Predicting the Ease of Human Category Learning Using Radial Basis Function Networks

Brett D. Roads
b.roads@ucl.ac.uk
Michael C. Mozer
mcmozer@google.com
Department of Computer Science and Institute of Cognitive Science,
University of Colorado Boulder, Boulder, CO 80309-0430, U.S.A.

Our goal is to understand and optimize human concept learning by predicting the ease of learning of a particular exemplar or category. We propose a method for estimating *ease values*, quantitative measures of ease of learning, as an alternative to conducting costly empirical training studies. Our method combines a psychological embedding of domain exemplars with a pragmatic categorization model. The two components are integrated using a radial basis function network (RBFN) that predicts ease values. The free parameters of the RBFN are fit using human similarity judgments, circumventing the need to collect human training data to fit more complex models of human categorization. We conduct two category-training experiments to validate predictions of the RBFN. We demonstrate that an instance-based RBFN outperforms both a prototype-based RBFN and an empirical approach using the raw data. Although the human data were collected across diverse experimental conditions, the predicted ease values strongly correlate with human learning performance. Training can be sequenced by (predicted) ease, achieving what is known as *fading* in the psychology literature and *curriculum learning* in the machine-learning literature, both of which have been shown to facilitate learning.

1 Introduction

Visual categorization is a critical skill in many professions, including radiology, dermatology, and satellite imagery analysis. The economic importance of visual categorization has motivated substantial research aimed at reducing the cost of training visual experts. One approach for improving training is to predict where learners are likely to make mistakes and adjust

B.R. is now at the Department of Experimental Psychology, University College London. M.M. is now at Google Brain.

Neural Computation **33, 376–397** (2021)
https://doi.org/10.1162/neco_a_01349

the training protocol appropriately (Nosofsky, Sanders, Zhu, & McDaniel, 2019). For example, a training protocol could train on easy exemplars only (Giguère & Love, 2013; Hornsby & Love, 2014; Patil, Zhu, Kopeć, & Love, 2014) or introduce easy exemplars first and then progress to harder ones (McLaren & Suret, 2000; Lindsey, Mozer, Huggins, & Pashler, 2013; Pashler & Mozer, 2013; Roads, Xu, Robinson, & Tanaka, 2018). Accurate assessment of a learner's knowledge also requires knowing something about stimulus difficulty: if an assessment is composed solely of easy exemplars, it will be challenging to distinguish between trained and untrained individuals. In this work, the relative ease of learning a particular exemplar is referred to as the *exemplar ease value*. The average exemplar ease value of all exemplars in a category is referred to as the *category ease value*. Exemplar and category ease values are collectively referred to as *ease values*. A variety of methods have been proposed for computing ease values, each with its own advantages and disadvantages. The primary objective of this work is to demonstrate a flexible and practical method for predicting ease values.

In this work, we leverage human similarity judgments in order to estimate ease values in a flexible and cost-effective manner. By collecting human similarity judgments, we can infer both a stimulus representation and the similarity function that operates over the representation, which we refer to as *psychological embedding* (Roads & Mozer, 2019). Modeling the stimulus representation enables us to anticipate how the arrangement of stimuli (in feature space) will affect ease of learning. For example, a lone exemplar—surrounded by exemplars belonging to a different category—is likely to be harder to learn. Given a psychological embedding, we explore two variants of a radial basis function network (RBFN) that make different ease value predictions. One variant is based on exemplars, the other on category prototypes. The RBFNs are compared to a third model, which is neutral to psychological theory and merely counts the frequency of different types of similarity judgments in order to determine ease values. Each model is evaluated by comparing the predicted ease values to empirically derived ease values. The proposed approach combines the positive qualities of existing methods while solving many of the practical limitations. For example, it applies to nonbinary classification and can gracefully handle overlapping categories.

2 Related Work

The most direct method of computing ease values is to derive them from the behavior of human participants performing the categorization task of interest. Stimuli that are easy to learn will be categorized correctly more often than stimuli that are difficult. The trial responses of multiple participants can be used to determine the error statistics associated with each exemplar or category. This approach can quickly become impractical for small-scale research programs. To obtain a high-power estimate of accuracy, multiple responses must be collected for each exemplar. While this may be feasible for small stimulus sets, the cost can become impractical for larger, real-world stimulus sets. These issues will bear on the experimental analyses presented later.

Another intuitive alternative is to obtain expert difficulty ratings for each stimulus (Evered, Walker, Watt, & Perham, 2014). Experts have learned which visual features are diagnostic and can draw on their experience to recall how difficult it was to learn a particular feature or category. Expert consultation fees can be (justifiably) expensive, making expert-based norms a nonoption for research on a tight budget. Furthermore, the rapid nature of visual expertise (Tanaka & Taylor, 1991) may also make it difficult for long-time experts to introspectively dissect material they have long since mastered.

In some cases, it is possible to collect norms that circumvent the cost of experts. By focusing on binary categorization tasks that rely on widely available knowledge, the general population can be recruited to norm stimuli. For example, Salmon, McMullen, and Filliter (2010) collected ratings for a set of images in which participants rated each stimulus on its "graspability." These norms can then been used by other researchers to create arbitrary category boundaries where a stimulus's distance from the category boundary determines its ease value (Khan, Mutlu, & Zhu, 2011; Lindsey et al., 2013). While this approach is relatively cost-effective, it does not work for real-world tasks where the concept of interest requires expert knowledge (e.g., malignant versus benign skin lesions).

Distance to the category boundary can also be used for artificial stimulus sets where the stimulus representation is known by design (Giguère & Love, 2013; Khan et al., 2011; Patil et al., 2014; Spiering & Ashby, 2008). Given linearly separable categories, it is possible to learn a hyperplane separator that defines a category boundary. Stimuli near the category boundary require more precise specification of the separator and are therefore less likely to be classified correctly given limited training data. Although distance to the category boundary is a simple and cost-effective approach, it becomes problematic when category distributions are highly overlapping or there are more than two categories.

Theories of human category learning provide an alternative approach for estimating ease values. If a model is capable of predicting the likelihood that a particular stimulus will be categorized correctly (Nosofsky, 1986; Kruschke, 1992; Love, Medin, & Gureckis, 2004; Nosofsky, Sanders, & McDaniel, 2018), then the model's predictions can be used to estimate ease values. Yet the predictive power of a category learning model comes at a cost. First, human category learning models rely on behavioral data to tune the model's free parameters. Consequently, behavioral data must be collected in a manner that is consistent with the model's assumptions, such as requiring an alternative forced-choice response format. Second, human category learning models also require a known stimulus representation.

The proposed approach addresses a number of critical limitations associated with the existing approaches. It works for any arbitrary category structure, not just binary tasks or linearly separable category structure. The approach is also designed to yield modular and reusable assets. In this work, the learned psychological embedding is used to compute ease values, but the psychological embedding makes relatively few assumptions, enabling it to be reused in other research projects. Finally, the proposed approach keeps costs down by collecting human data from widely available nonexperts and using a low-effort task. All of these features make the proposed approach a flexible and practical method for estimating ease values.

3 A Flexible and Practical Method for Estimating Ease Values

The proposed approach builds on the position that human category learning can be treated as density estimation (Ashby & Alfonso-Reese, 1995). One method for performing density estimation is to use radial basis function networks (RBFNs). At its simplest, an RBFN consists of three layers: an input layer, a hidden layer, and an output layer. A number of free parameters govern how activation propagates through the network. These free parameters belong to one of two network components: the similarity kernel (i.e., radial basis function) and the association weight matrix. These two components are implemented by seminal category learning models, such as the generalized context model (Nosofsky, 1986) and ALCOVE (Kruschke, 1992). These category learning models fit all free parameters using human training data.

In contrast to the standard approach in the human literature, we use a method that fits the free parameters associated with the similarity kernel using human similarity judgments. The remaining free parameters are fixed based on psychological theory. In the following three sections, we outline a simple class of RBFNs, detail how the free parameters are inferred, and describe two RBFN variants.

3.1 A Simple RBFN Model.

A standard method of representing stimuli is to treat each exemplar as a point in a multidimensional feature space. While an infinite number of potential visual and nonvisual features could be used, we assume that we can identify the subset of features that are most salient and relevant for the categorization task. The stimulus representation is denoted by $\mathbf{Z}$, where $\mathbf{z}_i$ indicates the D-dimensional feature vector of the ith stimulus. The input layer activations encode the query stimulus, $\mathbf{x} = \mathbf{z}_{query}$.

The similarity kernel $s\left(\mathbf{z}, \mathbf{z}'\right)$ specifies how similarity between two stimuli ($\mathbf{z}$ and $\mathbf{z}'$) decays as a function of distance in feature space. The form of the similarity kernel is constrained by existing psychological theory (Jones, Love, & Maddox, 2006; Jones, Maddox, & Love, 2006; Nosofsky, 1986;

Shepard, 1987). As done in previous work (Roads & Mozer, 2017, 2019; Roads et al., 2018), we integrate various psychological models into a general form to obtain

$$s\left(\mathbf{z}, \mathbf{z}'\right) = \exp\left(-\beta \left\|\mathbf{z} - \mathbf{z}'\right\|_{\rho,\mathbf{w}}^{\tau}\right) + \gamma, \tag{3.1}$$

where β, ρ, τ, and γ control the gradient of generalization. The norm $\left\|\mathbf{z} - \mathbf{z}'\right\|_{\rho,\mathbf{w}}$ denotes the weighted Minkowski distance:

$$\left\|\mathbf{z} - \mathbf{z}'\right\|_{\rho,\mathbf{w}} = \left(\sum_{j=1}^{D} w_j \left|z_j - z'_j\right|^{\rho}\right)^{\frac{1}{\rho}} \text{ where } w_j \geq 0 \text{ and } \sum_{j=1}^{D} w_j = D.$$

The parameter ρ controls the type of distance (e.g., $\rho = 2$ yields Euclidean distance), and D indicates the dimensionality of the embedding. The weights w correspond to attention weights and allow the similarity kernel to model differences in how individuals or groups attend to different dimensions in the psychological embedding. The weights sum to D so that when all the weights are equal ($w_j = 1$), we recover the standard (unweighted) Minkowski distance. While the remainder of this work assumes this similarity kernel, other differentiable similarity kernels could be substituted without loss of generality. For convenience, the parameters ρ, β, τ, and γ are denoted by the set variable $\boldsymbol{\theta}$.

The activation of the ith hidden unit $h_{,i}$ is determined by the similarity kernel (see equation 3.1),

$$h_i = s\left(\mathbf{x}, \mathbf{z}_i\right). \tag{3.2}$$

The vector $\mathbf{z}_i$ specifies the location (in feature space) of a particular basis function. An RBFN network typically has multiple basis functions. The basis function locations can be determined in a number of ways, which are discussed shortly.

The hidden layer is connected to the output layer via a fully connected association weight matrix W. The output layer has the same number of units as the number of categories in the categorization task. A normalizing softmax operation is applied to the raw output activations to yield output probabilities,

$$\mathbf{y} = \text{softmax}\left(\mathbf{h}W\right). \tag{3.3}$$

The ease value of the query stimulus is the probability that the query is correctly categorized.

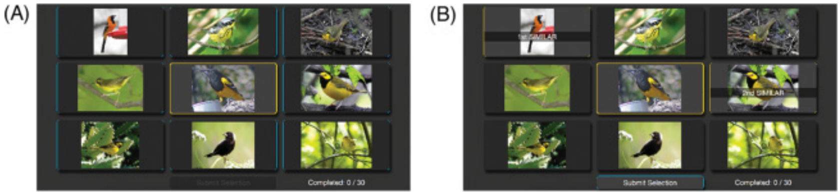

Figure 1: Sample displays shown to participants. The center image is the query image, and the surrounding images are the reference images. (A) Initially no reference examples are selected. (B) After participants make their selection, the two selected references are highlighted.

3.2 Joint Inference of a Psychological Embedding and Similarity Kernel. In a typical category learning research paradigm, the stimulus representation Z is determined independent of the category learning model. For example, one could use a set of hand-coded features, low-level computer vision features, or features from a pretrained deep neural network to determine the stimulus representation. The stimulus representation is then treated as fixed, and the remaining free parameters are fit using human training data.

In contrast, the proposed approach determines both the stimulus representation and the parameters of the similarity kernel at the same time, without using behavioral data from a human training experiment. This is achieved by applying an embedding algorithm to human similarity judgments in order to jointly infer the parameters of a similarity kernel and stimulus representation (Roads & Mozer, 2019). While many algorithms exist for determining a stimulus representation, such as metric multidimensional scaling, nonmetric multidimensional scaling, and t-distributed stochastic triplet embedding (Van Der Maaten & Weinberger, 2012), our approach leverages psychological theory to constrain the possible solutions. We refer to the inferred stimulus representation and similarity kernel as a *psychological embedding*. The procedure for collecting human similarity judgments and the mechanisms of the embedding algorithm are summarized below.

The first step to inferring a psychological embedding is to collect human similarity judgments for the set of stimuli. Inspired by approaches used in the computer vision community (Wah et al., 2014), human similarity judgments are collected by having novice participants view displays composed of nine randomly selected images arranged in a 3-by-3 grid (see Figure 1). Each display is composed of a query image (center image) and eight reference images (surrounding images). Participants are asked to select the two reference images most similar to the query image. When they make their selection, they also indicate which reference is most similar and second most similar. The ith judged display is denoted using a vector

$\mathcal{D}_i = (q_i, a_i, b_i, c_i, d_i, e_i, f_i, g_i, h_i)$, where q_i is a scalar indicating the query image and a_i to h_i are scalars indicating the reference images. In this arrangement, a_i and b_i represent the most similar and second most similar references respectively. For convenience, $\mathcal{R}_i$ indicates the set of reference images of $\mathcal{D}_i$. The set of all judged displays is indicated by $\mathcal{D}$.

Next, the set of all judged displays $\mathcal{D}$ is used to jointly infer a stimulus representation and similarity kernel. Given a set of observations, the likelihood of the data given the model parameters is

$$\mathcal{L} = \prod_i p\left(\mathcal{D}_i | \mathbf{Z}, \boldsymbol{\theta}\right). \tag{3.4}$$

In the case where participants select and rank two reference images, the likelihood of a single judged display is

$$p\left(\mathcal{D}_i | \mathbf{Z}, \boldsymbol{\theta}\right) = p\left(a_i | \mathbf{Z}, \boldsymbol{\theta}\right) p\left(b_i | a_i, \mathbf{Z}, \boldsymbol{\theta}\right). \tag{3.5}$$

Given a similarity kernel, the likelihood of participant selections is modeled in the same spirit as Luce's ratio of strengths formulation (Luce, 1959) and classic similarity choice models (Shepard, 1957, 1958; Nosofsky, 1985, 1986). The basic principle is that the probability of selecting a given reference is proportional to the similarity between the query and that reference:

$$p\left(\mathcal{D}_i | \mathbf{Z}, \boldsymbol{\theta}\right) = \frac{s\left(\mathbf{z}_{q_i}, \mathbf{z}_{a_i}\right)}{\sum_{r \in \mathcal{R}_i} s\left(\mathbf{z}_{q_i}, \mathbf{z}_r\right)} \frac{s\left(\mathbf{z}_{q_i}, \mathbf{z}_{b_i}\right)}{\sum_{r \in \mathcal{R}_{i \neg a_i}} s\left(\mathbf{z}_{q_i}, \mathbf{z}_r\right)}, \tag{3.6}$$

where $s\left(\mathbf{z}_i, \mathbf{z}_j\right)$ is the similarity kernel defined in equation 3.1.

After maximizing the log likelihood using gradient descent, we obtain a stimulus representation and a corresponding similarity kernel that models human-perceived similarity. During inference of the psychological embedding, we assume a single group and therefore set all attention weights to one.

3.3 Candidate RBFN Implementations. After inferring a psychological embedding, an RBFN requires two additional pieces of information: the basis function locations and the association weight matrix. Existing *exemplar* and *prototype* models provide guidance on setting these parameters (Kruschke, 1992; Minda & Smith, 2002; Nosofsky, 1986; Smith & Minda, 1998). Prototype and exemplar models have different assumptions regarding the encoding and storing of experience. Exemplar models assume that a separate memory is storage for each exemplar. Prototype models assume that only an average memory is stored for all exemplars belonging to the same category. Some research suggests that learners employ a prototype-based representation early in learning and transition to an exemplar-based representation as they approach mastery (Smith & Minda, 1998). If learners only encode category-level statistics, then a prototype model should make better predictions. If learners encode information about individual exemplars, then an exemplar model should make better predictions. Since both models provide plausible alternatives for computing ease values, we consider two different variants of an RBFN.

An exemplar model takes a fine-grained approach and places a basis function at the embedding location of each exemplar. The association weight matrix is fixed by assuming that each hidden unit is only connected to the output unit that corresponds to its category. This RBFN variant closely resembles the generalized context model (Nosofsky, 1986) except that there is no softmax free parameter that adjusts the determinism of human responses.

A prototype model takes a coarser approach and locates a basis function at the centroid of all exemplars belonging to a given category. In other words, a single basis function is used to represent each category. In a prototype model, we assume that the association weight matrix is the identity matrix that connects each category basis function to its appropriate output unit. The prototype implementation requires a bit more care to set up properly. Since an embedding is inferred on individual stimuli and the prototype basis function represents a category average, the parameters of the similarity kernel must be appropriately constrained and adjusted. First, the embedding algorithm is constrained to infer solutions where $\rho = 2$, $\tau = 2$, and $\gamma = 0$. Second, one multivariate gaussian is fit for each category. The fitted gaussians are then used as the corresponding basis function for each category. Since the fitted gaussians are not constrained to be spherical, the basis functions are not radial basis functions. However, the additional flexibility was allowed to give the prototype model the best chance at making good predictions.

Ease values are computed using a leave-one-out approach. Let us denote the set of all stimuli less the ith stimulus as $\mathcal{I}_{\neg i}$. An RBFN is fit using the set $\mathcal{I}_{\neg i}$ to determine the locations of the basis functions. The ease value of the ith stimulus is then determined by the probability that it is correctly classified. The inferred psychological embedding is a window into a novice's perception of the world. Unlike experts, they have not yet learned which visual features are diagnostic versus uninformative. The leave-one-out approach allows us to anticipate how the neighboring category structure will influence the ease of learning a particular stimulus.

3.4 An Alternative Count-Based Model. The proposed approach constrains inference using psychological theory. While well motivated by cognitive science research, the approach employs a number of theoretical assumptions. As an alternative, it is possible to create a model that is relatively unconstrained by theory but is trained on the same behavioral data.

Instead of inferring a psychological embedding, the human similarity judgments are treated as implicit categorization trials that used to assemble error statistics. This approach is similar to norming procedures used by other researchers (Hornsby & Love, 2014). This empirically based, atheoretic count-based model leverages the same data as the exemplar-based and prototype models but does so directly through the response counts and not through the intermediate representation of an embedding. The count-based model therefore serves as a control to ascertain the value of the inferred psychological embedding.

Each judged display tells us how often a given exemplar was judged to be similar to an exemplar of the same or a different category. For example, if a participant sees a query stimulus belonging to category j and selects two references that also belong to category j, this provides two votes that the query stimulus is easy. If a participant sees another query stimulus belonging to category j but selects two references that belong to category j and k, respectively, this provides one vote that the query stimulus is easy. By looping over all judged displays, a simple count matrix can be assembled that tracks how often a given exemplar is judged to be more similar to a reference of the same category versus a reference of a different category. After looping through all judged displays, the count matrix can be normalized such that each row sums to one. Each row in the normalized count matrix gives the probability that a particular exemplar will be judged as more similar to a reference of the same category versus a reference of a different category. These probabilities can be used to estimate the ease value.

4 Experiments

The above models encompass three intuitive methods for predicting difficulty. A good test of the proposed approach is to compare the predicted ease values to empirically derived ease values. Using two different human training experiments, we compute empirical ease values and compare them to the ease values predicted by three different models: an exemplar-, prototype-, and count-based model. The comparison will determine if the general approach is valuable and which model makes the most accurate predictions. For each experiment, we provide a brief description of the experimental design followed by a comparison of the predicted and observed ease values. In both experiments, the goal is to compare the predictive power of the three candidate models. The first experiment compares models based on their ability to predict generalization performance across three different points during training. The second experiment focuses on predicting generalization performance at the end of training.

A similar implementation of the proposed approach was previously demonstrated with a two-way, alternative forced-choice task (Roads et al., 2018). This work extends the previous work in three ways. First, ease values are predicted for a multiclass task rather than a binary classification

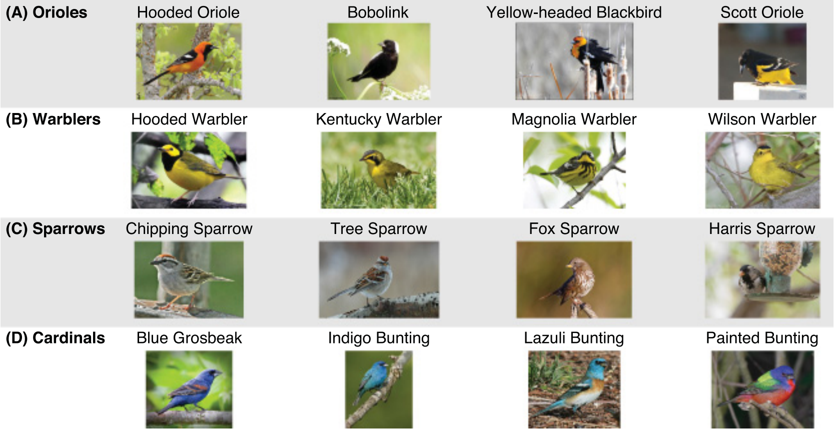

Figure 2: Example stimuli of the bird species used in this work. Each row contains four similar bird species, each of which belongs to the same or a similar taxonomic family. Validation experiment 1 used the 12 bird species in rows A to C. Validation experiment 2 used all 16 bird species.

task. Second, these values are compared to training data collected under a wider variety of conditions. Third, this work considers alternative models for computing ease values.

4.1 Validation Experiment 1. The goal of the first validation experiment is to predict the ease values associated with categorizing a set of 156 bird images representing 12 different categories. Empirical ease values are derived from experiment 1 of a previously conducted study (Roads & Mozer, n.d.a). The exemplar-, prototype-, and count-based models use similarity judgments that were previously collected for a similarity judgment database (Roads & Mozer, 2019). An abbreviated description of the human training study is included, with an emphasis given to details pertinent to the current work.

4.1.1 Methods.

Participants. Two sets of participants were used in this experiment. A psychological embedding was constructed from similarity judgments collected from 232 participants (Roads & Mozer, 2019). Human training data were collected from 160 participants. All participants were recruited from Amazon Mechanical Turk and paid at a rate of approximately $8.00 per hour for their time and effort.

Materials. A set of 156 bird images, representing 12 different species (see Figures 2A to 2C), was selected from the CUB-200 data set (Wah, Branson, Welinder, Perona, & Belongie, 2011). For each species, 13 exemplars were hand-picked by the first author. Exemplars were selected to make sure that the resolution of the image was sufficiently high, the bird was clearly visible in the image, and the bird exhibited the visual features characteristic of the species. To ensure a sufficiently challenging and representative task, species were selected such that there were three groups of four visually similar species.

Procedure. Empirical ease values and predicted ease values were derived using separate procedures. The procedure for deriving empirical ease values from training data is described first. The procedure for computing ease values from the candidate models is described second.

During the training experiment, participants completed trials at their own pace for approximately one hour. At the beginning of the experiment, the set of stimuli was randomly partitioned into a practice and, assessment set. The practice set was composed of seven exemplars from each category, while the assessment set was composed of the remaining six exemplars from each category. The practice and assessment sets were further partitioned into mini-sets of 12 exemplars containing one exemplar from each category. Each mini-set was used to create a mini-block composed of 12 trials.

At the highest level, the experiment was composed of three parts. Each part consisted of a practice phase, which took a fixed time of 15 minutes, followed by an assessment phase. During a practice phase, trials were arranged into mini-blocks consisting of exemplars from the practice set. If a participant made it through all practice mini-blocks, the sequence of practice mini-blocks was repeated. Each assessment phase was composed of a fixed number of trials. During each assessment phase, two mini-blocks (24 trials) were shown to participants. Once the exemplars were shown during the assessment phase, they were added to the practice set for use in the next practice phase. Critically, all trials presented during the assessment phase used unseen stimuli. On all trials (both practice and assessment), a query stimulus was presented along with a blank response box. Participants typed the name of the category corresponding to the query stimulus. After submitting their response, participants received corrective feedback. Participants were not scored based on capitalization, and answers within an edit distance of two were marked as correct.

In addition to this basic setup, some participants were assigned to conditions where they could request clues on some of the practice trials. The clue-enabled trials are referred to as *enriched trials*. In the *standard* condition, participants were never given enriched training trials. In the *enriched-3* condition, participants were given enriched training trials in all three parts of the experiment. In the *enriched-2* condition, participants were given enriched training trials in the first two parts of the experiment. In the *enriched-1* condition, participants were given enriched training trials only on the first part of the experiment. For the purpose of this work, we make no

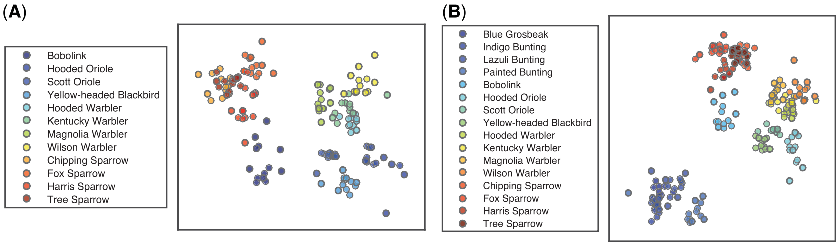

Figure 3: A two-dimensional psychological embedding of (A) the 156 unique bird images used in validation experiment 1 and (B) the 208 bird images used in validation experiment 2. Each point represents a unique stimulus and is colored according to category (i.e., species). Images judged to be similar are located closer together than images judged to be dissimilar.

distinction among the conditions and use all conditions in order to predict ease values. Each condition contained 40 participants.

Empirical ease values were computed using participant responses from the assessment trials only. For each query exemplar in the assessment trials, the exemplar ease value was determined by counting the total number of times the query exemplar was categorized correctly and dividing it by the total number of times that it was shown. Each category ease value was computed by counting the number of times a category was classified correctly and dividing it by the number of times the category was shown.

Predicted ease values were obtained by inferring a psychological embedding from 7520 2-choose-1 displays and 4733 8-choose-2 displays. Two separate embeddings were inferred: one for the exemplar model and a second for the prototype model. The psychological embedding used for the prototype model was constrained to have a gaussian similarity kernel in L_2 space in order to cohere with the later gaussian fitting procedure used by the prototype-based RBFN. The dimensionality of the embedding was determined using a cross-validation procedure (Roads & Mozer, 2019). For visualization purposes, a two-dimensional embedding of the stimuli is provided in Figure 3A. After inferring the respective embeddings, the exemplar and prototype models were used to generate predicted exemplar ease values. The raw similarity judgments were used by the count-based model to generate exemplar ease values. Predicted category ease values were computed by taking the average exemplar ease value of all exemplars within a particular category.

Each model was evaluated by computing the Spearman rank correlation coefficient between the predicted ease values and the empirical ease values. Separate correlations were computed for exemplar and category ease values. In order to determine the best model, the Steiger's (1980) method was

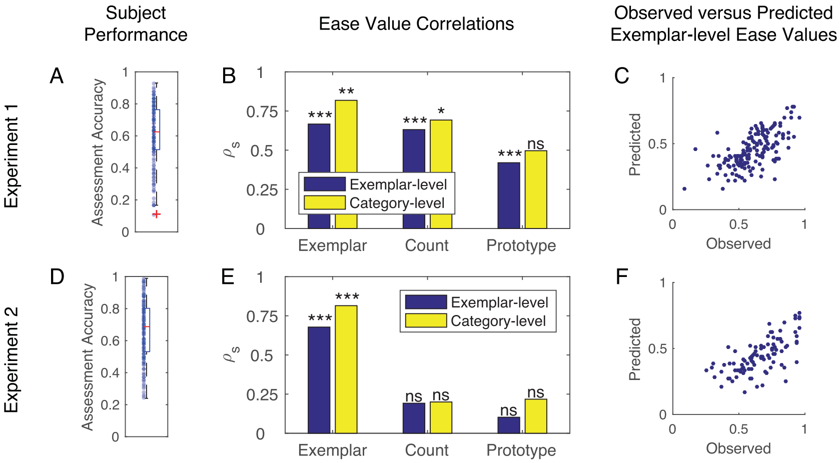

Figure 4: Results of experiments 1 and 2. Participant performance on assessment trials for (A) experiment 1 and (D) experiment 2. The data are plotted at the individual level and using a box plot. The red line indicates the median value, and the bottom and top edges of the box indicate the 25th and 75th percentiles. The whiskers extend to the most extreme data points not considered outliers, and the outliers are plotted individually using the + symbol. The Spearman's rank correlation coefficient between empirically derived ease values and predicted ease values for (B) experiment 1 and (E) experiment 2. Correlations were computed for exemplar ease values and category ease values. Symbols above the bars indicate the significance of the correlation. Scatter plots of observed versus predicted exemplar-level ease values using the exemplar model for (C) experiment 1 and (F) experiment 2.

used to test whether the difference between the different correlation scores was significant.

4.1.2 Results. The primary hypothesis of this work is that a simple RBFN can be used to predict ease values for a set of stimuli. Three models were proposed as candidates, with the hypothesis that they would perform differently.

On average, each stimulus appeared in an assessment trial 74 times ($SD = 6$), providing a relatively large sample size for estimating the empirical ease value of each stimulus. Across all assessment trials, accuracy was moderately high ($M = .61$, $SD = .19$), indicating that participants were able to learn parts of the categorization task (see Figure 4A).

Results indicated that simple models can be used to predict difficulty (see Figure 4B). However, results were inconclusive regarding the best model. Tests of the three pair-wise comparisons were conducted using Bonferroni adjusted alpha levels of .0167 per test (.05/3). Results indicated that the correlation was significantly lower for the prototype model ($\rho_s = .42$) than both the exemplar ($\rho_s = .67$), $T_2(155) = 3.96$, $p < .001$ and count model ($\rho_s = .63$), $T_2(155) = 2.97$, $p = .003$. The pairwise comparison between the correlations for the exemplar and count model was nonsignificant. The same testing procedure was applied to the correlations between empirical category-level difficulty and predicted category-level difficulty. The correlations for the exemplar, count, and prototype models were $\rho_s = .82$, $\rho_s = .69$, and $\rho_s = .50$, respectively. Results indicated that all category-level pairwise comparisons were nonsignificant.

4.1.3 Discussion. The moderately high correlation results of validation experiment 1 are encouraging because they show that a simple RBFN can explain both category- and exemplar-level classification. A simple RBFN can therefore be used to predict ease values. This is an exciting possibility given its potential usefulness in category training. However, the experiment did not clearly distinguish between the candidate models. One possible explanation is that the empirical observations of experiment 1 were derived from assessment trials that occurred near the beginning, middle, and end of the experiment. It is possible that prototype models describe early, novice-like knowledge representations, while exemplar models do a better job of capturing the minutia of expert knowledge representations (Smith & Minda, 1998). This perspective is consistent with computational models such as SUSTAIN (Love et al., 2004) and nonparametric Bayesian models of category learning (Sanborn, Griffiths, & Navarro, 2010), which develop the complexity of the knowledge state as needed. If this conjecture is true, we would expect prototype models to perform worse when empirical ease values are derived solely from later assessment trials.

4.2 Validation Experiment 2. In the second validation experiment, the goal is to predict the ease values associated with categorizing a set of 208 bird images representing 16 different categories. Empirical ease values are derived from human training data collected as part of a previously conducted study (Roads & Mozer, n.d.b). The exemplar-, prototype-, and count-based models use similarity judgments that were previously collected for a similarity judgment database. An abbreviated description of the human training study is included, with an emphasis given to details pertinent to the current work. Experiment 2 aims to replicate, strengthen, and extend the conclusions from experiment 1 via three key differences. First, the number of species in the training task increased from 12 in experiment 1 to 16 in experiment 2, with 13 instances of each species in each experiment. This increase makes the training task more challenging for human learners and places stronger constraints on ease predictions. Second, experiment 1 evaluated performance at three different time points during training with assessment stimuli randomized across participants, whereas experiment 2 places all assessment trials at the end of the experiment with the same assessment stimuli for all participants. This change allows us to focus on the time point we are most concerned with and to reduce variability in experiment 2 in order to obtain additional statistical power for discriminating between models. Third, experiment 1 was originally designed to study alternative training strategies, and while it is beneficial to have shown that ease predictions are robust over training strategy, strategy introduced a confound irrelevant to our current goal of discriminating among ease-prediction models. Consequently, experiment 2 allows us to determine which model best describes human generalization late in the learning trajectory.

4.2.1 Methods.

Participants. Two sets of participants were used for this experiment. A psychological embedding was constructed from similarity judgments collected from 342 participants and human training data were collected from 120 participants. All participants were recruited from Amazon Mechanical Turk and paid approximately $8.00 per hour for their time and effort.

Materials. A small data set of 208 bird images representing 16 species was collected from the CUB 200 image data set (Wah et al., 2011). Species were selected such that there were four groups of birds composed of four similar-looking species, roughly corresponding to four taxonomic families. For each species, 13 exemplars were hand-picked by the lead author in the same manner as described in validation experiment 1. The image data set was an expanded version of the image data set used in validation experiment 1, with four new species making up a new taxonomic family (see Figure 2D).

Procedure. Like experiment 1, two distinct procedures were used to derive empirical ease values and predicted ease values. We describe the procedure used to derive the empirical ease values first and that for computing ease values using the candidate models second.

The training experiment was split into a single practice phase immediately followed by an assessment phase. During the practice phase, participants completed 224 practice trials in approximately 40 minutes. During the assessment phase, they completed 96 trials in approximately 15 minutes. Including instructions and breaks, the entire experiment took approximately one hour.

The entire stimulus set was partitioned into a practice and assessment set. The practice set presented seven randomly selected exemplars from each of the 16 categories and the assessment set the remaining six exemplars from each category. The same partition was used for all participants.

During the practice phase, each exemplar in the practice set was shown twice. This phase was divided into four equal-length blocks in order to allow participants to rest if desired. The order of the practice trials was determined by the experiment condition. During the assessment phase, participants saw each exemplar from the assessment set once. The assessment trials were organized into six mini-blocks such that each mini-block showed one exemplar from each category. Every participant saw the same randomly generated assessment sequence.

Every trial displayed a query stimulus and a text box for typing a response. On practice trials, a submitted answer was graded, and corrective feedback was shown to participants. Participants chose when to advance to the next trial by clicking a button. Corrective feedback displayed the participant's response, as well as the correct response. On assessment trials, no feedback was provided, and participants clicked a button to advance to the next trial.

Participants were randomly assigned to one of three scheduling conditions that used a condition-specific scheduling policy for sequencing practice trials. For the purpose of this work, no distinction is made between the conditions.

Empirical ease values were from participant responses on all assessment trials. For each query exemplar in the assessment trials, the ease value was determined by computing the total number of times the query exemplar was categorized correctly and dividing it by the total number of times the query exemplar was shown. Since every participant used the same partition, empirical ease values could only be computed for 96 of the exemplars.

Predicted ease values were obtained, and each model was evaluated using the same method from validation experiment 1. In this experiment, embeddings were inferred from 7520 2-choose-1 displays and 8772 8-choose-2 displays. The dimensionality of the embedding was determined using a cross-validation procedure (Roads & Mozer, 2019). A two-dimensional embedding of the stimuli is provided in Figure 3B for visualization purposes. As before, Steiger (1980) was used to test if the difference between dependent correlation scores was significant.

4.2.2 Results. Having already confirmed that simple models can be used to predict ease values, the primary hypothesis of experiment 2 is that the exemplar model is better poised to explain behavior that occurs at the end of training. The primary hypothesis was confirmed, and the results indicated that the exemplar model was best at predicting exemplar and category ease values (see Figure 4C).

Each assessment stimulus appeared 180 times, providing a large sample size for estimating the empirical ease values. Across all assessment trials, accuracy was moderately high ($M = .67$, $SD = .18$), indicating that participants were able to learn parts of the categorization task (see Figure 4A).

Tests of the three pair-wise comparisons were conducted using Bonferroni adjusted alpha levels of .0167 per test (.05/3). Results indicated that the correlation was significantly higher for the exemplar model ($\rho_s = .68$) than both the count ($\rho_s = .19$), $T_2(95) = 5.37$, $p < .001$ and prototype models

($\rho_s = .10$), $T_2(95) = 6.21$, $p < .001$. The pairwise comparison between the correlations for the prototype and count models was nonsignificant.

The same testing procedure was applied to the correlations between empirical category-level difficulty and predicted category-level difficulty. Results indicated that the correlation was significantly higher for the exemplar model ($\rho_s = .81$) than both the count ($\rho_s = .20$), $T_2(15) = 3.64$, $p = .003$ and prototype models ($\rho_s = .22$), $T_2(15) = 4.85$, $p < .001$. The pairwise comparison between the correlations for the prototype and count model was nonsignificant.

4.2.3 Discussion. The results of the current experiment replicated the finding of validation experiment 1 that a simple RBFN can be used to predict ease values at both the exemplar and category levels. More important, the results of the current experiment convincingly demonstrate that the exemplar model is best able to predict ease values. The correlation between predicted and empirical ease values is higher when predicting category-level difficulty. This difference is likely the result of two factors. First, exemplar-level predictions are made at a finer level, requiring that a larger number of ease values be correctly ranked. Second, category-level predictions average across exemplar-level predictions, which means that category-level predictions leverage more data for each predicted ease value.

Interestingly, the results are consistent with the previously raised conjecture that prototype models are better suited for describing early learning behavior. The results of the two experiments are consistent with the idea that stimulus representations change over time. When predicting ease of learning on early as well as later trials, the exemplar and prototype RBFNs make equally good predictions. When predicting ease of learning on later trials, exemplar models make better predictions. Given these two experiments, the dominant strategy is to use an exemplar-based RBFN to predict ease values.

5 Discussion

A key challenge of cognitive modeling is determining an appropriate stimulus representation. Predicting ease values inherits this challenge. Many approaches for computing ease values sidestep this challenge by invoking simplifying assumptions (e.g., binary categorization) or indirectly probing the stimulus representation (e.g., expert ratings). This work attempts to tackle the challenge head-on by collecting similarity judgments that can be used to infer a psychological embedding. Modeling stimulus representations directly permits predictions for arbitrary stimulus sets, regardless of whether the task involves two or more than two categories. Equally important, directly modeling the stimulus representation produces a modular contribution that the rest of the scientific community can reuse.

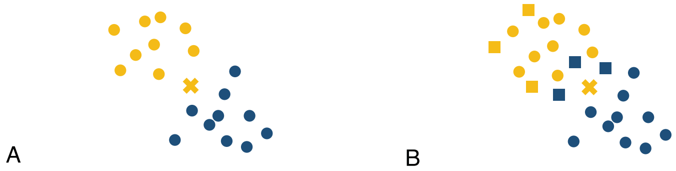

Figure 5: A cartoon depicting two different objective sets in a two-dimensional feature space. Each exemplar is represented by an icon where the color of the icon indicates the exemplar's category membership. (A) In the original objective set, the categories are nonoverlapping. (B) After adding six new exemplars (indicated by squares), the categories are now slightly overlapping. The exemplar denoted by an "x" will be easier to categorize in the first objective set compared to the second objective set.

Inferring a stimulus representation using similarity judgments has a number of additional advantages. Similarity judgments constitute a relatively low-effort task compared to a training task. Since participant compensation should be commensurate with task demands, these judgments are a relatively cheaper source of behavioral data. Furthermore, these judgments place fewer restrictions on the researchers collecting behavioral data. While training experiments typically require long sessions in order to produce useful data, similarity judgments can be collected in short or long sessions, giving researchers the flexibility to collect data in short sessions, potentially boosting participant recruitment.

In addition to the points already covered, the proposed approach is advantageous because it allows for extendability. In part, an ease value is challenging to obtain because it is defined relative to a learning objective. An ease value will depend on the stimulus itself, as well as all the other stimuli defined to be part of the objective set, that is, the set of all stimuli a learner is expected to know. To illustrate this point, consider a binary categorization task where each exemplar is characterized by two feature dimensions (see Figure 5). In the first scenario, the two categories are nonoverlapping (see Figure 5A). In a second scenario, the stimulus set has been expanded by six exemplars to create overlapping categories (see Figure 5B). The exemplar denoted by the gold "x" will likely be easier to categorize in the first scenario compared to the second one. This simple example illustrates how the ease value of an exemplar fundamentally depends on the stimulus set and the category membership of its (potentially confusing) neighbors.

If a researcher decides to expand (or shrink) the stimulus set, the corresponding ease values may change. If ease values were estimated empirically from error statistics, the corresponding behavioral data may no longer be useful. In contrast, similarity judgments are robust to changes in the stimulus set. Even if the set changes, past similarity judgments are still usable. Since researchers are often faced with immense uncertainty and tight budgets, it is reassuring to know that similarity judgments will hold their value even if research goals and experimental designs change.

The various capabilities of ease values mean that they can easily be used to implement a category training curriculum. When the difficulty of learning each exemplar is known, researchers can easily explore how different fading policies affect learning outcomes (Roads et al., 2018). For example, an easy-to-hard policy fades from exemplars with a high ease value to exemplars with a low ease value. More sophisticated curricula are also possible. For example, a researcher could implement a performance-dependent curriculum that selects an exemplar based on a learner's recency-weighted accuracy. Such a curriculum would select exemplars with ease values that are close to a learner's current abilities, making sure that the next trial is neither too easy nor too difficult. Since ease values provide exemplar-level information, they allow researchers to pursue previously inaccessible research questions.

6 Conclusion

The goal of this work is twofold: to specify a flexible and practical approach for computing ease values at both the exemplar and category levels. Flexibility is largely achieved by inferring a psychological embedding from human similarity judgments. Directly learning a stimulus representation enables researchers to tackle arbitrary categorization tasks. This freedom—along with the relatively low cost—makes the proposed approach easy to use with real-world data sets. The results of two separate validation experiments show that using an exemplar-based model is the dominant strategy for predicting empirical ease values. The first experiment computed empirical ease values using generalization behavior collected across a learning trajectory and was predicted equally well by a prototype and an exemplar model. A second experiment computed empirical ease values using generalization behavior at the end of training and was best predicted by an exemplar model. We hope this practical method will enable the development of better skill assessment protocols and more efficient training systems.

Acknowledgments

This research was supported by NSF grants SES-1461535, SBE-0542013, SMA-1041755, and DRL-1631428.

References

Ashby, F. G., & Alfonso-Reese, L. A. (1995). Categorization as probability density estimation. *Journal of Mathematical Psychology*, *39*(2), 216–233.

Evered, A., Walker, D., Watt, A. A., & Perham, N. (2014). Untutored discrimination training on paired cell images influences visual learning in cytopathology. *Cancer Cytopathology*, *122*(3), 200–210.

Giguère, G., & Love, B. C. (2013). Limits in decision making arise from limits in memory retrieval. *Proceedings of the National Academy of Sciences*, *110*(19), 7613–7618. http://www.pnas.org/content/110/19/7613.abstract

Hornsby, A. N., & Love, B. C. (2014). Improved classification of mammograms following idealized training. *Journal of Applied Research in Memory and Cognition 3*(2), 72–76. http://www.sciencedirect.com/science/article/pii/S2211368114000321

Jones, M., Love, B. C., & Maddox, W. T. (2006). Recency effects as a window to generalization: Separating decisional and perceptual sequential effects in category learning. *Journal of Experimental Psychology: Learning, Memory, and Cognition*, *32* 316–332.

Jones, M., Maddox, W. T., & Love, B. C. (2006). The role of similarity in generalization. In *Proceedings of the 28th Annual Meeting of the Cognitive Science Society* (pp. 405–410). Hillsdale, NJ: Erlbaum.

Khan, F., Mutlu, B., & Zhu, J. (2011). How do humans teach: On curriculum learning and teaching dimension. In J. Shawe-Taylor, R. S. Zemel, P. L. Bartlett, F. Pereira, & K. Q. Weinberger (Eds.), *Advances in neural information processing systems, 24* (pp. 1449–1457). Red Hook, NY: Curran. http://papers.nips.cc/paper/4466-how-do-humans-teach-on-curriculum-learning-and-teaching-dimension.pdf

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

Lindsey, R., Mozer, M. C., Huggins, W. J., & Pashler, H. (2013). Optimizing instructional policies. In C. J. C. Burges, L. Bottou, M. Welling, Z. Ghahramani, & K. Q. Weinberger (Eds.), *Advances in neural information processing systems, 26* (pp. 2778–2786). Red Hook, NY: Curran.

Love, B. C., Medin, D. L., & Gureckis, T. M. (2004). Sustain: A network model of category learning. *Psychological Review*, *111*(2), 309–332.

Luce, R. D. (1959). *Individual choice behavior: A theoretical analysis*. New York: Wiley.

McLaren, I. P. L., & Suret, M. B. (2000). Transfer along a continuum: Differentiation or association? In L. R. Gleitman & A. K. Joshi (Eds.), *Proceedings of the Twenty-Second Annual Conference of the Cognitive Science Society* (pp. 340–345). Cognitive Science Society.

Minda, J. P., & Smith, J. D. (2002). Comparing prototype-based and exemplar-based accounts of category learning and attentional allocation. *Journal of Experimental Psychology: Learning, Memory, and Cognition*, *28*, 275–292.

Nosofsky, R. M. (1985). Overall similarity and the identification of separable-dimension stimuli: A choice model analysis. *Perception and Psychophysics*, *38*(5), 415–432.

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

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), 328–353. doi:10.1037/xge0000369

Nosofsky, R. M., Sanders, C. A., Zhu, X., & McDaniel, M. A. (2019). Model-guided search for optimal natural-science-category training exemplars: A work in progress. *Psychonomic Bulletin and Review*, *26*(1), 48–76. https://doi.org/10.3758/s13423-018-1508-8

Pashler, H., & Mozer, M. C. (2013). When does fading enhance perceptual category learning? *Journal of Experimental Psychology: Learning, Memory, and Cognition*, *39*(4), 1162–1173. doi: 10.1037/a0031679

Patil, K. R., Zhu J., Kopeć, Ł., & Love, B. C. (2014). Optimal teaching for limited-capacity human learners. In Z. Ghahramani, M. Welling, C. Cortes, N. Lawrence, & K. Q. Weinberger (Eds.), *Advances in neural information processing systems*, *27* (pp. 2465–2473). Red Hook, NY: Curran.

Roads, B. D., & Mozer, M. C. (2017). Improving human-machine cooperative classification via cognitive theories of similarity. *Cognitive Science: An Multidisciplinary Journal*, *41*, 1394–1411. doi:10.1111/cogs.12400

Roads, B. D., & Mozer, M. C. (2019). Obtaining psychological embeddings through joint kernel and metric learning. *Behavior Research Methods*, *51*, 2180–2193. doi:10.3758/s13428-019-01285-3

Roads, B. D., & Mozer, M. C. (n.d.a). *Using enriched training environments for visual category training*. Manuscript submitted for publication.

Roads, B. D., & Mozer, M. C. (n.d.b). *Visual category training using structure-sensitive scheduling*. Manuscript submitted for publication.

Roads, B. D., Xu, B., Robinson, J. K., & Tanaka, J. W. (2018). The easy-to-hard training advantage with real-world medical images. *Cognitive Research: Principles and Implications*, *3*(38). doi:10.1186/s41235-018-0131-6

Salmon, J. P., McMullen, P. A., & Filliter, J. H. (2010). Norms for two types of manipulability (graspability and functional usage), familiarity, and age of acquisition for 320 photographs of objects. *Behavior Research Methods*, *42*, 82–95.

Sanborn, A. N., Griffiths, T. L., & Navarro, D. J. (2010). Rational approximations to rational models: Alternative algorithms for category learning. *Psychological Review*, *117*(4), 1144–1167.

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

Shepard, R. N. (1958). Stimulus and response generalization: Tests of a model relating generalization to distance in psychological space. *Journal of Experimental Psychology*, *55*(6), 509–523.

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

Smith, J. D., & Minda, J. P. (1998). Prototypes in the mist: The early epochs of category learning. *Journal of Experimental Psychology: Learning, Memory, and Cognition*, *24*(6), 1411–1436.

Spiering, B. J., & Ashby, F. G. (2008). Response processes in informationintegration category learning. *Neurobiology of Learning and Memory*, *90*(2), 330–338.

Steiger, J. H. (1980). Tests for comparing elements of a correlation matrix. *Psychological Bulletin*, *87*(2), 245–251.

Tanaka, J. W., & Taylor, M. (1991). Object categories and expertise: Is the basic level in the eye of the beholder? *Cognitive Psychology* *23*(3), 457–482. http://www.sciencedirect.com/science/article/pii/001002859190016H

Van Der Maaten, L. J. P., & Weinberger, K. Q. (2012). Stochastic triplet embedding. In *Proceedings of the IEEE International Workshop on Machine Learning for Signal Processing*. Piscataway, NJ: IEEE.

Wah, C., Branson, S., Welinder, P., Perona, P., & Belongie, S. (2011). *The Caltech-UCSD Birds-200-2011 Dataset* (Tech. Rep. No. CNS-TR-2011-001). Pasadena: California Institute of Technology.

Wah, C., Horn, G. V., Branson, S., Maji, S., Perona, P., & Belongie, S. (2014) Similarity comparisons for interactive fine-grained categorization. In *Proceedings of IEEE Computer Society on Computer Vision and Pattern Recognition.* Piscataway, NJ: IEEE.

Received January 25, 2019; accepted April 27, 2020.