Journal of Vision, 2019, *19(6)*, 20.

A Difficulty Predictor for Perceptual Category Learning

Luke A. Rosedahl & F. Gregory Ashby
Dynamical Neuroscience, University of California, Santa Barbara

Predicting human performance in perceptual categorization tasks in which category membership is determined by similarity has been historically difficult. This article proposes a novel biologically motivated difficulty measure that can be generalized across stimulus types and category structures. The new measure is compared to 12 previously proposed measures on four extensive data sets that each included multiple conditions that varied in difficulty. The studies were highly diverse and included experiments with both continuous- and binary-valued stimulus dimensions, a variety of different stimulus types, and both linearly- and nonlinearly-separable categories. Across these four applications, the new measure was the most successful at predicting the observed rank ordering of conditions by difficulty, and it was also most accurate at predicting the numerical values of the mean error rates in each condition.

Keywords: Categorization; Classification, Visual Category Learning, Difficulty; Information Integration;

Introduction

Humans are incredibly accurate at categorization. Whether deciding if your dog is hungry or whether a wine is a cabernet sauvignon or a merlot, humans are continually categorizing objects and events in their environment, often without conscious awareness. For the most part we perform incredibly well at this task, but when we fail – for example when a tumor is categorized as normal tissue – the consequences can be dire.

As machine learning and artificial intelligence methods progress, it is becoming ever more common to augment human performance in an effort to reduce categorization errors. Self-driving cars, parking assist, and auto-correct all exist to minimize human error and this trend is likely to continue in the future. If the goal is to increase human categorization performance, it is essential that we start explicitly looking for situations in which humans are likely to fail. There are a variety of factors that impact the difficulty of category learning, ranging from subjective factors, such as fatigue or motivation, to paradigm/environmental factors such as distractions or pressure (McCoy, Hutchinson, Hawthorne, Cosley, & Ell, 2014). But perhaps an even more fundamental factor is the difficulty of the task itself. Some category structures are fundamentally easier for humans to learn than others, but what is it that makes this learning easier? Intuitively we know it must be something to do with the structure of the categories, but what aspects of category structure affect difficulty and why?

One reason that this is still an open question is that the answer depends on the nature of the category-learning task. Rule-based (RB) category-learning tasks are those in which the category structures can be learned via some explicit reasoning process. In this case, categorization difficulty depends primarily on the complexity of the rule that must be learned (e.g., Feldman, 2000). Some prior studies have examined this issue (e.g., Salatas & Bourne, 1974). For example, rules based on two stimulus dimensions are more difficult to learn than rules based on one dimension, and among two-dimensional rules, disjunctions are more difficult than conjunctions. In prototype-distortion tasks, the category exemplars are created by randomly distorting a single category prototype, and difficulty increases with the amount of distortion (Posner & Keele, 1968). In an unstructured category-learning task, the stimuli are visually distinct and are as-

signed to each contrasting category randomly, and thus there is no rule- or similarity-based strategy for determining category membership. In this case, difficulty increases with the number of exemplars in each category.

On the other hand, predicting categorization difficulty is more problematic in information-integration (II) tasks, in which accuracy is maximized only if information from two or more stimulus components (or dimensions) is integrated at some predecisional stage. In II tasks, perceptual similarity determines category membership, and the optimal strategy is difficult or impossible to describe verbally. Explicit-rule strategies can be applied in II tasks, but they generally lead to suboptimal levels of accuracy because explicit-rule strategies make separate decisions about each stimulus component, rather than integrating this information.

Some previous work has tried to identify properties of II tasks that make learning difficult (Alfonso-Reese, Ashby, & Brainard, 2002), but the measures that were investigated were not derived from any theory of human category learning and they were only tested on some very limited category structures. This prompts the goal of this project: to develop a difficulty measure for II category learning based on the best current theories of human learning.

In the next section, we present a difficulty measure based on the most successful neurobiologically detailed model of II category learning – namely the procedural-learning component of COVIS (Ashby, Alfonso-Reese, Turken, & Waldron, 1998; Ashby & Waldron, 1999; Cantwell, Crossley, & Ashby, 2015). This model assigns a key role to the striatum, and as a result, we refer to the new difficulty measure as the Striatal Difficulty Measure (SDM). The COVIS procedural-learning model contains the most popular cognitive model of categorization – that is, the exemplar model – as a special case (Ashby & Rosedahl, 2017). Thus, the SDM is compatible with both models.

Methods

This section describes the SDM, overviews the other measures that the SDM is compared against, and describes the data sets that were used to compare all these measures.

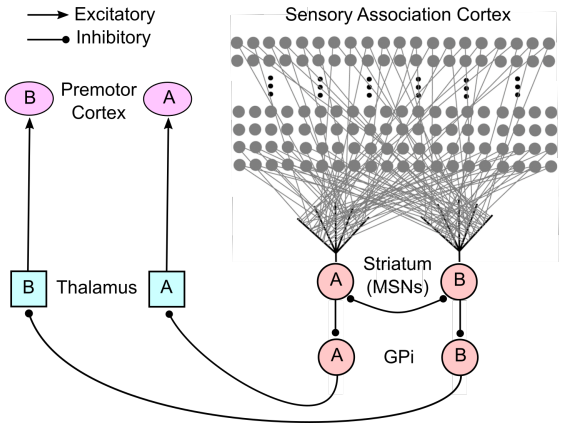

Figure 1. Architecture of the procedural-learning model of COVIS, which mimics the direct pathway through the basal ganglia. MSN = medium spiny neuron; GPi = internal segment of the globus pallidus.

Derivation of the Striatal Difficulty Measure (SDM)

The procedural-learning model of COVIS mimics the architecture of the direct pathway through the basal ganglia, which is illustrated in Figure 1. The computational version of this model is often called the striatal pattern classifier (SPC). The simplest version is a two-layer feedforward neural network that includes a large array of sensory cortical units in the input layer and a small set of striatal medium spiny units (MSNs) in the output layer – specifically, one MSN for each response alternative. Downstream units in the internal segment of the globus pallidus (GPi), the thalamus, and the premotor cortex are often omitted from the model since nothing that happens in these units can change the category response.

Initially, the sensory cortical and striatal layers are fully interconnected, with each unit in sensory cortex projecting to a unique synapse (on a spine) on each MSN. The strengths of these synapses are modified based on whether the feedback is positive or negative according to a biologically realistic form of reinforcement learning. On each trial, the most active MSN controls the response.

All versions of the SPC share similar properties. In particular, responding depends strongly on the summed similarity of the presented stimulus to the previously seen exemplars in each contrasting category. These similarity effects occur for several reasons. First, units in visual cortex respond maximally to some ideal stimulus and at a lower rate to stimuli

similar to the ideal stimulus. This is modeled via Gaussian tuning curves (mathematically identical to radial basis functions). Thus, if we let x_{iK} denote the i^{th} exemplar in category C_K, then on trials when x_{iK} is presented, activation in sensory unit j equals

$$A(x_{iK}, \mathbb{s}_j) = \exp\left[-d^2(x_{iK}, \mathbb{s}_j)/\gamma\right] \tag{1}$$

where $\mathbb{s}_j$ is the stimulus that maximally excites sensory unit j, $d(x_{iK}, \mathbb{s}_j)$ is the Euclidean distance between the perceptual representations of objects x_{iK} and $\mathbb{s}_j$, and γ captures how tightly sensory units are tuned. Thus, $A(x_{iK}, \mathbb{s}_j)$ increases with the similarity of the presented stimulus to $\mathbb{s}_j$.

Second, because of the nature of reinforcement learning, similarity effects in the SPC are consolidated at cortical-striatal synapses. In fact, Ashby and Rosedahl (2017) showed that under certain simplifying assumptions, the synaptic strength between sensory unit j and striatal unit K is proportional to the summed similarities of object $\mathbb{s}_j$ to all previously seen exemplars from category K. Since synaptic strength drives striatal activation, the probability of responding K on a trial when stimulus $\mathbb{s}_j$ is presented therefore increases with this sum.

Ashby and Rosedahl (2017) also showed that these summed similarities are mathematically identical to the summed similarities that are the basis of exemplar models of categorization (Nosofsky, 1986). So the exact same difficulty measure could be derived from exemplar theory. Although the two approaches are mathematically equivalent, they make very different cognitive assumptions. Exemplar theory assumes that each sum is computed from scratch on every trial. For example, to compute the summed similarity of the presented stimulus to the exemplars of category K, exemplar theory assumes that the subject activates the memory representation of every previously seen exemplar from category K, computes the similarity of the presented stimulus to each of these memory representations, and then sums all these similarities. Thus, exemplar theory predicts that as a subject gains experience at a specific classification task, more and more computation is required on each trial (because there are more terms in the sum). In contrast, the SPC assumes that the sums are encoded in the cortical-striatal synaptic strengths as a result of a reinforcement-learning process. Thus, the SPC assumes that no memory representations are retrieved during the categorization process.

On every classification trial, the SPC striatal units enter a winner-take-all competition to select the response. Therefore, the weaker the activation of the striatal unit corresponding to the correct category and the stronger the activation of the striatal units corresponding to incorrect categories, the more difficult the judgment. Activation is proportional to similarity, which suggests that task difficulty should increase with the simple ratio

$$D = \frac{S_B}{S_W}, \tag{2}$$

where S_B is between-category similarity and S_W is within-category similarity.

The SPC suggests specific forms for S_B and S_W. In particular, S_B should equal the similarity of every category exemplar to all exemplars in every contrasting category:

$$S_B = \sum_{K=1}^{R} \sum_{L\neq K}^{R} \sum_{i=1}^{n_K} \sum_{j=1}^{n_L} A(x_{iK}, x_{jL}), \tag{3}$$

where R is the number of contrasting categories, n_K is the number of exemplars in category K, n_L is the number of exemplars in each contrasting category L, and as in Eq. 1, $A(x_{iK}, x_{jL})$ is activation in the sensory unit that is maximally excited by stimulus x_{jL}. Similarly, S_W should equal the similarity of every exemplar to all exemplars in the same category:

$$S_W = \sum_{K=1}^{R} \sum_{i=1}^{n_K} \sum_{j\neq i}^{n_K} A(x_{iK}, x_{jK}). \tag{4}$$

Putting all this together produces the Striatal Difficulty Measure (SDM):

$$SDM = \frac{\sum_{K=1}^{R} \sum_{L\neq K}^{R} \sum_{i=1}^{n_K} \sum_{j=1}^{n_L} \exp\left[-d^2(x_{iK}, x_{jL})/\gamma\right]}{\sum_{K=1}^{R} \sum_{i=1}^{n_K} \sum_{j\neq i}^{n_K} \exp\left[-d^2(x_{iK}, x_{jK})/\gamma\right]}. \tag{5}$$

For completely overlapping categories this measure equals 1 because within-category similarity is equal to between category similarity. For infinitely separated categories (where the between-category similarity goes to 0), the measure equals 0.

Note that the only free parameter in Eq. 5 is γ, which is a

measure of how tightly tuned the subject's sensory system is to changes in the stimulus. Technically, γ could differ across stimulus dimensions, but in practice such differences would have to be extreme for SDM to change its predicted ordering of tasks by difficulty. Thus, a single value of γ will suffice in almost all applications. Furthermore, the numerical value of γ could be estimated from separate sensory discrimination data. As we will see however, the ordinal predictions of the SDM as to which of two (or more) conditions is most difficult, typically do not change when γ changes. So the actual numerical value of γ chosen does not appear to be critical. In the empirical applications considered below, we compute SDM by averaging across a wide range of γ values.

The SDM is closely related to a number of previously proposed difficulty measures. First, many machine-learning measures are based on an inverse of the Eq. 2 ratio:

$$D = \frac{D_W}{D_B}, \tag{6}$$

where D_W and D_B are some measures of the within- and between-category dissimilarities, respectively (e.g., Fukunaga, 2013). Most commonly, dissimilarity is defined as some increasing function of distance. Of these measures, perhaps the most similar to the SDM is the ratio of intra- to extra-class nearest-neighbor measure, which is often referred to as the N_2 measure (Lorena, Garcia, Lehmann, Souto, & Ho, 2018). The N_2 difficulty measure takes the form of Eq. 6 with

$$D_W = \sum_{K=1}^{R} \sum_{i=1}^{n_K} \min_{j \neq i} d(x_{iK}, x_{jK}), \tag{7}$$

As it should, note that this sum increases with the distance between category exemplars, so when incorporated into Eq. 6, the N_2 difficulty measure predicts that categories in which the exemplars are more widely distributed are more difficult to learn than categories in which the exemplars are tightly clustered. Analogously, the N_2 measure defines between-category separation as

$$D_B = \sum_{K=1}^{R} \sum_{i=1}^{n_K} \min_{\substack{j \\ L \neq K}} d(x_{iK}, x_{jL}). \tag{8}$$

Note that this sum increases with the distance between the category exemplars that are in contrasting categories, and thus, when incorporated in Eq. 6, the N_2 measure predicts that classification difficulty decreases with between-category separation.

Note that the SDM differs from the N_2 difficulty measure in two important ways. First, the SDM depends on all category exemplars, whereas N_2 assumes that only the nearest neighbors affect difficulty. Leading theories of human category learning assume that classification decisions depend on all previously seen category exemplars – not just the nearest neighbors (e.g., Estes, 1986; Medin & Schaffer, 1978; Nosofsky, 1986).

Second, N_2 depends on distance, whereas SDM depends on a nonlinear transformation of distance – namely, similarity. Considerable independent evidence suggests that human classification and generalization are determined primarily by similarity, rather than by distance (e.g., Shepard, 1987). This difference between SDM and N_2 changes the impact that stimulus spacing has on predicted difficulty. The Gaussian similarity function described in Eq. 1 has an inflection point at an intermediate distance. SDM therefore predicts that increasing distances for intermediately spaced stimuli will have a greater impact on difficulty than increasing the separation for either nearby or distant stimuli by the same amount. In contrast, defining difficulty in terms of distance, rather than similarity (e.g., as in the N_2 measure), predicts that all changes of a fixed distance should have equal effects on classification difficulty.

Previous Measures

To our knowledge, only one previous study has tried to predict human learning difficulty in II tasks. Alfonso-Reese et al. (2002) compared the ability of several different measures to predict the difficulty of five different category structures (shown in Figure 2). Included in this list were a measure of covariance complexity, a measure of class separation, and the error rate of an ideal observer. In contrast, many alternative difficulty measures have been proposed within the machine-learning literature – some that have the form of Eq. 6 and some that do not. Many of these were reviewed by Lorena et al. (2018), who divided the measures into six groups: feature overlapping measures, linearity measures, neighborhood measures (which includes N_2), network mea-

sures, dimensionality measures, and class balance measures.

The remainder of this article compares SDM to the measures examined by Alfonso-Reese et al. (2002) and to a variety of the machine-learning measures described by Lorena et al. (2018). All of these measures are compared in their ability to predict difficulty across a variety of different category structures. The structures are highly diverse, and include both continuous- and binary-valued stimulus dimensions, linearly- and nonlinearly-separable categories, and a variety of different stimulus types. As we will see, of all these measures, the SDM most accurately predicts human learning difficulty across all these very different conditions.

We will now provide a brief description of the difficulty measures used in this article. The equations are included for the more straightforward measures, whereas a qualitative description is provided for the others. More detailed descriptions of these latter measures can be found in Lorena et al. (2018).

Measures Considered by Alfonso-Reese et al. (2002)

The following measures were used by Alfonso-Reese et al. in their attempt to quantify procedural categorization difficulty.

Covariance Complexity (CC). Alfonso-Reese et al. (2002) used a covariance complexity (CC) measure proposed by Bozdogan (1990)

$$CC = \frac{1}{2}\text{rank}(\Sigma)\ln\left[\frac{\text{trace}(\Sigma)}{\text{rank}(\Sigma)}\right] - \frac{1}{2}\ln|\Sigma|, \tag{9}$$

where Σ is the common within-category variance-covariance matrix. Note that this measure is undefined if the contrasting categories are characterized by different within-category variance-covariance matrices.

Class Separation (C_{sep}). Following (Fukunaga, 2013), Alfonso-Reese et al. (2002) defined class separation (C_{sep}) as

$$C_{sep} = \text{trace}\left(\Sigma^{-1}S\right), \tag{10}$$

where Σ is the common variance-covariance matrix. The matrix S for a two category condition with categories A and B is defined as

$$S = \frac{1}{2}(\underline{\mu}_A - \underline{\mu})(\underline{\mu}_A - \underline{\mu})' + \frac{1}{2}(\underline{\mu}_B - \underline{\mu})(\underline{\mu}_B - \underline{\mu})' \tag{11}$$

where $\underline{\mu}$ is a vector which is the mean of $\underline{\mu}_A$ and $\underline{\mu}_B$. In the case where the two categories are characterized by different variance-covariance matrices Σ_A and Σ_B (e.g., as in the Ashby & Maddox, 1992 experiments),

$$\Sigma = \frac{1}{2}\Sigma_A + \frac{1}{2}\Sigma_B. \tag{12}$$

Error Rate of the Ideal Observer (eIO). This is the error rate that results from applying the optimal classification strategy.

Machine-Learning Measures

The following measures were designed for machine learning algorithms. More details on all the measures can be found in Lorena et al., 2018.

Volume of Overlapping Regions (VOR). The volume of overlapping regions (VOR) is a measure of feature overlap that depends on the amount of overlap of the category distributions on each stimulus dimension. Specifically, VOR is computed by finding the range of values on each dimension that are shared by both categories, multiplying these ranges together, and then normalizing.

Collective Feature Efficiency (CFE). Collective feature efficiency (CFE) is another measure of feature overlap that is based on the percentage of stimuli that can be correctly classified using bounds perpendicular to each stimulus dimension.

Error Rate of Nearest Neighbor Classifier (eNN). The error rate of nearest neighbor classifier (eNN) is the error rate of a classifier that assigns the stimulus to the category of its nearest neighbor among all other stimuli in the two categories.

Fraction of Borderline Points (FBP). The fraction of borderline points (FBP) is a function of the number

of stimuli that are connected to a stimulus belonging to the contrasting category in the minimum spanning tree constructed from the data.

Fraction of Hyperspheres Covering Data (T_1). The fraction of hyperspheres covering data measure (called T1 in Lorena et al., 2018) is constructed by first centering a hypersphere on each stimulus and setting the radius equal to the distance between that stimulus and the nearest stimulus from the contrasting category. All hyperspheres that are completely contained in another hypersphere are then removed and the measure is simply the fraction of hyperspheres that remain.

Average Density of the Network (Density). Several machine-learning difficulty measures are derived from the representation of the categories as a graph. Each category exemplar is represented as a node or vertex in the graph, and nodes are connected if their corresponding distance in stimulus space is less than some criterion value. Finally, edges that connect exemplars from contrasting categories are pruned.

The average density of the network (density) is the number of edges in the graph divided by the maximum possible number of edges in a graph with the same number of nodes. Thus, if the graph has N edges and n nodes, then

$$density = \frac{N}{n(n-1)/2}. \tag{13}$$

Clustering Coefficient (ClsCoef). The clustering coefficient (ClsCoef) is a measure of network average local density. First, for each node, define its neighborhood as the set of all nodes that are directly connected. The ClsCoef is the mean density of each of these neighborhoods.

Note that ClsCoef is smaller for less dense networks or for structures where the categories overlap (leading to many non-connected stimuli from opposing classes within the neighborhood of any given stimulus).

Hub Score (Hubs). The hub score (Hubs) is another network measure equal to the number of connections a node has weighted by the number of connections of each of its neighbors.

This leads to stimuli that are connected to many other stimuli that are also highly connected having a large score. Less dense categories and a higher degree of overlap between categories will both cause this measure to predict higher difficulty.

Data Analysis

We compared the efficacy of the SDM to all of the other measures described in the previous section at predicting human categorization performance in four different published studies. The studies all used different stimulus types and included categorization conditions that differed in difficulty. The data sets from these four studies included five category structures from Alfonso-Reese et al. (2002), six classic structures from Shepard, Hovland, and Jenkins (1961), three structures from Ashby and Maddox (1992), and three from Ell and Ashby (2006). Each of these studies used different stimulus types. Shepard et al. (1961) used binary-valued stimulus dimensions, whereas the other studies used continuous-valued dimensions. Alfonso-Reese et al. (2002) and Shepard et al. (1961) used stimuli that varied on three dimensions, whereas the stimuli used by Ashby and Maddox (1992) and Ell and Ashby (2006) varied on two stimulus dimensions. Alfonso-Reese et al. (2002) and Ell and Ashby (2006) used linearly separable categories, Ashby and Maddox (1992) used nonlinearly separable categories, and Shepard et al. (1961) included both linearly and nonlinearly separable categories.

Our primary analysis focused on the ability of each difficulty measure to correctly rank order the observed classification error rates from each condition of these four studies. Some of the measures increase with predicted classification difficulty (e.g., CC, eIO, VOR), whereas the others decrease with predicted difficulty (e.g., C_{sep}, Density, ClsCoef). For measures in this latter group, we generated a predicted rank ordering by inverting the order of the measure. So for example, the condition with the smallest C_{sep} was ranked as most difficult and the condition with the largest C_{sep} was ranked as least difficult.

For each category structure, the SDM was calculated by randomly selecting 300 stimuli from the category distributions and averaging across 10 such sets to determine the SDM value for a single γ. This process was repeated for all

values of γ ranging from 5–50 in 5 step intervals (so 5, 10, 15, ..., 45, 50) and the final difficulty score was the average of the scores for all values of γ. In practice, the value of γ can be found by fitting previous results using the same stimuli, but here we are interested in *a priori* difficulty predictions of SDM, rather than in its ability to account for difficulty post hoc by adjusting the value of γ. The machine-learning measures were computed using the R package provided by Lorena et al. (2018).

Results

Alfonso-Reese, Ashby, and Brainard (2002)

Alfonso-Reese et al. (2002) compared the ability of the CC, eIO, and C_{sep} difficulty measures to rank order human performance on the five different classification tasks described in Figure 2. A fourth measure was also included (orientation of the optimal bound), but because it failed to make any differential predictions for the majority of the category structures they analyzed, it was excluded from comparison here. In all tasks, the stimuli were bar graphs that displayed the numerical values of blood pressure, white blood cell count, and serum potassium level of a hypothetical patient. The subject's task was to use these three values to diagnose the patient with either disease A or B.

Table 1 shows the observed rank ordering of the tasks according to the mean percent errors of subjects during the last block of training, along with the predicted rank ordering according to the SDM, the eight measures selected from Lorena et al. (2018), and the three measures from Alfonso-Reese et al. (2002). Also shown (in the rightmost column) is Spearman's rank correlation for each model that measures the ordinal agreement between the predicted and observed orderings. Note that the SDM, N_2, T_1, and Density measures performed best and that the former three measures all made identical ordinal predictions – mispredicting only one pair of conditions (conditions 3 and 5).

It should be noted that due to the similar error rates between conditions 3, 4, and 5 (32.1%, 29.6%, and 30.0% respectively), it is unclear whether there is any real difficulty difference among these conditions.

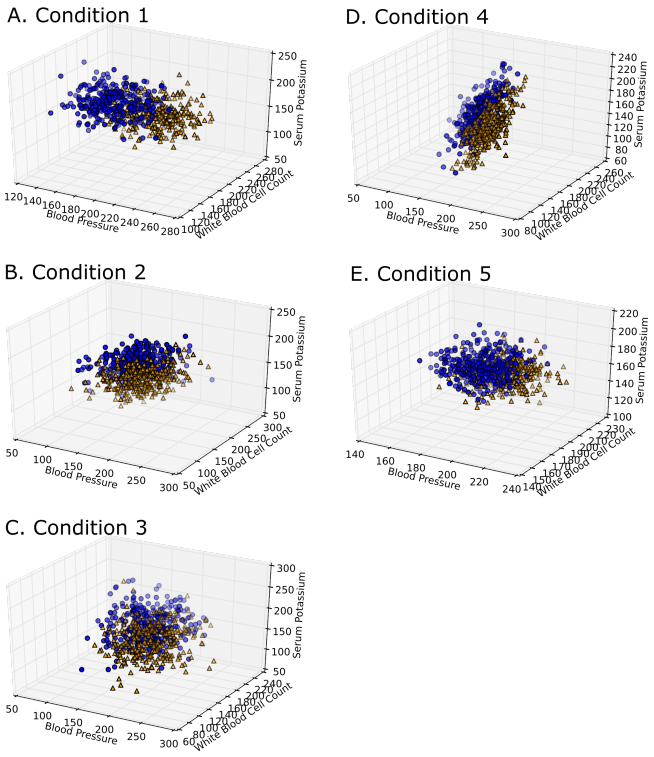

Figure 2. Alfonso-Reese et al. (2002) category structures.

Table 1
Predicted and Observed Difficulties for the Alfonso-Reese et al. (2002) Category Structures

Measure	**Difficulty**	***r***
ClsCoef	C5>C1>C2>C4>C3	-.40
C_{sep}	C5>C1>C2>C3=C4	-.31
eIO	C5>>C1>C2>C3=C4	-.31
VOR	C2>C3>C4>C5>C1	.30
CC	C3=C4>C2>C1=C5	.47
CFE	C2>C3>C5>C4>C1	.40
eNN	C5>C3>C4>C1>C2	.80
FBP	C5>C3>C4>C1>C2	.80
Hubs	C3>C4>C5>C1>C2	.80
Density	C3>C4>C5>C2>C1	.90
T_1	C5>C3>C4>C2>C1	.90
N_2	C5>C3>C4>C2>C1	.90
SDM	C5>C3>C4>C2>C1	.90
Observed Ordering	**C3>C5>C4>C2>C1**	
Percent Errors	**32.1>30.0>29.6>18.3>13.5**	

A natural question is whether the good performance of the SDM depends on the specific numerical value chosen for γ. To investigate this question, we examined how the ordinal predictions of the SDM change as a function of γ. The re-

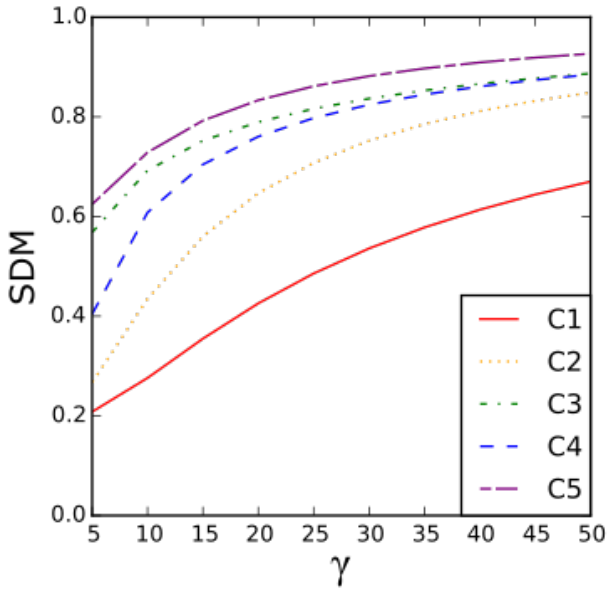

Figure 3. Predicted difficulty in the Alfonso-Reese et al. (2002) conditions as a function of the SDM γ parameter.

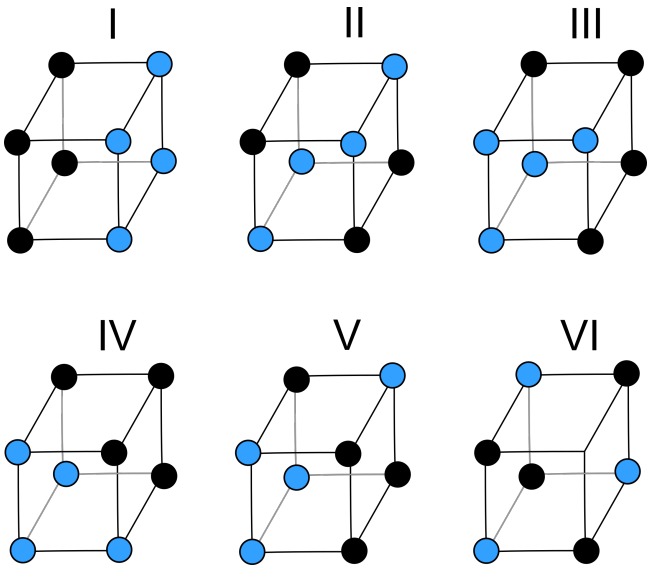

Figure 4. Shepard, Hovland, and Jenkins (1961) category structures. Black dots represent stimulus coordinates of category A exemplars and blue dots represent stimulus coordinates of category B exemplars.

sults are shown in Figure 3, which shows the predicted value of the SDM in each condition across a wide range of different γ values. The rank ordering in Table 1 was computed from the mean SDM from each of these curves. Note that none of the curves cross, which means that the ordinal predictions of the SDM are invariant across different values of γ. We performed similar analyses for each of the other empirical applications considered below, and in every case, none of the curves crossed. Thus, at least for the empirical applications considered in this article, the ordinal predictions of the SDM do not depend on the specific numerical value chosen for γ.

Shepard, Hovland, and Jenkins (1961)

Shepard et al. (1961) compared categorization performance for six category structures created from stimuli that varied across trials on three binary-valued dimensions. Each stimulus was a geometric object that varied in shape (triangle versus square), size (small versus large) and color (black versus white). The category structures are described abstractly in Figure 4.

These six tasks have been replicated many times with a variety of different stimulus types and are perhaps the most widely used category structures for testing new theories of categorization. For example, ALCOVE (Kruschke, 1992; Nosofsky, Gluck, Palmeri, McKinley, & Glauthier, 1994), the context model (Nosofsky, 1984), the generalized context model (Nosofsky, 1986), COVIS (Ashby et al., 1998; Edmunds & Wills, 2016), and SUSTAIN (Love & Medin, 1998) have all been shown to account for the consensus difficulty ordering of VI > III = IV = V > II > I (e.g., Nosofsky et al., 1994; Smith, Minda, & Washburn, 2004). These demonstrations all required estimating a large number of free parameters however, and for this reason, we did not include any of these models in the analyses included here. For example, Nosofsky (1984) estimated 18 free parameters when he showed that the context model was consistent with the Shepard et al. (1961) difficulty order. On the other hand, it is important to note that after this parameter-estimation process, the resulting models also provide good fits to the learning curves – an ability that is beyond the scope of the SDM. The SDM is not proposed as a model of categorization or category learning. Rather, we propose the SDM as a measure that makes *a priori* predictions of categorization difficulty.

Class separation is undefined with some of these categories because the within-category variance-covariance matrix is singular. As a result, we compared all other measures

to the consensus ordering from the six conditions. Values of 0 and 100 were used for each binary-valued dimension to approximately equate the range of stimulus values to those used in the other experiments. Results are shown in Table 2. Note that SDM performs better than all the previous top performers – correctly ordering the difficulty of all conditions except type II. Three measures that performed poorly on the Alfonso-Reese et al. (2002) data outperform SDM here: VOR, CFE, and FBP. However, note that two of these measures (VOR and CFE) predict no difference between category structure VI and structures III, IV, and V. In contrast to this prediction, many studies have shown that the type VI categories are, by far, the most difficult for people to learn (Nosofsky et al., 1994; Smith et al., 2004).

Table 2
Predicted and Observed Difficulties for the Shepard et al. (1961) Category Structures

Measure	**Difficulty**	***r***
ClsCoef	IV>I>III>V>II>VI	-.39
Hubs	II>V>IV>III>I>VI	-.33
T_1	I=II=III=V=VI>IV	-.14
eNN	I=II=III=IV=V=VI	0.0
eIO	I=II=III=IV=V=VI	0.0
N_2	I=II=III=IV=V=VI	0.0
CC	V>III>IV>I=II=VI	.29
Density	VI>V>II>I=III>IV	.31
SDM	VI>II>V>III>IV>I	.58
CFE	II=III=IV=V=VI>I	.70
VOR	III=IV=V=VI>I=II	.88
FBP	VI>V>III=IV>II>I	.95
Observed Ordering	**VI>V=IV=III>II>I**	
Percent Error	**14.3>7.5=6.5=6.1>3.2>1.0**	

Note. The difficulty ordering for the covariance complexity measure was computed by Alfonso-Reese et al. (2002). The error rates used here are from Nosofsky et al. (1994).

The reduced performance of SDM on these data, relative to the data of Alfonso-Reese et al. (2002) is driven by two factors: the better than predicted human performance on type II categories and the failure of SDM to predict exactly equal performance on category types III, IV, and V. Note that for the type II categories, perfect performance is possible with the (explicit) disjunction rule: Respond B if the stimulus is at level 1 on dimensions 1 and 3 *or* if the stimulus is at level 2 on both of these dimensions; otherwise respond A. Thus, one possibility is that category types I and II are best described as RB tasks, in which case the SDM should not be expected to apply. Also, of course, the decision to set the observed difficulties of types III, IV, and V equal in Table 2 is because previous studies have generally not agreed on the ordering of these types and any differences that have been reported were small. SDM could be generalized to predict equal difficulties by requiring, for example, that the predicted difficulties of two tasks exceed some criterion before a strict ordering is predicted.

Ashby and Maddox (1992)

Ashby and Maddox (1992) trained participants on the three category structures described in Figure 5. Each category was created by drawing 800 random samples from a bivariate normal distribution. In all experiments, the two category distributions had different variance-covariance matrices, so in each case the optimal decision boundary was nonlinear (i.e., quadratic). The three experiments included separate conditions (with separate subjects) that used the coordinate values of the random samples shown in Figure 5 to create two different stimulus types: rectangles that varied across trials in height and width, and circles with a radial line that varied across trials in circle size and line orientation.[1]

The results are shown in Table 3. The observed accuracies and difficulties were based on performance during the last 300 trials. Note that accuracy was highest in Experiment 3, second highest in Experiment 1, and lowest in Experiment 2, so the observed difficulty ordering was E2 > E1 > E3. This same ordering held for both stimulus types, so in these experiments at least, difficulty depended on category structure, but not on the type of stimuli that were used.

SDM was one of four measures to correctly rank order the three experiments by difficulty, joined by the C_{sep}, Density, and Hubs measures. The three measures that outperformed SDM for the Shepard et al. (1961) categories (VOR, CFE, and FBP) and two of the three measures that performed as well as SDM on the Alfonso-Reese et al. (2002) categories

[1]Experiments 1 and 2 included a third condition in which the stimuli were two connected line segments that varied across trials in length. However, Ashby and Maddox (1992) did not include those stimuli in their Experiment 3, and so those conditions are not considered here. Even so, the difficulty ordering for the excluded conditions was the same as for the other conditions, so the only effect of including the line-segment data would be to slightly change the Experiment 1 percent correct listed in Table 3.

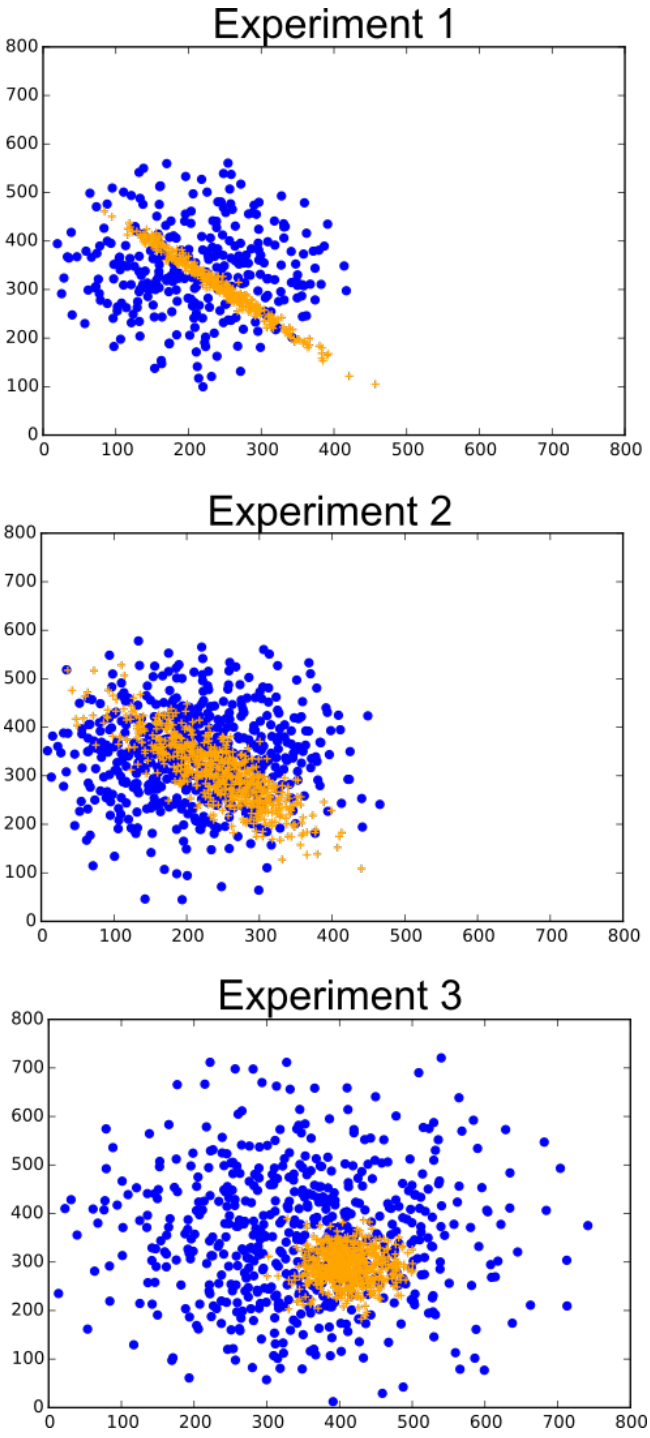

Figure 5. Ashby and Maddox (1992) category structures. In each case, the categories were created by random sampling from a bivariate normal distribution. In each experiment, the distributions had different variance-covariance matrices.

(T_1 and N_2) all failed to properly rank order the experiments.

Table 3
Predicted and Observed Difficulties for the Ashby and Maddox (1992) Category Structures

Measure	**Difficulty**	***r***
ClsCoef	E3>E2>E1	-.50
CFE	E1>E2>E3	.50
CC	E1>E2>E3	.50
eNN	E2>E3>E1	.50
FBP	E2>E3>E1	.50
T_1	E2>E3>E1	.50
N_2	E2>E3>E1	.50
VOR	E1>E2>E3	.50
eIO	E2>E1=E3	.87
C_{sep}	E2>E1>E3	1.0
Density	E2>E1>E3	1.0
Hubs	E2>E1>E3	1.0
SDM	E2>E1>E3	1.0
Observed Ordering	**E2>E1>E3**	
Percent Errors	**35>25>14**	

Ell and Ashby (2006)

Ell and Ashby (2006) studied the effects of category separation on categorization performance by training participants on category structures that varied on the distance between the category means but were identical in all other aspects. The categorization stimuli were Gabor disks that varied across trials on spatial frequency and orientation. The category structures are described in Figure 6. As in the Ashby and Maddox (1992) experiments, the stimuli comprising each category were random samples from a bivariate normal distribution. However, in these experiments, both category distributions had identical variance-covariance matrices, so in each case the optimal boundary was linear (as in the Alfonso-Reese et al., 2002 conditions). Therefore, the different conditions varied only in category separation.

As expected, performance improved substantially with category separation. Thus, any measure sensitive to separation will correctly order these conditions by difficulty. The results, based on the last block of performance are shown in Table 4. Note that all measures (including SDM) correctly rank order the conditions by difficulty, except for CC, which predicts equal performance in the three conditions. This is because the CC measure is sensitive only to the complexity of the variance-covariance matrices that describe the contrasting categories. Because the categories in the Ell

and Ashby (2006) experiments all had identical variance-covariance matrices, the CC measure incorrectly predicts equal performance in the three conditions.

Table 4
Predicted and Observed Difficulties for the Ell and Ashby (2006) Category Structures

Measure	**Difficulty**	***r***
CC	L=M=H	0.0
C_{sep}	L>M>H	1.0
CC	L>M>H	1.0
CFE	L>M>H	1.0
Density	L>M>H	1.0
eNN	L>M>H	1.0
FBP	L>M>H	1.0
Hubs	L>M>H	1.0
T_1	L>M>H	1.0
eIO	L>M>H	1.0
N_2	L>M>H	1.0
VOR	L>M>H	1.0
SDM	L>M>H	1.0
Observed Ordering	**L>M>H**	
Percent Errors	**47>21>1**	

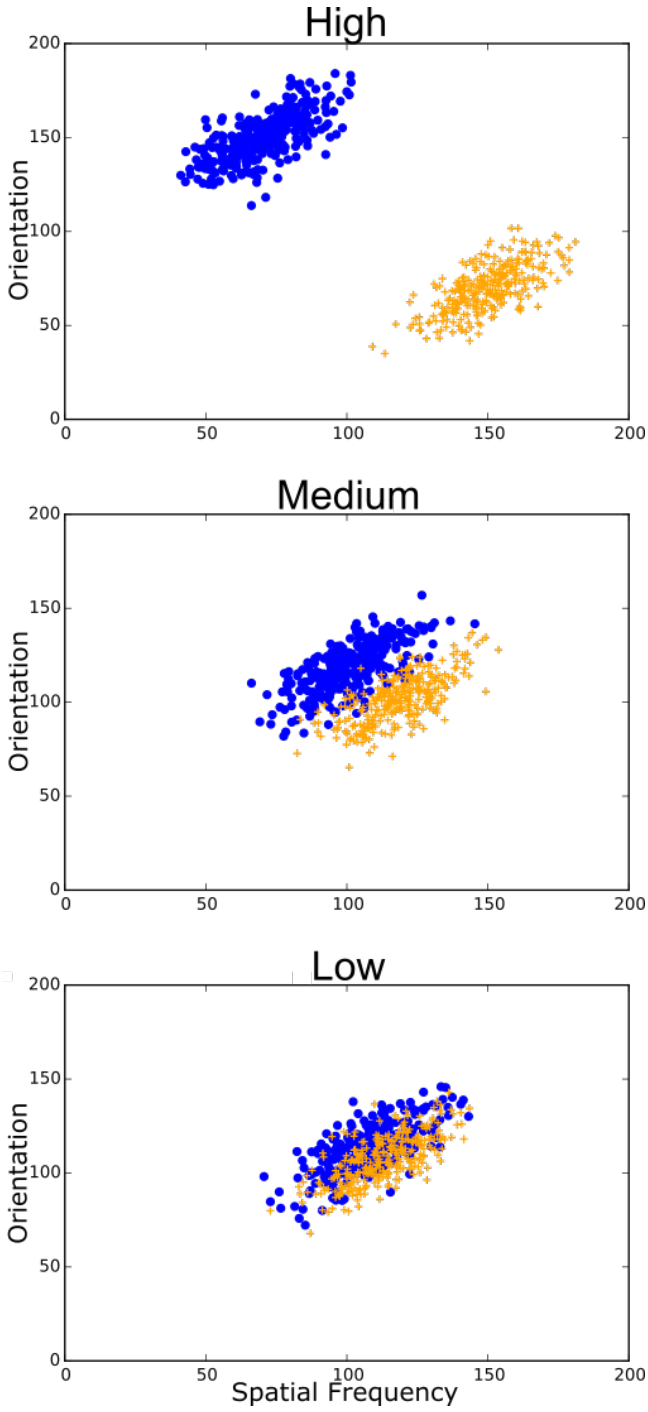

Figure 6. Ell and Ashby (2006) category structures. In each case, categories were created by random sampling from a bivariate normal distribution. All distributions had identical variance-covariance matrices. The three conditions varied the inter-mean distance to create high, medium, and low class separation.

Comparing Across All Experiments

The SDM performed best across all the data sets examined so far. Even so, these results must be interpreted with caution because of the small number of category structures examined in each application. Because of these small numbers, the Spearman's rank correlations reported in Tables 1 – 4 are based on small sample sizes. This section attempts to alleviate this concern by comparing performance across all the data sets examined above.

First, we summarized the rank-order performance of each measure by computing its mean Spearman's *r* in all four applications described above (i.e., across Tables 1 – 4). Results are shown in Table 5. Note that, overall, SDM performed best, followed by FBP and Density and then distantly by VOR.

Table 5
Average Spearman's r Across All Category Structures

Measure	Average Spearman's *r*
ClsCoef	-.07
CC	.32
eIO	.39
C_{sep}	.56
T_1	.57
eNN	.58
N_2	.60
Hubs	.62
CFE	.65
VOR	.67
Density	.80
FBP	.81
SDM	.87

The rank orderings considered so far only examine ordinal predictions of the difficulty measures. However, each measure makes a quantitative prediction about the difficulty of any particular category structure. And in all applications considered above, we have an empirical quantitative estimate of difficulty – namely, the average error rate of the human learners. So a more ambitious question is to ask how well the various measures predict the observed error rates.

Before proceeding, however, there are several complications to consider. First, the quantitative value of difficulty predicted by each measure is not average error rate, but rather some other statistic. For example, in the case of SDM, the statistic is described by Eq. 5. Suppose we call the numerical value of difficulty predicted by a measure D and the observed average error rate of human learners E. Then the various measures all predict that

$$E = f(D), \tag{14}$$

where f is some strictly increasing function (and therefore order preserving). However, none of the measures specify the form of f. This is why we focused on predicted rank orderings (because the predicted rank ordering is the same for any increasing function f). We will use the same strategy here, but in addition, we will also compare the ability of the most successful measures to predict the observed value of average error rate in all conditions and experiments considered above, under the assumption that f is linear. However, it is important to note that, in general, there is no reason to expect f to be linear.

A second complication is that the four applications considered above each included different amounts of training and different instructions to the subjects. The measures are blind with respect to these factors. Thus, they predict the same quantitative value of difficulty regardless of whether subjects received 100 or 1000 trials of training. Obviously, we expect average error rates to be lower in the latter case, so a mispredicted average error rate by a measure in a specific experiment does not necessarily mean that the measure is flawed. For this reason, the results in this section should be interpreted with caution. Despite these misgivings however, we believe that comparing the quantitative predictions of the measures across all experiments is a useful exercise. First, there is no reason to expect these issues to plague one measure any more than the others. Thus, even if all the predictions are inaccurate, it could still prove useful to compare the accuracy of different measures. Second, the most likely effect of these complications should be to reduce the accuracy of prediction. Thus, whereas it might be problematic to interpret results if all measures make inaccurate predictions, the opposite scenario is less troubling. In particular, accurate predictions by a measure are most likely to occur because that measure is a valid predictor of classification difficulty, rather than because of either complication. With those caveats in mind, we can proceed to the analysis.

For each category structure in the four data sets, we computed the numerical value of difficulty predicted by each of the measures (excluding category separation since it is not defined for the SHJ data set) and then compared these to the observed mean (across subjects) error rates. We evaluated the accuracy of these predictions in two ways – by computing the Spearman's rank correlation and the Pearson's squared correlation between predicted difficulty and the observed error rates. The results are shown in Table 6.

Table 6
Spearman's rank correlation and Pearson's squared correlation between predicted difficulty and mean observed error rate across all category structures considered in this article.

Measure	Spearman's r	Pearson's r^2
CFE	.05	.18
FBP	.12	.09
Hubs	.18	.02
VOR	.25	.02
CC	.39	.13
ClsCoef	.54	.31
eIO	.78	.73
Density	.82	.76
N_2	.83	.76
T_1	.83	.83
eNN	.87	.83
SDM	.93	.87

Note that SDM performs best according to both measures, with a Spearman's r of .93 and a Pearson's r^2 of .87. The nearest neighbor classifier (eNN) is second best, followed by the hyperspheres (T_1), N_2, and density measures. Thus, despite the complications described above, the SDM accounts for an impressive 87% of the variance in the mean error rates across study.

Figure 7 plots mean error rate in each study along with predicted difficulty for the six best performing measures. Also shown are the best-fitting regression lines and the squared Pearson correlation. Note that the high r^2 for the SDM suggests that the function f from Eq. 14 is fairly linear in these applications.

The performance of the SDM can be even further improved by selecting the single best performing value of γ = 10. In this case the measure accounts for 91% of the variance. This could presumably be increased even further by using values of γ tailored to each stimulus type (because each stimulus has a different visual representation, the neural tuning curves will differ across stimulus types, and therefore γ should also differ). Even so, there are two different reasons that we chose to base the r^2 in Table 6 on the mean value of SDM across a wide range of γ values. First, none of the other measures include a free parameter, so to keep the comparisons fair, neither should the SDM. Second, the goal of this article is to develop a difficulty measure that makes accurate *a priori* predictions of difficulty.

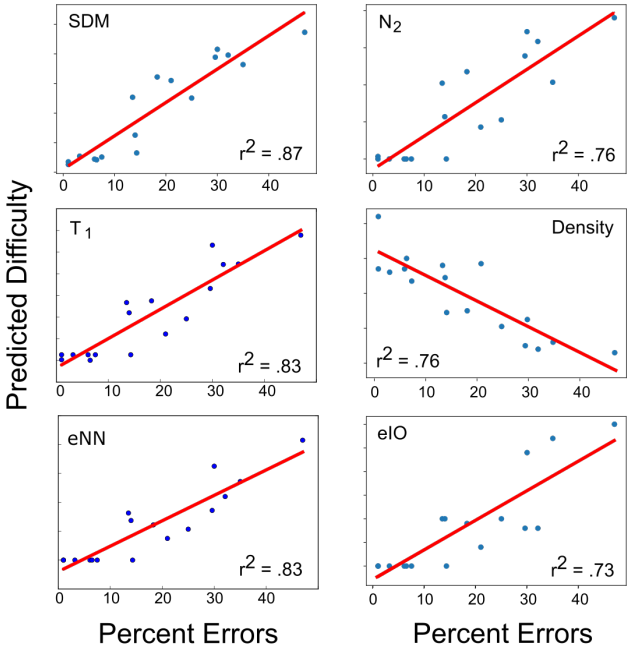

Figure 7. Scatterplots of predicted difficulty for six different measures against mean observed error rate for all category structures from the four applications considered in this article. Also shown are the best-fitting regression line and resulting Pearson r^2. Note that in the case of Density, the ordinate is "Predicted Ease of Classification."

General Discussion

Across a wide range of category-learning data sets, the SDM outperformed several difficulty measures that have been used previously on human data (CC, eIO, and C_{sep}), as well as eight previously used measures from the machine-learning literature (VOR, CFE, FBP, eNN, T_1, Density, ClsCoef, and Hubs). All of these measures were compared on four extensive data sets that each included multiple conditions that varied in difficulty. The studies were highly diverse and included experiments with both continuous- and binary-valued stimulus dimensions, a variety of different stimulus types, and both linearly- and nonlinearly-separable categories. Across these four applications, the SDM was the most successful measure at predicting the observed rank ordering of conditions by difficulty with an average Spearman's r of .87, and it was also the most accurate measure of the six tested at predicting the numerical values of the mean error rates in each condition (accounting for 87% of the variance

in error rates across all conditions).

The only real failure in the ordinal predictions of the SDM is that the Shepard et al. (1961) type II categories turn out to be easier for humans to learn than the SDM predicts. However, as noted earlier, the optimal strategy for the type II categories has a straightforward verbal description (i.e., as a logical disjunction). This is also true for the type I categories. Therefore, types I and II are best characterized as rule-based tasks, whereas types III, IV, V, and VI seem more like information-integration tasks. Multiple systems theories of human category learning (e.g., COVIS; Ashby & Valentin, 2017) predict that rule-based and information-integration tasks are learned in qualitatively different ways, and it for this reason that the SDM was developed specifically to predict difficulty only in information-integration tasks.

Another possibility however, is that none of the Shepard et al. (1961) categories are learned procedurally because the stimuli vary on only three binary-valued dimensions. For example, Feldman (2000, 2004) showed that the difficulty of the Shepard et al. conditions is perfectly predicted by the Boolean complexity of the rule that describes category membership. If so, then the SDM should not be expected to accurately predict the difficulty of any Shepard et al. conditions. Whether or not any of these conditions are learned procedurally is an open question. Even so, there is evidence that categories in which the stimuli vary on four binary-valued dimensions are learned procedurally when Boolean complexity is high (Waldron & Ashby, 2001). Also, of course, in almost all real-world information-integration categories, objects vary on continuous- rather than binary-valued perceptual dimensions.[2] Thus, the Shepard et al. (1961) conditions are not representative of real-world categorization tasks. More research on how people learn the Shepard et al. categories is clearly needed. In any case, our hypothesis is that the SDM will accurately predict the difficulty of any categories learned procedurally.

One difference between the SDM and all other measures considered in this article is that the SDM has a free parameter (i.e., γ), whereas the other measures do not. This is because the SDM was constructed to predict difficulty for human learners, whereas all other measures are meant to predict difficulty of an optimal classifier (i.e., an ideal observer). The optimal classifier operates noise free, whereas even the best human learner must deal with perceptual noise. The γ parameter measures that noise (e.g., note from Figure 3 that difficulty increases with γ).

In the current applications, SDM-predicted difficulty did not depend much on γ (e.g., see Figure 3). Even so, the inclusion of γ in the measure allows the SDM to make some unique predictions, relative to the other measures. For example, adding a noise mask to the stimulus display should increase the number of visual neurons that respond and therefore increase γ. Thus, the SDM predicts that adding a noise mask increases difficulty. Similarly, SDM predicts that uniformly contracting the entire stimulus space will also increase difficulty. In contrast, none of the other measures predict that either of these manipulations will have any effect on difficulty because adding a mask or uniformly contracting the space should not affect the performance of the optimal classifier.

A future research project that might be worth pursuing would be to add a noise-sensitive parameter to some or all of the other measures considered here. This might improve their ability to predict human difficulty, although Figure 3 suggests that this improvement might have little effect on the measures' ordinal predictions. Such a project is well outside the scope of the current article however, because the computational implementation of a noise-sensitive parameter would likely be unique to each measure. For example, none of the other measures depend on radial basis functions or tuning curves, so they include no structure that would allow a parameter identical to γ to be added.

The success of the SDM in the applications considered in this article, relative to all other measures, suggests that the SDM might be used to improve computer-assisted classification. With access to the SDM, a computer would be in the best possible position to determine when humans would be in most need of computer assistance.

[2] One commonly cited counterexample is that animals either have wings or they do not. However, this binary categorization is the result of a decision. Perceptually there is enormous variability in the structures that might be labeled wings. For example, consider the differences among eagles, penguins, and seahorses.

Conclusions

Overall the SDM has the potential to be a valuable tool in both experimental design and human performance enhancement. A future research goal should be to generalize the SDM to account for many other factors that are known to affect human category learning, including fatigue (Maddox et al., 2009), stress (Ell, Cosley, & McCoy, 2011), and the retinal location of the stimulus during training versus testing (Rosedahl, Eckstein, & Ashby, 2018). The SDM can then be used to improve human-computer partnerships for important categorization tasks such as radiologists scanning x-rays for tumors, TSA agents examining bag scans for banned items, and more.

Acknowledgments

This research was supported by the National Science Foundation Graduate Research Fellowship under Grant No. 1650114 and by NIMH grant 2R01MH063760.

References

Alfonso-Reese, L. A., Ashby, F. G., & Brainard, D. H. (2002). What makes a categorization task difficult? *Perception & Psychophysics*, *64*(4), 570–583.

Ashby, F. G., Alfonso-Reese, L. A., Turken, A. U., & Waldron, E. M. (1998). A neuropsychological theory of multiple systems in category learning. *Psychological Review*, *105*(3), 442-481.

Ashby, F. G., & Maddox, W. T. (1992). Complex decision rules in categorization: contrasting novice and experienced performance. *Journal of Experimental Psychology: Human Perception and Performance*, *18*(1), 50-71.

Ashby, F. G., & Rosedahl, L. (2017). A neural interpretation of exemplar theory. *Psychological Review*, *124*(4), 472-482.

Ashby, F. G., & Valentin, V. V. (2017). Multiple systems of perceptual category learning: Theory and cognitive tests. In *Handbook of categorization in cognitive science (second edition)* (pp. 157–188). Elsevier.

Ashby, F. G., & Waldron, E. M. (1999). On the nature of implicit categorization. *Psychonomic Bulletin & Review*, *6*(3), 363-378.

Bozdogan, H. (1990). On the information-based measure of covariance complexity and its application to the evaluation of multivariate linear models. *Communications in Statistics-Theory and Methods*, *19*(1), 221–278.

Cantwell, G., Crossley, M. J., & Ashby, F. G. (2015). Multiple stages of learning in perceptual categorization: Evidence and neurocomputational theory. *Psychonomic Bulletin & Review*, *22*, 1598–1613.

Edmunds, C., & Wills, A. J. (2016). Modeling category learning using a dual-system approach: A simulation of shepard, hovalnd and jenkins (1961) by covis. In *Cogsci.*

Ell, S. W., & Ashby, F. G. (2006). The effects of category overlap on information-integration and rule-based category learning. *Perception & Psychophysics*, *68*(6), 1013-1026.

Ell, S. W., Cosley, B., & McCoy, S. K. (2011). When bad stress goes good: increased threat reactivity predicts improved category learning performance. *Psychonomic bulletin & review*, *18*(1), 96–102.

Estes, W. K. (1986). Array models for category learning. *Cognitive Psychology*, *18*(4), 500–549.

Feldman, J. (2000). Minimization of boolean complexity in human concept learning. *Nature*, *407*(6804), 630.

Feldman, J. (2004). How surprising is a simple pattern? quantifying "eureka!". *Cognition*, *93*(3), 199–224.

Fukunaga, K. (2013). *Introduction to statistical pattern recognition, second edition*. Elsevier.

Kruschke, J. K. (1992). Alcove: an exemplar-based connectionist model of category learning. *Psychological review*, *99*(1), 22.

Lorena, A., Garcia, L., Lehmann, J., Souto, M., & Ho, T. (2018). How complex is your classification problem? a survey on measuring classification complexity. *arXiv*.

Love, B. C., & Medin, D. L. (1998). Sustain: A model of human category learning. In *Aaai/iaai* (pp. 671–676).

Maddox, W. T., Glass, B. D., Wolosin, S. M., Savarie, Z. R., Bowen, C., Matthews, M. D., & Schnyer, D. M. (2009). The effects of sleep deprivation on information-integration categorization performance. *Sleep*, *32*(11), 1439–1448.

McCoy, S. K., Hutchinson, S., Hawthorne, L., Cosley, B. J., & Ell, S. W. (2014). Is pressure stressful? the impact of pressure on the stress response and category learning. *Cognitive, Affective, & Behavioral Neuroscience*, *14*(2), 769–781.

Medin, D. L., & Schaffer, M. M. (1978). Context theory of classification learning. *Psychological Review*, *85*(3), 207–238.

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

Nosofsky, R. M. (1986). Attention, similarity, and the identification–categorization relationship. *Journal of Experimental Psychology: General*, *115*(1), 39–57.

Nosofsky, R. M., Gluck, M. A., Palmeri, T. J., McKinley, S. C., & Glauthier, P. (1994). Comparing modes of rule-based classification learning: A replication and extension of shepard, hovland, and jenkins (1961). *Memory & cognition*, *22*(3), 352–369.

Posner, M. I., & Keele, S. W. (1968). On the genesis of abstract ideas. *Journal of experimental psychology*, *77*(3p1), 353–363.

Rosedahl, L. A., Eckstein, M. P., & Ashby, F. G. (2018). Retinal-specific category learning. *Nature Human Behaviour*, *2*(7), 500.

Salatas, H., & Bourne, L. (1974). Learning conceptual rules: Iii. processes contributing to rule difficulty. *Memory & Cognition*, *2*(3), 549–553.

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

Shepard, R. N., Hovland, C. I., & Jenkins, H. M. (1961). Learning and memorization of classifications. *Psychological monographs: General and applied*, *75*(13), 1.

Smith, J. D., Minda, J. P., & Washburn, D. A. (2004). Category learning in rhesus monkeys: a study of the shepard, hovland, and jenkins (1961) tasks. *Journal of Experimental Psychology: General*, *133*(3), 398.

Waldron, E. M., & Ashby, F. G. (2001). The effects of concurrent task interference on category learning: Evidence for multiple category learning systems. *Psychonomic Bulletin & Review*, *8*(1), 168-176.