediger, 1997; Shiffrin, Huber, & Marinelli, 1995) and decreases one's ability to discriminate between targets and lures from those categories in forced-choice tests (Konkle, Brady, et al., 2010). Other prominent examples are the list length effect, the finding that memory gets worse as the overall size of the study list is increased (Strong, 1912); as well as the von Restorff effect, which is the finding of superior memory for an item that is dissimilar to the other items on the study list (Hunt, 1995; von Restorff, 1933).

Summed-similarity exemplar models naturally capture many of these study-list context effects. To take a simple example, consider the effect of category size on false-alarm rates to lures. Note that as category size increases, the summed similarity of a test probe to exemplars on the study list will tend to increase. Thus, such models provide a natural account of the fundamental finding that false-alarm rates tend to be greater for test lures that are members of large-size categories than small-size ones.

The summed-similarity signal can be conceptualized as being composed of two components. For old test probes, one component is the self-match of the test probe to its own representation in memory. And for both old and new test probes, the second component is inter-item similarity: the similarity of a test probe to the memory representations of other exemplars besides itself. In general, summed-similarity exemplar models predict higher recognition rates for old test probes (hit rates) than for new test probes (false-alarm rates) because the self-match of an old test probe to its own exemplar trace provides a significant boost to the summed-similarity signal. However, the models also predict high false-alarm rates for new test probes that are highly similar to old exemplars stored in memory (such as category prototypes), because such items will also give rise to a large-magnitude summed-similarity signal. These ideas are illustrated schematically in Figure 1. The middle panel depicts a set of 5 study items in memory. The left panel and the right panel depict an example target (old) and lure (new) test probe, respectively. The color and thickness of the arrows depict the similarity of the test probes to the individual items in memory. The target test probe has strong self-similarity to its representation in memory and has moderate interitem similarity to a second study item from the same category. Its summed similarity would be high, so the observer would tend to recognize it as old. The lure test probe has moderate similarity to two study items from its category, although it is not an exact match to either. It would yield at least a moderate summed-similarity signal, so observers might false-alarm to this type of lure.

Numerous previous studies have been reported that illustrate that the type of summed-similarity exemplar model described above can indeed yield very accurate predictions of individual-item hit and false-alarm rates in recognition-memory experiments (e.g., Nosofsky, 1988, 1991; Nosofsky et al., 2011; Shin & Nosofsky, 1992). A major limitation of the original versions of such models, however, is that they do not predict any variations in self-similarity. In the standard versions of the models, exemplars are represented as points in a multidimensional psychological space, and similarity is a decreasing function of distance in the space. Because the distance between an exemplar and itself is always zero, all objects were presumed to have identical self-similarities. For this reason, the standard versions of the model fail in some cases to capture crucial variation in hit rates among items. One such example is a well-known phenomenon from the face-recognition literature in which face targets with highly distinctive features (such as scars) often have higher hit rates than do more typical faces (e.g., Busey & Tunnicluff, 1999; Valentine & Ferrara, 1991). Because distinctive faces are presumably located in isolated regions of multidimensional similarity space, standard summed-similarity exemplar models predicted that distinctive faces would have *lower* summed similarity than typical ones, so they failed to predict the hit-rate advantage for old distinctive faces.

To address this limitation, Nosofsky and Zaki (2003) proposed and tested a hybrid-similarity exemplar model (for related ideas, see Lee & Navarro, 2002). As in standard versions of the models, test probes are assumed to give rise to a global activation of memory via a summing of similarity to studied exemplars, where similarity is assumed to be a decreasing function of distance in a multidimensional similarity space. However, unlike in the standard versions, a key idea was that items with highly distinctive features may give rise to a greater degree of *self-match* than do items without such features (see below for formal details). This high degree of self-match can potentially dominate the summed-similarity signal, leading to high recognition hit rates for old items with such features. Nosofsky and Zaki (2003) demonstrated that the hybrid-similarity exemplar model provided good quantitative accounts of individual-item old-new recognition in experiments in which artificial distinctive features were manually added to highly controlled stimuli embedded in low-dimensional similarity spaces. In one experiment, for example, they used color patches varying in lightness, saturation and hue, and added highly distinctive punctuation marks to a small subset of the to-be-remembered items. Hit rates increased dramatically for the items with the highly distinctive features, and the hybrid-similarity model accounted in quantitative detail for a variety of intricate effects in the old-new recognition data observed in the experiments.

Despite the successes briefly reviewed above, a fundamental limitation of experimental tests of the cognitive-modeling approach to individual-item memorability is that the experiments all involved use of highly simplified perceptual stimuli and artificial category structures, rather than the real-world, high-dimensional objects that are the focus of modern work. To begin to address this limitation, in recent research Nosofsky and Meagher (2022) and Meagher and Nosofsky (2023) applied the hybrid-similarity exemplar model to predict individual-item old-new recognition for a set of rock-image stimuli. They used multidimensional-scaling methods (Lee, 2001; Shepard, 1980) to embed the rock stimuli in a high-dimensional similarity space. In addition, they collected ratings of the extent to which individual rock images contained highly distinctive features that made them stand out from other items in the set. Combining these sources of information, Nosofsky and Meagher (2022) and Meagher and Nosofsky (2023) were able to apply the hybrid-similarity exemplar model to account reasonably well for hit- and false-alarm rates associated with the individual rock images in their experiments.

Motivation for the Present Study

The main purpose of the present work was to pursue Meagher and Nosofsky's (2023) recent investigations of the hybrid-similarity exemplar model in this real-world high-dimensional rocks domain with still more challenging tests. A critical new question concerned the extent to which the model could capture the *context-dependent* nature of memorability. We conducted two experiments in which participants first studied a list of rock images organized into categories and were then tested on their ability to judge whether test probes from the categories were old or new. Following previous research, a key variable that we manipulated was category size: some categories were composed of large numbers of studied items and other categories were composed of small numbers of studied items (see also Konkle, Brady, et al., 2010). In addition, due to built-in intrinsic variation in the stimulus set, the studied categories also varied in their degrees of within- and between-category similarity. Finally, the rock images varied in the extent to which they contained distinctive features that were likely to make them stand out from other objects in the set. Each of these variables provides an example of forms of study-list context that we hypothesize are likely to strongly influence memorability. The goal was to test the ability of the hybrid-similarity model to capture quantitatively how old-new recognition performance varied as a function of all these variables. As we noted earlier, such models have been demonstrated to capture increases in false-alarm rates with increasing category size because increasing the number of similar studied items increases the summed similarity between the probe and the contents of memory (Hintzman, 1988; Nosofsky et al., 2011; Shiffrin et al., 1995). However, to our knowledge, no models have simultaneously accounted for the joint effects of category size, within- and between-category similarity, and presence of distinctive features on variations in memorability for individual old and new items.

We organize the remainder of our article as follows. First, we report the detailed method of Experiment 1. Next, we provide a presentation of the formal hybrid-similarity exemplar model itself, because the experimental results are best understood within the framework of the model. We then report the empirical and modeling results from Experiment 1. To anticipate, the hybrid model captures numerous aspects of the old-new recognition data quite well, including the effects of category size, within- and between-category similarity, and presence of distinctive features on overall hit and false-alarm rates. Nevertheless, it falls short in the goal of predicting fine-grained differences in the hit rates of the individual items. In an attempt to address this limitation, we then explore a variety of extensions of the core version of the hybrid-similarity model. These extensions formalize a variety of processes that may influence old-new recognition, but that are not included in the core model. Examples of some of the processes include potential roles of: i) differential attention to diagnostic dimensions, ii) contributions of judgments of category goodness to memory judgments, and iii) the role of output interference on old-new recognition. Although some of these extensions yield significant improvements in the model's overall account of the data, even these extended models fall short in our goal of predicting the detailed individual-item hit rates.

Finally, to further investigate the individual-item hit-rate issue, we conduct a second experiment that is closely related to the first, but that uses a different encoding task during the study phase. As explained below, in Experiment 1 participants engaged in a category-matching task during the initial study phase. By comparison, in Experiment 2 they simply observed the set of study items and were instructed to try to remember them. As it turned out, the modified encoding task resulted in substantially worse overall recognition performance in Experiment 2 compared to Experiment 1, and the formal modeling suggests that participants gave far less weight to the distinctive features of the rock images in recognizing old items. The overall differences in the patterns of results across the two experiments then motivate us to conduct modeling analyses on subgroups of participants across the two experiments who behaved similarly to one another, and we use the model to capture the basis for these individual differences. The results suggest that some participants relied heavily on distinctiveness in recognizing old target items, whereas other participants were far less sensitive to this variable.

In sum, we will see both some impressive successes of the hybrid-similarity model in capturing the data across our experiments, but also limitations that point to directions of future research. We conclude with a discussion of such future directions for potentially improving the ability of the model to predict the context-dependent nature of individual-item memorability: Scientific advancement entails not only demonstrations of formal-modeling success but clear acknowledgement of places where the formal models are falling short.

Experiment 1

Design

The stimulus materials were a set of 240 rock images organized into 24 categories of 10 samples each. Most of the categories were ones defined in the geologic sciences. However, to strengthen the intended category-size manipulation, we thought it important to use category structures in which within-category similarities tended to clearly exceed between-category similarities. This structural constraint is not always satisfied for geologically-defined categories (e.g., Nosofsky, Sanders, Gerdom, et al., 2017). Hence, in constructing our stimulus set, we did not include outlier samples that we judged to be more similar to members of contrast categories than to members of their own category. In addition, in several cases we created our own "psychological" rock categories (see Method section for details). A listing of the set of 24 rock categories along with a set of descriptors for the categories used in the experiment is provided in Table 1. Illustrations of the complete set of rock images and their organization into the categories are available at https://osf.io/a958r/.

Each subject saw a single list of 75 study items. For each subject, each of the 24 categories was randomly assigned to a different category-size condition. Within the single study list, there were five size-1 categories, 5 size-2, 5 size-4, and 5 size-8 categories. In addition, there were 4 size-0 categories that didn't appear on the study list. For each individual subject, the rock-image samples chosen for study in each size condition were chosen randomly from each of the 24 categories. On each trial of the study phase, a rock image was presented on the computer screen together with a table listing the 24 category names along with a descriptor of each category (see Table 1). Subjects chose the category whose descriptor they believed best characterized the rock image. Following the response, corrective feedback was then provided. The purpose of this manipulation was to keep subjects engaged in the task and to potentially strengthen the intended category-size manipulation by having subjects encode the images in terms of the category descriptors. Following the study phase there was an 82-trial test phase in which we presented two old targets and two new lures from each of the categories (1 target and 1 lure from the size-1 categories; and 2 lures from the size-0 categories) and subjects judged whether each test probe was old or new. The order of presentation of the test items was fully randomized for each subject and no corrective feedback was provided.

Fitting the hybrid-similarity exemplar model to the data requires that it be provided with an input space that specifies the coordinates of each object along a set of psychological dimensions. As in previous work (e.g., Nosofsky et al., 2018), we used multidimensional scaling (MDS) techniques based on the modeling of similarity-judgment data to derive this space. The details of this new similarity-scaling work are reported in Appendix A. As in previous work (e.g., Meagher & Nosofsky, 2023; Nosofsky, Sanders, & McDaniel, 2018), based on a combination of overall fit and interpretability of the derived dimensions, we settled on use of an 8-dimensional scaling solution for the present project (see Appendix A for details).

Recall that in the original work that tested the hybrid-similarity exemplar model, Nosofsky and Zaki (2003) experimentally manipulated the presence of distinctive features by manually adding the distinctive features to a highly simplified set of artificial stimuli. As a proxy for the presence of analogous types of features in the present set of images of naturally occurring, real-world objects, an independent group of subjects provided a set of "distinctiveness" ratings (see also Meagher & Nosofsky, 2023). The subjects were instructed to judge, on a scale from 1 to 9, the extent to which each item possessed a highly distinctive feature not present in the other items in the set (see Appendix B for details). The mean distinctive feature ratings for the rock images will be used as a second fundamental form of input to the hybrid-similarity model.[1] The top row of Figure 2 provides examples of rock images that received high distinctive-feature ratings, and the bottom row provides examples of rock images that received low distinctive-feature ratings.

Importantly, by making reference to the independently derived MDS solution for the rock images and to the set of distinctive-feature ratings, our goal will be to apply the hybrid-similarity exemplar model to the old-new recognition data with a minimum of parameter estimation.

Method

Subjects

This study was approved by the Indiana University Institutional Review Board, protocol #21800. The study was not preregistered. The subjects were 228 undergraduates from Indiana University who received credit towards an introductory psychology-course requirement. All subjects reported having normal or corrected-to-normal vision and normal color vision. As an overall measure of individual-subject performance, we computed for each subject their hit-minus-false-alarm rate [P(H)-P(FA] during the test phase. Based on visual inspection of a histogram of the individual-subject P(H)-P(FA) values, we judged that subjects with P(H)-P(FA) values less than 0.10 were poor-performing outliers and we decided to exclude these subjects from our subsequent formal-modeling analyses. This led us to exclude 25 subjects, leaving 203 for the formal analyses. This sample size yields power 0.999 for detecting a medium-sized effect (d = 0.5) in within-subject statistical analyses with α=.05.

Stimuli and Apparatus

The overall structure of the stimulus set was described in the Design section. Recall that we constructed some categories to be “psychological” ones that did not correspond to ones defined in the geologic sciences. Specifically, without use of physical and chemical tests, samples of shale and slate cannot be discriminated. Hence, we defined a “gray slate” category composed of gray samples of both slate and shale and a “colored shale” category composed of colored samples of slate and shale. Second, it is extremely difficult to discriminate between amphibolite and gabbro based solely on visual inspection. Hence, we used light-shopping techniques to make dark gray in color all samples of gabbro and added a blue tint to all samples of amphibolite.

The stimuli were presented on a 23-inch LCD computer screen. The stimuli were displayed on a white background. Each rock picture was approximately 2.1 in. wide and 1.7 in. tall. Participants sat approximately 20 in. from the computer screen, so each rock picture subtended a visual angle of approximately 6.0^{o} x 4.9^{o}. Although we cannot precisely estimate the camera-distance of each individual rock picture, Nosofsky et al. (2018) had selected all pictures in consultation with an expert in geology education such that they made clearly visible the salient characteristic features of the rocks. The experiments were programmed in MATLAB and the Psychophysics Toolbox (Brainard & Vision, 1997). All subjects were tested individually in private, sound-attenuated cubicles.

Procedure

The general procedure was described in the Design section. On each trial of the study phase, a rock image appeared toward the center-right of the screen, and the listing of the 24 numbered categories along with their descriptions appeared on the left of the screen. The computer keys were labelled with responses 1-24 and on each trial the subject pressed the key that indicated their category response. These responses were self-paced. Following the response, the rock image and category-description table remained on the screen, and corrective feedback was presented for 3 s. The order of presentation of the 75 study images was fully randomized for each individual participant. The test phase followed shortly after the study phase. The test list was constrained such that one of the old targets from each category was always the first one from that category that had been presented on the study list; the second old target from each category was chosen randomly. We observed no effect of category-based serial position in the data and do not discuss this variable further. Both lures from each category were randomly sampled from the available items that had not served as study-list items.

Subjects were instructed that on each trial a rock image would be presented and that they should inspect it carefully. If they judged that the rock was an old one that had been presented during the study phase, then they should press the OLD key (J) on the keyboard; whereas if they judged that the rock was a new one that was not presented during the study phase, they should press the NEW key (F) on the keyboard. Subjects were asked to make their old-new judgments in less than 7 s and a warning screen appeared if they did not meet this requirement. The order of the 82 test-phase items was fully randomized for each individual subject.

All computer code related to the conducting of the experiment and the formal-modeling analyses is available from the first author upon request.

Hybrid-Similarity Exemplar Model

In this article we focus on a core version of the hybrid-similarity exemplar model that uses a minimum of parameter estimation. According to the model, each test item *i* is presumed to give rise to a familiarity signal F_i. Following Nosofsky (1988), the probability that the test item *i* is judged to be "old" is then given by

$$\mathrm{P(Old}|i) = F_i / (F_i + k), \qquad (1)$$

where *k* is a response-criterion parameter.

Here, familiarity F_i is computed by summing the hybrid-similarity (η_{ij}) of test item *i* to each of the individual study exemplars *j*:

$$\mathrm{F_i} = \sum_j \eta_{ij} \qquad (2)$$

These hybrid-similarities (η_{ij}) are computed as follows. First, the continuous-distance of *i* to *j* is computed from the MDS solution derived for the exemplars (see Appendix A) by using a Euclidean metric,

$$d_{ij}=[\sum|x_{im} - x_{jm}|^2]^{1/2}, \qquad (2)$$

where x_{im} is the value of item *i* on dimension *m*. Following Shepard (1987), the continuous-dimension similarity of item *i* to exemplar *j* (s_{ij}) is then assumed to be an exponential decay function of the distance between *i* and *j*,

$$s_{ij} = \exp(-c \cdot d_{ij}), \qquad (3)$$

where *c* is an overall sensitivity parameter that describes the rate at which similarity declines with distance.

The hybrid similarities (η_{ij}) are computed by multiplying the continuous-dimension similarities by factors associated with matching or mismatching distinctive features. The hybrid-similarity ideas are motivated by Tversky's (1977) classic feature-contrast model of similarity, which assumes that common features boost similarity (including self-similarity); and that features that differ between two items reduce similarity. Thus, the self-similarity of item *i* to itself is assumed to be *boosted* by the factor $\exp(\beta \cdot \delta_i)$, where β is a freely estimated self-match scaling parameter and δ_i is the mean distinctive-feature rating for item *i*. The intuition, for example, is that a rock with a highly distinctive feature such as a fossil is likely to produce a strong match to its trace in memory. Note that because the continuous-distance between an item and itself is always equal to zero (Equation 2), in the purely continuous model the self-similarity of an item to itself is always equal to one (Equation 3). Thus, in the hybrid-similarity model, the hybrid self-similarity of item *i* to itself is equal to

$$\eta_{ii} = \exp(\beta \cdot \delta_{\mathrm{i}}) \cdot s_{ii} = \exp(\beta \cdot \delta_{\mathrm{i}}) \qquad (4)$$

By contrast, for the present paradigm, the hybrid-similarity of item *i* to a mismatching exemplar *j* is assumed to be *reduced* by the factor exp(-α·δ$_{\mathrm{i}}$), where α is a freely estimated mismatch-scaling parameter. For the present paradigm, our assumption is that to the extent that item *i* is judged to have a highly distinctive feature, it is unlikely that the feature would be shared by some mismatching exemplar in the stimulus set, so the natural assumption is that the distinctive feature would tend to reduce similarity between item *i* and a mismatching exemplar *j*.[2] Thus, for this paradigm, the hybrid-similarity of item i to mismatching exemplar j is given by[3]

$$\eta_{ij} = \exp(-\alpha \cdot \delta_{\mathrm{i}}) \cdot s_{ij} = \exp(-\alpha \cdot \delta_{\mathrm{i}}) \cdot \exp(-c \cdot d_{ij}). \qquad (5)$$

As noted above, the factors β and α in the above hybrid-similarity equations are freely estimated scaling parameters that provide measures of the extent to which distinctive features boost self-similarity and reduce inter-item similarity, respectively. An important special-case of the model arises when $\alpha = \beta = 0$, in which case the hybrid model reduces to the purely continuous-dimension summed-similarity exemplar model tested by Nosofsky (1988; Shin & Nosofsky, 1992; see also Nosofsky, 1991) in early investigations. We will refer to this special-case version as the “baseline” model; it is a 2-parameter model that uses only the response-criterion parameter *k* (Equation 1) and the sensitivity-parameter *c* (Equation 3). We will refer to the hybrid-similarity extension described above as the “core” model; it is a 4-parameter model that estimates *k*, *c*, and the distinctive-feature match and mismatch parameters: β (Equation 4) and α (Equation 5).

Experimental and Formal Modeling Results

Model-Fitting Method

We fitted the baseline and core versions of the model to the *individual-trials* data of the individual participants by using a maximum-likelihood criterion. As one means of comparing fits of models with differing numbers of free parameters, we use the Bayesian Information Criterion (BIC; Schwarz, 1978), given by

$$\text{BIC} = -2\ln(L) + P*\ln(N), \qquad (6)$$

where *L* is the (maximum) likelihood of the data given the model, *P* is the number of free parameters in the model, and *N* is the number of data observations on which the fit is based. The latter term on the right side of Equation 6 is a penalty term that penalizes a model for its number of free parameters. The model that achieves a smaller BIC is considered to provide a more parsimonious account of the data.

As noted above, the core model makes use of only 4 free parameters. For simplicity, in the main body of our article, we report results in which these parameters were held fixed across all subjects. The purpose is to demonstrate that even a very low-parameter version of the hybrid model captures many of the fundamental phenomena that we observed in the experiment. We fitted the model to the data by using the Hooke and Jeeves (1961) parameter-search algorithm and used multiple starting configurations in the parameter searches. We also conducted hierarchical Bayesian modeling to account for individual differences across participants. The hierarchical Bayesian modeling is reported in Appendix C. An advantage of the hierarchical Bayesian approach is that it allows the model to account for individual differences, but at the expense of requiring the estimation of a very large number of free parameters. The posterior predictions of the hierarchical Bayesian models for the aggregate data were essentially identical to the ones we report in the main body of the article and none of our conclusions were changed.

We emphasize that in the ensuing presentation of summary results and predictions, the 4 parameters of the core model are held fixed across all summary data sets. For purposes of comparison we also show the best-fitting predictions from the baseline model in which $\alpha = \beta = 0$. The maximum-likelihood parameters from the core and baseline versions of the model are reported in Table 2.

In Table 3 we provide a listing of summary fit statistics for the core and baseline models. The various summary-fit statistics are discussed below. At the outset, we note that the core model provides a substantially better BIC fit to the individual-trials data of the participants than does the baseline model (see Table 3). We will see a number of reasons for this superiority of the core model compared to the baseline model in our ensuing presentation.

Survey of Findings

Organization of Model-Fitting Results. We now proceed through a survey of summary empirical findings along with the formal-modeling accounts of those findings. For each main finding, we illustrate predictions from both the baseline and core versions of the model to clarify cases in which the distinctive-feature scaling parameters are and are not contributing to the accounts of the data. Summary statistics related to each main finding are reported in Table 3. The summary statistics are the negative log-likelihood (-lnL), which provides a measure of the absolute fit of the model to the individual-trials data; the BIC, which provides a measure of penalty-corrected fit; and listings of the proportion of variance accounted for (R^2) as well as the correlation (r) between the observed and predicted probabilities for each summary finding. Separately reporting R^2 and r is important because there are some cases in which r is high (suggesting that the model is providing a good account of the pattern in the data); but in which R^2 is low (because the model systematically over-predicts or under-predicts the absolute magnitudes of the observed probabilities).

Category-Size Effects. Figure 3 shows that false-alarm rates for new lures increased dramatically with increases in category size. This trend is well predicted by both the baseline and core models, although there is slight over-prediction of false-alarm rates at small category sizes and slight under-prediction at large category sizes. The observed trend is consistent with past studies that have also manipulated category size in old-new recognition experiments. The explanation in terms of the models is straightforward: As category size increases, summed similarity of novel test probes to the study items increases, so false-alarm rates increase. The figure also shows that there was a much smaller increase in hit rates with increases in category size, a trend that is again well captured by both models. The models predict smaller increases in hit rates than in false-alarm rates because the perfect match of an item to itself dominates the summed-similarity signal, so there is a relatively smaller contribution from the summed inter-item similarities.[4] Clearly, the models also capture the overall magnitude of the hit and false-alarm rates for the different category-size conditions.

Hit and False-Alarm Rates for Individual Categories. In Figure 4 we plot the observed against predicted false-alarm and hit rates for each of the individual 24 categories themselves, averaged across the different size conditions. Both models do a good job of capturing the false-alarm rates associated with the 24 categories (r = 0.84). Apparently, some categories have greater degrees of within-category similarity than do others, and the greater the degree of within-category similarity, the greater is the summed similarity. There is not nearly as much variation in the hit rates as for the false-alarm rates in these data. Nevertheless, for the baseline model, the correlation between predicted and observed hit rates is r = 0.00, whereas for the hybrid model the correlation is r = 0.50. Thus, without making use of the distinctive-feature scaling parameter β, the summed-similarity exemplar model fails to account for any of the variance in the category-level hit rates. Apparently, some categories contain members with more distinctive features than do other categories, and the mechanisms in the hybrid model allow it to account for the higher overall hit rates observed in those categories.

d' Values for the Individual Categories. In Figure 5 we plot the observed-against-predicted category-level *d'* values for each of the 24 categories, which provide a rough measure of the extent to which observers could discriminate the old members of each category from the new members. Both models do a good job of capturing this measure of performance, but with a quantitative advantage for the hybrid model (see Table 3). Note that to capture cases in which the *d'* value is high, the model needs to simultaneously predict a relatively high hit rate and a relatively low false-alarm rate for the category. According to the hybrid model, this tends to occur when the items in the category contain highly distinctive features: the self-matches of target items on their distinctive features will tend to boost their hit rates, but the mismatches of lure items on their distinctive features will tend to reduce their false-alarm rates.

False-Alarm Rates for the Size-0 Categories. Figure 6 plots the observed-against-predicted false-alarm rates for lures from the size-0 categories. For both versions of the model there is a significant correlation between the observed and predicted false-alarm rates (mean r = 0.57), but both versions systematically over-predict the overall magnitude of these false-alarm rates. One reason why the correlation arises is because there are differing degrees of between-category similarity across the 24 categories: Items from size-0 categories that are similar to members of *other* studied categories will tend to have higher false-alarm rates. We discuss below possible reasons why the current versions of the model tend to over-predict the overall magnitude of the false-alarm rates for the size-0 categories.

Distinctiveness-Bin Analysis. In our next analysis, we divided the 240 individual rock images into four equally sized distinctiveness-rating bins and computed the observed and predicted mean hit rates and false alarm rates for each bin. The results are shown in Figure 7. As can be seen, the mean hit rates increase with increases in rated distinctiveness, whereas the false-alarm rates decrease. To confirm the statistical significance of these observations, we computed for each individual subject the least-squares regression line relating the choice probabilities to the bin number (1-4). For the hit rates, the mean regression slope was significantly greater than zero (m_H = 0.027, $t(202)$ = 6.59, p < .0001); whereas for the false-alarm rates, the mean regression slope was significantly less than zero (m_{FA} = -0.052, $t(202)$ = -9.84, p < .0001). As can be seen from inspection of the right panel of the figure, the hybrid-similarity model does an excellent job of capturing both effects (r_H = 0.95, r_{FA} = 0.92). It predicts the increase in hit rates because it formalizes the idea that items with highly distinctive features produce stronger matches to their traces in memory (Equation 4). By contrast, as can be seen in the left panel of Figure 7, when the distinctive-feature match parameter β is held fixed at zero, then the baseline model fails to predict any increase in hit rates as a function of rated distinctiveness ($r_H = 0.00$). We note as well that the hybrid model also correctly predicts the decreasing false-alarm rates with increases in rated distinctiveness. It does so for two reasons: First, those items with high values of rated distinctiveness tend to lie in isolated regions of the derived MDS space for the rocks, resulting in low values of summed similarity. Second, the presence of a highly distinctive feature reduces even more the similarity of a lure item to the studied exemplars (Equation 5).

Individual-Item Hit and False-Alarm Rates. Finally, in an ambitious analysis, we computed the observed and predicted false-alarm and hit rates associated with the individual 240 items, averaged across the different category-size conditions of the experiment. Sample size for each individual item is relatively small because each participant was tested with only a small subset of items from each category and each tested item was presented only once; therefore, these results need to be interpreted with caution. As shown in Figure 8, both versions of the model capture much of the variation in the individual-item false-alarm rates ($r = 0.63$). However, in line with the previous results we have reported, the baseline model fails to capture any of the variability in the hit rates ($r = 0.02$). The hybrid model is starting to capture some of the individual-item variation in the hit rates ($r = 0.32$), but it too is clearly falling far short in meeting this goal.

To gain some sense of a “noise ceiling” associated with the individual-item data, we conducted the following analysis. In each of 1000 simulations, we randomly divided the subjects into two equal-sized groups. For each group, we then computed the false-alarm rate for each of the 240 items and the hit rate for each of the 240 items. We then computed Spearman-Brown-corrected split-half correlations for the resulting false-alarm and hit rates. Across the 1000 simulations, the average Spearman-Brown corrected correlation was r = 0.851 for the false-alarm rates and r = 0.562 for the hit rates. Thus, there appears to be clear room for improvement with respect to predicting performance at the individual-item level.

Extended Models

In an effort to improve the hybrid model's account of the individual-item data, we explored a wide variety of extensions of the core model. One class of extensions made allowance for the possibilities that psychological familiarity may be nonlinearly related to summed similarity or that psychological distinctiveness may be nonlinearly related to judged distinctiveness on the rating scale. A second class of extensions made allowance for more flexible computations of interitem similarity owing to selective-attention weighting of the psychological dimensions or to the possibility that certain dimensions that are diagnostic of classification are missing from the MDS solution derived from the similarity-judgment data. Another extension formalized how output interference during the test phase may have impacted the participants' old-new recognition judgments. And in a final extension, we considered the possibility that old-new recognition in the present paradigm may be mediated not only by a global-familiarity signal but also by the extent to which the individual test probes were judged to be good or bad examples of their respective categories.

As we will report, some of these extensions led to substantial improvements in the BIC fit of the model to the complete set of individual-trials data. However, it turned out that none yielded major improvements in the fit to the individual-item hit-rate data. Instead, depending on the extension, the improvements could be traced to improved fits to the individual-item false-alarm rates; improved fits to the category-size effects; or improved fits to how the overall pattern of hit rates and false-alarm rates evolved as the test-phase progressed. Because our main interest in these extended-model investigations involved a search for an improved account of the individual-item hit rates, and none achieved that goal, we have decided to place the detailed report of the extended-model investigations in Appendix E.

Discussion

Using a minimum of parameter estimation, the core version of the global-familiarity hybrid-similarity exemplar model was able to provide good accounts of a wide variety of benchmark phenomena involving the context-dependent nature of memorability in this rock-category domain. The phenomena included overall effects of category size, within- and between-category similarity, and the presence of distinctive features on both category-level false-alarm rates for lures and hit rates for old study items. By comparison, a baseline version of the exemplar model that did not take account of distinctive-feature information failed to capture any of the variability in the hit rates across categories or individual items.

Although the hybrid model took significant steps in the right direction, it nevertheless fell short in providing a satisfactory account of the *individual-item* hit rates and false-alarm rates. In other words, although the model succeeded in capturing fundamental benchmarks in various forms of averaged data, it fell short in capturing the detailed variability in the individual-item data. Furthermore, we explored a number of extended versions of the core model in the hope of improving the fits to those data. Although some of these extensions yielded markedly improved overall BIC fits and improved fits to the individual-item false-alarm rate data, the predictions of the individual-item hit rates remained essentially unchanged (see Appendix E).

To investigate these issues further, we decided to conduct a second experiment using the same materials and the same design structure as in Experiment 1. An important aspect of our Experiment-1 procedure was the category-description encoding task in which the participants engaged during the study phase. Various investigators have argued that engaging in categorization or analogical reasoning is likely to have an important influence on the types of memory representations that are formed (Davis, Love, & Preston, 2012; De Brigard, Brady, Ruzic, & Schacter, 2017; Ichien, Alfred, Baia, et al., 2023; Nosofsky, 1991; Sakamoto & Love, 2004). Although one of our extended models incorporated category-goodness judgments in its machinery, it is possible that the category-description encoding task influenced the memory representations and recognition judgments in a manner that was not well captured by this extended model.

Therefore, in Experiment 2 we decided to use a more neutral encoding task in which participants simply observed the study items one by one and were instructed to try to do their best to remember them. The instructions made no mention that the rock images were organized into categories. One question of interest was whether we would still observe the same strong category-size effects under these alternative study conditions as we observed in Experiment 1. A second question was whether with these alternative study instructions the core model might yield a better account of the individual-item hit-rate and false-alarm-rate data.

To preview, although we once again observed strong category size effects that were well described by the model, and the model continued to provide good accounts of a variety of other benchmark results, its ability to account for the variability in the individual-item hit-rate data was even worse than in Experiment 1. Moreover, overall recognition performance was substantially worse than in Experiment 1, and we did not observe an effect of mean rated distinctiveness on the individual-item hit rates. We will suggest that the latter two results are likely to be interconnected and we will use the hybrid-similarity model to characterize these interrelated effects.

Experiment 2

Design

The design of Experiment 2 was the same as that of Experiment 1, except for the changed study-phase instructions.

Method

Subjects

The subjects were 312 undergraduates from Indiana University who received credit towards an introductory psychology-course requirement. All subjects reported having normal or corrected-to-normal vision and normal color vision. Recall that, based on visual inspection of the histogram of individual-subject P(H)-P(FA) values, we had excluded from analysis 25 of 228 subjects from Experiment 1 who were outliers on this performance measure. The excluded subjects constituted .1096 of the Experiment-1 sample. Visual inspection of the Experiment-2 histogram of P(H)-P(FA) values did not reveal a clear breakpoint of outlier subjects. Because we were interested in making performance comparisons across the two experiments, we decided to exclude the same proportion of low-performing subjects from our formal analyses of the Experiment-2 data. This led us to exclude 36 subjects whose P(H) – P(FA) scores were less than .04, which is .1154 of the sample (the proportions didn’t match exactly across the two experiments because of ties in the data).   The resulting sample size of 276 analyzed subjects yields power 0.999 for detecting a medium-sized effect (*d* = 0.5) in within-subject statistical analyses with α=.05.

Stimuli, Apparatus, and Procedure

The stimuli, apparatus, and procedure were the same as in Experiment 1, except for the changed study-phase instructions. The participants were provided with neutral instructions that they would be viewing a series of 75 rock images and that they should observe the rocks carefully and try to remember them. They were also informed that after the study phase, they would be tested on their memory for the rocks. The instructions provided no information that the rock images were organized into different categories.

Experimental and Formal Modeling Results

Summary Fits

The maximum-likelihood parameters from the core hybrid model and the baseline model are reported in Table 2. The summary fits of the core and baseline models are reported in Table 4.

Whereas the core hybrid model had yielded a substantially better BIC fit than the baseline model to the Experiment-1 data, the improvement in fit to the Experiment-2 data is modest. In further contrast to the results from Experiment 1, inspection of the best-fitting free parameters in Table 2 reveals that in Experiment 2 the self-match scaling parameter of the core model takes on a value near zero, β = 0.019. In addition, inspection of the various summary-fit statistics for the different forms of summary data reveals little difference between the core and baseline models.

For simplicity, in our ensuing survey of experimental results, we will show the predictions from only the core model; the predictions from the baseline model were extremely similar in all cases.

Survey of Findings

Category-Size Effects. Figure 9 shows that false-alarm rates again increased dramatically with increases in category size, with a smaller increase in hit rates. The model again does a reasonably good job of capturing these category-size effects, although it slightly underestimates their magnitude. The model also captures reasonably well the absolute magnitudes of the hit and false-alarm rates. Hence, prior knowledge that the rock images were organized into categories and the requirement of engaging in a categorization task during study was not needed for producing the robust category-size effects. Instead, they appear to arise from the similarity structure of the to-be-remembered items themselves.

Note that whereas the false-alarm rates in Experiment 2 were in roughly the same ballpark as in Experiment 1, there was a substantial drop in the hit rates (compare Figure 9 to Figure 3). The model accounts for the drop in the hit rates from Experiment 1 to Experiment 2 by virtue of changes in its best-fitting parameters (see Table 2). First, the value of the sensitivity parameter (c), which measures the overall ability of the observers to discriminate between old and new test probes, is lower in Experiment 2 than in Experiment 1. Second, the value of the distinctive-feature match parameter (β) takes on a near-zero value in Experiment 2.

We had not predicted a priori this pattern of reduced discriminability across the two experiments. However, a reasonable interpretation is that the changed encoding-task instructions of Experiment 2 led observers to process the stimuli less deeply than did the category-matching task of Experiment 1 (Craik & Lockhart, 1972). Shallower processing would result in worse overall ability to discriminate between old and new items and might also be expected to reduce observers' ability to make use of distinctive features for purposes of remembering old items. We will see converging evidence for the latter possibility in some of our ensuing analyses of the data.

Hit and False-Alarm Rates for Individual Categories. In Figure 10 we plot the observed against predicted false-alarm and hit rates for the 24 categories, averaged across the different size conditions. Once again, the model does a reasonably good job of capturing the variation across the category-level false-alarm rates, $r = 0.77$. However, its fit to the category-level hit-rate data is worse than in Experiment 1, $r = 0.32$. An explanation for the worse fit to the hit-rate data is that the presence of distinctive features apparently played a smaller role overall in influencing observers' memory for old items in Experiment 2 than it did in Experiment 1: In Experiment 1, the correlation between the mean category-level distinctiveness ratings and the mean category-level hit rates was $r = 0.542$, whereas in Experiment 2 this correlation was only $r = 0.077$. Because the presence of distinctive features apparently had a much smaller overall influence on category-level hit rates in Experiment 2, it is not surprising that the distinctive-feature self-match scaling parameter dropped to its near-zero value in Experiment 2 ($\beta = 0.019$).

d' Values for the Individual Categories. The plot of observed-against-predicted category-level *d'* values is provided in Figure 11. Given the lower hit rates in Experiment 2 compared to Experiment 1, it is not surprising that the overall *d'* values are much lower in the present experiment (compare Figure 11 to Figure 5). Nevertheless, the model continues to provide a good account of the variation in *d'* values across the 24 categories, $r = 0.819$.

False-Alarm Rates for the Size-0 Categories. Figure 12 provides a plot of the observed-against-predicted false-alarm rates for the size-0 categories. These results are extremely similar to the ones from Experiment 1 (compare Figure 12 to Figure 6). The predictions from the model are significantly correlated (r = 0.65) with the observed data, but the model tends to over-predict the magnitude of most of the size-0 false-alarm rates.[5]

Distinctiveness-Bin Analysis. Perhaps the most telling difference between the pattern of results across Experiments 1 and 2 is provided in Figure 13, which plots the mean hit and false-alarm rates as a function of the four distinctiveness-rating bins. On the one hand, as expected, false-alarm rates again decrease with increases in rated distinctiveness, and the mean regression slope (m_{FA} = -0.073) is significantly less than zero, $t(275) = -16.86$, $p < .0001$. By contrast, whereas hit rates had significantly increased with rated distinctiveness in Experiment 1, the function is basically flat in Experiment 2, and the mean regression slope (m_H = 0.005) is not significantly different from zero, $t(275) = 1.12$, $p > .10$. The Figure-13 results provide converging evidence for our earlier suggestion that the alternative encoding conditions of Experiment 2 led to less overall reliance on use of distinctive features for the purpose of recognizing old study items. As can be seen from inspection, the summed-similarity exemplar model provides good fits to the Figure-13 data. It predicts the decreasing false-alarm rates as distinctive-feature ratings increase for the same reasons as explained previously. Its prediction of a very-slightly decreasing function for the hit-rate data arises because of the near-zero setting of its self-match scaling parameter (β = 0.019) and the large value of its mismatching distinctive-feature scaling parameter (α = 0.089).

Individual-Item Hit and False-Alarm Rates. The predictions of the individual-item false-alarm and hit rates are shown in Figure 14. The correlation of the model's predictions with the false-alarm rate data is basically the same as in Experiment 1, $r = .65$; but the correlation with the individual-item hit-rate data is worse, $r = .19$. The Spearman-Brown corrected split-half correlations were *r* = .892 for the false-alarm rates and *r* = .623 for the hit rates. Thus, once again, the core version of the hybrid model fails to provide a satisfactory account of the individual-item data.

Extended Models

We again fitted the set of extended models to the individual-trials data. The BIC fits of the extended models as well as their fits to the various forms of summary data are reported in Appendix E. The results were similar to those reported for Experiment 1. Several of the extended models yielded quite improved BIC fits to the data. Depending on the specific extension, the locus of the better fits tended to involve improved accounts of the false-alarm-rate data or the output-interference effects. However, none yielded a satisfactory account of the individual-item hit-rate data.

Exploratory Subgroup-Modeling Analyses of the Experiments 1 and 2 Results

As argued above, one interpretation of the differing pattern of results across Experiments 1 and 2 is that the category-description encoding task of Experiment 1 led to greater overall use of distinctive features for purposes of recognizing old target items. However, individual-participant differences in cognitive processing might exist within each of the experiments as well. If this line of reasoning is correct, then we might be able to identify certain subgroups of participants who behaved similarly across the two experiments and the core model might provide a similar characterization of those participants' performance.

There is a wide variety of ways in which subgroups might be formed for this exploratory analysis. Here, we decided to use the distinctiveness-bins regression approach that we used to analyze the data from Figures 7 and 13. For Experiment 1, we defined Subgroup 1 to be the set of participants whose regression slopes for hits exceeded the mean slope of .027, whereas Subgroup 2 was composed of the participants whose regression slopes were less than .027. For Experiment 1, there were 102 participants in Subgroup 1 and 101 participants in Subgroup 2. We used the same cutoffs for forming the subgroups of Experiment 2: Among the 276 participants in Experiment 2, only 111 had regression slopes for hits greater than .027 (Subgroup 1), whereas 165 participants had regression slopes less than .027 (Subgroup 2).

We then fitted the core hybrid-similarity model to the data of each subgroup of each experiment, using the same model-fitting method as described previously. The maximum-likelihood parameters for the fit of the model to each subgroup of each experiment are reported in Table 5. Plots of observed-against-predicted data are shown in Figure 15 for the distinctiveness-bins analysis; and in Figure 16 for the category-level false-alarm and hit rates.

Inspection of Figure 15 reveals that for Subgroup 1 in both experiments, hit rates increased with increases in rated distinctiveness; whereas for Subgroup 2 in both experiments, hit rates decreased with increases in rated distinctiveness. This pattern is unsurprising given that the subgroups were selected with this criterion variable in mind. Nevertheless, the result suggests the possibility that although the different encoding-task instructions across the experiments influenced the proportion of participants who made use of distinctive features to recognize old items, a substantial number of participants behaved similarly in this regard across the experiments. Likewise, a substantial number of participants had in common the reverse mode of processing, in which the presence of a highly distinctive feature in a target did *not* promote its recognition. The parameter estimates reported in Table 5 are consistent with this interpretation. For Subgroup 1 of both experiments, the magnitude of the self-match distinctive-feature scaling parameter (β) is large in magnitude, whereas the estimate of the mismatch scaling parameter (α) is near zero. By contrast, Subgroups 2 of both experiments show the reverse pattern of distinctive-feature scaling-parameter estimates.

Inspection of Figure 16 reveals the interesting result that prediction of the category-level hit-rate data is improved when the model is applied to the subgroups data rather than to the data of each experiment as a whole (compare Figure 16 to Figures 4 and 10). For Subgroup 1 of Experiment 1, the correlation between the predicted and observed individual-category hit rates rises to r = 0.75 (compared to r = 0.50 when all the Experiment-1 participants were modeled as a whole). And for Subgroup 1 of Experiment 2, the correlation rises to r = 0.71 (compared to r = 0.32 when all the Experiment-2 participants were modeled as a whole). The improved predictions reflect that across both experiments, Subgroup 1 tended to have higher hit rates for categories that had members with highly distinctive features: the correlation between the category-level hit rates of Subgroup 1 across the two experiments was r = 0.88. The model captures this pattern by virtue of the high setting of the distinctive-feature self-match parameter for Subgroup 1. Interestingly, for Subgroup 2 of Experiment 2, the model's predictions also improve: the correlation between predicted and observed category-level hit rates rises to r = 0.61 (compared to r = 0.32 when all Experiment-2 participants were modeled as a whole). The reason for the improvement is different, however, than for the Subgroup-1 participants. For Subgroup 2 of Experiment 2, category-level hit rates tend to be higher for categories that have greater within-category similarity, not for categories that have members with highly distinctive features: the correlation between the category-level hit rates of the Experiment-2/Subgroup-2 participants and the Experiment-2/Subgroup-1 participants was only r = 0.19. Finally, as is apparent from inspection of the upper-right panel of Figure 16, there was very little variability in hit rates across the categories for Subgroup 2 of Experiment 1, so here the correlation between model-predicted and observed hit rates is low. In sum, although there remains significant room for further improvement, the core model's predictions of the hit-rate data improve noticeably when subgroups of participants within the experiments are identified who showed similar performance patterns.[6]

General Discussion

Summary and Interpretations

In this project we sought to provide challenging tests of a hybrid-similarity exemplar model for predicting the context-dependent nature of memorability in a rock-image domain. The domain seems to provide a good representative of a set of real-world natural categories involving the kinds of complex, high-dimensional stimuli that are the focus of other modern work in the memorability literature. An advantage of testing the model's performance in this domain is that extensive similarity-scaling work has already taken place to position the rock images in a high-dimensional psychological space (with additional scaling work having taken place in the present project). This derived psychological space serves as a fundamental source of input for applying the model to the prediction of individual-item old-new recognition data. By making reference to this derived MDS solution, as well as to a newly collected set of distinctive-feature ratings, we were able to apply the exemplar model to the prediction of old-new recognition in this real-world domain with a minimum of parameter estimation.

In short, we found that the hybrid-similarity model provided good quantitative accounts of a wide variety of benchmark experimental effects observed in our Experiment 1. These included good accounts of category-size effects on both false-alarm and hit rates and good predictions of the absolute magnitude of those category-size false-alarm and hit rates; reasonably accurate quantitative predictions of false-alarm and hit rates for the 24 individual categories themselves (averaged across the different category-size conditions); and accurate predictions of the measured category-level *d'* values across the 24 individual categories. In addition, the model provided a good overall account of how hit rates and false-alarm rates varied with changes in rated distinctiveness of the items. In particular, it predicted the sharply decreasing false-alarm rates that arose with increases in rated distinctiveness, yet the rising hit rates that arose with increases in this variable. When considered at the level of the 240 individual items in our stimulus set, the model did a fair job of accounting for the variation in the false-alarm rates, but fell short in its account of the individual-item hit rates. We formulated a wide variety of extended models to try to improve the core hybrid-similarity model's fit to the individual-item hit rates, but none provided a satisfactory account. We consider below future directions for improving the model's account of the individual-item recognition data.

Whereas in Experiment 1 we used a category-description task during the study phase of the recognition task, in Experiment 2 we provided neutral study instructions that made no mention of the existence of categories. Although we continued to observe robust category-size effects in the second experiment (that were again well captured by the model), overall recognition performance in Experiment 2 was substantially worse than in Experiment 1, mainly due to much lower hit rates. In addition, whereas there was a strong correlation between rated distinctiveness and the magnitude of the hit rates in Experiment 1, that correlation dropped precipitously in Experiment 2. A possible interpretation is that shallower depth of processing during the Experiment-2 study phase (e.g., Craik & Lockhart, 1972) led to less reliance on use of distinctive features for recognizing old items, which would help explain the greatly reduced hit rates observed in the second experiment.

A core idea in the present modeling approach is that the presence of unique distinctive features is assumed to boost self-similarity but to decrease interitem similarity. These effects result in an interesting dynamic with respect to the prediction of how hit rates may vary with the presence of distinctive features. Whether or not hit rates are predicted to increase depends on whether the boost to self-similarity exceeds the reduction in inter-item similarity in the overall summed-similarity computation. In Experiment 1 of the present article, we tended to see boosts in hit rates with presence of distinctive features, captured by the relatively large magnitude of the self-match distinctive-feature scaling parameter β in the model. By contrast, in Experiment 2, we tended not to see such boosts, and the estimate of the self-match scaling parameter was very small in magnitude. These interpretations were supported by follow-up modeling analyses in which we identified subgroups of participants across the two experiments who behaved similarly to one another. We applied the hybrid-similarity model separately to the subgroups. One subgroup appeared to take advantage of distinctive features in target probes to bolster their recognition of those probes; the performance of this subgroup was reasonably well captured by a version of the hybrid-similarity model in which the setting of the distinctive-feature self-match parameter was large in magnitude. A second subgroup did not have higher hit rates to target probes that had high distinctive-feature ratings; for this subgroup, the estimate of the distinctive-feature self-match parameter was small in magnitude. Future research is needed to test systematically how variations in study-phase instructions may impact participants' use of distinctive features for making their old-new recognition judgments, as well as to study the basis for the individual-participant differences in the use of this cognitive process.

In any case, just as occurred in Experiment 1, although the hybrid-similarity model provided good accounts of a wide variety of benchmark results in different forms of averaged data in Experiment 2, it again fell short in its quantitative fits to the variability in the individual-item hit-rate data. And, once again, the various extended models that we formulated did little to improve the model in this regard. In the remainder of our General Discussion, we discuss possible future research directions that might improve this current limitation of the hybrid-similarity model.

Limitations and Future Research Directions

Formalizing "Distinctiveness". The construct of "distinctiveness" is a classic one in the cognitive psychology of memory and extremely thoughtful theoretical accounts of the construct have been put forward in the literature (e.g., Craik & Jacoby, 1979; Craik & Tulving, 1975; Hunt, 1995; Hunt & McDaniel, 1993; Schmidt, 1991). Nevertheless, precise formalizations remain elusive. Nosofsky and Zaki's (1993) attempt at formalization was to suggest that distinctive items contain features that go "outside" a continuous-dimension similarity space in which most of the other items are embedded. There appear to be two different ways that this form of distinctiveness may arise and enhance memory. First, an item may be distinctive with respect to general-world context and the lifetime of experience that an observer brings with them to the laboratory experiment. This form of distinctiveness may play an important role in the intriguing studies that have demonstrated differential memorability of individual items embedded in very long study lists with no obvious category-based organization (e.g., Bainbridge, 2019; Bainbridge, Isola, & Oliva, 2013; Isola, Xiao, Parikh, et al., 2013; Khosla et al., 2015). But another form of distinctiveness arises within the context of the laboratory experiment itself depending on the precise composition of the to-be-remembered study lists.

A fundamental limitation of the current research approach is that it probably failed to account fully for this latter form of distinctiveness, because the extent to which an object is judged to have a "distinctive" feature can itself be a highly context-dependent phenomenon. For example, a feature is likely to be judged as highly distinctive in study contexts in which it rarely occurs but to be lacking in distinctiveness in study contexts in which it is common. In the present research, the distinctiveness ratings were obtained in a separate, generic context in which participants were exposed to the complete set of 240 items from which the experimental materials were sampled. But in the recognition experiment itself, any given subject was exposed to only a random subset of these items in which individual category sizes varied dramatically across different subjects. Hence, the ratings obtained in the distinctiveness-judgments task may have been a poor proxy for what individual participants experienced as psychologically distinctive features in the recognition experiment. For example, an item from a category with high within-category similarity may be unlikely to receive a high distinctiveness rating in the generic ratings experiment; however, it might have high psychological distinctiveness if it were the sole item presented from that category in the recognition experiment.

The above line of reasoning would help explain why Meagher and Nosofsky (2023) had previously demonstrated a case in which the hybrid-similarity model yielded quite good predictions of individual-item hit rates for the present types of rock images. In that previous study, distinctive-feature ratings were obtained for a set of 540 rock images, *all* of which were used as either study or test items in the independently conducted old-new recognition experiment. As was the case in Experiment 1 of the present study, the rock images were organized into categories and participants had engaged in a category-learning task during the study phase. Unlike in the present study, however, the size of all categories was the same and all participants experienced the same set of study items and the same set of target and lure test probes. Hence, there was likely a much greater match between the distinctive-feature rating-task context and the recognition-task context in that earlier study than in the present one.   To test this context-match hypothesis, in future research we plan to obtain distinctiveness ratings for the objects in contexts that come closer to matching the context that participants experience when they engage in the old-new recognition task itself. For example, for any given assignment of the size variable to the individual 24 categories in the recognition experiment, that same assignment could be made for a corresponding group of participants who provide distinctive-feature ratings.

A related limitation of the present research approach is that we relied on use of only generic ratings of presence of distinctive features as a source of input to the model. Although our instructions provided examples of the types of features that might be considered as distinctive ones (see Appendix B), there may be large individual differences in the interpretation of what constitutes a distinctive feature. In addition, across different domains, the presence of certain types of distinctiveness may have a strong influence on memory, whereas other forms of distinctiveness may not. For example, in their study of memorability for visual images belonging to everyday-object categories, Konkle et al. (2010) found that forms of low-level perceptual distinctiveness had a minor influence on memory performance, whereas forms of distinctiveness that gave rise to different conceptual kinds and sub-categories had major effects. Future research for testing computational cognitive models of memorability might probe these issues by asking the distinctiveness question from varied perspectives.

The conception advanced by Nosofsky and Zaki (2003) in proposing the hybrid-similarity exemplar model was that the distinctive features of objects lie outside the continuous-dimension space in which the objects are embedded, causing them to "stand out" from the other objects in the set. Using current psychological-scaling methods, the presence of such features is reflected only indirectly in the continuous-dimension MDS solutions for objects by locating such objects in more isolated parts of the continuous-dimension space. A crucial goal of future work is to instead develop techniques that can identify the specific distinctive features that play a role in influencing memorability and incorporate them as part of the feature space itself. A potentially fruitful route to achieving this goal may be to merge the impressive, modern deep-learning approaches to accounting for memorability (Bylinskii et al., 2022; Dubey et al., 2015; Khosla et al., 2015; Needel & Bainbridge, 2022) with the present type of cognitive-modeling approach. The deep-learning networks might be used to extract specific distinctive features that could then be used as a source of input to the cognitive model.

Alternative Cognitive-Modeling Approaches. Finally, in the present research we pursued the modeling of individual-object old-new recognition from the perspective of only a single cognitive model, namely the hybrid-similarity exemplar model. Other extremely influential cognitive models of old-new recognition have of course also been proposed in the literature, although it goes far beyond the scope of this particular research study to test these models as well. One such class of models supposes that recognition judgments are based on evidence that depends not only on "familiarity" but also on "recollection", where "recollection" is conceived of as being a process that involves the retrieval of specific details associated with the prior presentation of an item (for reviews and alternative approaches to formalizing such dual-process models, see, e.g., Wixted, 2007; Yonelinas, 2002). A second class of models springs from the

“Retrieving Effectively from Memory” (REM) framework of Shiffrin and Steyvers (1997), with a dynamic model of recognition memory that incorporates REM principles having been proposed recently by Cox and Shiffrin (2017). This framework advances a cognitive-Bayesian approach to modeling old-new recognition. According to REM, observers compute the likelihood that test probes are represented in a set of memory traces of old study items as opposed to being members of the class of “new” items, and judge that a test probe is old if this likelihood exceeds .5. Conceivably, there may be close psychological connections between cases in which the hybrid-similarity exemplar model predicts a high self-match similarity value; in which dual-process models posit that the recollection process is playing a major role; and in which REM predicts a high old-new likelihood-ratio associated with a specific memory trace. To our knowledge, however, the dual-process models and REM models have not yet been formalized to account for memorability and old-new recognition at the level of individual visual items embedded in high-dimensional psychological spaces. Such directions are also exciting ones for future research.

Constraints on Generality

The participants in this study were sampled from undergraduate students enrolled in introductory psychology courses at Indiana University who received credit towards a course requirement. This target population seems reasonable for these initial theoretical investigations that test the adequacy of the hybrid-similarity model to capture with quantitative accuracy the wide variety of fundamental benchmark phenomena involving the context-dependent nature of memorability examined here. Clearly, however, the study falls far short in testing for systematic differences across alternative populations of participants, which was not the goal of this initial theoretical investigation. Having obtained some preliminary support for the proposed theoretical approach, an important direction of future research is to test for systematic differences in the patterns of context-dependent memorability across alternative populations of participants and test the ability of the hybrid-similarity model to provide a theoretical characterization of any such observed differences.

Another constraint on generality is that we modeled context-dependent memory in only a single stimulus domain, namely that of rock categories. We believe that this domain was a sensible one for our initial investigation because we were able to take advantage of the extensive similarity-scaling work conducted in recent years for embedding the rock images in a high-dimensional feature space (Meagher & Nosofsky, 2023; Nosofsky et al., 2017, 2018, 2020; Sanders & Nosofsky, 2020). However, future work is needed for testing the model in alternative stimulus domains as well. Other investigators have made impressive progress in deriving high-dimensional psychological-scaling solutions for large numbers of objects from more everyday-object categories (e.g., Hebart, Zheng, et al., 2020; Hout, Goldinger, & Brady, 2014), so extending the tests of cognitive models of memorability to such domains is a promising direction as well.

References

Bainbridge, W. A. (2019). Memorability: How what we see influences what we remember. In *Psychology of Learning and Motivation—Advances in Research and Theory* (Vol. 70, pp. 1–27). Academic Press Inc. https://doi.org/10.1016/bs.plm.2019.02.001

Boehm, U., Marsman, M., Matzke, D., & Wagenmakers, E. J. (2018). On the importance of avoiding shortcuts in applying cognitive models to hierarchical data. *Behavior research methods*, *50*, 1614-1631.

Brainard, D. H., & Vision, S. (1997). The psychophysics toolbox. *Spatial vision*, *10*(4), 433-436.

Busey, T. A., & Tunnicliff, J. L. (1999). Accounts of blending, distinctiveness, and typicality in the false recognition of faces. *Journal of Experimental Psychology: Learning, Memory, and Cognition*, *25*(5), 1210.

Bylinskii, Z., Goetschalckx, L., Newman, A., & Oliva, A. (2022). Memorability: An image-computable measure of information utility. *Human Perception of Visual Information: Psychological and Computational Perspectives*, 207-239.

Colloff, M. F., Wade, K. A., & Strange, D. (2016). Unfair lineups make witnesses more likely to confuse innocent and guilty suspects. *Psychological Science*, *27*(9), 1227-1239.

Cox, G. E., & Shiffrin, R. M. (2017). A dynamic approach to recognition memory. *Psychological review*, *124*(6), 795.

Craik, F. I. M., & Jacoby, L. L. (1979). Elaboration and distinctiveness in episodic memory. In L. Nilsson (Ed.), *Perspectives on memory research: Essays in honor of Uppsala University's 500th anniversary* (oo, 145-166). Hillsdale, NJ: Erlbaum.

Craik, F. I. M, & Lockhart, R. S. (1972). Levels of processing: A framework for memory research. *Journal of Verbal Learning and Verbal Behavior*, *11*, 671-684.

Craik, F. I., & Tulving, E. (1975). Depth of processing and the retention of words in episodic memory. *Journal of experimental Psychology: general*, *104*(3), 268.

Criss, A.H., Malmberg, K. J., & Shiffrin, R. M. (2011). Output interference in recognition memory. *Journal of Memory and Language*, *64*, 316-326

Davis, T., Love, B. C., & Preston, A. R. (2012b). Striatal and hippocampal entropy and recognition signals in category learning: simultaneous processes revealed by model-based fMRI. *Journal of Experimental Psychology: Learning, Memory, and Cognition*, *38*(4), 821.

De Brigard, F., Brady, T. F., Ruzic, L., & Schacter, D. L. (2017). Tracking the emergence of memories: A category-learning paradigm to explore schema-driven recognition. *Memory & cognition*, *45*, 105-120.

Dubey, R., Peterson, J., Khosla, A., Yang, M. H., & Ghanem, B. (2015). What makes an object memorable?. In *Proceedings of the ieee international conference on computer vision* (pp. 1089-1097).

Fox, J., & Osth, A. F. (2023). Modeling the continuous recognition paradigm to determine how retrieval can impact subsequent retrievals. *Cognitive Psychology*, 147, 101605.

Gillund, G., & Shiffrin, R. M. (1984). A retrieval model for both recognition and recall. *Psychological review*, *91*(1), 1.

Hebart, M. N., Zheng, C. Y., Pereira, F., & Baker, C. I. (2020). Revealing the multidimensional mental representations of natural objects underlying human similarity judgements. *Nature human behaviour*, *4*(11), 1173-1185.

Heit, E. (1994). Models of the effects of prior knowledge on category learning. *Journal of Experimental Psychology: Learning, Memory, and Cognition*, *20*(6), 1264.

Hintzman, D. L. (1988). Judgments of frequency and recognition memory in a multiple-trace memory model. *Psychological review*, *95*(4), 528.

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

Hout, M. C., Goldinger, S. D., & Brady, K. J. (2014). MM-MDS: A multidimensional scaling database with similarity ratings for 240 object categories from the massive memory picture database. *PloS one*, *9*(11), e112644.

Hunt, R. R. (1995). The subtlety of distinctiveness: What von Restorff really did. *Psychonomic Bulletin & Review*, *2*, 105-112.

Hunt, R. R., & McDaniel, M. A. (1993). The enigma of organization and distinctiveness. *Journal of memory and language*, *32*(4), 421-445.

Ichien, N., Alfred, K. L., Baia, S., Kraemer, D. J., Holyoak, K. J., Bunge, S. A., & Lu, H. (2023). Relational and lexical similarity in analogical reasoning and recognition memory: Behavioral evidence and computational evaluation. *Cognitive psychology*, *141*, 101550.

Kahana, M. J., & Sekuler, R. (2002). Recognizing spatial patterns: A noisy exemplar approach. *Vision research*, *42*(18), 2177-2192.

Khosla, A., Raju, A. S., Torralba, A., & Oliva, A. (2015). Understanding and predicting image memorability at a large scale. In *Proceedings of the IEEE international conference on computer vision* (pp. 2390-2398). IEEE, Santiago, Chile. https://doi.org/10.1109/ICCV.2015.275

Konkle, T., Brady, T. F., Alvarez, G. A., & Oliva, A. (2010). Conceptual distinctiveness supports detailed visual long-term memory for real-world objects. *Journal of experimental Psychology: General*, *139*(3), 558.

Kramer, M. A., Hebart, M. N., Baker, C. I., & Bainbridge, W. A. (2023). The features underlying the memorability of objects. *Science advances*, *9*(17), eadd2981.

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

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

Meagher, B. J., & Nosofsky, R. M. (2023). Testing formal cognitive models of classification and old-new recognition in a real-world high-dimensional category domain. *Cognitive Psychology,* 145, 101596.

Needell, C. D., & Bainbridge, W. A. (2022). Embracing new techniques in deep learning for estimating image memorability. *Computational Brain and Behavior*, 1–17.

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

Nosofsky, R. M. (1988). Exemplar-based accounts of relations between classification, recognition, and typicality. *Journal of Experimental Psychology: Learning, Memory, and Cognition*, *14*(4),

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), 3–27.

Nosofsky, R. M., Little, D. R., Donkin, C., & Fific, M. (2011). Short-term memory scanning viewed as exemplar-based categorization. *Psychological Review*, *118*(2), 280–315.

Nosofsky, R. M., & Meagher, B. J. (2022). Retention of exemplar-specific information in learning of real-world high-dimensional categories: Evidence from modeling of old-new item recognition. *Proceedings of the 44th Annual Conference of the Cognitive Science Society*.

Nosofsky, R. M., & Osth, A. F. (2024). Hybrid-similarity exemplar model of context-dependent memorability. *2024 Proceedings of the Cognitive Science Society*.

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. (2018). Tests of an exemplar-memory model of classification learning in a high-dimensional natural-science category domain. *Journal of Experimental Psychology: General*, *147*(3), 328.

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*(2), 530–556.

Nosofsky, R. M., Sanders, C. A., Meagher, B. J., & Douglas, B. J. (2020). Search for the missing dimensions: Building a feature-space representation for a natural-science category domain. *Computational Brain and Behavior, 3*, 13–33.

Nosofsky, R. M., & Zaki, S. R. (2002). Exemplar and prototype models revisited: response strategies, selective attention, and stimulus generalization. *Journal of Experimental Psychology: Learning, memory, and cognition, 28*(5), 924.

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), 1194–1209.

Osth, A. F., & Dennis, S. (in press). Global matching models of recognition memory. *The Oxford Handbook of Human Memory*.

Osth, A. F., Jansson, A., Dennis, S., & Heathcote, A. (2018). Modeling the dynamics of recognition memory testing with an integrated model of retrieval and decision making. *Cognitive Psychology*, 104, 106-142.

Robinson, K. J., & Roediger III, H. L. (1997). Associative processes in false recall and false recognition. *Psychological Science*, *8*(3), 231-237.

Sakamoto, Y., & Love, B. C. (2004). Schematic influences on category learning and recognition memory. *Journal of Experimental Psychology: General*, *133*(4), 534.

Sanders, C. A., & Nosofsky, R. M. (2020). Training deep networks to construct a psychological feature space for a natural-object category domain. *Computational Brain & Behavior*, *3*(3), 229-251.

Schmidt, S. R. (1991). Can we have a distinctive theory of memory?. *Memory & cognition*, *19*, 523-542.

Schwarz, G. (1978). Estimating the dimension of a model. *The annals of statistics*, 461-464.

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. https://doi.org/10.1126/science.3629243

Shiffrin, R. M., Huber, D. E., & Marinelli, K. (1995). Effects of category length and strength on familiarity in recognition. *Journal of Experimental Psychology: Learning, Memory, and Cognition*, *21*(2), 267.

Shiffrin, R. M., & Steyvers, M. (1997). A model for recognition memory: REM—retrieving effectively from memory. *Psychonomic bulletin & review*, *4*, 145-166.

Shin, H. J., & Nosofsky, R. M. (1992). Similarity-scaling studies of dot-pattern classification and recognition. *Journal of Experimental Psychology: General*, *121*(3), 278.

Strong Jr, E. K. (1912). The effect of length of series upon recognition memory. *Psychological Review*, *19*(6), 447.

Turner, B. M., Sederberg, P. B., Brown, S. D., & Steyvers, M. (2013). A method for efficiently sampling from distributions with correlated dimensions. *Psychological methods*, *18*(3), 368.

Tversky, A. (1977). Features of similarity. *Psychological review*, *84*(4), 327.

Valentine, T., & Ferrara, A. (1991). Typicality in categorization, recognition and identification: Evidence from face recognition. *British Journal of Psychology*, *82*(1), 87-102.

von Restrorff, H. (1933). Ober die Wirkung von Bereichsbildungen im Spurenfeld. *Psychologische Forschung*, *18,* 299-342.

Watanabe, S. (2013). A widely applicable Bayesian information criterion. *The Journal of Machine Learning Research*, *14*(1), 867-897.

Wixted, J. T. (2007). Dual-process theory and signal-detection theory of recognition memory. *Psychological review*, *114*(1), 152.

Yonelinas, A. P. (2002). The nature of recollection and familiarity: A review of 30 years of research. *Journal of memory and language*, *46*(3), 441-517.

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

Author Notes

The following materials associated with this article are available on the OSF website https://osf.io/a958r/: the complete sets of individual-subject data from Experiments 1 and 2; the complete set of rock-image stimuli used in the experiments along with their category assignments; the multidimensional scaling solution for the rock images; the set of distinctive-feature ratings for the rock images; the set of goodness-of-category-example ratings obtained for the rock images; and a file that indexes the specific rock images that are new to this study or that had already been used in previous studies.

This work was supported in part by Discovery Project Grant DP240101264 from the Australian Research Council.

A preliminary report of a subset of the results and modeling analyses in the present article was published in the 2024 Proceedings of the Cognitive Science Society (Nosofsky & Osth, 2024).

Correspondence concerning this article should be addressed to Robert Nosofsky, 1101 E. Tenth Street, Department of Psychological and Brain Sciences, Indiana University, Bloomington, IN 47405. E-mail: nosofsky@iu.edu.

Footnotes

[1] In future research, we hope to develop techniques that can identify the specific distinctive features that are present in the different images. We discuss these future-direction issues at length in our General Discussion.

[2] By comparison, in alternative experimental paradigms, distinctive features might be explicitly added to foils such that they match the distinctive features of targets stored in memory (e.g., Colloff, Wade, & Strange, 2016; Nosofsky & Zaki, 2003). In this case, such distinctive features would be assumed to boost the similarity of the foil to the target.

[3] An extended version of the hybrid-similarity model for the present paradigm allows distinctive features of the stored exemplars themselves to also reduce similarity. In this extended version, $\eta_{ij} = \exp(-\alpha_1 \cdot \delta_i) \cdot \exp(-\alpha_2 \cdot \delta_j) \cdot \exp(-c \cdot d_{ij})$. We found that adding this free parameter led to minuscule improvements in the fit of the model to the data.

[4] Konkle et al. (2010) observed a closely related result in a recognition-memory experiment involving everyday-object categories. Observers viewed object images with a different number of exemplars presented from each category. Rather than using Yes-No testing for individual items, observers were tested by indicating which of two exemplars (a target vs. a lure) they had previously studied. Forced-choice accuracy declined with increases in category size. The present model predicts this qualitative effect because the ratio of target summed-similarity to lure summed-similarity decreases as category size increases.

5. The reader will note a consistent outlier point in the Figures 6 and 12 plots in which the observed false-alarm rate for the size-0 category way exceeds its predicted false-alarm rate. In both cases, the point corresponds to the category of granite. In general, the models under-predicted the magnitude of the granite false-alarm rate for the other category sizes as well. We hypothesized that this systematic under-prediction might reflect that participants had substantial pre-experimental familiarity with exemplars of granite and this pre-experimental familiarity might have contributed to their high "old" response rates for members of this category (e.g., Heit, 1994). To test this hypothesis, we collected judgments of pre-experimental familiarity for the different rock categories from a separate group of subjects and included these judgments in a mixture model. The mixture model led to only small improvements in the fit of the model to the entire set of old-new recognition data.

6. We conducted the subgroup analyses for the individual-item hit-rate data as well. The qualitative pattern of results was the same as for the category-level hit-rate data. Again, however, the individual-item hit-rate data were far noisier than the category-level hit-rate data, so the absolute magnitude of the correlations between the model's predictions and the observed data remained low.

Appendix A

Derivation of the Multidimensional-Scaling Solution for the Rock Images

Among the 240 rock images used in our experiment, 199 had been used in previous studies of rock classification and recognition. Previous multidimensional-scaling work based on analysis of similarity-judgment data had been used to embed those 199 images in an 8-dimensional MDS solution (for details, see, e.g., Meagher & Nosofsky, 2023). There were 41 new rock images used in the present experiment. Among these 41 new images, 20 were the "gray gabbro" and "blue amphibolite" images that we produced by using light-shopping techniques. The remaining 21 new images were miscellaneous images that we gathered using web searches. These were used to fill in some of the categories for which we removed outlier images that we had judged to be more similar to members of contrast categories than to members of their own categories. A complete listing of the 199 old images used in previous work and the 41 additional new images used in the present study is provided in the supplementary materials (https://osf.io/a958r/).

To apply the formal models, we needed to embed the 41 new images within the same 8-dimensional MDS space as the 199 old images were embedded. To accomplish this embedding, we conducted a new similarity-judgment experiment and analyzed the similarity-judgment data with the same type of MDS model used previously. There were 258 subjects who participated in the similarity-judgment experiment. The experiment had a total of 280 trials. On each trial, a pair of rock images was presented side by side and centered on the computer screen. Following the previous MDS-scaling work, subjects judged the similarity of the pair using a scale from 1 (highly dissimilar) to 9 (highly similar). There were six different types of pairs and they are listed along with their frequencies of presentation in Table A1. The order of presentation of all pairs was randomized, as was the left-right placement of each image in the presented pair.

Table A1. Pair Types, Frequencies of Presentation, and Mean Ratings in the Similarity-Judgment Experiment

Pair Type	Frequency	Mean Rating
1. two randomly chosen samples of "gray gabbro"	20	7.39
2. two randomly chosen samples of "blue amphibolite"	20	7.87
3. a randomly chosen sample of "gray gabbro" and a randomly chosen stimulus outside of the gabbro category	20	3.52
4. a randomly chosen sample of "blue amphibolite" and a randomly chosen stimulus outside of the amphibolite category	20	2.42
5. a randomly chosen "other-new" stimulus and a second randomly chosen stimulus from the same category	50	5.29
6. a randomly chosen "other-new" stimulus and a randomly chosen stimulus from a different category	150	2.99

The mean similarity judgments for each pair type are reported in the final column of Table A1. The mean ratings show a sensible pattern, with "within-category" pair types (1, 2, 5) tending to receive high ratings and "between-category" pair types (3, 4, 6) low ratings. Within-category pair types 1 and 2 received higher ratings than within-category pair type 5, reflecting that the explicit manipulation of a common tint for the “gray gabbro” and “blue amphibolite” categories apparently led to particularly high within-category similarity.

The next step was to conduct a parameter search for the 8-dimensional coordinates of the 41 new rock images that minimized the sum of squared deviations between the predicted and observed individual-trial similarity ratings. Following previous work (e.g., Meagher and Nosofsky, 2023), the MDS model assumed a linear decreasing relation between judged similarity (*s*) and Euclidean distance (*d*) in the MDS space, $s = b - md$. The parameters b=9.00 and m=0.80 were held fixed at the values that provided a best fit of the MDS model to the original set of 480 rock images used in the Meagher and Nosofsky (2023) experiment. The model accounted for 54.2% of the variance in the individual-trial ratings for the new rock images from the present experiment. The complete set of coordinate parameters for the 240 rock images used in the present experiment are available at https://osf.io/a958r/. This MDS solution is provided as input for all the different versions of the model fitted to the old-new recognition experiment.

Appendix B

Distinctive-Feature Ratings Study

The participants in this study were 98 undergraduates from Indiana University who received credit towards an introductory psychology course requirement. The participants provided distinctive-feature ratings for the 240 rock images that were used in the recognition experiment. The images were presented one by one in the center of the computer screen in a randomized order for each individual participant. The details of the stimulus displays and apparatus were the same as already described for the old-new recognition experiment in the main text. The mean distinctiveness ratings for the 240 rock images are provided in a table in the OSF website associated with this article (https://osf.io/a958r/).

The participants received the following set of instructions:

```
In this experiment you will be presented with a set of pictures of rocks.
We would like you to rate the extent to which each rock picture contains
  a highly distinctive feature that distinguishes it from other rocks in the set.
For example, if you were viewing faces, then a face with a
  highly prominent scar would have a highly distinctive feature.

Likewise, rocks may also contain highly distinctive features.
For example, a rock may contain a fossil, a bright crystal, a giant hole,
   and so forth.
Please note that a feature should be judged as “distinctive” only if it rarely
   occurs.
If many rocks contain that same feature, then it should not be judged as
“distinctive”.

Rocks that do not contain any distinctive features should receive ratings of 1 or 2
Rocks that have medium distinctive features should receive ratings of 4, 5 or 6
Rocks that contain highly distinctive features should receive ratings of 8 or 9

Most rocks do not contain distinctive features, so most of your ratings should be
   low.
Please reserve your high ratings for rocks that contain truly distinctive features.

Examples are shown at the bottom of the screen.
```

Appendix C

Hierarchical Bayesian Modeling of the Core and Baseline Models

In this appendix, we describe the results of applying both the base model as well as the core hybrid similarity model to data using hierarchical Bayesian methods (e.g., Boehm, Marsman, Matzke, & Wagenmakers, 2018). Using this method, parameters are simultaneously estimated at both the group- and participant-levels. Parameters were estimated using differential-evolution Markov chain Monte Carlo techniques (Turner, Sederberg, Brown, & Steyvers, 2013). We used non-informative prior distributions for each of the group-level parameters. The means of the *k* and *c* parameters were sampled from truncated normal distributions with a lower bound of zero. Specifically, the mean of *k* was sampled from a distribution with a mean of 1 and a standard deviation of 1, while the *c* parameter was sampled from a distribution with a mean of 2 and a standard deviation of 1. The standard deviation parameters were sampled from a gamma distribution with a shape of 1 and a rate of 3. The α and β parameters were sampled on a log scale – the means of these parameters were drawn from normal distributions with a mean of 0 and a standard deviation of 5. The standard deviations of these parameters were drawn from a gamma distribution with shape 1 and rate 1.

The number of chains was always equal to 3 times the number of participant level parameters. After 6,000 burn-in iterations, 1 in 20 MCMC iterations were collected as samples of the posterior distribution until 500 were collected from each chain. The models were compared using the widely applicable information criterion (WAIC: Watanabe, 2013).

While we do not visually show the fits of the model to the data here, the group-averaged posterior predictive distributions from the hierarchical Bayesian implementations strongly resembled the group-level fits in the main text. The model selection results can be seen in Table

C1. In both Experiment 1 and 2, the core hybrid similarity model outperforms the base model by a considerable margin. Table C1 also lists the correlations between the data and the model's predictions of the hit rates (HR) and false alarm rates (FAR). For the model's predicted hit and false alarm rates, we averaged across all of the posterior predictive samples for either the category or the item level to get a stable estimate of the predicted HR or FAR. One can see that these correlations bear a strong resemblance to the group-level fit from the main text, suggesting that variation in individual participant parameters is insufficient to strongly improve performance at the category- or item-level.

Table C1

	Core (4)	Base (2)
Experiment 1		
WAIC	16879	16990
Category *r*: HR	.48	.00
Category *r*: FAR	.82	.82
Category *r*: FAR (size-0)	.61	.62
Item *r*: HR	.33	.00
Item *r*: FAR	.65	.65
Experiment 2		
WAIC	27282	27401
Category *r*: HR	.33	.40
Category *r*: FAR	.76	.74
Category *r*: FAR (size-0)	.70	.66
Item *r*: HR	.22	.24
Item *r*: FAR	.66	.63

Appendix D

Goodness-of-Example Ratings Experiment

The participants in this study were 69 undergraduates from Indiana University who received credit towards an introductory psychology course requirement. The participants provided goodness-of-example ratings for the 240 rock images that were used in the recognition experiment. We excluded 9 of the participants from further analysis who had correlations of less than $r = .30$ with the averaged ratings for the 240 rock images, leaving a sample size of 60 participants. The details of the stimulus displays and apparatus were the same as already described for the old-new recognition experiment in the main text.

The experiment was organized into 24 blocks in which ratings were provided for the members of each individual category, one category per block. The ordering of categories was randomized for each individual participant. On each block, participants were first given the category name along with the description of that category that had been used in the old-new recognition experiment. Next, the 10 members of the category were presented in the center of the computer screen, one at a time in a random order, and participants were instructed to simply observe them. Next, the 10 members were presented again, one at a time in a new random order, and participants were instructed to rate how good an example each rock image was of its category on a scale from 1 (very bad example) to 9 (excellent example).

The mean goodness-of-example ratings for the 240 rock images are provided in a table in the OSF website associated with this article (https://osf.io/a958r/).

The detailed instructions that participants received in the experiment are provided below:

```
In this experiment you will be presented
  with rock images belonging to 24 different categories.
During each block, we will first name and describe a category
  and then present you with 10 examples from that category.
Next, we will present those 10 examples again, one at a time.
Your job is to rate how good an example each rock image
  is of its category on a scale from
  1 (very bad example) to 9 (excellent example).
If you feel that all the rocks are excellent examples of their
  category then you can give high ratings to all of them.
Please reserve low ratings for rock images that truly
  seem to you to be bad examples of their category.
```

Appendix E

Extended Versions of the Core Hybrid-Similarity Model

In this appendix we provide a report of the various extended versions of the core hybrid-similarity model that we fitted to the Experiment-1 and Experiment-2 data. A report of summary-fit statistics from the models is provided in Table E1 for Experiment1 and in Table E2 for Experiment 2.

Nonlinear relation between familiarity and summed similarity. In the first extension, we make allowance for the possibility that psychological familiarity is non-linearly related to summed similarity by assuming that

$$F_i = [\ \sum_j \ \eta_{ij}\ ]^{\gamma}, \qquad \text{(E1)}$$

where γ is a power-transform parameter. The use of γ has a long history in applying the summed-similarity exemplar model to both categorization and old-new recognition data and in some cases it has played a crucial role (for example discussion, see Nosofsky & Zaki, 2002). In the present case, however, extending the core model with γ yields only a slightly better overall BIC fit, and the component fits to the summary empirical findings remain essentially the same as before (see Tables E1 and E2). Despite the small improvement in BIC fit, we continue to allow γ to vary in our remaining extended-model explorations, because it may take on a more important role when combined with other modifications to the core model.

Nonlinear relation between psychological and rated distinctiveness. In our second extension, we made allowance for the possibility that the psychological distinctiveness of a rock image is non-linearly related to its rated distinctiveness. Again, this type of assumption is commonly made when incorporating rating-scale data into the machinery of formal computational models, because the psychological properties of the rating scale are generally unknown. Following Meagher and Nosofsky (2023), we assumed specifically that psychological distinctiveness (δ_P) was related to rated distinctiveness (δ_R) using the transformation

$$\begin{aligned} \delta_P &= \delta_0 + (\delta_R - \delta_0)^v, \quad \text{if } \delta_R > \delta_0 \\ \delta_P &= \delta_0 - (\delta_0 - \delta_R)^u, \quad \text{if } \delta_R < \delta_0 \end{aligned} \qquad \text{(E2)}$$

where δ_0 is a "reference value" on the rating scale, and *v* and *u* are power-transform parameters. The transformation provides flexibility in the type of nonlinear function that may relate psychological distinctiveness to the direct ratings and allows for the possibility that the shape of the nonlinear function changes for low vs. high ratings. As reported in Table E1, however, in Experiment 1, making allowance for these transformations led to virtually no improvement in BIC compared to the core model and the account of the individual-item false-alarm and hit rates was unchanged. There was somewhat greater improvement for some of the summary-fit measures in Experiment 2 (see Table E2) but the overall account of the hit-rate data remained unsatisfactory.

Dimensional Attention Weighting. Another common extension of the summed-similarity exemplar makes allowance for attention-weighting of the component dimensions of the space (Nosofsky, 1986, 1991). In this extension, the distance between test-item *i* and exemplar *j* is given by

$$d_{ij}=[\sum w_m \cdot |x_{im} - x_{jm}|^2]^{1/2}, \qquad \text{(E3)}$$

where the attention weights w_m are constrained to vary between 0 and 1 and to sum to 1. As reported in Tables E1 and E2, making allowance for the attention-weight parameters in the distance function yields an improved BIC compared to the core model. However, inspection of the tables reveals that the improvement in model fit stems mostly from an improved account of the false-alarm-rate data, not the hit-rate data.

Missing Dimensions and Within-Category Similarity. A more unusual extension to the exemplar model in application to the present rocks domain was introduced by Nosofsky et al. (2019, 2020). This extension is motivated by the assumption that the MDS solution for the rock stimuli derived from similarity-judgment data provides an incomplete characterization of the full set of dimensions that compose the rock stimuli. In particular, Nosofsky et al. (2019, 2020) found evidence that in classification-learning tasks, participants latch on to additional dimensions that are highly diagnostic for classifying the rocks but that were not revealed in the similarity-scaling studies themselves. Because of these "missing" dimensions, the standard MDS solution derived from the similarity-judgment data tends to underestimate the magnitude of within-category similarities compared to between-category similarities. To "repair" this systematic source of error in the MDS solution, Nosofsky et al. (2019) introduced a "within-category sensitivity" parameter (c_w) to the modeling: The similarity between any two exemplars within the same category was assumed to be given by

$$s_{ij} = \exp(-c_w \cdot d_{ij}), \qquad \text{(E4)}$$

where $c_w < c$ in Equation 3. We fitted this model to the present old-new recognition data. As reported in Tables E1 and E2, although it involves the addition of only two free parameters (γ and c_w) compared to the core version of the exemplar model, the improvements in BIC are quite sizeable. Nevertheless, even this model fails to yield any improvement to the predictions of the individual-item hit and false-alarm rates. Instead, most of the improvement is due to improved quantitative predictions of the various category-size effects.

Output Interference. In still another analysis, we investigated the potential role of output interference on the present old-new recognition results. In various previous studies, researchers have found that old-new recognition performance changes systematically over the course of the test phase. It is often the case that one's ability to discriminate between old and new test probes declines. In Figure E1 we plot mean hit and false-alarm rates as a function of ten-trial blocks during the test phase in Experiments 1 and 2. As is apparent from inspection, hit rates showed a decline across test blocks; the decline in false-alarm rates was somewhat smaller in magnitude. Although the serial position of individual items on the test list was randomized across participants in our experiment, it is possible that variability in these placements could have contributed to variability in the individual-item hit rates. Researchers have proposed a variety of models to capture the role of output interference in old-new recognition and categorization (Criss et al., 2011; Fox & Osth, 2023; Osth, Jansson, Dennis, & Heathcote, 2018; Zaki & Nosofsky, 2007), but the present experiment was not designed to discriminate among them. Here we report results from a representative from this class of models proposed by Nosofsky and Zaki (2007). The gist of the model is that the individual items that appear on the test list are stored along with the original study exemplars in memory, although with reduced memory strengths. For each test probe, the observer sums its similarity to all the old study exemplars and to the previously presented test items. The observer compensates for the growing levels of summed similarity that accrue during the test phase by using an increasingly strict response-criterion k in the decision rule (see Equation 1). As reported in Tables E1 and E2, this output-interference model yields a much improved BIC compared to the core version of the model. Not surprisingly, the locus of the improvement is in capturing how mean hit rates and false-alarm rates varied with serial position on the test list (see Tables E1 and E2 and a plot of the observed and predicted results in Figure E1). Nevertheless, as reported in Tables E1 and E2, the output-interference model provided no improvement in capturing any of the other summary findings in the data.

Category Goodness. In a final modeling extension, we considered the possibility that other factors beyond summed similarity might contribute to the participants' old-new recognition decisions. Recall that the encoding task during the study phase of the experiment was for participants to choose the category description that best matched each of the presented study items. It seems plausible that participants' judgments of the "category goodness" of the presented test items might have influenced their old-new recognition decisions. For example, some participants may have been more likely to provide "old" responses to test items that they judged to be better examples of the described categories. To explore this possibility, we conducted a new study in which participants were presented with the 24 categories one at a time along with their category descriptions. The participants were then asked to provide goodness-of-example (*GoE*) ratings for the individual rock images that were the members of the categories (see Appendix D for details). We then tested an extended mixture model that combined the summed hybrid-similarity familiarity measure (F_i) with the mean GoE_i ratings for the test items *i*. Let GoE_0 denote a "reference" value on the *GoE* rating scale. The formal mixture model assumed that the evidence in favor of an "old"response for item *i* [E_{old}(i)] was given by

$$\begin{aligned} E_{old}(i) &= F_i + w_{pos} \cdot F_i \cdot (GoE_i - GoE_0)^v, \quad \text{if } GoE_i > GoE_0 \\ E_{old}(i) &= F_i + w_{neg} \cdot F_i \cdot (GoE_0 - GoE_i)^u, \quad \text{if } GoE_i < GoE_0 \end{aligned} \tag{E5}$$

where F_i is the familiarity measure from the summed hybrid-similarity model; and w_{pos}, w_{neg}, v, u, and *ref* are freely estimated parameters. As reported in Tables E1 and E2, the mixture model yields a slightly improved BIC compared to the core hybrid-similarity model and there is some small improvement in the ability of this model to fit the individual-item hit-rate data. However, the mixture model still falls far short of providing a satisfactory account of these data.

Table 1. Listing of Category Numbers, Names and Descriptions and their Associated Mean False-Alarm Rate and Hit-Rate Percentages

FAR,HR	Cat #	Name and Description
37 76	1.	Andesite. Fine-grained with small embedded fragments.
44 75	2.	Basalt. Dark gray/black and fine-grained, not shiny.
55 80	3.	Gabbro. Dark gray and coarse grained.
65 89	4.	Granite. Light-colored and coarse-grained.
49 91	5.	Obsidian. Shiny glassy black, scalloped surfaces.
18 86	6.	Pegmatite. Long black or green crystal bands.
40 86	7.	Peridotite. Green and coarse grained.
35 85	8.	Pumice. Porous, holes.
60 87	9.	Amphibolite. Blue, coarse-grained, rough.
51 82	10.	Anthracite. Shiny black, rough layered surface.
46 87	11.	Gneiss. Straight stripes.
20 87	12.	Marble. Embedded veins.
27 84	13.	Migmatite. Curved swirly stripes.
28 79	14.	Phyllite. Shimmery oily surface.
43 76	15.	Quartzite. Milky splotches on colored background.
44 88	16.	Slate. Gray, flat sheet-like layers.
28 79	17.	Breccia. Embedded large angular fragments.
17 85	18.	Chert. Tan, scalloped surfaces.
30 91	19.	Conglomerate. Embedded large rounded fragments.
36 75	20.	Micrite. Tan, fine-grained homogenous.
28 82	21.	Gypsum. Single translucent crystal.
36 87	22.	Rock Salt. Multiple cubic crystals.
24 88	23.	Sandstone. Sandy texture, stripes.
32 82	24.	Shale. Colored, flat sheet-like layers.

FAR = mean false-alarm-rate percentage
HR = mean hit-rate percentage

Table 2. Maximum-Likelihood Parameter Estimates from the Core Hybrid-Similarity Model and the Baseline Model in Experiments 1 and 2.

Experiment 1

Model	*c*	*k*	β	α
Core Hybrid Model	1.094	0.442	0.147	0.016
Baseline Model	1.254	0.256	---	---

Experiment 2

Model	*c*	*k*	β	α
Core Hybrid Model	0.883	0.704	0.019	0.089
Baseline Model	0.960	0.695	---	---

Note. Parameters are held fixed at zero in cells without entries.

Table 3. Experiment 1: Summary Fit Measures for the Core and Baseline Models.

	Model			
Summary Fit Measure	Core (4)		Baseline (2)	
-ln L	8811.5		8860.7	
BIC	17661.8		17740.8	
	R^2	r	R^2	r
Cat-Size FA	.88	.99	.93	.99
Cat-Size Hit	.47	.83	.36	.86
Cat FA	.67	.84	.69	.84
Cat Hit	.24	.50	.00	.00
Cat d’	.77	.89	.69	.84
Size-0 FA	.02	.57	.03	.56
Binned FA	.85	.92	.81	.91
Binned Hit	.83	.95	.00	.00
Item FA	.39	.63	.37	.64
Item Hit	.09	.32	.00	.04

Notes. Values in parentheses are the number of free parameters for the model. The -ln L and BIC rows are fits with respect to the individual-trials data. For the remaining summary-fit measures, the two values reported in each model column are the proportion of variance accounted for (R^2) and the correlation between the model-predicted and observed data values (r).

Fit Measures

-ln L = negative natural log-likelihood
BIC = Bayesian Information Criterion
Cat-Size FA = False-Alarm rate as a function of category size
Cat-Size Hit = Hit rate as a function of category size
Cat FA = False-alarm rate for the 24 categories
Cat Hit = Hit rate for the 24 categories
Cat d’ = d’ values for the 24 categories
Size-0 FA = False-alarm rate for the 24 categories in the size-0 condition
Binned FA = False-alarm rate as a function of distinctiveness bin
Binned Hit = Hit rate as a function of distinctiveness bin
Item FA = False-alarm rate for the 240 individual items
Item Hit = Hit rates for the 240 individual items

Table 4. Experiment 2: Summary Fit Measures for the Core and Baseline Models.

	Model			
Summary Fit Measure	Core (4)		Baseline (2)	
-ln L	14008.5		14031.3	
BIC	28057.1		28082.7	
	R^2	*r*	*R*^{2}	*r*
Cat-Size FA	.93	.99	.97	.99
Cat-Size Hit	.54	.99	.67	.99
Cat FA	.57	.77	.59	.77
Cat Hit	.09	.32	.16	.40
Cat d’	.71	.87	.64	.80
Size-0 FA	.39	.65	.33	.59
Binned FA	.77	.89	.67	.89
Binned Hit	.00	.00	.00	.00
Item FA	.43	.65	.39	.62
Item Hit	.03	.19	.05	.23

Notes. Values in parentheses are the number of free parameters for the model. The -ln L and BIC rows are fits with respect to the individual-trials data. For the remaining summary-fit measures, the two values reported in each model column are the proportion of variance accounted for (*R*^{2}) and the correlation between the model-predicted and observed data values (*r*).

Fit Measures
-ln L = negative natural log-likelihood
BIC = Bayesian Information Criterion
Cat-Size FA = False-Alarm rate as a function of category size
Cat-Size Hit = Hit rate as a function of category size
Cat FA = False-alarm rate for the 24 categories
Cat Hit = Hit rate for the 24 categories
Cat d’ = d’ values for the 24 categories
Size-0 FA = False-alarm rate for the 24 categories in the size-0 condition
Binned FA = False-alarm rate as a function of distinctiveness bin
Binned Hit = Hit rate as a function of distinctiveness bin
Item FA = False-alarm rate for the 240 individual items
Item Hit = Hit rates for the 240 individual items

Table 5. Maximum-Likelihood Core-Model Parameters in Fits to Subgroup Data

	c	k	α	β
Exp. 1, Subgroup 1 (102)	0.995	0.751	0.000	0.285
Exp. 1, Subgroup 2 (101)	1.159	0.221	0.098	0.010
Exp. 2, Subgroup 1 (111)	0.839	1.396	0.000	0.238
Exp. 2, Subgroup 2 (165)	0.813	0.662	0.176	0.000

Note. Values in parentheses are number of subjects in each subgroup. Subgroup 1 subjects had distinctive-bin regression slopes greater than .027 (mean slope in Exp. 1). Subgroup 2 subjects had distinctive-bin regression slopes less than .027.

Table E1. Experiment 1: Summary Fit Measures for the Extended Models.

	Model							
Summary				Core+				Core+
Fit			Core+	NonLin	Core+	Core+	Core+	Cat.
Measure	Core	Base.	γ	Dist.	Weights	cw	Output	Good.
	(4)	(2)	(5)	(8)	(12)	(6)	(8)	(10)
-ln L	8811.5	8860.7	8800.8	8789.4	8758.8	8707.5	8764.8	8774.7
BIC	17661.8	17740.8	17650.2	17656.5	17634.2	17473.4	17607.3	17646.7
Cat-Size FA	.88 .99	.93 .99	.85 .99	.86 .99	.90 .99	.99 .99	.85 .99	.84 .99
Cat-Size Hit	.47 .83	.36 .86	.16 .82	.26 .77	.39 .84	.53 .82	.22 .82	.16 .86
Cat FA	.67 .84	.69 .84	.69 .85	.71 .85	.79 .80	.71 .85	.70 .86	.69 .85
Cat Hit	.24 .50	.00 .00	.25 .54	.23 .48	.27 .55	.29 .54	.23 .52	.28 .58
Cat d′	.77 .89	.69 .84	.75 .89	.77 .88	.80 .92	.74 .87	.75 .89	.75 .88
Size-0 FA	.02 .57	.03 .56	.00 .58	.01 .59	.07 .57	.37 .61	.01 .60	.00 .60
Binned FA	.85 .92	.81 .91	.83 .92	.88 .94	.91 .97	.89 .94	.84 .92	.80 .89
Binned Hit	.83 .95	.00 .00	.75 .96	.90 .95	.77 .96	.90 .96	.59 .95	.76 .95
Item FA	.39 .63	.37 .64	.41 .64	.42 .65	.47 .68	.42 .65	.43 .66	.43 .66
Item Hit	.09 .32	.00 .04	.09 .33	.10 .32	.10 .35	.10 .32	.10 .35	.15 .40
Test-Blk FA	.00 .00	.00 .00	.00 .00	.00 .00	.00 .00	.00 .00	.00 .39	.00 .00
Test-Blk Hit	.00 .00	.00 .00	.00 .00	.00 .00	.00 .00	.00 .00	.67 .92	.00 .00

Notes. Values in parentheses under each model name are the number of free parameters for that model. The -ln L and BIC rows are fits with respect to the individual-trials data of the participants. For the remaining summary-fit measures, the two values reported in each cell are the proportion of variance accounted for (R^2) and the correlation between the model-predicted and observed data values (*r*).

(Table E1 Continued on next page)

Table E1, Continued

Models

Core = core hybrid-similarity (sensitivity, response-criterion, beta, alpha)
Base. = baseline (sensitivity, response-criterion)
Core + γ = (adds response-scaling parameter to core)
Core + NonLin Dist. = core + nonlinear distinctiveness
Core + Weights = allows attention-weight of dimensions
Core + cw = includes within-category sensitivity
Core + Output = adds output-interference process
Core + Cat. Good = adds category-goodness mixture process

Fit Measures

-ln L = negative natural log-likelihood
BIC = Bayesian Information Criterion
Cat-Size FA = False-Alarm rate as a function of category size
Cat-Size Hit = Hit rate as a function of category size
Cat FA = False-alarm rate as a function of category, averaged across the different size conditions
Cat Hit = Hit rate as a function of category, averaged across the different size conditions
Cat d’ = d’ values computed from the individual category hit and false-alarm rates
Size-0 FA = False-alarm rate as a function of category in the size-0 condition
Binned FA = False-alarm rate as a function of distinctiveness bin
Binned Hit = Hit rate as a function of distinctiveness bin
Item FA = False-alarm rate for the individual items
Item Hit = Hit rates for the individual items
Test-Blk FA = Mean false-alarm rate as a function of test block
Test-Blk Hit = Mean hit rate as a function of test block

Table E2. Experiment 2: Summary Fit Measures for the Extended Models.

	Model							
Summary				Core+				Core+
Fit			Core+	NonLin	Core+	Core+	Core+	Cat.
Measure	Core	Base.	γ	Dist.	Weights	cw	Output	Good.
	(4)	(2)	(5)	(8)	(12)	(6)	(8)	(10)
-ln L	14008.5	14031.3	14004.9	13950.2	13899.9	13944.2	13924.1	13972.4
BIC	28057.1	28082.7	28060.9	27980.6	27920.2	27948.6	27928.4	28045.1
Cat-Size FA	.93 .99	.97 .99	.92 .99	.92 .99	.96 .99	.99 .99	.90 .99	.93 .99
Cat-Size Hit	.54 .99	.67 .99	.47 .99	.47 .99	.58 .99	.73 .99	.51 .99	.49 .99
Cat FA	.57 .77	.59 .77	.58 .77	.60 .78	.72 .87	.60 .78	.61 .80	.62 .80
Cat Hit	.09 .32	.16 .40	.10 .33	.14 .42	.16 .43	.13 .38	.09 .33	.18 .44
Cat d’	.71 .87	.64 .80	.70 .86	.70 .86	.76 .91	.78 .88	.68 .88	.70 .87
Size-0 FA	.39 .65	.33 .59	.39 .65	.40 .67	.55 .76	.45 .65	.42 .70	.45 .70
Binned FA	.77 .89	.67 .89	.76 .89	.90 .95	.88 .97	.79 .90	.79 .91	.76 .89
Binned Hit	.00 .00	.00 .00	.00 .00	.00 .00	.00 .00	.00 .00	.00 .00	.00 .00
Item FA	.43 .65	.39 .62	.42 .65	.45 .67	.44 .67	.42 .65	.46 .68	.43 .66
Item Hit	.03 .19	.05 .23	.03 .20	.08 .29	.06 .25	.10 .32	.03 .20	.07 .27
Test-Blk FA	.00 .00	.00 .00	.00 .00	.00 .00	.00 .00	.00 .00	.27 .58	.00 .00
Test-Blk Hit	.00 .00	.00 .00	.00 .00	.00 .00	.00 .00	.00 .00	.81 .95	.00 .00

Notes. Values in parentheses under each model name are the number of free parameters for that model. The -ln L and BIC rows are fits with respect to the individual-trials data of the participants. For the remaining summary-fit measures, the two values reported in each cell are the proportion of variance accounted for (R^2) and the correlation between the model-predicted and observed data values (*r*).

(Table E2 Continued on next page)

Table E2, Continued

Models

Core = core hybrid-similarity (sensitivity, response-criterion, beta, alpha)

Base. = baseline (sensitivity, response-criterion)

Core + γ = (adds response-scaling parameter to core)

Core + NonLin Dist. = core + nonlinear distinctiveness

Core + Weights = allows attention-weight of dimensions

Core + cw = includes within-category sensitivity

Core + Output = adds output-interference process

Core + Cat. Good = adds category-goodness mixture process

Fit Measures

-ln L = negative natural log-likelihood

BIC = Bayesian Information Criterion

Cat-Size FA = False-Alarm rate as a function of category size

Cat-Size Hit = Hit rate as a function of category size

Cat FA = False-alarm rate as a function of category, averaged across the different size conditions

Cat Hit = Hit rate as a function of category, averaged across the different size conditions

Cat d’ = d’ values computed from the individual category hit and false-alarm rates

Size-0 FA = False-alarm rate as a function of category in the size-0 condition

Binned FA = False-alarm rate as a function of distinctiveness bin

Binned Hit = Hit rate as a function of distinctiveness bin

Item FA = False-alarm rate for the individual items

Item Hit = Hit rates for the individual items

Test-Blk FA = Mean false-alarm rate as a function of test block

Test-Blk Hit = Mean hit rate as a function of test block

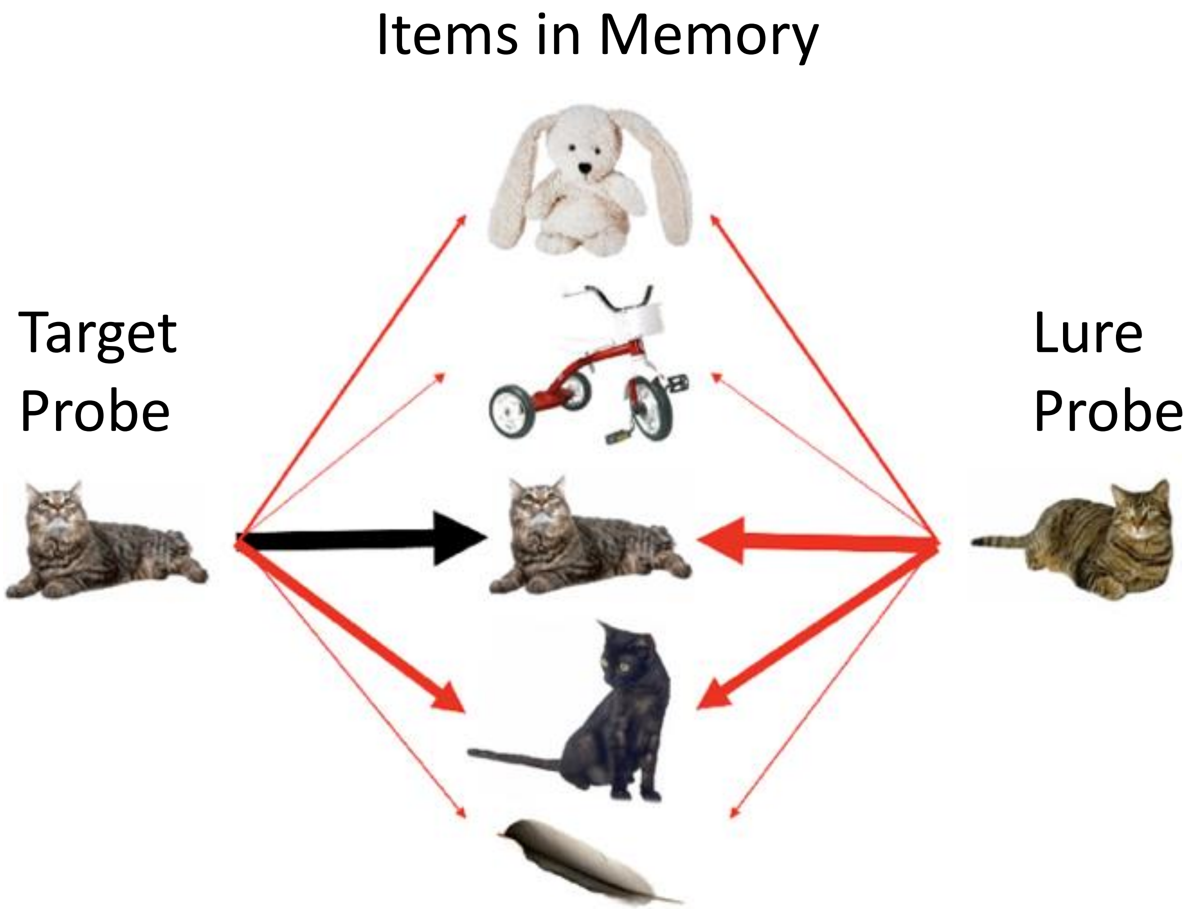


Figure 1. Schematic illustration of how self-similarity and inter-item similarity contribute to summed similarity in the exemplar model of recognition. The thick solid-black arrow illustrates the similarity of the target probe to its own representation in memory. The thick red arrows illustrate the varying degrees of high inter-item similarity of the target and lure probes to some of the other items in memory. Thinner red arrows illustrate lower degrees of inter-item similarity.

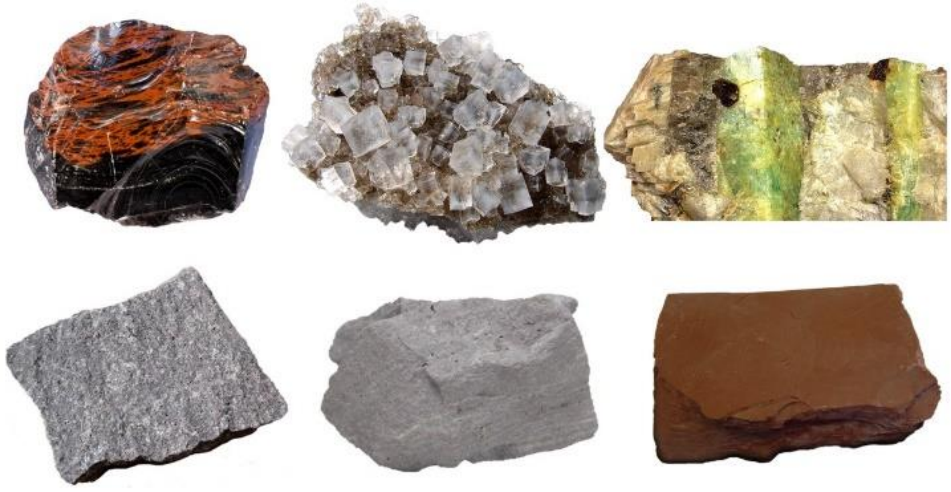


Figure 2. Examples of rock images that received high (top row) vs. low (bottom row) distinctive-feature ratings.

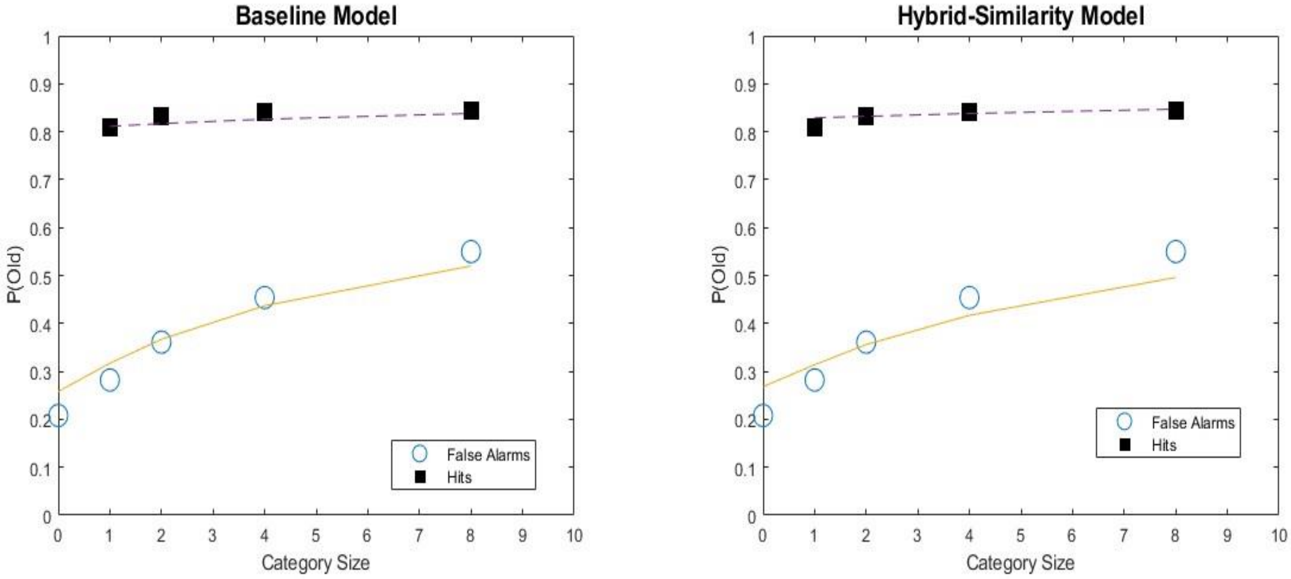


Figure 3. Experiment 1: Mean observed and predicted false-alarm and hit rates as a function of category size.

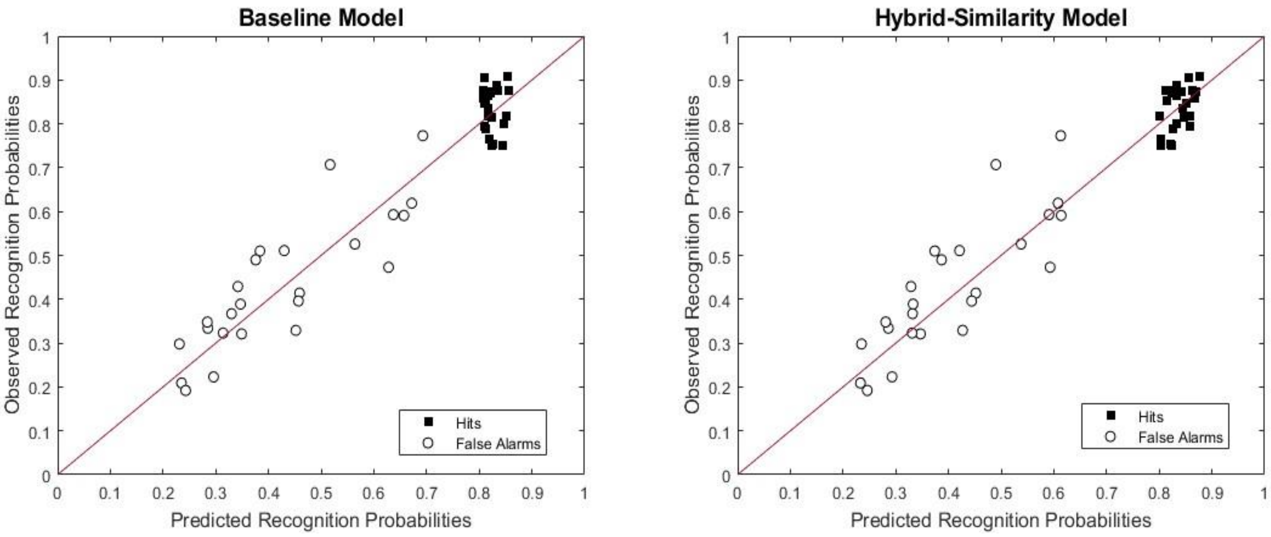


Figure 4. Experiment 1: Mean observed against predicted hit and false-alarm rates for each of the 24 categories.

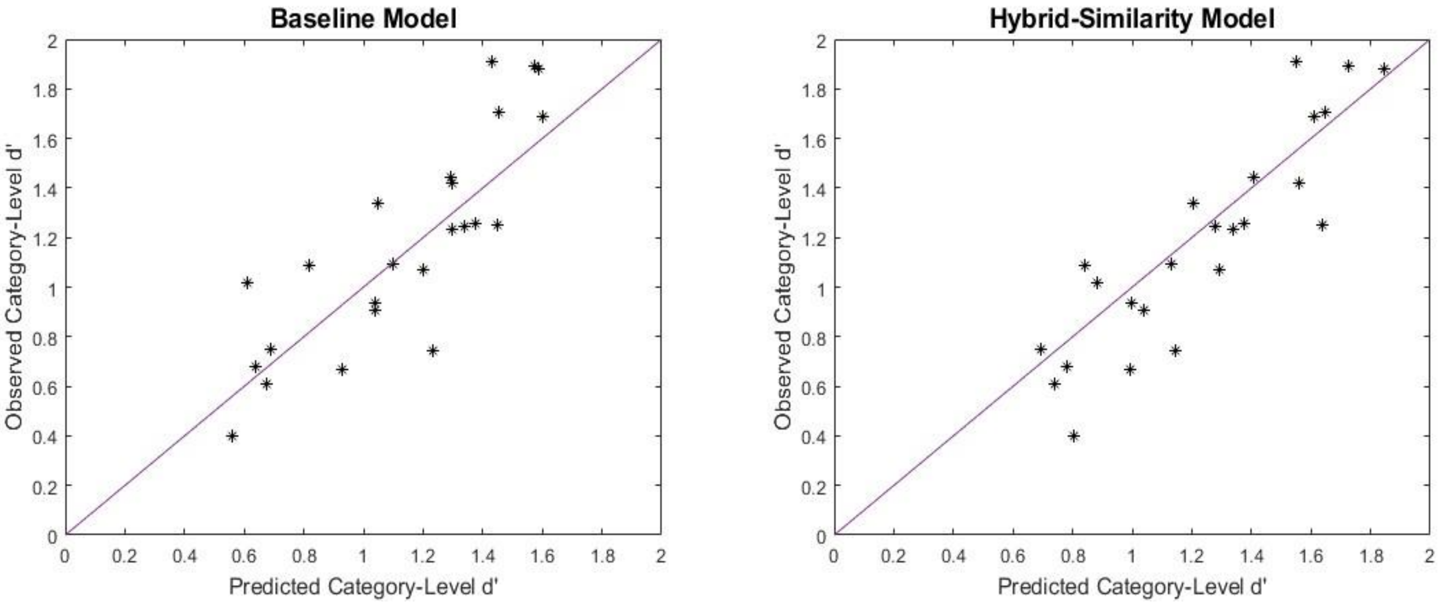
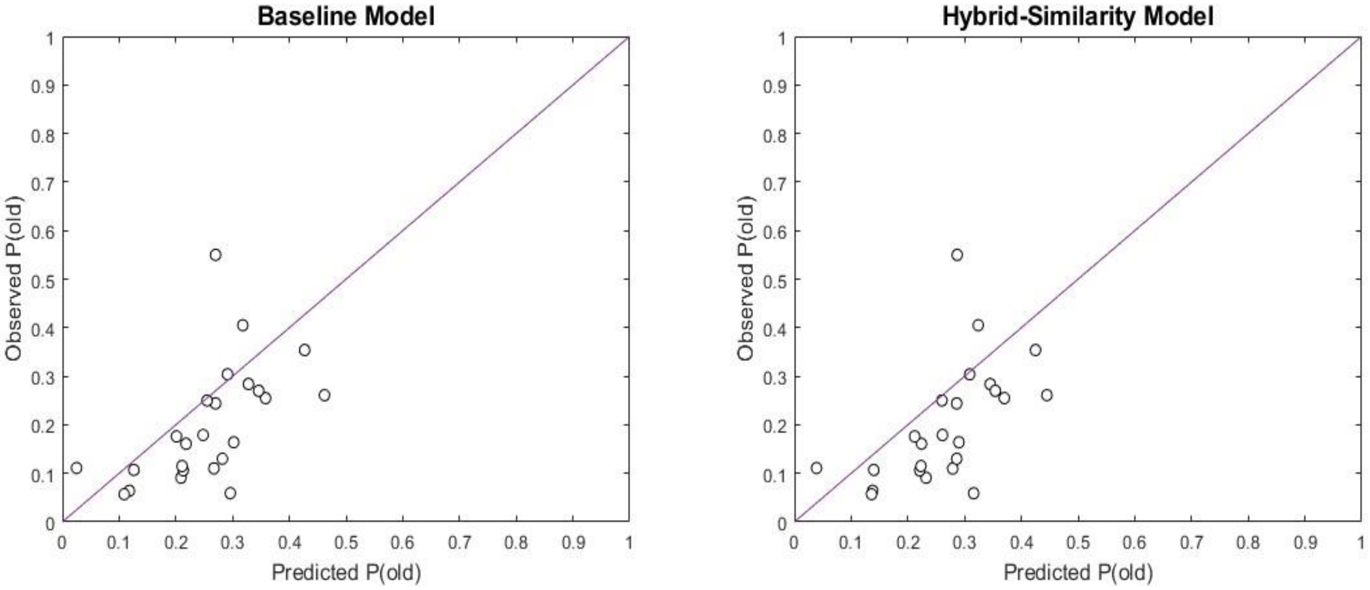


Figure 5. Experiment 1: Mean observed-against-predicted category-level d’ values for the 24 categories.


Baseline Model

Hybrid-Similarity Model

Figure 6. Experiment 1: Mean observed-against-predicted false-alarm rates for the 24 size-0 categories.

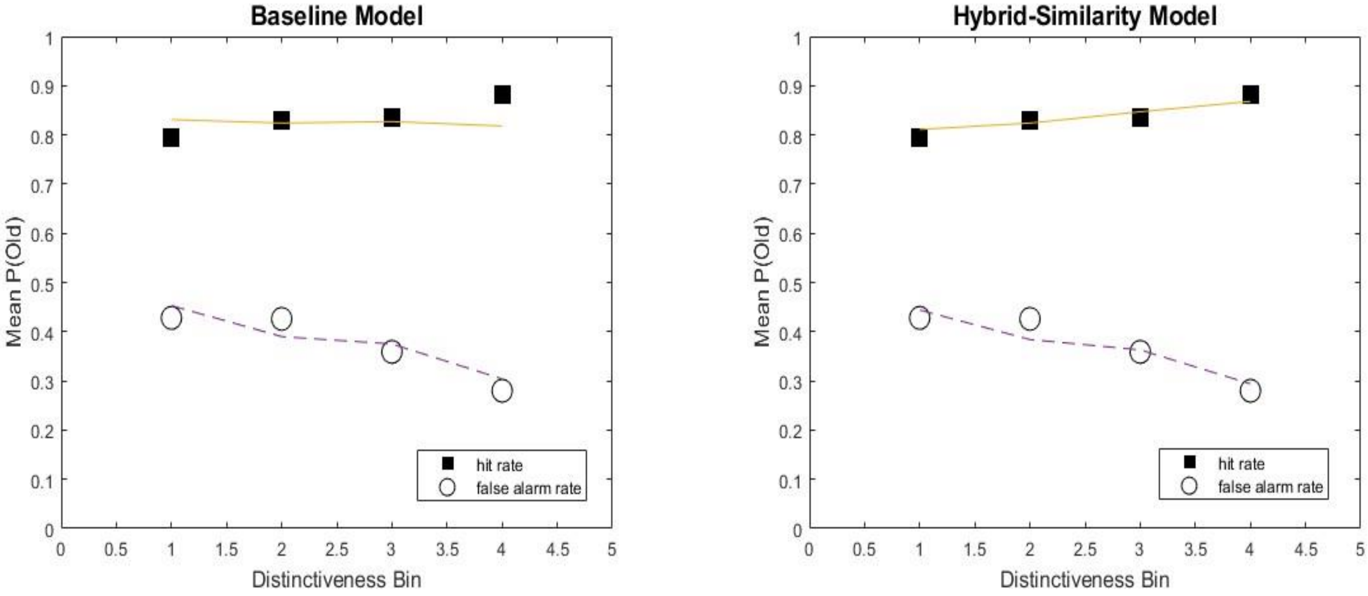


Figure 7. Experiment 1: Mean observed and predicted hit and false-alarm rates as a function of distinctiveness-rating bin.

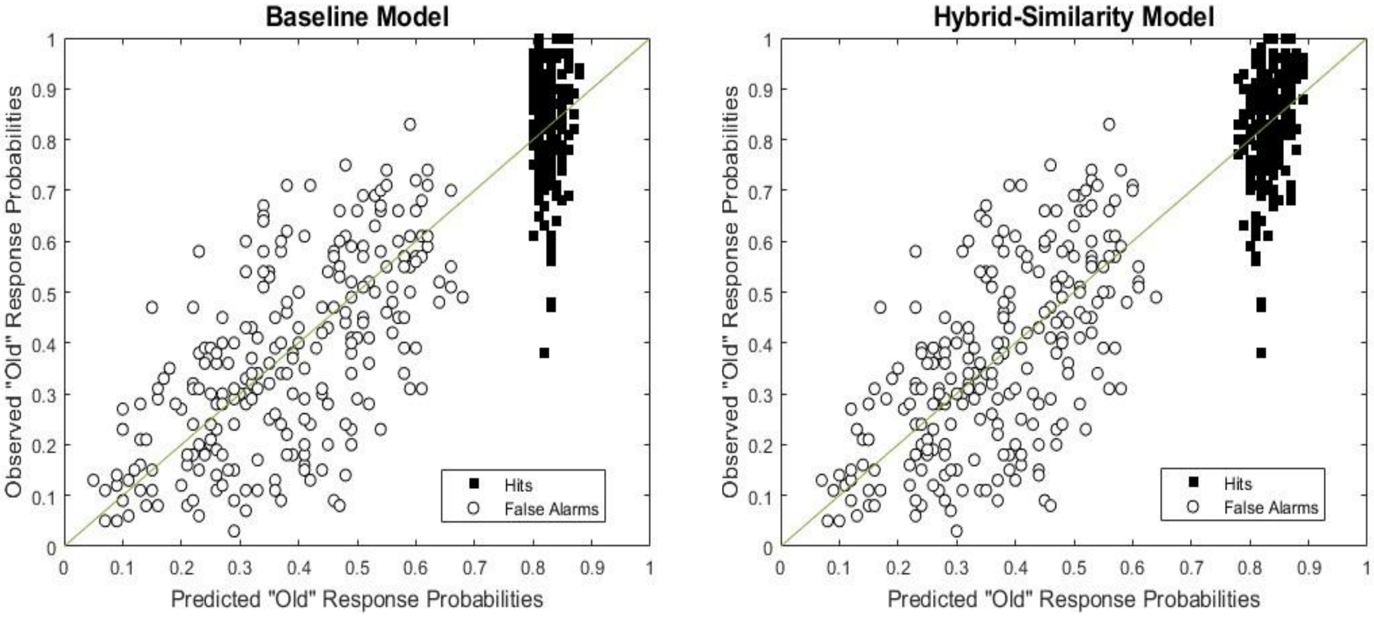


Figure 8. Experiment 1: Mean observed against predicted hit and false-alarm rates for the individual 240 items.

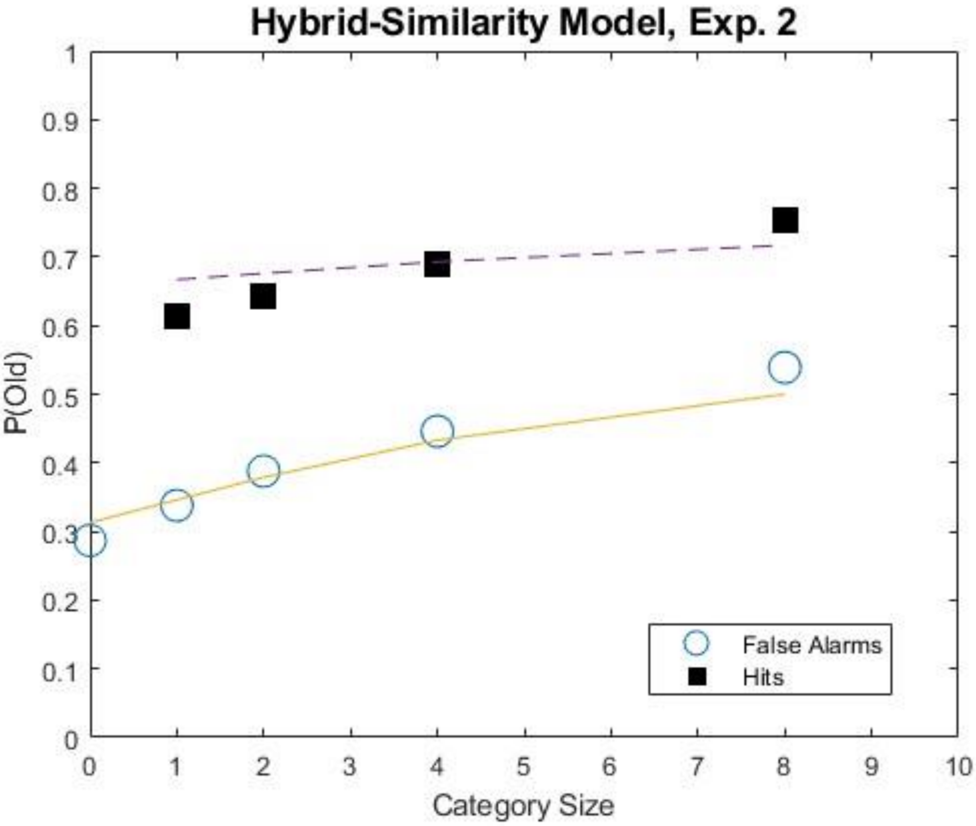


Figure 9. Experiment 2: Mean observed and predicted false-alarm and hit rates as a function of category size.

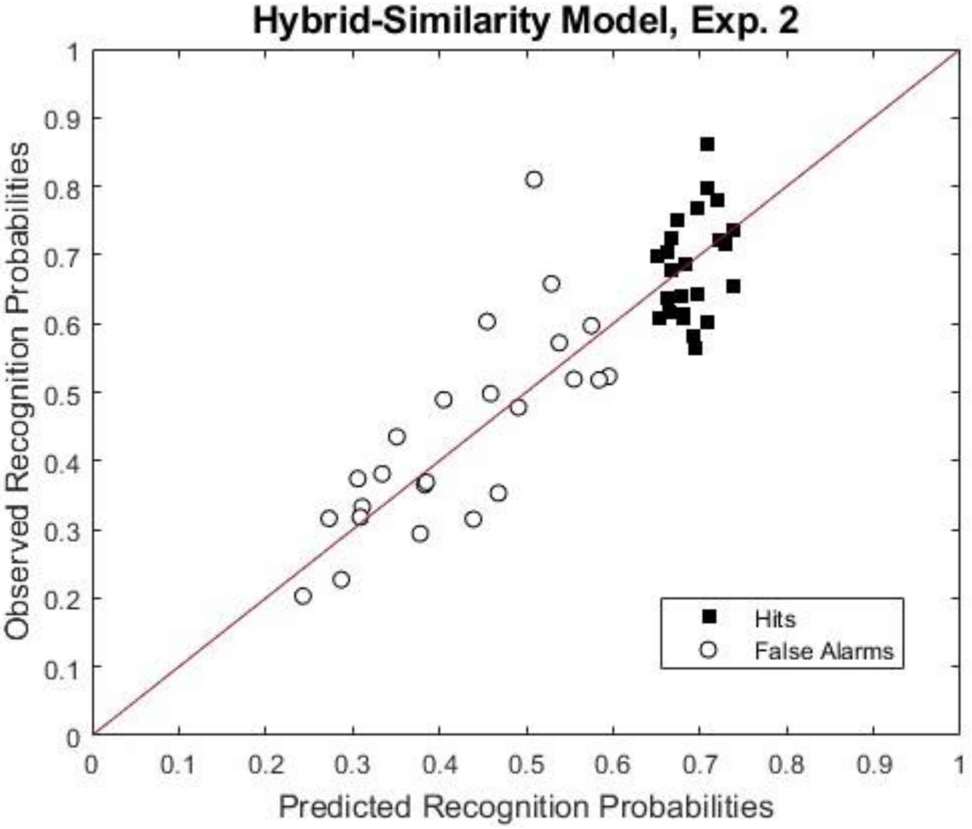


Figure 10. Experiment 2: Mean observed against predicted hit and false-alarm rates for each of the 24 categories.

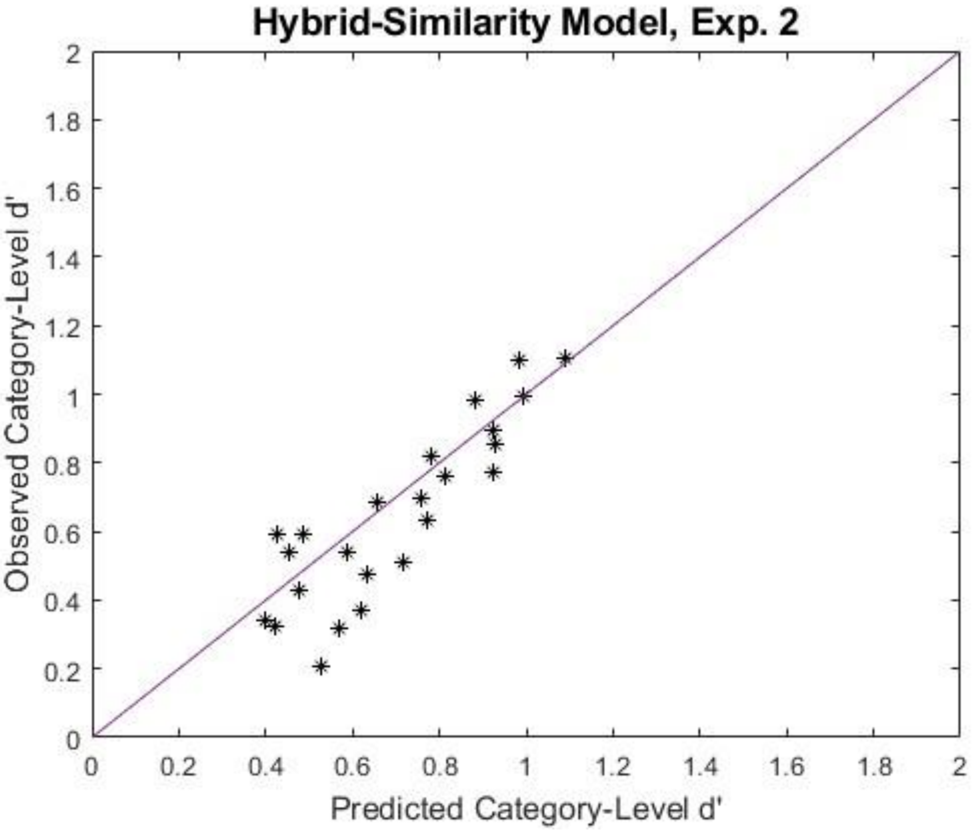


Figure 11. Experiment 2: Mean observed-against-predicted category-level d’ values for the 24 categories.

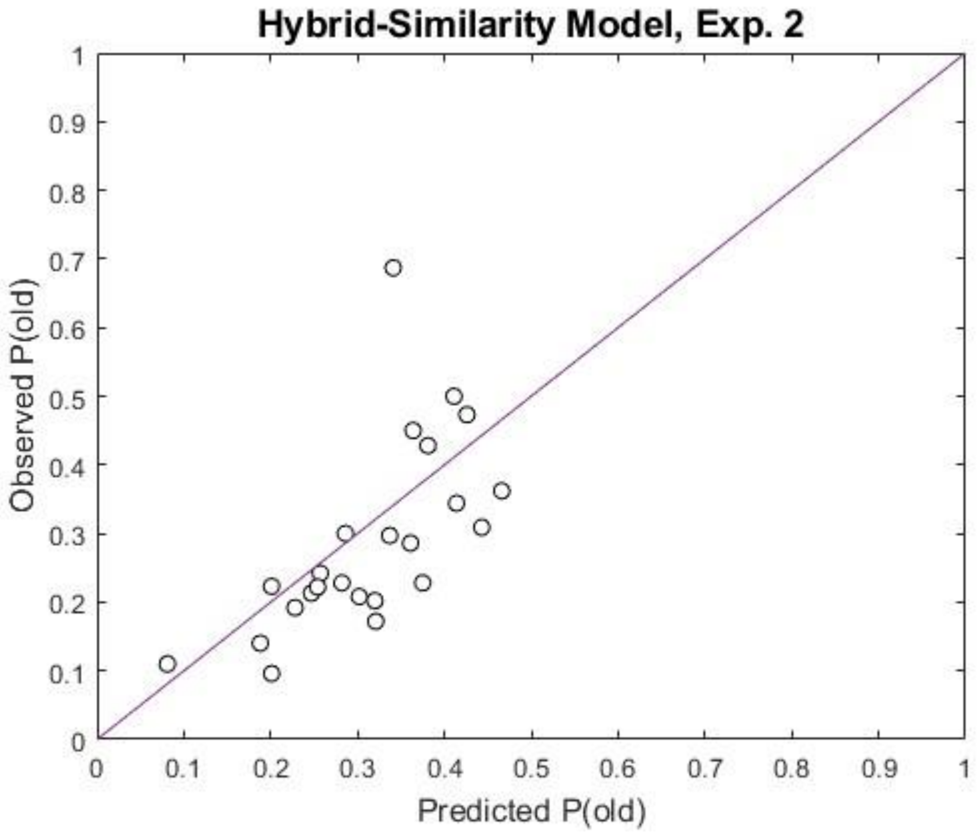


Figure 12. Experiment 2: Mean observed-against-predicted false-alarm rates for the 24 size-0 categories.

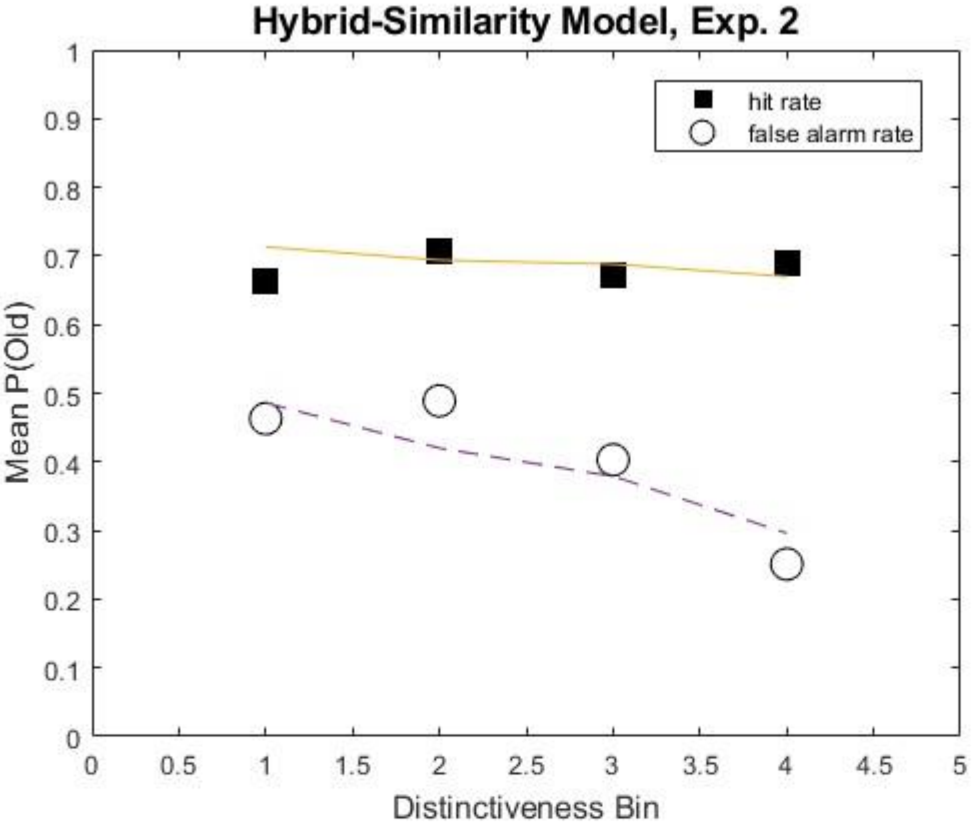


Figure 13. Experiment 2: Mean observed and predicted hit and false-alarm rates as a function of distinctiveness-rating bin.

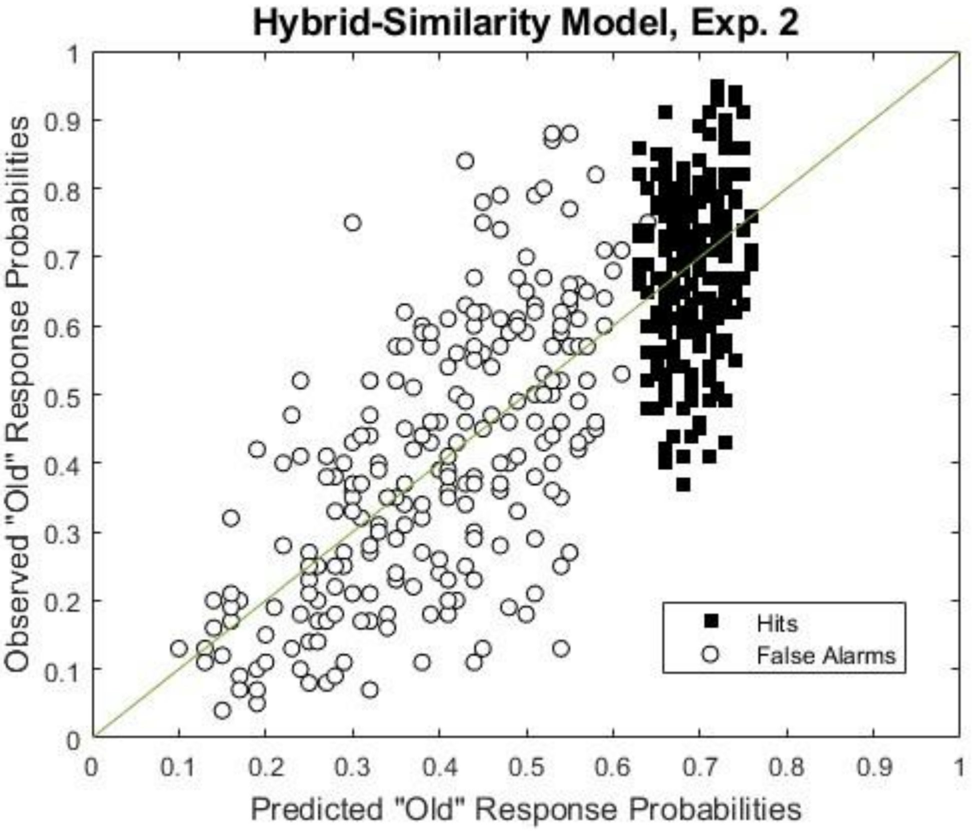


Figure 14. Experiment 2: Observed against predicted hit and false-alarm rates for the individual 240 items.

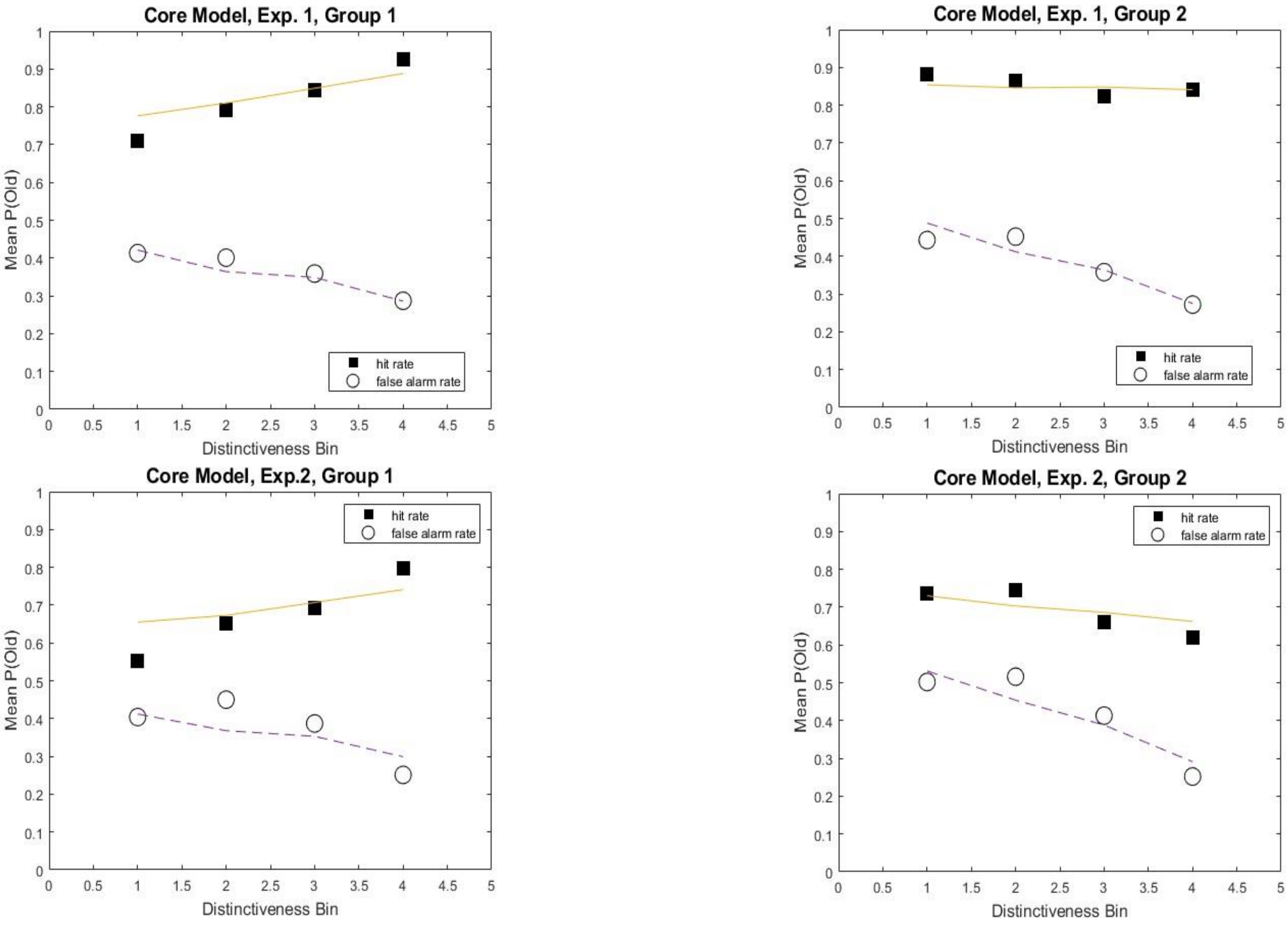


Figure 15. Experiment 1 and 2 subgroup modeling analyses. Plots of observed and core-model predicted hit- and false-alarm rates as a function of distinctiveness-rating bin.

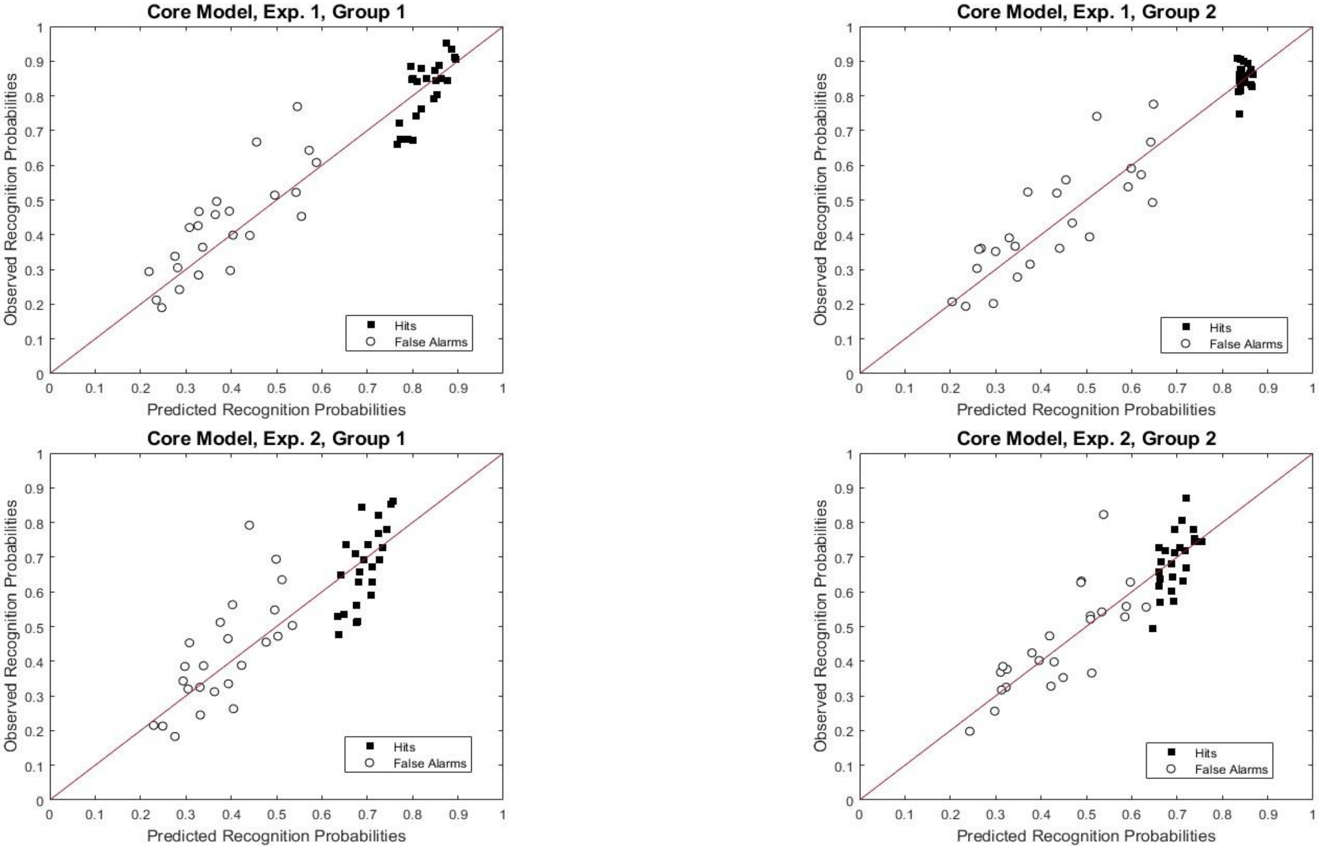


Figure 16. Experiment 1 and 2 subgroup modeling analyses. Plots of observed against core-model predicted category-level hit- and false-alarm rates for each of the 24 categories.

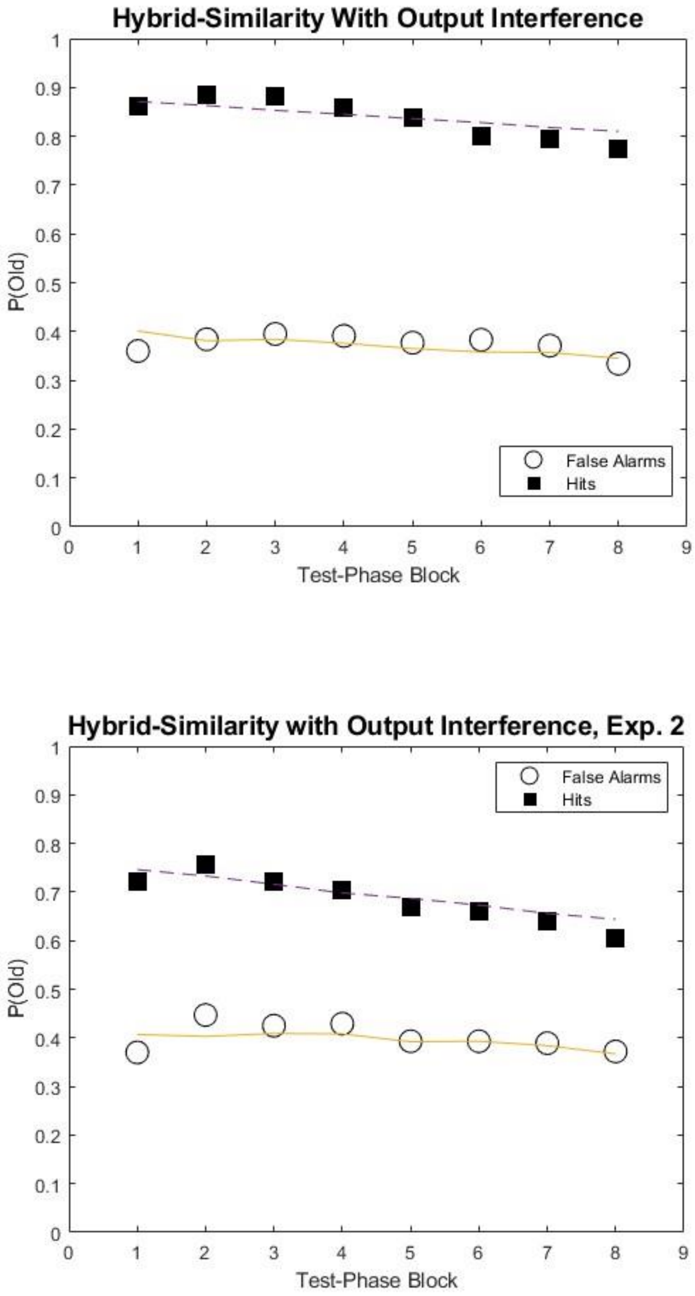


Figure E1. Mean observed and predicted hit and false-alarm rates as a function of 10-trial blocks during the test phase. The predictions are from the extension of the hybrid-similarity model that incorporates output interference. Top panel = Exp. 1, bottom panel = Exp. 2.