frontiers | Frontiers in Psychology

TYPE Brief Research Report
PUBLISHED 09 December 2024
DOI 10.3389/fpsyg.2024.1477514

OPEN ACCESS

EDITED BY
Ulrich Hoffrage,
Université de Lausanne, Switzerland

REVIEWED BY
Kenneth Kurtz,
Binghamton University, United States
Yoshihisa Fujita,
Kyoto University, Japan

*CORRESPONDENCE
Troy M. Houser
✉ thouser@uoregon.edu

RECEIVED 07 August 2024
ACCEPTED 25 November 2024
PUBLISHED 09 December 2024

CITATION
Houser TM (2024) A boundedly rational model for category learning.
Front. Psychol. 15:1477514.
doi: 10.3389/fpsyg.2024.1477514

A boundedly rational model for category learning

Troy M. Houser[1,2]*

[1]Department of Psychology, University of Oregon, Eugene, OR, United States, [2]Institute of Neuroscience, University of Oregon, Eugene, OR, United States

The computational modeling of category learning is typically evaluated in terms of the model's accuracy. For a model to accurately infer category membership of stimuli, it has to have sufficient representational precision. Thus, many category learning models infer category representations that guide decision-making and the model's fitness is evaluated by its ability to accurately choose. Substantial decision-making research, however, indicates that noise plays an important role. Specifically, noisy representations are assumed to introduce an element of stochasticity to decision-making. Noise can be minimized at the cost of cognitive resource expenditure. Thus, a more biologically plausible model of category learning should balance representational precision with costs. Here, we tested an autoencoder model that learns categories (the six category structures introduced by Roger Shepard and colleagues) by balancing the minimization of error with minimization of resource usage. By incorporating the goal of reducing category complexity, the currently proposed model biases category decisions toward previously learned central tendencies. We show that this model is still able to account for category learning performance in a traditional category learning benchmark. The currently proposed model additionally makes some novel predictions about category learning that future studies can test empirically. The goal of this paper is to make progress toward development of an ecologically and neurobiologically plausible model of category learning that can guide future studies and theoretical frameworks.

KEYWORDS

category learning, autoencoder (AE) neural networks, concept learning, generalization (psychology), RULEX, rate distortion theory, efficient coding theory

Introduction

Conceptual knowledge is a defining characteristic of human intelligence. A powerful way that conceptual knowledge is used is by generalizing it to novel situations, enabling efficient and adaptive behavior (Shepard, 1957, 1987, 1994). For example, when we go to a new grocery store, we can generalize previously acquired knowledge about grocery store layouts to infer that the cheese will be close to the milk. A concept is a mental representation of a category (Goldstone et al., 2018). Thus, the concept of a snake refers to the mental representation of a subjectively constructed category labeled *snake*. Given that categories are constructed by individuals to organize their personal experiences, there are numerous possibilities for *how one* might categorize. Despite considerable advancements in the field, there remains a lack of consensus among researchers regarding the psychological nature of categories. In what follows, we introduce a boundedly rational theoretical framework and novel extension of a previously posited process-level computational model that can capture key aspects of human category learning and memory. The guiding notion is that concepts are boundedly rational representations of categories.

Bounded rationality when acquiring category knowledge

Humans make decisions based on internal representations of external variables (Gershman and Daw, 2017; Niv, 2019), but how such variables are encoded and subsequently decoded to make a decision remains an open question. In real-world decision making, biological systems often have to infer latent states (e.g., categories). Many cognitive models of categorization decisions assume veridical internal representations of categories (Nosofsky, 1986; Nosofsky et al., 1994a,b). Substantial work in reinforcement learning and magnitude discrimination suggests that some amount of noise is inevitable in internal representations (Azeredo da Silveira et al., 2021; Barretto-García et al., 2023; Li et al., 2017; Prat-Carrabin and Woodford, 2022, 2024; Spitzer et al., 2017). This is to say that it is likely infeasible for biological systems to encode and decode information without error. According to the principle of efficient coding (Barlow, 2013), biological systems should seek to maximize representational precision *while minimizing resource consumption*.

The category learning model proposed by Kurtz (2007), called the DIVergent Autoencoding (DIVA), has made important advances in making the modeling of category judgements more biologically realistic. DIVA is a neural network model that utilizes an autoencoder architecture. Autoencoders traditionally learn stimulus mappings in an unsupervised fashion. They have three main components: (1) an encoder, (2) a bottleneck, and (3) a decoder. The encoder takes input data and transforms it to a low dimensional space (the bottleneck). The bottleneck is a form of data compression, or dimensionality reduction, often employed in statistical methods like principal component analysis (PCA) or t-distributed stochastic neighbor embedding (t-SNE). It forces the model to extract out statistical regularities in the data, effectively shedding the irrelevant information, and therefore minimizing resource expenditure. Then these compressed representations are decompressed by the decoder, which transforms them back into their original dimensions, so as to reconstruct the input. Decoding is not trivial, as it is decoding *from the bottleneck*. In other words, the decoder attempts to reconstruct the input after getting rid of some of its original signal, consistent with the notion from efficient coding theory that biological systems have to balance representational precision with resource expenditure. Low reconstruction error indicates that the bottleneck extracted regularities well. Given that an autoencoder's function is to reconstruct the original input, it is typically not an architecture used to model supervised learning, which attempts to make discrete decisions. However, DIVA makes use of a divergent output layer that enables it to make categorical decisions. We discuss this feature below.

However, the traditional autoencoder can have trouble with generalizing because it can overfit to the data (Monshizadeh et al., 2021), by simply reconstructing learned exemplars rather than a category's central tendency (Bozkurt et al., 2021). Reconstructing a category's central tendency should facilitate broader generalization abilities. To circumvent this issue, we use a variational autoencoder (VAE; Kingma and Welling, 2019).

Rather than deterministically mapping inputs to the bottleneck component, VAEs map inputs to probability distributions, thereby adding a stochastic element and enabling generation of diverse outputs. Moreover, rather than sampling directly from these learned distributions [$z \sim \mathcal{N}(\mu, \sigma^2)$], which would be computationally intractable, VAEs use the "reparameterization trick" (Kingma et al., 2015). The reparameterization trick expresses the latent probability distributions as deterministic functions of their first two moments: $z = \mu + \sigma \cdot \varepsilon$, where ε is noise (which is a random sample from a 0 mean Gaussian with unit variance, see Kingma et al., 2015; Kingma and Welling, 2019). This trick makes the sampling procedure differentiable, which in turn allows the model parameters (μ and σ) to be updated through gradient descent optimization. The loss function that gets optimized is also unique for VAEs. It is a sum of two forms of loss, which is the key theoretical contribution that making DIVA variational makes. The loss function for VAEs is the sum of reconstruction error and the discrepancy between prior and posterior distributions for a sampled latent variable z. Reconstruction error is equivalent to distortion in rate distortion theory. It is a measure proportional to the mean squared error between the input and the reconstruction of the input produced by the decoder. The discrepancy between prior and posterior distributions is known as the Kullback-Leibler divergence (Cover and Thomas, 1991) and it functions as a regularizer, constraining decoded representations to be biased toward their prior distribution. This is a desirable property as it entails that, for example, a category representation acquired across numerous experiences cannot be substantially altered from a single outlier exemplar. In other words, the Kullback-Leibler divergence minimizes resources spent on encoding specific exemplars by penalizing higher discrepancies between the input and the central tendency of previous inputs.

It is known that allocated cognitive resources differs between people and can even fluctuate from moment to moment. Therefore, we made use of the β-VAE, which incorporates a non-negative parameter (β) that scales the Kullback-Leibler divergence (Higgins et al., 2017). By scaling the Kullback-Leibler divergence, the bias toward the central tendency of experience can be made more or less prominent. It is conceptually related to cognitive capacity (Bates and Jacobs, 2020), given that less reliance on priors means one can efficiently encode more specific information. Specifically, autoencoders by their very nature try to reconstruct an input, which may make them susceptible to overfitting to the identity of a stimulus (Steck, 2020). In the extreme case that an autoencoder learns to memorize every training stimulus, it would resemble the famous exemplar model (Nosofsky, 1986, 1987) of categorization. However, in the case of categories with many exemplars, this becomes computationally infeasible and thus a tradeoff must be maintained between precision of memories and resource expenditure. Because the Kullback-Leibler divergence functions as a regularizer, constraining representations to resemble prior representations, the VAE additionally minimizes the resource expenditure. Thus, by scaling the Kullback-Leibler divergence, β induces more or less reliance on the prior, effectively tilting the balance of precision and complexity toward one or the other. The relationship between β-VAEs and rate distortion theory has previously been made mathematically concrete (Alemi et al., 2017a,b).

Finally, we make the β-VAE divergent, as in DIVA and for reasons which we expound upon next. Traditional autoencoders utilize a single decoder to decode *n*-categories, or use multiple autoencoders for each category (Oja, 1989). Such approaches to category learning do not capture differences in category learning driven by learning conditions, such as the nature and number of contrasting categories. In the former case, it is difficult to apply to supervised learning and in the latter case, this is because each category is modeled independently (Kurtz, 2007). To solve this issue, Kurtz (2007) proposed a single (shared) hidden layer of units and *n* decoders, or *category channels,* in DIVA in order to obtain reconstruction errors for each category. Comparing reconstruction errors then allows one to test the following assumption, namely that using the model's low-dimensional representation of one category to reconstruct the current stimulus is better than using the model's representation of another category to reconstruct the current stimulus. Moreover, by maintaining a shared hidden layer, DIVA and the extension proposed here are plausible models of multitask learning (Ben-David and Schuller, 2003; Caruana, 1996, 1997, 1994), which has recently been revealed to naturally facilitate generalization and abstraction (Driscoll et al., 2024; Garner and Dux, 2023; Sanh et al., 2022; Wards et al., 2023) and may be related to mixed selectivity in the brain (Jeffrey et al., 2020; Kaufman et al., 2022; Rigotti et al., 2013), including the hippocampus (Bernardi et al., 2020; Kira et al., 2023) and the prefrontal cortex (Dang et al., 2021; Parthasarathy et al., 2017), both of which are involved in concept learning. Given the shared layer, the current model claims that the bottleneck component constitutes a space of multiple psychological spaces superimposed upon each other, which is distinct from predictions made by autoencoder models with a single decoder. This means that the current model will yield different reconstructions under different learning conditions (i.e., it utilizes interdependent encoding techniques). By allocating a unique output channel for each category, divergent autoencoder architectures can model supervised learning by obtaining reconstruction errors for each category. For a schematic and relevant terms of the model proposed here see Figure 1.

To test the viability of the currently proposed model, which we call BR-DIVA (for *Boundedly-Rational*-DIVA, see below), we compare its ability to capture a classic benchmark of category learning to the original DIVA model and consider unique predictions by making DIVA variational. The aim of the current paper is to guide future research by positing a few category learning predictions that follow logically from computational principles.

Model features

The VAE model proposed here is a neural network model with three layers composed of three, two, and six neuron-like units, respectively. The number of units per layer were selected based on the stimulus set used in the current study. Because the stimuli are three-dimensional, the input layer is composed of three units and the output layer is composed to 3 x *n-categories* units. To be comparable to DIVA, which used two hidden layer units, we fixed the number of units in the hidden layer, or bottleneck to two. More details on the stimulus set are provided below. Input and output layers are fully connected with the hidden layer (i.e., the bottleneck). These connections denote the associations between input stimuli, internal cognitive representations, and reconstructions and are learned by iterative updating of weights that scale each connection strength. Unit weights are learned via standard backpropagation (Rumelhart et al., 1986) and activations are passed through a sigmoid function. Weights are updated in proportion to the learning rate. Unit weights are initialized with random values between default values of ± 0.5, which is convention for neural network research (Kolen and Pollack, 1990) and used in the paper introducing DIVA (Kurtz, 2007).

Activations spread from input to hidden layer units. The hidden layer is comprised of two neuron-like units, which is what gives it its status as a bottleneck. That is, by projecting three-dimensional inputs (see below) onto a two-dimensional space, the encoder is forced to reduce the input's dimensionality. Then the hidden layer projects to the output layer, which has dimensionality equal to the dimensionality of the input stimulus for each channel, which is why the output layer has 6 units (3 units for each category; see below for explanation of the stimuli).

To optimize model fit, a loss function gets minimized. The loss function is the sum of two terms: (1) reconstruction error, and (2) weighted Kullback-Leibler divergence. To obtain the measure of reconstruction error, squared differences between each category channel's output node activations and the input are calculated and scaled with a sensitivity parameter that controls the amount of attention paid to each feature. Summing these differences within each category channel yields a reconstruction error for each category. These measures are then added to the Kullback-Leibler divergence that itself gets scaled by the regularization parameter β. For additional details on how parameter settings relate to category learning, see (Kurtz, 2007, 2015). Here, we fix the sensitivity and learning rate parameters to 1 for brevity [as was done in the original DIVA simulations (Kurtz, 2007)]; and to elucidate the differences between DIVA and BR-DIVA models. DIVA also makes use of an attention breadth parameter that specifies how much attention is allocated to specific dimensions vs. all dimensions; however, to facilitate ease of comparison, this parameter was also fixed to 1 for both models.

To demonstrate the plausibility of the current model's ability to capture human category learning, we test its ability to simulate category learning on the seminal "Six Problems" introduced by Shepard et al. (1961).

The Six Problems

Shepard et al. (1961) tested the difficulty of categorization judgments depending on how the same 8 stimuli were grouped. Specifically, participants were shown three-dimensional stimuli, where each dimension denotes a binary feature (e.g., color, size, and shape). These eight stimuli can be grouped into two groups in 70 different ways, but only six of these are structurally distinct. By "structurally distinct," we mean that a grouping is not different simply by swapping out features. For example, a Type 1 grouping assigns all four stimuli with one color value (say, black) to Category A and all four stimuli with the other color value (say, white) to Category B. Grouping the stimuli using the same kind

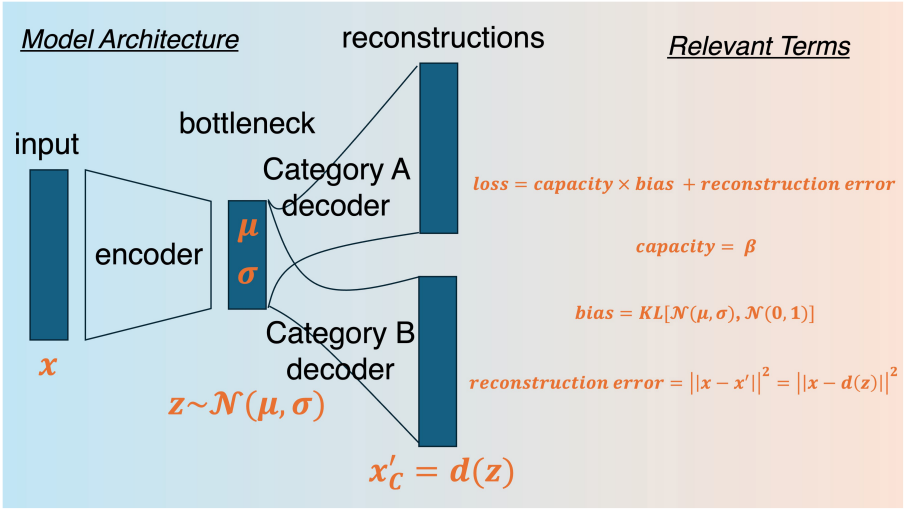

FIGURE 1
BR-DIVA model architecture. From the far left side, the model begins by taking in an input vector and projecting it onto a hidden layer (bottleneck). Then decoders for each category samples from the hidden layer space to reconstruct the input. Relevant terms reveals the loss function that gets optimized, which is a sum of reconstruction error and capacity-weighted bias. Capacity is simply a freely estimated parameter and bias is the Kullback Leibler divergence between prior and posterior distributions at the hidden layer. Reconstruction error is the squared absolute difference between input and reconstructed representations.

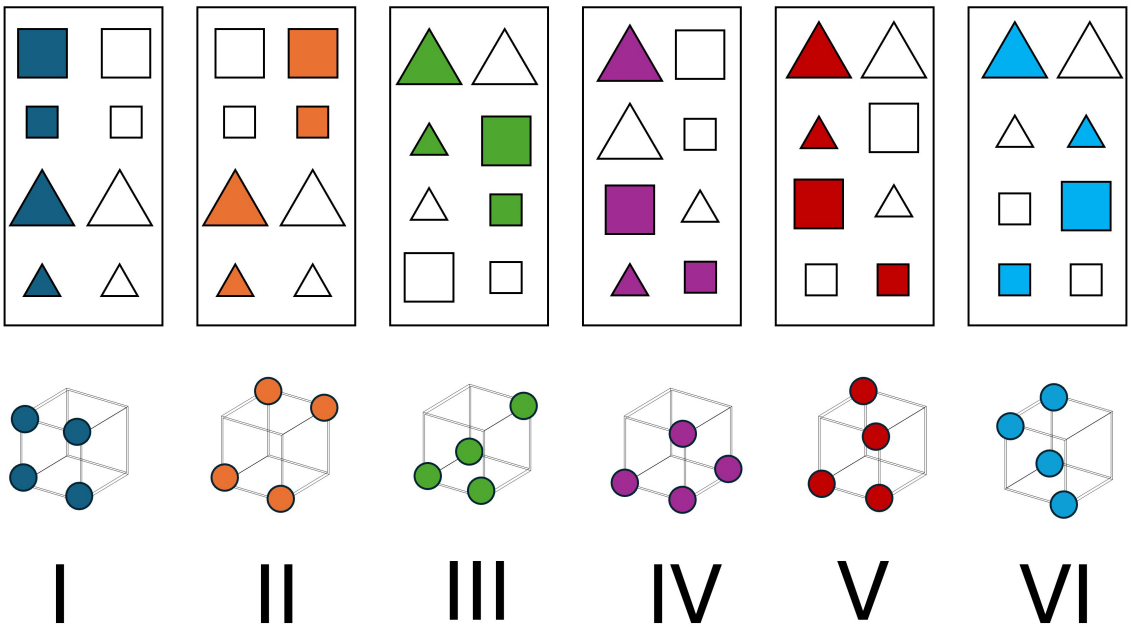

FIGURE 2
Six Problems. Every category structure implemented in the seminal paper by Shepard et al. (1961). Within each panel, each stimulus on the left belongs to one category and all the stimuli on the right belong to another category. Below the top panels is a 3-dimensional representation of the each category structure.

of unidimensional rule, simply for a different dimension (i.e., grouping all small stimuli into A and all large stimuli into B) is a technically unique grouping but not structurally distinct.

The six types of groupings differ in the number of dimensions one must attend to in order to achieve optimal performance (Type 1: one dimension, Type 2: two dimensions, and Types 3–6: three dimensions). Type 1 adheres to a unidimensional rule-based structure, such that all stimuli with one value on a dimension (e.g., color in Figure 2) belong to Category A while all stimuli with the other value on the same dimension belong to Category B. Type 2 is an exclusive-OR (XOR) problem, where two dimensions are relevant. In Figure 2, Category A stimuli can be white and square or orange and triangle. Types 3, 4, and 5 can all be characterized as rule-plus-exception structures, where a single dimension defines

category assignments for three of the category's four stimuli and thus the fourth stimulus for each category must be memorized. Type 6 is the most difficult because it lacks any within-category similarity structure, meaning one must memorize each of the eight stimulus-response associations to perform optimally. Figure 2 shows an example for each of the six types.

The main findings (i.e., that performance follows difficulty level; Type 1 > Type 2 > Types 3–5 > Type 6) from the Six Problems introduced in Shepard et al. (1961) have been replicated many times, with larger sample sizes, diverse stimulus sets, and across species (Kurtz et al., 2013; Nosofsky et al., 1994a; Smith et al., 2004).

Method

We simulated $n = 100$ participants that performed each of the Six Problems. The model was constructed as a stateful list processor (see Wills et al., 2017) and used the *slpDIVA* function (DIVA model) from the R package *catlearn* (Wills et al., 2017) as a starting point. The current model begins each simulation with randomly initiated weights. A binary three-dimensional input, representing one of the eight stimuli from the Six Problems (i.e., a trial), serves as the first layer and is mapped to 2 probability distributions (i.e., the bottleneck) via matrix multiplication with a set of input weights. These distributions are reparameterized via the reparameterization trick (Kingma et al., 2015). Reparameterized means of these distributions are the hidden unit activation levels which are then projected to two three-dimensional output layers via a set of output weights for each category. The output weights represent input reconstructions. A category judgment, which gets a 1 or 0 for accuracy, is whichever category has less reconstruction error. One simulation is 20 blocks of category learning, where a single block is one iteration through all eight stimuli, presented to the model in random order. We tested both the BR-DIVA and original DIVA model in order to test for any additional benefit of making DIVA variational.

We ran the above procedure for each of 50 different β values, from 0.01 to 100 in evenly spaced increments on a logarithmic scale. By fixing the parameters common to both the original DIVA model and the currently proposed BR-DIVA model, we can succinctly evaluate the contribution that bounded rationality makes to the divergent autoencoding architecture of category learning.

We conducted statistical analysis on the simulated performances from both BR-DIVA and DIVA. All analyses were done on accuracy (proportion correct), though plots show error rate (proportion incorrect) to facilitate easy comparison with previous studies studying the Six Problems. To test the extent to which BR-DIVA's category learning reflects the order of difficulty observed in the Six Problems, we ran a simple linear regression, predicting aggregated performance (overall mean accuracy) from problem type and β parameter value. We ran the same tests for performance from DIVA model (without the β parameter predictor). We ran *post-hoc* paired samples *t*-tests when necessary. To compare performance to empirical data, we obtained public datasets deposited in the R package *sixproblems*. These datasets are from Nosofsky et al. (1994a) and Lewandowsky (2011), and we will refer to these datasets as nosofsky94 and lewandowsky11 for simplicity. We briefly describe these datasets below.

After comparing overall performance, we evaluated differences in performance over time (learning curves) between BR-DIVA and DIVA. We conducted simple linear regression models predicting accuracies from block and type for both models.

Nosofsky94 is comprised of 120 participants. Each participant performed two problem types and each problem type was administered an equal number of times. Thus, there were 40 participants assigned to each problem type. The order of problem type assignment to each participant was counterbalanced. The first two blocks comprised one showing of each of the eight stimuli and all subsequent blocks comprised two showings of each of the eight stimuli. Participants continued the task until reaching a criterion of four consecutive sub-blocks of eight stimuli with perfect accuracy or for a maximum of 25 blocks.

Lewandowsky11 is comprised of 113 participants, who each did all six problem types in counterbalanced order. Each problem type was studied for a maximum of 12 blocks, where each block featured 2 showings of each of the eight stimuli. Study was terminated if accuracy was perfect for two consecutive blocks.

To compare learning curves predicted by BR-DIVA with observed data, we ran a mixed effects linear regression model using the *lmer* function from R's lmerTest package. This model predicted accuracy from problem type (3–5), block, and their interaction. We also included subject IDs and which dataset the data came from Nosofsky94 or Lewandowsky11 as random effects. For effects of problem type, Type 5 was entered into the model as the reference group. Thus, positive coefficients for Types 3 and 4 indicate higher accuracy than Type 5, and vice versa.

Results

Order of difficulty

The relative ease of acquisition of category knowledge across the Six Problems introduced in Shepard et al. (1961) was tested in the boundedly rational model proposed here. We first ran a simple linear regression, predicting average proportion of correct responses (across simulated subjects and blocks) from type (1–6) and β. Please note that all βs with associated *p*-values below are referring to regression coefficients and not the model parameter. This reveals significant main effects of all types ($\beta_{1-2} = -0.09$, $p < 0.001$; $\beta_{1-3} = -0.09$, $p < 0.001$; $\beta_{1-4} = -0.08$, $p < 0.001$; $\beta_{1-5} = -0.1$, $p < 0.001$; $\beta_{1-6} = -0.4$, $p < 0.001$). Moreover, visual inspection of Figure 3A tells us that performance follows the order of difficulty typically observed. Supplementary Figure 1 additionally shows that BR-DIVA, like the original DIVA, can capture the revised ordering of the Six Problems, as elucidated in Kurtz et al. (2013). Further, the BR-DIVA model performance remains relatively stable across all tested β values, at least at the aggregated level (Figure 3A). Paired-samples *t*-tests showed that BR-DIVA predicts worse accuracy than DIVA on Type 2 [$t_{(99)} = -3.17$, $p = 0.002$] and, more prominently, Type 5 [$t_{(99)} = -5.76$, $p < 0.001$], and predicted significantly better accuracy than DIVA on Type 4 [$t_{(99)} = 4.20$, $p < 0.001$]. All other *ps* > 0.402. Overall sums of squared differences between error probabilities as observed

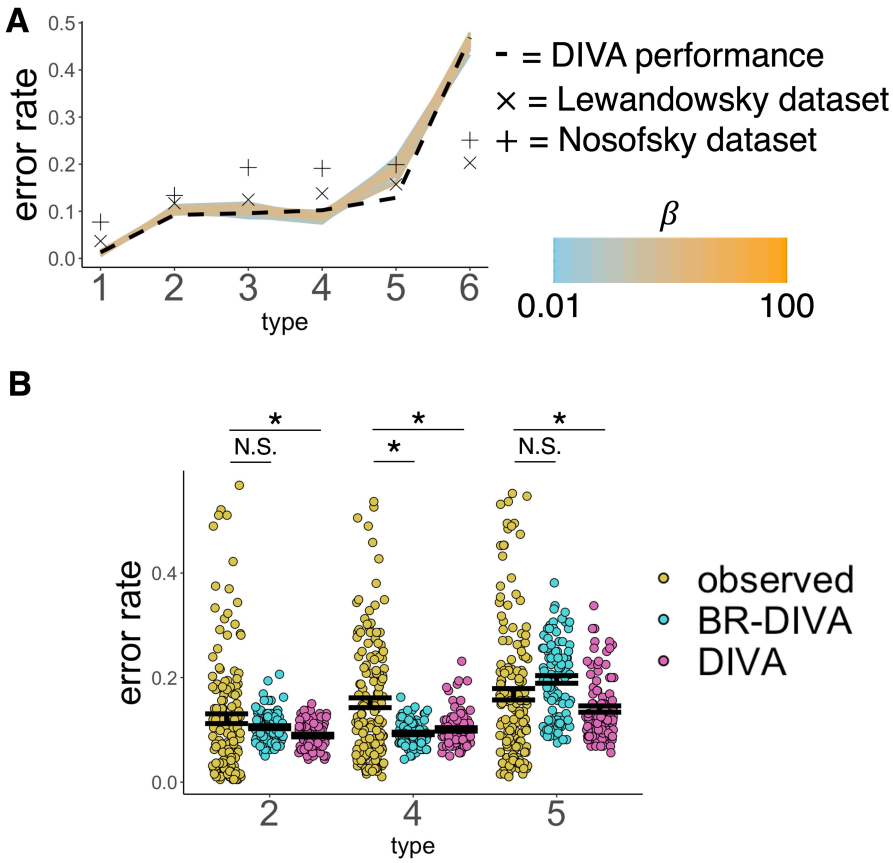

FIGURE 3
Overall model performance on the Six Problems. **(A)** X-axis denotes problem type and the y-axis denotes the overall mean performance. Colored lines are BR-DIVA predictions assuming β values ranging from 0.01 (light blue) to 100 (orange). The dashed line are predictions made by DIVA. Xs are empirically observed performances from Lewandowsky (2011) and +s are empirically observed performances from Nosofsky et al. (1994a). **(B)** Overall mean accuracy predicted by both BR-DIVA (blue) and DIVA (pink) and observed performance from participants from both Lewandowsky (2011) and Nosofsky et al. (1994a). Dots represent individual participants or simulated participants. Error bars are ±SEM. * <0.05. NS > 0.05.

in Nosofsky94/Lewandowsky11 and both BR-DIVA and DIVA are reported in Supplementary Table 1.

To determine whether these unique predictions made by BR-DIVA better reflect empirical performance than DIVA, we compared performance to that reported in Nosofsky et al. (1994a) and Lewandowsky (2011). We ran six two-samples *t*-tests, comparing simulated performances by BR-DIVA and DIVA on Types 2, 4, and 5 with subject averages from Nosofsky et al. and Lewandowsky on the same problems. We collapsed across both datasets, but running the analyses on each dataset separately support the same conclusions. Both models predict significantly better accuracy on Type 4 than is actually observed [BR-DIVA: $t_{(251)} = 5.63$, $p < 0.001$; DIVA: $t_{(251)} = 4.18$, $p < 0.001$]. Intriguingly, however, while DIVA predicts significantly more categorization accuracy than is actually observed for both Types 2 [$t_{(251)} = 2.49$, $p = 0.014$) and 5 [$t_{(251)} = 2.80$, $p = 0.005$], BR-DIVA's predictions statistically match observed performances [Type 2: $t_{(251)} = 1.50$, $p = 0.135$; Type 5: $t_{(251)} = -1.38$, $p = 0.170$]. Thus, at the aggregate level, BR-DIVA makes many of the same predictions as DIVA with respect to the Six Problems, as is to be expected given that BR-DIVA is a variational version of DIVA. However, BR-DIVA makes aggregate predictions for Types 2 and 5 that are statistically similar to what is empirically observed in people whereas DIVA does not (assuming all shared parameters are the same across models; Figure 3B).

Learning curves

To obtain a finer-grained perspective of category learning, we next looked at the learning curves for BR-DIVA. We found that BR-DIVA learns at a similar rate to DIVA for Types 1, 2, 3, and 4, and that learning is relatively stable across different values for β. For Type 5, BR-DIVA and DIVA clearly make different predictions (by the final block, BR-DIVA's best performance, across β*s*, was 96% accuracy, which DIVA surpasses on the 13th block; Figure 4A). Moreover, DIVA's learning curve for Type 6 appears to fluctuate more erratically than BR-DIVA's performance. To follow-up on these observations, we ran two linear regression models, predicting model accuracy on either Type 5 or Type 6 from block (1–20), model (BR-DIVA, DIVA), and their interaction. Please note that all β*s* with associated *p*-values below are referring to regression coefficients and not the model parameter.

The regression model predicting Type 5 performance showed only a main effect of block ($\beta = 0.02$, $p < 0.001$; all other *ps* > 0.177), meaning both models successfully learned the category

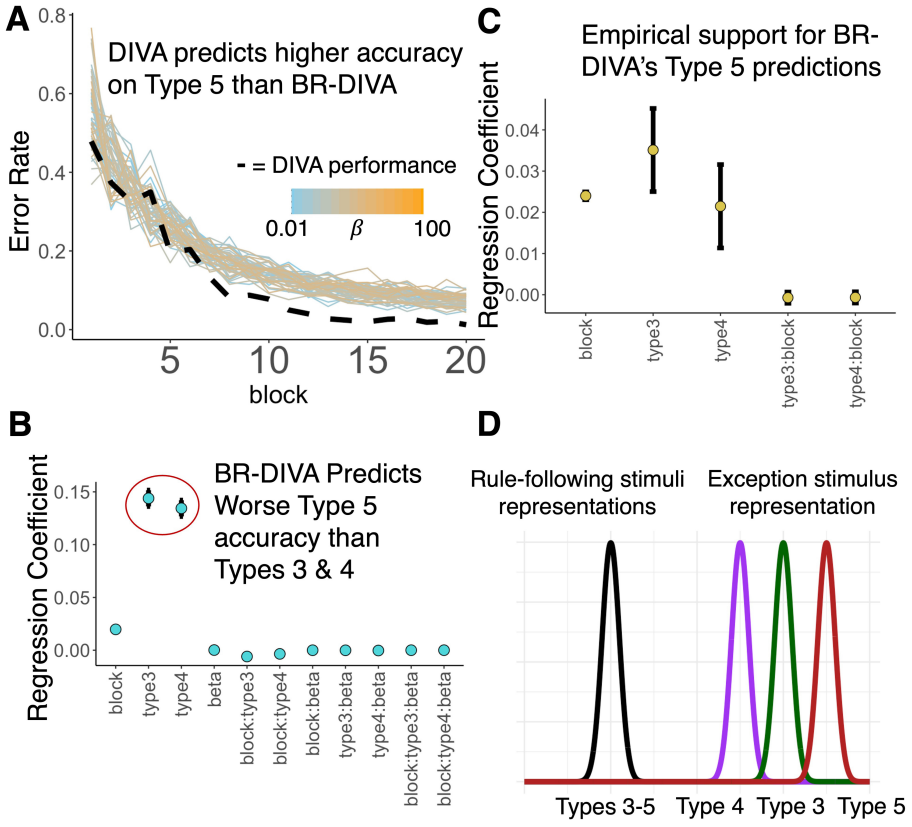

FIGURE 4
BR-DIVA and observed data suggest Type 5 is harder to learn than Types 3 and 4. **(A)** Learning curves predicted by BR-DIVA (colored lines) and DIVA (dashed line). Colored lines are BR-DIVA predictions assuming β values ranging from 0.01 (light blue) to 100 (orange). **(B)** Coefficients for predictors denoted along the x-axis from linear regression model predicting accuracy across blocks for the BR-DIVA model. Both circled coefficients are significantly different from 0 (to reiterate, the effects of problem type, Type 5 was entered into the model as the reference group). **(C)** Coefficients for predictors denoted along the x-axis from a mixed effects model predicting accuracy across blocks for data obtained from Lewandowsky (2011) and Nosofsky et al. (1994a). Regression coefficients for Type 3 and Type 4 are significantly different from zero. Error bars are ±SEM. **(D)** Conceptual schematic explaining why BR-DIVA predicts worse performance on Type 5 than Types 3 and 4 (which also explains its better Type 4 than Type 3 prediction). The black distribution on the left represents a prior distribution for a category representation represented in the bottleneck layer. This distribution is assumed to be learned by rule acquisition, as Types 3–5 all adhere to a rule-plus-exception category structure. Given that this distribution is a category representation, then learning the exception stimulus for each of these types will require this distribution to expand to incorporate the exception. As such, learning the exception stimulus should be a function of its distance from the prior distribution (in the plot, distance along the x-axis). Type 4's exception is closest to its rule-followers, Type 3's exception is second-closest to its rule-followers, and Type 5's exception is furthest from its rule-followers.

structure over time. Similarly, the regression model predicting Type 6 performance showed a main effect of block ($\beta = 0.02$, $p < 0.001$), but also a marginal effect of model [$\beta(BRDIVA - DIVA) = 0.02$, $p = 0.064$]. Supplementary Figure 2 shows learning curve predictions for BR-DIVA at all tested βs and DIVA.

Given the consistent differences between Type 5 performance between BR-DIVA and DIVA (Figures 3B, 4A), we ran an additional test to try and formulate a specific prediction that could guide future empirical research. Given that many studies on the Six Problems focus on Types 1, 2, 4, and 6 only (Kurtz et al., 2013; Love, 2002; Love and Markman, 2003; Minda et al., 2008; Rabi and Minda, 2016; Rehder and Hoffman, 2005a), likely because Types 3, 4, and 5 tend to be lumped together due to similar performance on these problems (Nosofsky et al., 1994a; Shepard et al., 1961), it is perhaps notable that BR-DIVA predicted worse performance on Type 5 than DIVA and that BR-DIVA captured the empirical data for this category structure better. Therefore, we ran an additional linear regression model, predicting BR-DIVA accuracies from Types 3, 4, and 5 from block (1–20), Type (3–5), β*s*, and all interactions. Indeed, this model showed that Type 5 accuracy was significantly lower than both Types 3 ($\beta_{3-5} = 0.14$, $p < 0.001$) and 4 ($\beta_{4-5} = 0.14$, $p < 0.001$). This model also revealed significant Type 3 x block ($\beta_{3,block-5,block} = -0.006$, $p < 0.001$) and Type 4 x block ($\beta_{3,block-5,block} = -0.003$, $p < 0.001$) interactions, such that learning curves were steeper for Type 5. See Figure 4B for all model predictor effects.

To test the extent to which these unique predictions made by BR-DIVA are reflected in the real world, we ran a linear mixed effects model, predicting correct responses by participants from two previously collected datasets (Lewandowsky, 2011; Nosofsky et al., 1994a) from type (3–5), block, and their interaction. We also included subject IDs and which dataset the data came from as random effects. As was expected, there was a main effect of block ($\beta_{block} = 0.02$, $p < 0.001$); however, consistent with the predictions

made by BR-DIVA, there were also main effects of Type 3 ($\beta_{3-5} = 0.04$, $p < 0.001$) and Type 4 ($\beta_{4-5} = 0.02$, $p = 0.034$). Interactions between block and Types 3 and 4 were not statistically significant (both $|\beta s| < 0.001$, both $ps > 0.604$). See Figure 4C for all model predictor effects. Figure 4D shows a schematic meant to visualize a plausible explanation for these results, which is further expounded upon in the discussion. Additionally, Supplementary Figure 3 shows the low-dimensional representations of each category for BR-DIVA, as well as inter-item distances in the low-dimensional space, which reveals that BR-DIVA represents Type 5 exception stimuli as further from rule-following stimuli than for Types 3 and 4 exception stimuli. Supplementary Figure 4 provides further evidence for this notion that Type 5 difficulty is a function of its inter-item distances by visualizing error rates across blocks split into rule-following and exception stimuli. Whereas, for Types 3 and 4 exception stimuli are learned at a pace similar to their rule-following stimuli, Type 5 shows that exception stimuli error rates remain higher than rule-following error rates until roughly the 15th block. Notably, however, this interpretation is incomplete as Supplementary Figure 3 shows that low dimensional representations of Type 4 exception stimuli are further from rule-following stimuli than Type 3′s exception stimuli.

Discussion

In this brief report, we simulated performance on the canonical Six Problems known to elucidate general category learning behavior (Shepard et al., 1961) using an autoencoder model that applies principles of efficient coding (Barlow, 2013) to encode information in a boundedly rational manner. We showed that this model—BR-DIVA—captures the classical order of difficulty observed on the Six Problems (Nosofsky et al., 1994a; Shepard et al., 1961). Beyond these findings, the boundedly rational model proposed here predicted lower accuracy on Type 5 than what is predicted by the autoencoding model it is based on. Importantly, we found that this unique prediction is more aligned with empirical data than the base model. We discuss and speculate on this finding next.

Type 5 is more difficult than Types 3 and 4

The classical Six Problems of category learning introduced in Shepard et al. (1961) produced substantial excitement about Types 1, 2, 4, and sometimes 6. Many studies that use the Six Problems only focus on this subset (Kurtz et al., 2013; Love, 2002; Minda et al., 2008; Rabi and Minda, 2016; Rehder and Hoffman, 2005b). Since the findings from Shepard and colleagues, there has been a tendency to lump performance on Types 3–5 together, as if they were the same category structures. Indeed, they do all adhere to a rule-plus-exception design (Nosofsky et al., 1994b); however, it is perhaps notable that the boundedly rational model put forth in the current paper consistently predicted worse performance on Type 5 than Types 3 and 4. This prediction did not reach statistical significance in the model on which the boundedly rational model is based on (i.e., DIVA). When comparing boundedly-rational-DIVA and DIVA to empirically observed performance differences between Type 5 and Types 3 and 4, we found that the data is more consistent with the boundedly-rational-DIVA's predictions.

One possible explanation for this discrepancy is in terms of information gain, which expresses the amount of information gained about a signal by observing another variable (Mathy, 2010). For example, by learning the weather one is likely better able to gauge what clothes a random person will be wearing. Thus, knowing the weather reduces one's uncertainty about what clothes people will be wearing. In terms of the Six Problems, information gain is relevant because it denotes the amount of information a given stimulus supplies about the categories. This notion is particularly important for rule-plus-exception category structures because it is assumed that people will learn the unidimensional rule first (Figure 4D, black distribution), in which case learning of the exception stimulus (Figure 4D, colored distributions) is a function of how distinct it is from the rule-following stimuli (Figure 4D, distance between black and colored distributions). In other words, learning a rule first to categorize stimuli will induce a bias toward the rule-following stimuli. As such, the more distinct (i.e., the more informative or the further from the bias) the exception stimulus is, the harder it will be to learn it. Consistent with this interpretation, the exception stimulus in Type 5 has a larger average distance from Type 5′s rule-following stimuli than Types 3 or 4. This within-category distance measure is proportional to a commonly used metric known as *structure ratios* (Conaway and Kurtz, 2017). This interpretation is also in line with Nosofsky et al. (1994b)'s RULEX model, which suggests that people test simple rules first and gradually hypothesize more complex rules if the simpler ones fail. In Supplementary Figure 3, the hidden unit activations for each of the eight stimuli in Type 5 from a representative simulation are plotted and visualized based on both category and whether the stimulus adhered to a unidimensional rule or not. Interestingly, this figure shows that exception stimuli are represented as further from rule-following stimuli within the same category (e.g., compare inter-item distances between red triangles and red circle, and between blue triangles and blue circle). Supplementary Figure 3 also shows that these distances are significantly greater than rule-to-exception stimulus distances for Types 3 and 4. Thus, the low-dimensional representations of stimuli are consistent with the interpretation visualized in Figure 4D. Together, this highlights the importance of priors during the process of learning categories and that, at least some, category structures' difficulty is a function of balancing representational precision with complexity.

Limitations

The current work is not meant to encompass all categorization phenomena. Indeed, the current work only tested one category learning paradigm (i.e., classification), comprised of relatively simple stimuli. The simplicity of the stimuli actually limits the amount of dimensionality reduction that could be performed by BR-DIVA in the current work, given that stimuli were three-dimensional and the bottleneck layer was two-dimensional. This could also be why there was no difference across simulations with different β*s*. Future work will need to test for BR-DIVA's applicability to higher dimensional, naturalistic, and continuous

stimuli, in addition to other paradigms, such as inference training and function learning. The current work was meant to take the first steps toward more broader applications, and thus, we generated BR-DIVA predictions and compared them with empirical data. Future studies will need to pit BR-DIVA against leading computational models of categorization such as SUSTAIN (Love et al., 2004), and prototype and exemplar models (Minda and Smith, 2002; Nosofsky, 1986, 1987, 1992; Smith and Minda, 2000, 2002). Moreover, while we did observe differences between Type 5 and Types 3 and 4 in empirical data, the analysis revealing this difference was targeted by using a subset of the overall dataset (only Types 3–5). As such, and in combination with many previous studies showing minimal performance discrepancies between these category structures, it is likely that this effect is quite subtle and future studies will need to test this prediction explicitly before any conclusive interpretations can be made.

Data availability statement

The original contributions presented in the study are included in the article/Supplementary material, further inquiries can be directed to the corresponding author.

Author contributions

TH: Writing – review & editing, Writing – original draft.

Funding

The author(s) declare that no financial support was received for the research, authorship, and/or publication of this article.

Conflict of interest

The author declares that the research was conducted in the absence of any commercial or financial relationships that could be construed as a potential conflict of interest.

Publisher's note

Supplementary material

The Supplementary Material for this article can be found online at: https://www.frontiersin.org/articles/10.3389/fpsyg.2024.1477514/full#supplementary-material

References

Alemi, A. A., Fischer, I., Dillon, J. V., and Murphy, K. (2017a). "Deep variational information bottleneck," in *5th International Conference on Learning Representations, ICLR 2017 - Conference Track Proceedings*. arXiv:1612.00410v2.

Alemi, A. A., Poole, B., Fischer, I., Dillon, J. V., Saurous, R. A., and Murphy, K. (2017b). *Information Theoretic Analysis of Deep Latent Variable Models. arXiv [Preprint]*. arXiv:1711.00464v3.

Azeredo da Silveira, R., Sung, Y., and Woodford, M. (2021). Optimally imprecise memory and biased forecasts. *SSRN Electr. J.* 2021:3731244. doi: 10.2139/ssrn.3731244

Barlow, H. B. (2013). Possible principles underlying the transformations of sensory messages. *Sens. Commun.* 3:13. doi: 10.7551/mitpress/9780262518420.003.0013

Barretto-García, M., de Hollander, G., Grueschow, M., Polanía, R., Woodford, M., and Ruff, C. C. (2023). Individual risk attitudes arise from noise in neurocognitive magnitude representations. *Nat. Hum. Behav.* 7:4. doi: 10.1038/s41562-023-01643-4

Bates, C. J., and Jacobs, R. A. (2020). Efficient data compression in perception and perceptual memory. *Psychol. Rev.* 2020:rev0000197. doi: 10.1037/rev0000197

Ben-David, S., and Schuller, R. (2003). "Exploiting task relatedness for multiple task learning," in *Lecture Notes in Artificial Intelligence (Subseries of Lecture Notes in Computer Science). Vol. 2777* (Berlin; Heidelberg: Springer).

Bernardi, S., Benna, M. K., Rigotti, M., Munuera, J., Fusi, S., and Daniel Salzman, C. (2020). The geometry of abstraction in the hippocampus and prefrontal cortex. *Cell* 183:31. doi: 10.1016/j.cell.2020.09.031

Bozkurt, A., Esmaeili, B., Tristan, J. -B., Brooks, D., Dy, J., and van de Meent, J.-W. (2021). "Rate-regularization and generalization in variational autoencoders," in *Proceedings of the 24th International Conference on Artificial Intelligence and Statistics*, 130. Retrieved from: https://par.nsf.gov/biblio/10280434

Caruana, R. (1994). "Learning many related tasks at the same time with backpropagation," in *NIPS 1994: Proceedings of the 7th International Conference on Neural Information Processing Systems* (Cambridge, MA: MIT Press).

Caruana, R. (1996). "Algorithms and applications for multitask learning," in *Conference on Machine Learning* (San Francisco, CA: Morgan Kaufmann Publishers Inc.), 12.

Caruana, R. (1997). Multitask learning. *Machine Learn.* 28:34. doi: 10.1023/A:1007379606734

Conaway, N., and Kurtz, K. J. (2017). Similar to the category, but not the exemplars: a study of generalization. *Psychon. Bullet. Rev.* 24:1. doi: 10.3758/s13423-016-1208-1

Cover, T. M., and Thomas, J. A. (1991). *Elements of Information Theory*. New York, NY: Wiley.

Dang, W., Jaffe, R. J., Qi, X. L., and Constantinidis, C. (2021). Emergence of non-linear mixed selectivity in prefrontal cortex after training. *J. Neurosci.* 41:20. doi: 10.1523/JNEUROSCI.2814-20.2021

Driscoll, L. N., Shenoy, K., and Sussillo, D. (2024). Flexible multitask computation in recurrent networks utilizes shared dynamical motifs. *Nat. Neurosci.* 27:6. doi: 10.1038/s41593-024-01668-6

Garner, K. G., and Dux, P. E. (2023). Knowledge generalization and the costs of multitasking. *Nat. Rev. Neurosci.* 24:653. doi: 10.1038/s41583-022-00653-x

Gershman, S. J., and Daw, N. D. (2017). Reinforcement learning and episodic memory in humans and animals: an integrative framework. *Ann. Rev. Psychol.* 68:33625. doi: 10.1146/annurev-psych-122414-033625

Goldstone, R. L., Kersten, A., and Carvalho, P. F. (2018). "Categorization and concepts," in *Stevens' Handbook of Experimental Psychology and Cognitive Neuroscience* (New York, NY: Wiley).

Higgins, I., Matthey, L., Pal, A., Burgess, C., Glorot, X., Botvinick, M., et al. (2017). "β-VAE: learning basic visual concepts with a constrained variational framework," in *5th International Conference on Learning Representations, ICLR 2017 - Conference Track Proceedings*. Available at: https://openreview.net/forum?id=Sy2fzU9gl

Jeffrey, J. W., Palmer, S. E., and Freedman, D. J. (2020). Nonlinear mixed selectivity supports reliable neural computation. *PLoS Comput. Biol.* 16:1007544. doi: 10.1371/journal.pcbi.1007544

Kaufman, M. T., Benna, M. K., Rigotti, M., Stefanini, F., Fusi, S., and Churchland, A. K. (2022). The implications of categorical and category-free mixed selectivity on representational geometries. *Curr. Opin. Neurobiol.* 77:102644. doi: 10.1016/j.conb.2022.102644

Kingma, D. P., Salimans, T., and Welling, M. (2015). "Variational dropout and the local reparameterization trick," in *Advances in Neural Information Processing Systems. Vols. 2015-January* (New York, NY: Curran Associates, Inc.).

Kingma, D. P., and Welling, M. (2019). An introduction to variational autoencoders. *Found. Trends Machine Learn.* 12:56. doi: 10.1561/2200000056

Kira, S., Safaai, H., Morcos, A. S., Panzeri, S., and Harvey, C. D. (2023). A distributed and efficient population code of mixed selectivity neurons for flexible navigation decisions. *Nat. Commun.* 14:2. doi: 10.1038/s41467-023-37804-2

Kolen, J. F., and Pollack, J. B. (1990). Back propagation is sensitive to initial conditions. *Compl. Syst.* 1990:4.

Kurtz, K. J. (2007). The Divergent Autoencoder (DIVA) model of category learning. *Psychon. Bullet. Rev.* 14:560–576. doi: 10.3758/BF03196806

Kurtz, K. J. (2015). Human category learning: toward a broader explanatory account. *Psychol. Learn. Motivat.* 63:77–114. doi: 10.1016/bs.plm.2015.03.001

Kurtz, K. J., Levering, K. R., Stanton, R. D., Romero, J., and Morris, S. N. (2013). Human learning of elemental category structures: revising the classic result of Shepard, Hovland, and Jenkins (1961). *J. Exp. Psychol.* 39:a0029178. doi: 10.1037/a0029178

Lewandowsky, S. (2011). Working memory capacity and categorization: individual differences and modeling. *J. Exp. Psychol.* 37:a0022639. doi: 10.1037/a0022639

Li, V., Castañón, S. H., Solomon, J. A., Vandormael, H., and Summerfield, C. (2017). Robust averaging protects decisions from noise in neural computations. *PLoS Comput. Biol.* 13:e1005723. doi: 10.1371/journal.pcbi.1005723

Love, B. C. (2002). Comparing supervised and unsupervised category learning. *Psychon. Bullet. Rev.* 9:829–835. doi: 10.3758/BF03196342

Love, B. C., and Markman, A. B. (2003). The nonindependence of stimulus properties in human category learning. *Mem. Cogn.* 31:790–799. doi: 10.3758/BF03196117

Love, B. C., Medin, D. L., and Gureckis, T. M. (2004). SUSTAIN: a network model of category learning. *Psychol. Rev.* 111:309–332. doi: 10.1037/0033-295X.111.2.309

Mathy, F. (2010). Assessing conceptual complexity and compressibility using information gain and mutual information. *Tutor. Quant. Methods Psychol.* 6, 16–30. doi: 10.20982/tqmp.06.1.p016

Minda, J. P., Desroches, A. S., and Church, B. A. (2008). Learning rule-described and non-rule-described categories: a comparison of children and adults. *J. Exp. Psychol.* 34:a0013355. doi: 10.1037/a0013355

Minda, J. P., and Smith, J. D. (2002). Comparing prototype-based and exemplar-based accounts of category learning and attentional allocation. *J. Exp. Psychol.* 28, 275–292. doi: 10.1037//0278-7393.28.2.275

Monshizadeh, M., Khatri, V., Gamdou, M., Kantola, R., and Yan, Z. (2021). Improving data generalization with variational autoencoders for network traffic anomaly detection. *IEEE Access* 9, 2169–3536. doi: 10.1109/ACCESS.2021.3072126

Niv, Y. (2019). Learning task-state representations. *Nat. Neurosci.* 22, 1544–1553. doi: 10.1038/s41593-019-0470-8

Nosofsky, R. M. (1986). Attention, similarity, and the identification-categorization relationship. *J. Exp. Psychol.* 115, 39–61. doi: 10.1037//0096-3445.115.1.39

Nosofsky, R. M. (1987). Attention and learning processes in the identification and categorization of integral stimuli. *J. Exp. Psychol.* 13, 87–108. doi: 10.1037//0278-7393.13.1.87

Nosofsky, R. M. (1992). *Exemplar, Prototypes, and Similarity Rules. From Learning Theory to Connectionist Theory: Essays in Honor of William K. Estes, Vol. 1*. Lawrence Erlbaum Associates, Inc.

Nosofsky, R. M., Gluck, M. A., Palmeri, T. J., Mckinley, S. C., and Glauthier, P. (1994a). Comparing modes of rule-based classification learning: a replication and extension of Shepard, Hovland, and Jenkins (1961). *Mem. Cogn.* 22, 352–362. doi: 10.3758/BF032 00862

Nosofsky, R. M., Palmeri, T. J., and McKinley, S. C. (1994b). Rule-plus-exception model of classification learning. *Psychol. Rev.* 101, 53–79. doi: 10.1037/0033-295X.101.1.53

Oja, E. (1989). Neural networks, principal components, and subspaces. *Int. J. Neural Syst.* 18:475. doi: 10.1142/S0129065789000475

Parthasarathy, A., Herikstad, R., Bong, J. H., Medina, F. S., Libedinsky, C., and Yen, S. C. (2017). Mixed selectivity morphs population codes in prefrontal cortex. *Nat. Neurosci.* 20, 1770–1779. doi: 10.1038/s41593-017-0003-2

Prat-Carrabin, A., and Woodford, M. (2022). Efficient coding of numbers explains decision bias and noise. *Nat. Hum. Behav.* 6, 1142–1152. doi: 10.1038/s41562-022-01352-4

Prat-Carrabin, A., and Woodford, M. (2024). Imprecise probabilistic inference from sequential data. *Psychol. Rev.* 131, 1161–1207. doi: 10.1037/rev0000469

Rabi, R., and Minda, J. P. (2016). Category learning in older adulthood: a study of the Shepard, Hovland, and Jenkins (1961) Tasks. *Psychol. Aging* 31, 185–197. doi: 10.1037/pag0000071

Rehder, B., and Hoffman, A. B. (2005a). Eyetracking and selective attention in category learning. *Cogn. Psychol.* 51, 1–41. doi: 10.1016/j.cogpsych.2004.11.001

Rehder, B., and Hoffman, A. B. (2005b). Thirty-something categorization results explained: selective attention, eyetracking, and models of category learning. *J. Exp. Psychol.* 31, 811–829. doi: 10.1037/0278-7393.31.5.811

Rigotti, M., Barak, O., Warden, M. R., Wang, X. J., Daw, N. D., Miller, E. K., et al. (2013). The importance of mixed selectivity in complex cognitive tasks. *Nature* 497, 585–590. doi: 10.1038/nature12160

Rumelhart, D. E., Hinton, G. E., and Williams, R. J. (1986). Learning representations by back-propagating errors. *Nature* 323, 533–536. doi: 10.1038/323533a0

Sanh, V., Webson, A., Raffel, C., Bach, S. H., Sutawika, L., Alyafeai, Z., et al. (2022). "Multitask prompted training enables zero-shot task generalization," in *ICLR 2022 - 10th International Conference on Learning Representations*. arXiv:2110.08207v3.

Shepard, R. N. (1957). Stimulus and response generalization: a stochastic model relating generalization to distance in psychological space. *Psychometrika* 22, 325–345. doi: 10.1007/BF02288967

Shepard, R. N. (1987). Toward a universal law of generalization for psychological science. *Science* 237, 1317–1323. doi: 10.1126/science.3629243

Shepard, R. N. (1994). Perceptual-cognitive universals as reflections of the world. *Psychon. Bullet. Rev.* 1, 2–28. doi: 10.3758/BF03200759

Shepard, R. N., Hovland, C. I., and Jenkins, H. M. (1961). Learning and memorization of classifications. *Psychol. Monogr.* 75, 1–42. doi: 10.1037/h0093825

Smith, J. D., and Minda, J. P. (2000). Thirty categorization results in search of a model. *J. Exp. Psychol.* 26, 3–27. doi: 10.1037//0278-7393.26.1.3

Smith, J. D., and Minda, J. P. (2002). Distinguishing prototype-based and exemplar-based processes in dot-pattern category learning. *J. Exp. Psychol.* 28, 800–811. doi: 10.1037//0278-7393.28.4.800

Smith, J. D., Minda, J. P., and Washbum, D. A. (2004). Category learning in rhesus monkeys: a study of the Shepard, Hovland, and Jenkins (1961) tasks. *J. Exp. Psychol.* 133, 398–414. doi: 10.1037/0096-3445.133.3.398

Spitzer, B., Waschke, L., and Summerfield, C. (2017). Selective overweighting of larger magnitudes during noisy numerical comparison. *Nat. Hum. Behav.* 1:e0145. doi: 10.1038/s41562-017-0145

Steck, H. (2020). "Autoencoders that don't overfit towards the identity," in *Advances in Neural Information Processing Systems. Vols. 2020-December* (New York, NY: Curran Associates, Inc.).

Wards, Y., Ehrhardt, S. E., Filmer, H. L., Mattingley, J. B., Garner, K. G., and Dux, P. E. (2023). Neural substrates of individual differences in learning generalization via combined brain stimulation and multitasking training. *Cerebr. Cortex* 33, 11679–11694. doi: 10.1093/cercor/bhad406

Wills, A. J., O'Connell, G., Edmunds, C. E. R., and Inkster, A. B. (2017). Progress in modeling through distributed collaboration: concepts, tools and category-learning examples. *Psychol. Learn. Motivat.* 66, 79–115. doi: 10.1016/bs.plm.2016.11.007