Cognition 222 (2022) 104997

Contents lists available at ScienceDirect

Cognition

journal homepage: www.elsevier.com/locate/cognit

A neural network model of the effect of prior experience with regularities on subsequent category learning

Casey L. Roark [a,b,*], David C. Plaut [a,b], Lori L. Holt [a,b,c]

[a] Department of Psychology, Carnegie Mellon University, 5000 Forbes Ave, Pittsburgh, PA 15213, USA
[b] Center for the Neural Basis of Cognition, 5000 Forbes Ave, Pittsburgh, PA 15213, USA
[c] Neuroscience Institute, Carnegie Mellon University, 5000 Forbes Ave, Pittsburgh, PA 15213, USA

ARTICLE INFO

Keywords:
Neural network
Perception
Category learning
Statistical regularities

ABSTRACT

Categories are often structured by the similarities of instances within the category defined across dimensions or features. Researchers typically assume that there is a direct, linear relationship between the physical input dimensions across which category exemplars are defined and the psychological representation of these dimensions. However, this assumption is not always warranted. Through a set of simulations, we demonstrate that the psychological representations of input dimensions developed through long-term prior experience can place very strong constraints on category learning. We compare the model's behavior to auditory, visual, and cross-modal human category learning and make conclusions regarding the nature of the psychological representations of the dimensions in those studies. These simulations support the conclusion that the nature of psychological representations of input dimensions is a critical aspect to understanding the mechanisms underlying category learning.

1. Introduction

Categorization is thought to be at the heart of many complex processes, such as object recognition (Richler & Palmeri, 2014) and speech perception (Holt & Lotto, 2010), and appears to be dependent on distributional regularities across exemplars that define a category. For instance, infants form sound and object categories based on the statistical distributions they experience in the input (Eimas, 1975; Maye, Werker, & Gerken, 2002; Smith, Jayaraman, Clerkin, & Yu, 2018; Werker, Yeung, & Yoshida, 2012). Adults are also sensitive to statistical structure of novel categories (Folstein, Gauthier, & Palmeri, 2010; Goudbeek, Cutler, & Smits, 2008; Pierrehumbert, 2003). Learners can approximate category distributions even from complex, non-Gaussian distributions (Gifford, Cohen, & Stocker, 2014) and are sensitive to statistical structure both within- and between-categories (Gureckis & Goldstone, 2008). The statistical structure in the sensory world is reflected in psychological and neural representations (Drucker, Kerr, & Aguirre, 2009; Lewicki, 2002; Schwartz & Simoncelli, 2001; Tijsseling & Gluck, 2002).

Learners' sensitivity to category-specific regularities has led researchers to investigate the importance of distributional regularities on category learning (e.g., Ashby, Alfonso-Reese, Turken, & Waldron, 1998; Aslin & Newport, 2014; Carvalho, Chen, & Yu, 2021). One influential theory of category learning that suggests that the neural and computational mechanisms supporting category learning are determined by the distributional regularities of those categories (Ashby et al., 1998). Specifically, Ashby and colleagues suggest that optimal learning of rule-based (RB) categories, which requires selective attention to individual dimensions and can be learned via hypothesis testing, relies on an explicit categorization system, supported by prefrontal cortex and the head of the caudate nucleus in the striatum (Ashby & Ell, 2001; Ashby & Waldron, 2000). In contrast, optimal learning of information-integration (II) categories, which requires pre-decisional integration across multiple dimensions learned via procedural learning mechanisms, relies on an implicit categorization system, supported by the putamen and body and tail of the caudate nucleus in the striatum (Ashby & Waldron, 1999, 2000). Thus, the relationship of the categories to the component dimensions is thought to be fundamental to the mechanisms of category learning. Consistent with this view, proponents of dual systems theory have generally demonstrated that RB categories requiring selective attention to relatively simple visual dimensions (e.g., orientation and spatial frequency in Gabor patches) are learned better and faster than II categories requiring integration over the same dimensions (see Ashby & Maddox, 2011 for a review).

* Corresponding author at: Department of Communication Science & Disorders, University of Pittsburgh, 5012 Forbes Ave, Pittsburgh, PA 15260, USA.
E-mail addresses: casey.l.roark@gmail.com (C.L. Roark), plaut@cmu.edu (D.C. Plaut), loriholt@cmu.edu (L.L. Holt).

https://doi.org/10.1016/j.cognition.2021.104997
Received 8 October 2020; Received in revised form 16 December 2021; Accepted 18 December 2021

From a mechanistic point of view, it is the distribution of stimuli within internal perceptual representations that influences category learning. Critically, the dimensions of this *perceptual* space may not necessarily be aligned with the *input* dimensions that are explicitly manipulated by experimenters. Previous investigations of the effect of category distributions on learning are driven by the – often implicit – assumption that exemplar distributions defined across input dimensions are linearly mapped to perceptual dimensions (Ashby & Soto, 2015; Johannesson, 2001). As such, experimental design and interpretation rely on the assumption of congruence between input and learners' internal perceptual dimensions. For some of the most well-studied dimensions in the visual domain, such as line length and orientation or spatial frequency and orientation of lines in a Gabor patch, the assumption of alignment between input dimensions and perceptual dimensions is likely valid, as representations of simple visual input dimensions are known to be orthogonal (Everson et al., 1998). However, for more complex visual objects, representations may reflect more abstract, latent dimensions, rather than veridical representations of the physical dimensions (Fleming & Storrs, 2019).

Recent applications of the dual systems theory in the auditory domain make clear that the experimenter assumption of alignment between input and perceptual dimensions may be problematic (Roark & Holt, 2019; Scharinger, Henry, & Obleser, 2013). In the auditory domain, it is likely that the representations of many dimensions are not independent (Garner, 1974), reflected in interdependent coding of even basic acoustic dimensions (Wang, 2007). When input dimensions defining categories are perceptually interdependent, statistically equivalent category input distributions can lead to very different learning challenges depending on whether the input distributions align (or misalign) with perceptual representations (Roark & Holt, 2019). Further, real-world auditory categories are often defined by many input dimensions, making the nature of perceptual representations of those dimensions difficult to determine. For example, there are at least 16 dimensions contributing to consonant voicing distinctions in speech (Lisker, 1986) and 20 contributing to perception of fricatives (McMurray & Jongman, 2011).

The nature of dimensions is important to consider because it affects what learners are able to do with those dimensions. For interdependent, integral dimensions, processing stimuli in a holistic manner is easy and selectively attending to individual dimensions is difficult (Foard & Kemler Nelson, 1984; Garner, 1974, 1976; Kemler Nelson, 1993). These kinds of constraints on processing can persist even with expertise-level training—for instance, even color experts are not able to optimally selectively attend to the integral visual dimensions of brightness and saturation (Burns & Shepp, 1988).

Yet, how the perceptual representation of information influences category learning is not well understood. Some researchers have directly addressed the correspondence between physical and psychological dimensions. One approach to ensure that experimenter assumptions are aligned with psychological reality is to approximate perceptual space using multidimensional similarity (MDS) models prior to category learning (Nosofsky, 1992; Shepard, 1980). While this approach avoids making the explicit assumption about the alignment between input and psychological representations, it makes the concept of 'dimensions' more difficult to define and, as a result, the nature of the psychological representations is not well understood. Others have more explicitly addressed the assumption of the alignment between physical and psychological dimensions, either by determining that the assumption is not problematic if the relationship between the input and perceptual dimensions is monotonic (Ashby & Gott, 1988) or by directly estimating the mapping between input and perceptual dimensions (Crossley & Ashby, 2015). While these approaches avoid the explicit assumption by computationally estimating the dimensions, it is not clear that this is applicable to all combinations of dimensions, as these researchers focused on relatively straightforward visual dimensions (e.g., orientation and width of a bar; line length and orientation).

Here, we test the influence of the alignment between input dimensions and psychological dimensions in an abstract dimensional space that could reflect multiple combinations of dimensions without the need to approximate the representation beforehand. We specifically capitalize on a dual systems approach to category learning, using distinctions between RB and II categories as a testbed for modeling the influence of perceptual representations on category learning. Because there is often an implicit assumption that the specific properties of experimenter-manipulated dimensions align with perceptual representations, there is the possibility that what are transparently 'rule-based' or 'information integration' distributions in input space may not be best described this way in terms of the underlying perceptual space. Therefore, understanding the nature of the categorization problem requires understanding perceptual representations.

We emphasize that the alignment between input dimensions and perceptual dimensions is likely to have a broad influence on category learning, in no way specific to the theoretical commitments of this dual systems theory (Ashby et al., 1998) or challenges to it (e.g., Kalish, Newell, & Dunn, 2017; Lewandowsky, Yang, Newell, & Kalish, 2012; Newell, Dunn, & Kalish, 2011). Our examination of this issue in the context of RB and II category learning from the dual systems perspective is a choice of convenience for the sake of its ease of exposition in our modeling efforts.

In the current investigation, we present a neural network model that demonstrates that so-called 'rule-based' and 'information-integration' categories in input space may not be reflected as such in perceptual space, and that this has dramatic consequences for category learning. We gave the model extensive, long-term experience with five kinds of structured regularities in a theoretical sensory environment, which it learned to reflect in its stable, 'adult-like' internal perceptual representations. We then examined how differences in the regularities the model experienced during this training phase influenced how the model learned identically structured 'rule-based' and 'information-integration' categories as defined in the input space. Our results demonstrate that the underlying perceptual representations developed across long-term experience place strong constraints on novel category learning. We also compare the model's behavior to human behavior from prior perceptual category learning studies across different sensory modalities.

2. Methods

Our approach involved two training phases. First, during the representation learning phase, we trained the model on a particular relationship within a two-dimensional space using an autoencoder. This training phase is meant to simulate a learners' lifetime of experience with a pair of sensory dimensions. Second, during the category learning phase, the model was subjected to a category learning experiment where this trained dimensional space was mapped to discrete category outputs. These two stages enable examination of how long-term experience shapes representations and, subsequently, how those representations influence category learning.

2.1. Model architecture

There are two components to the model architecture (Fig. 1): the lower level supports representation learning, in which perceptual representations are gradually shaped through extensive pre-experimental, task-independent experience that models long-term experience in the sensory world; the higher level supports category learning, in which the evoked representations of different stimuli are relatively rapidly associated with particular behavioral responses within an experimental context.

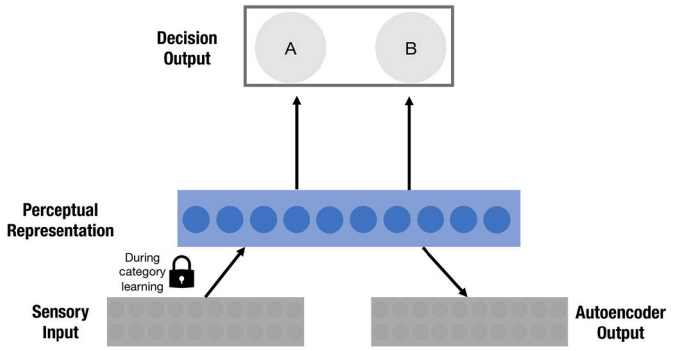

Fig. 1. Model architecture.

2.2. Representation learning

Our core assumption is that perceptual representations are tuned to capture the statistical structure of the ensemble of long-term perceptual experience, such that common features and feature combinations are coded in more detail than less common features and combinations. Although there are many ways of implementing this type of statistical learning, we adopted the approach of an *autoencoder* (Hinton, 1989), in which a neural network learns to reconstruct its inputs via one or more smaller, "bottleneck" layers of hidden units, because this allowed the same computational principles to apply to both representation and category learning.

Thus, in the model, representation learning over two physical input dimensions *x* and *y* was implemented by an autoencoder that received structured sensory input that it learned to recreate over an equal-sized output layer via a smaller single hidden layer (Fig. 1). Specifically, a 20 unit 'sensory input' layer was connected to a ten-unit 'perceptual representation' hidden layer which was connected to a 20 unit 'autoencoder output' layer. Ten of these 20 units reflected the physical *x*-dimension value and ten reflected the physical *y*-dimension value. For each dimension, a particular value was represented as a normalized Gaussian distribution centered on that value; the activation of the 10 units sampled this distribution uniformly over the full range of the dimension (such that their activations always summed to 1). This encoding allows for graded input, which reflects population encoding of information in sensory cortex. Activations in the sensory input layer also had a small amount of uniform noise (range = 0.1) to reflect a small amount of noise in the perceptual encoding of a stimulus. The goal of the network at this stage was to recreate the non-noisy input in the output layer. This training experience thus formed the perceptual representations of the network in the hidden layer. The number of input/output units and units in the hidden layer was determined based on our prior experience with these kinds of models. This was the only number of units that we implemented.

We trained the model on five separate training environments, reflective of different statistical relationships that might exist in the sensory world (Fig. 2): 1) no correlation or other relationship between two dimensions (Independent), 2) a positive relationship between two dimensions (Positive), 3) a negative relationship (Negative), 4) the *x*-dimension is represented in more detail than the *y*-dimension (X-Dimension), and 5) the *y*-dimension is represented in more detail than the *x*-dimension (Y-Dimension). To ensure that, in each condition, the model had experience within the entire space, the model also experienced inputs drawn from a uniform distribution (Fig. 2, gray points). The input of the model during training was biased such that 90% of the stimuli were drawn from the biased representation distribution (Independent, Positive, Negative, X-Dimension, Y-Dimension) and 10% were drawn from the uniform distribution. Table 1 shows the means, variance, and covariance of the representation training and uniform distributions.

These environments are not meant to capture any specific natural

Table 1
Representation learning distribution information.

Distribution Type	*M* (x, y)	σ^2 (x, y)	Covariance
Independent	0.5, 0.5	0.023, 0.023	0
Positive	0.5, 0.5	0.029, 0.029	0.028
Negative	0.5, 0.5	0.029, 0.029	−0.028
X-Dimension	0.5, 0.5	0.057, 0.00032	0
Y-Dimension	0.5, 0.5	0.00032, 0.057	0
Uniform	0.5, 0.5	0.10, 0.10	0

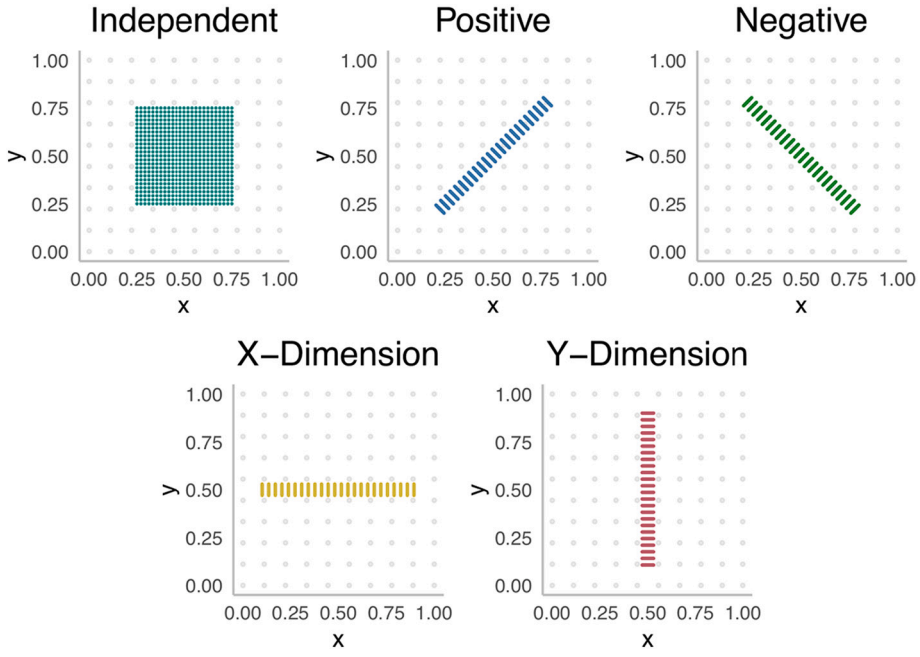

Fig. 2. Stimulus distributions for representation learning.
Note. Each biased distribution (colored points) also has the same uniform distribution (gray points).

signal statistics, but rather reflect clear alternative scenarios to demonstrate how these simple relationships might be encoded in the perceptual system and ultimately affect category learning. The representation training phase is meant to reflect long-term experience with statistical regularities in perceptual environments that amount to a lifetime of experience.

To simulate the gradual encoding of long-term statistical regularities into adult-like stable representations, we trained the network for 50,000 epochs of batch learning (i.e., all exemplars presented once before the model updates its weights) across the 624 stimuli within each training distribution (625 for Independent distribution), using back-propagation to minimize reconstruction error, with a learning rate of 0.0001, no momentum, and a bound of 1.0 on the length of the weight change vector. The hidden and output units in all parts of the network used a sigmoid activation function. These learning parameters are intentionally conservative and were chosen solely to ensure that representation learning was stable and effective.

2.3. Category learning

To simulate short-term training of novel category distinctions in an experimental context, in the category learning phase, the model weights from the sensory input layer to the hidden layer were frozen, reflecting a long-term consistency in experience and the resulting development of robust psychological representations (e.g., adult-like representations). To measure the network's categorization decision, a two-unit decision output layer was connected to the perceptual representation hidden layer (Fig. 1). The activation within these units reflects the model's choice between the two categories.

2.3.1. Category distributions

For each of the five representation environments, the model was separately trained on four category learning problems (Fig. 3A). The category distributions were created by sampling a bivariate Gaussian distribution using the *mvnorm* function in the MASS package in R (Venables & Ripley, 2002). We sampled for a single category using normalized coordinates (0–1) and then manipulated and rotated that distribution to create all other categories. Each of the category learning problems was identical in terms of statistical structure (category variance and overlap between categories; Table 2). The key difference is the rotation of the categories in physical input space, such that the category distinction requires different reliance on the physical input dimensions. These category environments were designed to reflect two rule-based (RB) problems that can be learned using a single input dimension (RB-X dimension, RB-Y dimension) and two information-integration (II) problems that require integration across the two dimensions (II-Positive and II-Negative). The naming scheme of the categories reflects the dimensions across which the categories can be distinguished (X-Dimension, Y-Dimension, Positive axis or Negative axis). For instance, learning RB-X categories requires learning that the categories can be distinguished based on the *x*-dimension and that the *y*-dimension is not informative of category membership. Critically, as a consequence of the representation learning phase, the input dimensions (i.e., the experimenter-defined dimensions) do not necessarily align with the model's internal perceptual representations.

Test stimuli were created using an identical procedure, sampling only 50 exemplars per category (Fig. 3B, Table 2). Due to the probabilistic nature of the sampling, the means and variances vary slightly but are very similar to the training distributions. The dimensions that are relevant for category identity are identical in the training and test environments.

2.3.2. Training procedure

We trained the category learning network with back-propagation using two distinct training paradigms. The first training paradigm – *batch* learning – like the representation learning paradigm, is conservative and stable in order to most clearly illustrate effects of representation on learning. This learning paradigm is meant to be a more abstracted version of the model's behavior to better understand the constraints of existing representations on the learnability of categories. This model is not meant to perfectly reflect human behavior or the way in which humans update their representations during learning. The second paradigm – *online* learning – is a closer approximation to experience in actual human category learning experiments, as the network updates its weights after each stimulus presentation. During category learning, weights in the representation network were held fixed, on the assumption that most experiments are too brief to substantially affect underlying perceptual representations. For all four category types (II-Negative, II-Positive, RB-X, and RB-Y), exemplars were presented randomly without replacement in training and test. Models were trained on all 200 stimuli from each category learning environment (100 stimuli per category) and tested on a separate set of 100 stimuli from each category environment (50 stimuli per category). We trained and tested 10 simulated subjects on each of the combinations of training paradigm (batch, online), representation distribution (Independent, Positive, Negative, X-Dimension, Y-Dimension), and category problem (RB-X, RB-Y, II-Positive, II-Negative) to get a sense of the variability in the behavior of the model. For both paradigms, after training, the model was tested on the 100 test stimuli while keeping the weights fixed (i.e., providing no feedback to the model). The hidden and output units in all parts of the network used a sigmoid activation function.

2.3.3. Training paradigm 1: batch learning

We trained the category learning network using a batch learning paradigm to understand the learnability of the categories with repeated exposures. All 200 category stimuli were presented to the model and then the model updated its connection weights using a learning rate of 0.01 and no momentum. For each simulated subject, the model was tested after each weight update (called an *epoch*).

2.3.4. Training paradigm 2: online learning

In separate runs, we trained the category learning network using an online learning paradigm to approximate human behavior during category learning, as the network updated its weights after each stimulus presentation. During online learning, the network was trained using a learning rate of 0.5 and no momentum. For each simulated subject, the model was tested after a single sweep through all 200 exemplars.

3. Results

3.1. Categorization accuracy

We present the results from batch and online learning together. We determined the categorization accuracy of the model by examining the percent of category exemplars for which each output activation was within 0.45 of its correct (target) activation of 0 or 1, assessed after each epoch of batch learning (Fig. 4A) and after presentation of all exemplars in online learning (Fig. 5A).[1] The results were similar across the two training methods (Table 3). There were specific patterns of accuracy for the different category problems that largely depended on the nature of the representation distribution. The data are available through the Open Science Framework repository at osf.io/w64nu (Roark, Plaut, & Holt, 2020).

[1] We used an activation criterion of 0.45 (rather than 0.5) to minimize spurious responding caused when activations are very close to 0.5. This provides a more conservative estimation of the model's category knowledge as accuracy would be near zero if the model was guessing. In general, performance below 50% indicates that the model failed to reliably learn the categories and performance can be reliably below 50% if activations of both category units for some stimuli fall within the range 0.45–0.55.

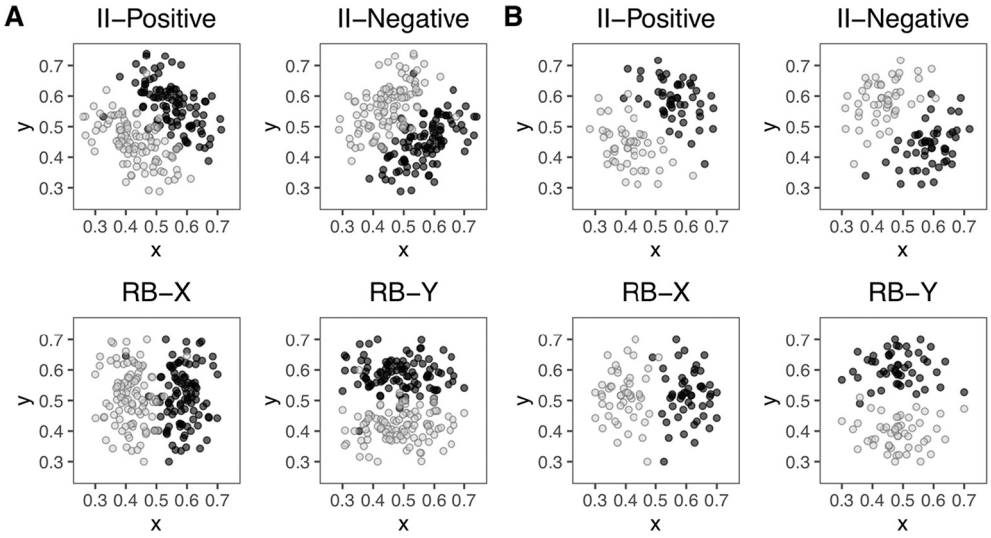

Fig. 3. Category input distributions.
Note. A. Training distributions. B. Test distributions. Individual points reflect members of different categories (black is Category A, gray is category B).

Table 2
Category distribution information.

	M (x, y)	σ^2 (x, y)	Covariance
	Training		
II-Negative: Category A	(0.57, 0.45)	(0.0059, 0.0053)	0.0030
II-Negative: Category B	(0.45, 0.57)	(0.0053, 0.0059)	0.0030
II-Positive: Category A	(0.55, 0.57)	(0.0053, 0.0059)	−0.0030
II-Positive: Category B	(0.43, 0.45)	(0.0059, 0.0053)	−0.0030
RB-X: Category A	(0.58, 0.51)	(0.0026, 0.0086)	0.00028
RB-X: Category B	(0.42, 0.51)	(0.0026, 0.0086)	−0.00028
RB-Y: Category A	(0.49, 0.58)	(0.0086, 0.0026)	−0.00028
RB-Y: Category B	(0.49, 0.42)	(0.0086, 0.0026)	0.00028
	Test		
II-Negative: Category A	(0.59, 0.44)	(0.0046, 0.0047)	0.0020
II-Negative: Category B	(0.44, 0.59)	(0.0047, 0.0046)	0.0020
II-Positive: Category A	(0.56, 0.59)	(0.0047, 0.0046)	−0.0020
II-Positive: Category B	(0.41, 0.44)	(0.0046, 0.0047)	−0.0020
RB-X: Category A	(0.62, 0.52)	(0.0026, 0.0066)	0
RB-X: Category B	(0.39, 0.52)	(0.0026, 0.0066)	0
RB-Y: Category A	(0.48, 0.61)	(0.0066, 0.0026)	0
RB-Y: Category B	(0.48, 0.39)	(0.0066, 0.0026)	0

For the Independent distribution, all four category types were learned quickly to a high degree of accuracy. During batch learning, the model demonstrated slightly higher accuracy for the RB categories than the II categories. In the final epoch of batch learning, accuracies were the following: II-Negative 82.3%, II-Positive 81.9%, RB-X 91.9%, and RB-Y 86.8%. The results were very similar after online learning with highest accuracy for the two RB categories (RB-X 94.5%, RB-Y 92.4%) and slightly lower accuracy for the two II categories (II-Negative 83.8%, II-Positive 88.7%).

For Positive and Negative distributions, the model learned one of the II categories very well and failed to learn the other. When the model was trained to represent a Negative relationship across the input dimensions, the II-Negative categories were learned very well (83.1% batch, 83.8% online), the RB-X and RB-Y categories were learned at an intermediate level (RB-X: 58.9% batch, 74.6% online; RB-Y: 58.2% batch, 72.6% online), and the II-Positive categories were learned very poorly (0% batch, 49.5% online). When the model was trained to represent a Positive relationship across the input dimensions, this pattern was reversed; the II-Positive categories were learned very well (83.1% batch, 92.7% online), the RB-X and RB-Y categories were learned at an intermediate level (RB-X: 61.3% batch, 70.7% online; RB-Y: 58.1% batch, 70.7% online), and the II-Negative categories were learned very poorly (0% batch, 41.6% online).

When the model was trained to represent the *x*-dimension or *y*-dimension in more detail, the patterns were similar. For X-Dimension representations, the RB-X categories were learned the best (88.3% batch, 90.9% online), the two II categories had intermediate accuracies (II-Negative: 61.1% batch, 62.3% online; II-Positive: 57.8% batch, 69.0% online), and the RB-Y categories were learned the worst (0% batch, 43.8% online). The pattern was reversed for the Y-Dimension representations. For the Y-Dimension representations, the RB-Y categories were learned the best (87.5% batch, 91.2% online), the two II categories had intermediate accuracies (II-Negative: 55.9% batch, 72.5% online; II-Positive: 60.6% batch, 67.2% online), and the RB-X categories were learned the worst (0% batch, 49.0% online).

In summary, for all distribution types, the category that was aligned with the long-term regularity experienced in the representation training phase was learned the best, the category that was misaligned with the long-term regularity was learned the worst, and the two other category types were learned at intermediate levels.

3.2. Real-valued error (Loss)

When examining the patterns of real-valued error produced by the model – sometimes termed *loss* – for different representation types and categories, the error patterns mirror the accuracy patterns, with higher accuracy reflected as low error (Figs. 4B, 5B). Overall, the model had lower error with more training.

For both batch and online training of the Independent distribution, the error was lower for the RB categories than the II categories. For Negative distribution, the II-Positive categories are difficult for the model and have the highest error, whereas the II-Negative categories have the lowest error. For the Positive distribution, the II-Negative categories are difficult for the model and have the highest error, whereas the II-Positive categories have the lowest error. For X-Dimension, RB-Y categories have the highest error rate and RB-X categories have the lowest. For Y-Dimension, RB-X categories have the highest error rate and RB-Y categories have the lowest.

3.3. Analysis of representations

To understand why the model was successful or failed at learning, it is useful to probe further into its behavior. We assessed the model's representations by examining the pattern of error in the uniform distribution from the representation learning phase and the categorization

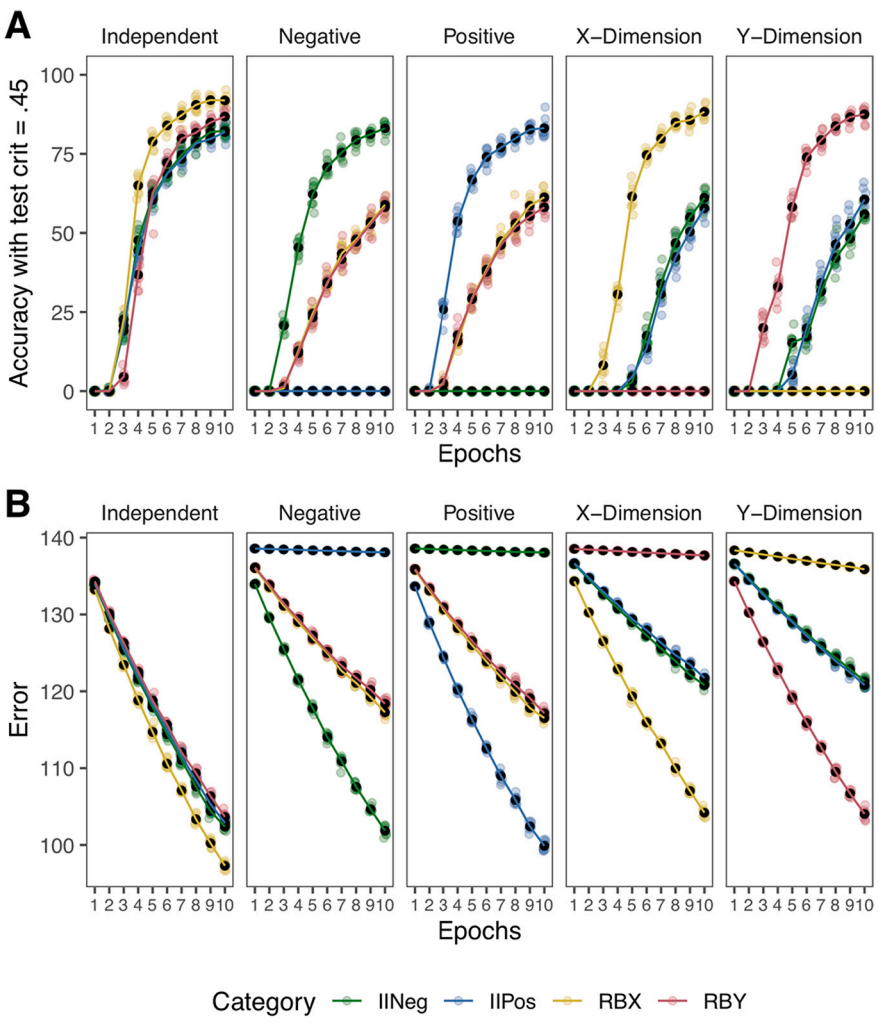

Fig. 4. Batch learning model accuracy and error (loss). Note. A. Model accuracy and B. Model error (loss) across epochs for batch training for the five distribution types and four category learning environments. Individual runs of the model are shown as colored points, the mean performance is shown as a black point, and the error bars reflect SEM. Note that performance can be reliably below 50% if activations of both category units for some stimuli fall within the range 0.45–0.55.

response behavior for different stimuli.

3.3.1. *Representation learning*

First, it is useful to confirm that the model learned the distributions over its long-term experience (e.g., 50,000 epochs) during the representation learning phase. We tested the model on all training stimuli and plotted the pattern of reconstruction error for all stimuli from the uniform distribution that accompanied each of the representation distributions (Fig. 6). These patterns demonstrate that the network clearly learned the distribution with which it had most experience, having the lowest error in areas that it had the most experience and higher error elsewhere. Critically, the error patterns are specific to the nature of the bias in the representation distribution.

3.3.2. *Category learning*

Next, we asked more specifically about what the model learned about the stimulus space during category learning. As specified in the methods section, we tested the model after all 200 stimuli were presented in online learning. Here, we examine the stimulus-specific response patterns for each different representation distribution (Independent, Positive, Negative, X-Dimension, Y-Dimension) and category type (II-Positive, II-Negative, RB-X, RB-Y).

The patterns of responses across stimuli after online learning demonstrate that the long-term experience of the model influenced the way that it learned the categorization tasks (Fig. 7). For the Independent distribution, the model learned to separate the categories quite well, especially when the stimuli were far from the boundary between categories. The model was especially confused for these boundary stimuli for the II-Negative and II-Positive categories (responding with proportion of category A responses near 0.5, indicating that the model made an equal number A and B responses).

For the rest of the distributions, as discussed in the category learning results section, there was a clear benefit for the category distinction that aligns with the representation distribution (Negative + II-Negative, Positive + II-Positive, X-Dimension + RB-X, Y-Dimension + RB-Y). The categories that were extremely difficult for the model to learn were orthogonal to the representation distribution (Negative + II-Positive, Positive + II-Negative, X-Dimension + RB-Y, Y-Dimension + RB-X).

Interestingly, the bias created by a specific representation training distribution was also evident in the response pattern in the categories that were neither aligned nor orthogonal to that distribution. Take the X-Dimension distribution, for instance. The model separated the RB-X categories along the *x*-dimension, as would be expected if the model was responding optimally. The model also demonstrated this same *x*-dimension bias in responding for the II-Negative and II-Positive categories. That is, even though these categories require separation along both dimensions (minor or major axes), the model responded with an *x*-dimension bias. This led to intermediate accuracy for these categories because this strategy, while suboptimal, sometimes aligns with feedback leading to an intermediate level of category learning. In direct contrast, the orthogonal category (RB-Y) is learned very poorly because the prior experience results in perceptual representations that largely collapse this dimension, and thus, it cannot be used to separate the categories.

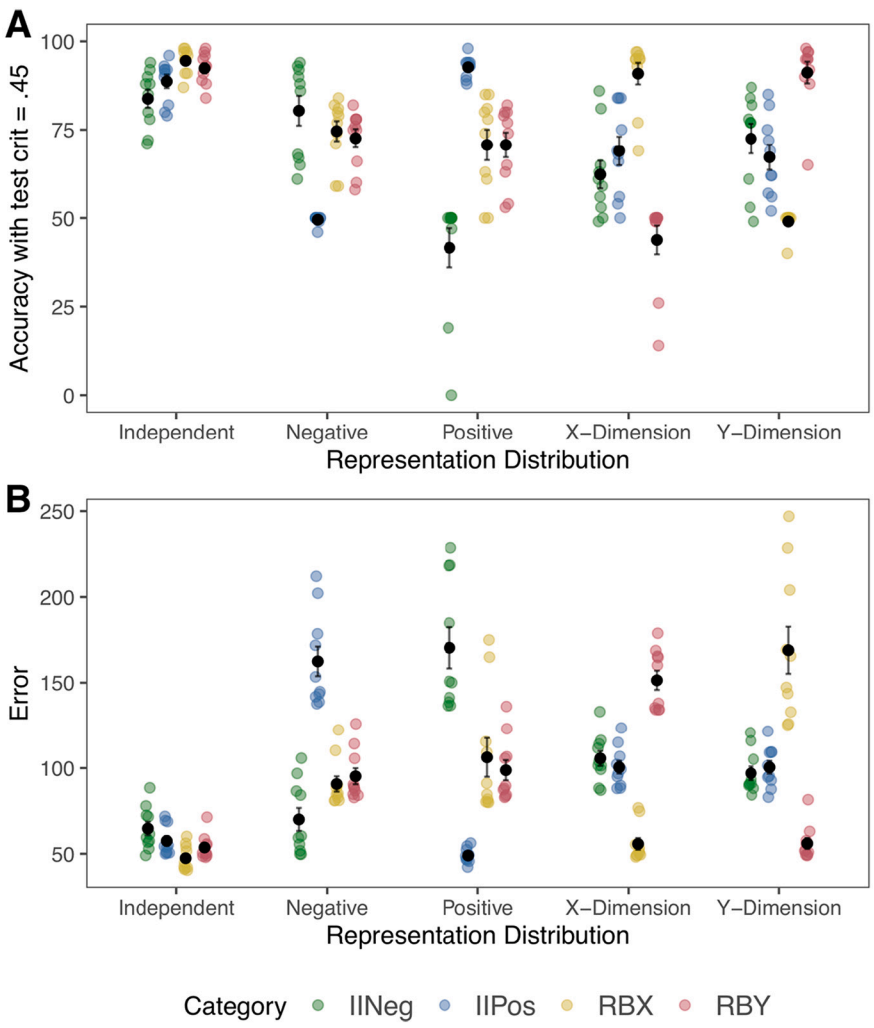

Fig. 5. Online learning model accuracy and error (loss). Note. A. Model accuracy and B. Model error (loss) after the single pass through all 200 category exemplars for online learning for the four distribution types and four category learning environments. Individual runs of the model are shown as colored points, the mean performance is shown as a black point, and the error bars reflect *SEM*.

Table 3
Accuracy results across training methods

Representation distribution	II-Negative	II-Positive	RB-X	RB-Y
		Batch Learning		
Independent	82.3 [81.3, 83.3]	81.9 [80.3, 83.5]	91.9 [90.9, 92.9]	86.8 [85.5, 88.1]
Negative	83.1 [81.8, 84.5]	0 [n/a, n/a]	58.9 [57.5, 60.3]	58.2 [56.5, 59.9]
Positive	0 [n/a, n/a]	83.1 [81.0, 85.2]	61.3 [59.5, 63.1]	58.1 [56.7, 59.5]
X-Dimension	61.1 [59.2, 63.0]	57.8 [56.0, 59.6]	88.3 [87.0, 89.6]	0 [n/a, n/a]
Y-Dimension	55.9 [54.7, 57.1]	60.6 [58.3, 62.9]	0 [n/a, n/a]	87.5 [86.0, 89.0]
		Online Learning		
Independent	83.8 [78.0, 89.6]	88.7 [84.4, 93.0]	94.5 [91.9, 97.1]	92.4 [89.2, 95.6]
Negative	80.4 [70.9, 90.0]	49.5 [48.6, 50.4]	74.6 [68.1, 81.0]	72.6 [66.7, 78.5]
Positive	41.6 [29.1, 54.1]	92.7 [90.6, 94.8]	70.7 [60.9, 80.5]	70.7 [62.8, 78.6]
X-Dimension	62.3 [53.4, 71.2]	69.0 [59.8, 78.2]	90.9 [84.0, 97.8]	43.8 [34.6, 53.0]
Y-Dimension	72.5 [63.0, 82.0]	67.2 [59.2, 75.2]	49.0 [46.7, 51.3]	91.2 [84.2, 98.2]

Note. Mean accuracy with 95% confidence intervals across ten simulated subjects for batch learning (after final epoch) and online learning.

There are several relevant patterns in these results. First, categories that are orthogonal with the distributional representations show at-chance performance (Fig. 7, values near 0.50 shown in white). Second, when trained on the II categories with X-Dimension or Y-Dimension representations, the model separates the categories based on a single dimension, instead of two. For example, the II-Negative/X-Dimension panel demonstrates that the categories are separated based on the *x*-dimension. Similarly, when trained on the RB categories with Positive or Negative representations, the model separates the categories based on two dimensions, instead of one. For example, the RB-X/Negative panel demonstrates that the categories are separated based on both *x* and *y* dimensions across the negative axis. These results demonstrate that the model does not struggle to learn in the same way across different types of categories. Instead, the reason that the model struggles is directly related to how the representation distribution relates to the categories being learned. The model is struggling because it is applying its representational bias during category learning and this bias cannot be overcome based on the feedback received during category learning.

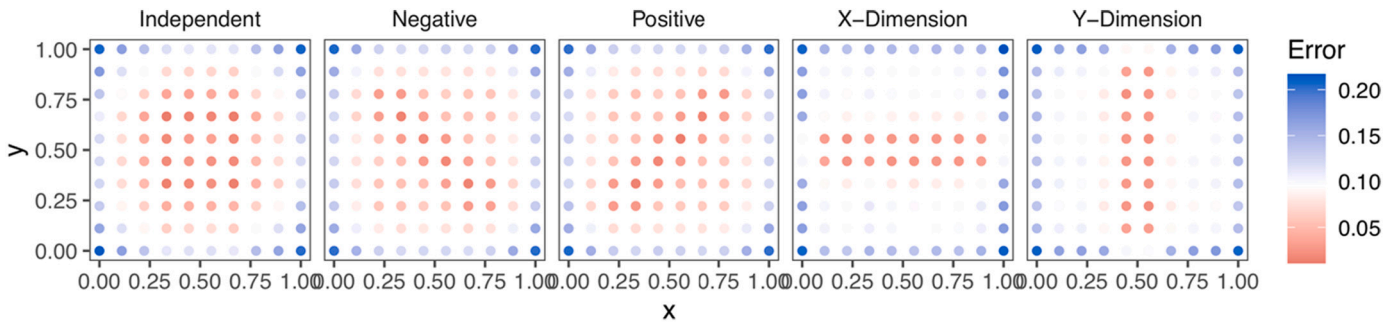

Fig. 6. Error in reconstructing the uniform distribution after representation learning.

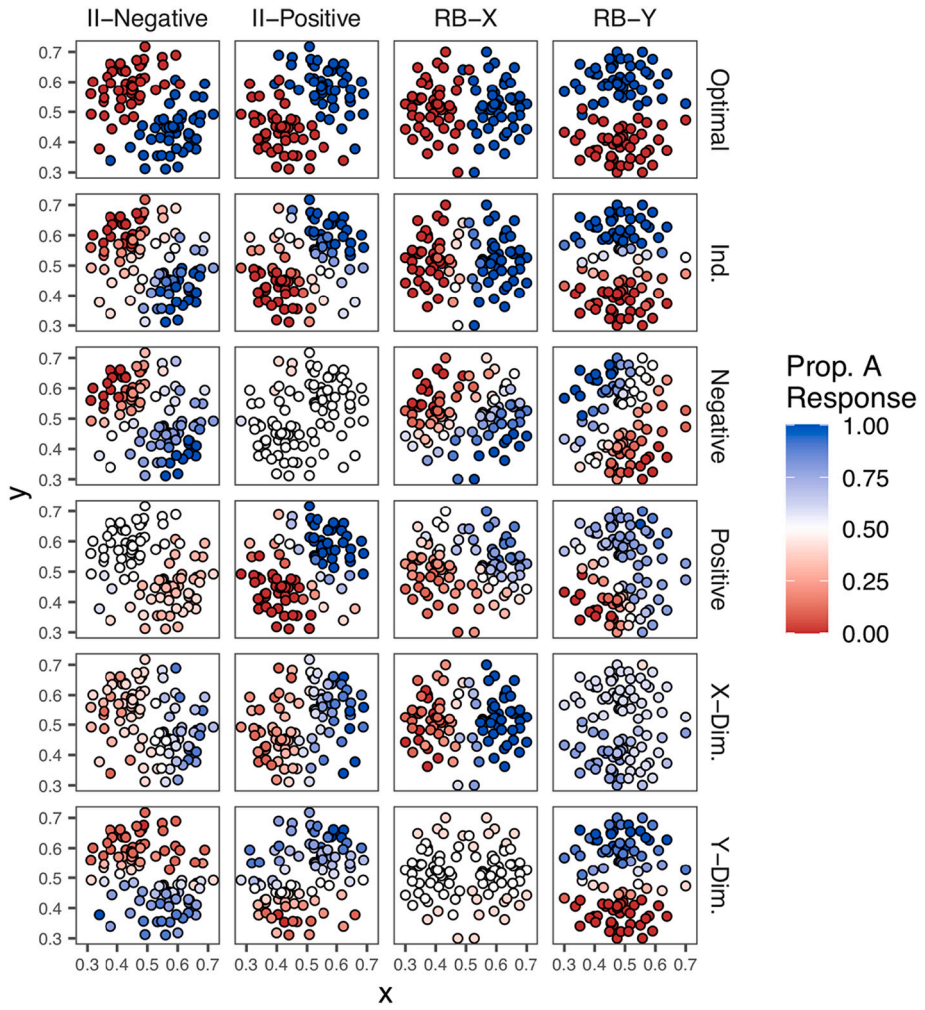

Fig. 7. Model category responses in test after online learning.
Note. The values reflect the proportion of category A responses for each stimulus. The Optimal row reflects the ground-truth category identities for reference.

3.4. Summary and interpretation of results

We investigated the impact of long-term experience with different statistical regularities that result in different perceptual representations on category learning. We observed specific patterns of results depending on the type of category being learned and the nature of learned perceptual representations.

These results support a theoretical framework (Fig. 8) that demonstrates that perceptual representations and category distribution structure interact to affect learning outcomes. Specifically, long-term experience with some statistical regularity (e.g., a negative or positive correlation, Fig. 8A) results in enhanced representation along the axis of high variability in experience and reduced representation along the orthogonal axis (Fig. 8B). As a result of this experience, categories that are statistically identical in input space (Fig. 8C) are not identical in perceptual space (Fig. 8D). Categories that *align* with perceptual representations (here, negative correlation, Fig. 8 top) are easier to learn than categories that *conflict* with perceptual representations (here, positive correlation, Fig. 8 bottom).

Our results support this framework. When two input dimensions were Independent, the model learned to accurately represent the value on each dimension regardless of the value on the other – that is,

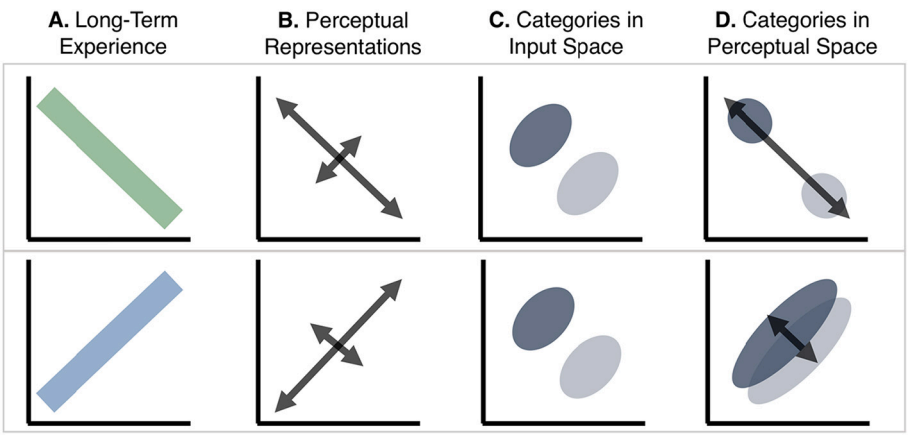

Fig. 8. Theoretical framework.
Note. A. The theoretical framework is demonstrated with the example of long-term experience with a negative (top) or positive (bottom) correlation between two dimensions. B. This experience stretches perceptual representations along the axis of experience and shrinks representation along the orthogonal axis. C-D. As a result, categories that are statistically identical in input space are no longer identical in perceptual space. When the category distinction is aligned with experience (negative axis, top), categories are easily distinguishable. When the distinction conflicts with experience (positive axis, bottom), categories are more difficult to distinguish.

independently – potentially by devoting separate hidden units to each dimension. As a result, we found that the model demonstrated a bias for learning RB over II categories because, for the former, category learning can easily learn the relevant weights from only the relevant subset of hidden units, analogous to "selective attention" to that dimension. By contrast, II category learning requires sensitivity the precise relationship between the two dimensions, which was more difficult to learn. This pattern is consistent with previous findings that generally demonstrate an advantage for RB over II categories for simple visual dimensions (Ashby & Maddox, 2011).

When the two sensory dimensions are interdependent (Positive or Negative distributions), the learnability of categories depended on the alignment with the representations. Categories were easier to learn when they required a distinction along the axis that was more strongly reflected in learned representations (i.e., the axis of high variance). For example, because the model experienced high variability along the Negative axis for the Negative training distribution, the model was better at learning categories that can be distinguished along this axis, leading to better performance for II-Negative categories than II-Positive categories. We found the reverse pattern for the Positive representation distribution. In both cases, the RB categories were learned at intermediate levels.

Finally, when the representation training distributions reflected enhanced encoding of one dimension relative to the other (X-Dimension and Y-Dimension distributions), we found differences in how well the RB categories were learned. As with the other training environments, we found that when the model has experience with higher variance along one dimension, it more faithfully represented this dimension in the perceptual representation hidden layer. As a result, we found that after training with high variability on the X-Dimension, RB-X categories were easier for the model to learn than RB-Y categories, with performance for the II categories at intermediate levels. We observed the opposite pattern for the Y-Dimension representations.

Further, examining the pattern of responses to the category stimuli revealed that the reason the model succeeded or struggled to learn was because it applied its representational bias during category learning. For example, when the model was trained on the X-Dimension distribution, the pattern of responses for the RB-X category and the two II categories demonstrated that the model was using the *x*-dimension to separate the categories. The model failed to learn the categories that were completely orthogonal to their representation (e.g., RB-Y categories for the X-Dimension distribution).

In sum, these results demonstrate the potential for existing perceptual representations to impact category learning, especially when the physical dimensions or experimenter-defined dimensions do not align with the dimensions of representations. In general, having extensive experience with variation along a dimension makes it easier to distinguish categories that vary along that dimension and more difficult to distinguish categories orthogonal to that dimension.

4. Comparison with human behavior

While the model's behavior can be explained by the theoretical framework, it is not yet clear how this relates to human behavior. In this section, we compare the model's behavior and human behavior across prior studies of category learning in multiple sensory modalities. When we can observe the pattern of accuracy in humans across several category learning types in the same sensory space, we are able to draw conclusions about the nature of human perceptual representations across particular dimension pairs. This kind of comparison is especially useful in cases in which the underlying cognitive or neural representations of dimensions are not well understood, as with complex auditory dimensions.

We searched the literature for category learning experiments that examine these four kinds of category distributions – unidimensional RB along both dimensions and II distributions with categories distinguished along the positive axis and the negative axis. Very few studies make a complete comparison of these four category types. It is more typical for experimenters to choose one RB distribution and one II distribution to compare. However, several studies have trained participants on all four types – one experiment with auditory dimensions (Roark & Holt, 2019), one experiment with visual dimensions (Ell, Ashby, & Hutchinson, 2012), and two experiments using cross-modal stimuli with auditory and visual dimensions (Smith et al., 2014). We will compare the model's behavior to the human behavior in each of these experiments. We should note at the outset, though, that quantitative aspects of the input distributions to which the model was exposed were designed to illustrate the impact of these distributions and are unlikely to match the relationships among actual real-world dimensions precisely.

4.1. *Roark and Holt (2019): Auditory dimensions*

In Roark and Holt (2019), participants learned categories based on the auditory dimensions of center frequency (CF) and modulation frequency (MF) of nonspeech tones. As in the simulations, they trained participants on four category problems – RB-CF, RB-MF, II-Positive, or II-Negative with feedback (four blocks of 96 trials each).[2]

[2] The terminology of II-Positive and II-Negative in the current manuscript reflects the axis that is important for distinguishing the categories. Terminology of the previous studies discussed here (Ell et al., 2012; Roark & Holt, 2019; Smith et al., 2014) are all based on the axis of an optimal decision boundary that separates the categories. As such, we have relabeled the II-Positive and II-Negative categories when discussing these three studies to match the terminology of the current manuscript.

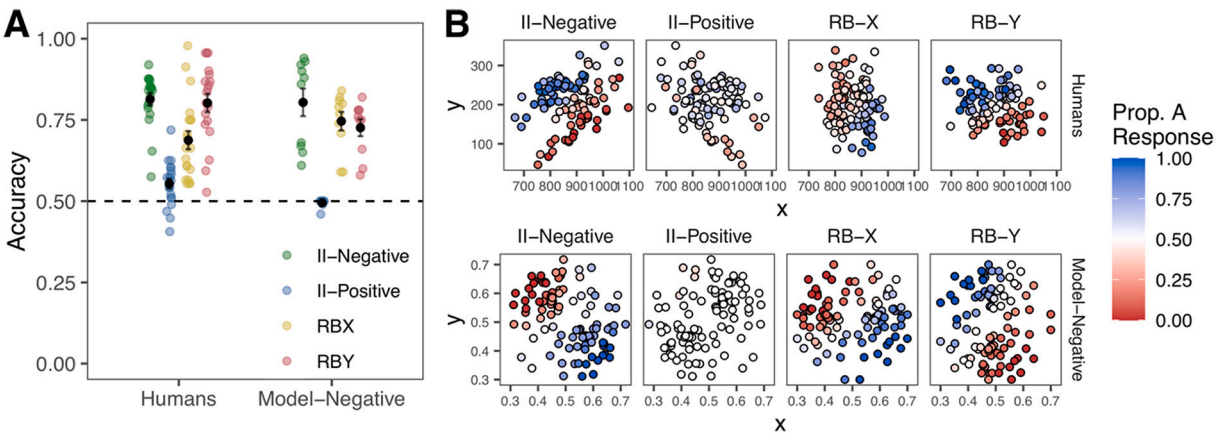

Fig. 9. Comparison of human and model behavior.
Note. A. Human categorization performance in the generalization test in Roark and Holt (2019) compared with model performance after extensive training with the Negative distribution. Individual points reflect individual subjects or simulation runs with the means in black. Error bars reflect *SEM*. B. Heat map of proportion of category A responses for each category distribution from Roark and Holt (2019) participants (top) and the corresponding test distributions for the model after training with the Negative distribution (bottom).

Roark and Holt (2019) found that the category problems with the highest accuracy were the II-Negative and RB-MF, with RB-CF learned at more moderate levels, and II-Positive learned at the lowest levels (Fig. 9A). This overall pattern most closely aligns with the model's behavior for the Negative distribution, indicating that these acoustic dimensions may have a representation that reflects a long-term negative relationship between CF and MF. Further, the model's response behavior for the Negative distribution is similar to human performance (Fig. 9B). The model and human participants excel at separating the II-Negative categories, fail to separate the II-Positive categories reliably, and demonstrate a bias to separate the RB-X and RB-Y categories in a way that reflects usage of *both* dimensions, rather than just one. This pattern of human learning was not predicted by an existing literature that has focused more on whether categories require one or multiple dimensions and could be classified as 'RB' or 'II' categories. The pattern of human categorization accuracy is consistent with the model's behavior and illustrates that the nature of perceptual representations influences learning outcomes.

4.2. *Ell et al. (2012): Visual dimensions*

In Ell et al. (2012) Experiment 2, participants learned categories based on the visual dimensions of saturation and brightness, two of the defining features of color perception. As in our simulations, they trained participants on four category problems – RB-Saturation, RB-Brightness, II-Positive, or II-Negative with feedback (nine blocks of 80 trials each).

By the end of training, participants performed similarly on all four category learning problems. However, there were differences in early learning which may give clues about which category distinctions are better in alignment with the way humans represent the visual dimensions. In the first block, RB-Brightness had higher accuracy than RB-Saturation and II-Positive but was not significantly different from II-Negative. None of the other comparisons were statistically different, but there were few subjects in each condition, and this was not the main comparison of interest to these authors. However, the general pattern in which one RB category is learned better than another aligns with the model's behavior for the X-Dimension or Y-Dimension distributions. Therefore, this may reflect a situation where brightness may have a more veridical or detailed representation relative to saturation. Although examination of the visualization of the data from Ell et al. (2012) indicates that there may be some differences among the four category problems, the statistical analyses do not indicate a difference. It would be necessary to examine this same kind of category learning with a larger sample to truly understand the nature of the representation of these dimensions.

Additionally, whereas in the current set of simulations, the performance for the worst-performing category problem is around chance levels, participants in Ell et al. (2012) were able to learn all four category problems to a similar extent by the end of 720 trials. As mentioned, the current model simulations used fairly extreme training distributions to demonstrate a first-pass confirmation that the nature of the representations can have a strong impact on learning outcomes. However, it is likely the case that to match human behavior and representations more closely, the training distributions would need to be less extreme.

4.3. *Smith et al. (2014): Cross-modal dimensions*

In Smith et al. (2014) Experiments 1 and 2, participants learned categories with one visual and one auditory dimension. The dimensions varied across the two experiments, but the results are very similar, so we discuss them together. The auditory dimension was duration of three 100 Hz tones in Experiment 1 and frequency of a pure tone in Experiment 2. The visual dimension was pixel density in both experiments.

Because the purpose of these experiments was not to compare accuracy of the two RB and two II tasks, Smith et al. (2014) did not compare accuracy across the four tasks. Instead, their goal was to contrast RB and II category learning and so they compared the average accuracies for the two RB tasks to the average accuracies for the two II tasks. This comparison stems from their investigation into the differences between RB and II category learning but distorts our ability to compare the statistical outcomes to the current set of model simulations.

However, we can observe the pattern in the reported means from their experiments to assess the descriptive pattern of results within the four category learning problems. These descriptive results indicate that for Experiment 1, the two RB problems are learned better than the two II problems (RB-Auditory: 91.1%, RB-Visual: 94.5%, II-Negative: 74.6%, II-Positive: 74.3%), which aligns with the model's behavior with the Independent distribution, reflecting a situation where the two sensory dimensions are encoded independently. This pattern may be expected because it is likely that cross-modal dimensions are encoded in distinct and separate sensory representations.

In contrast, in Experiment 2, there was slightly higher accuracy for the RB-Auditory problem compared to the RB-Visual problem (88.5% accuracy compared to 77.8%). However, performance on each was better than for the two II problems (II-Negative: 68.2%, II-Positive: 70.0%). This exact pattern is not represented directly in the model's behavior. However, it is still mostly aligned with a version of Independent representations in which one dimension might be represented

slightly more faithfully than the other dimension or perhaps a hybrid between Independent and X- or Y-Dimension representations. Though there are some limitations in our ability to compare the effects to the model behavior directly, it seems reasonable that one of these dimensions may be more salient than the other, which may have influenced learning outcomes.

5. Discussion

The current set of simulations demonstrates that the nature of long-term experience in a sensory environment can shape the representations of input dimensions in a way that, in turn, drastically impacts category learning behavior. Depending on the nature of the representations that are shaped by experience, some category learning problems are easily learnable, whereas others are more difficult. The simulation results demonstrate that it is critical to consider the constraints that the perceptual system and existing representations place on learning to understand the mechanisms of perceptual category learning. The nature of the learning problem may differ substantially depending on the perceptual representations across the very same input dimensions.

As with all models, the current model incorporates specific assumptions. We address our assumptions about the input training space and training paradigm and discuss potential implications for the interpretation of the results. First, our model used relatively simple and somewhat extreme training spaces that are clearly highly abstract relative to the way sensory information is distributed in the real world. While there was a small amount of noise in the input to the model to reflect modest perceptual noise in the encoding process, there was no noise in the actual distributions. Future elaboration of this model should include a simulation of the kind of variability and noise that exists in real-world sensory environments. Additionally, the model was applied to a two-dimensional input space. The world beyond simple experiments has many more dimensions, some of which are relevant whereas others are irrelevant for category distinctions. A future version of this model should seek to understand how multiple dimensions may be represented independently and, in conjunction, what the effects on higher-level cognition might be.

Our approach also involved freezing the hidden layer weights during category learning such that no changes could be made to the weights due to category learning. This design reflects a situation where the representations of sensory dimensions are not changed with additional short-term experience during category learning. It is possible that representations could continue to change as a result of category learning. Prior work has demonstrated that categorization training can affect representations in many ways, including by creating new dimensions (Goldstone, Lippa, & Shiffrin, 2001), increasing discriminability across category-relevant dimensions (Feldman, 2021; Schyns, Goldstone, & Thibaut, 1998), decreasing discriminability of within-category distinctions (Goldstone, 1998), and affecting neural representations of dimensions at early levels of processing (Ester, Sprague, & Serences, 2019). Other work suggests that in the presence of long-term perceptual biases, like those created in the representation learning phase, short-term experience may not substantially affect existing representations (Roark & Holt, 2020). It is important to acknowledge that experience is a continuous cycle and that a lifetime or a single experiment may influence our representations of the sensory world. Future work should focus on the interaction of long-term and short-term regularities and clarify how and when representations change with experience.

Finally, the current model used an autoencoder and trained representations to reflect sensory regularities based on the distributional statistics alone (i.e., without feedback). Such an approach does not fully reflect the complexities of human learning or sensory experience. Future work might expand the model to compare representation training methods. We suspect that representation training with feedback may impact subsequent category learning behavior even more strongly than with the self-supervised autoencoder paradigm used here.

Indeed, the relative importance of unsupervised versus supervised learning in human category learning is an area of active debate. Some have suggested that sensory regularities experienced across the long-term may be learned via unsupervised learning mechanisms, as modeled in the representation learning phase (Frost, Armstrong, & Christiansen, 2019; Saffran & Kirkham, 2018), and that speech perception and speech category learning can be modeled with similar unsupervised approaches as those used here (i.e., autoencoder; Elman & Zipser, 1988; Nixon & Tomaschek, 2021; Getz, Nordeen, Vrabic, & Toscano, 2017; Toscano & McMurray, 2010). However, others have suggested that learning complex categories (e.g., speech categories) may necessarily involve feedback of some sort and may not be possible with passive exposure to statistical distributions alone (Feldman, Griffiths, Goldwater, & Morgan, 2013; Lim, Fiez, & Holt, 2019; Nixon, 2020). Recent work has also examined how individuals learn about category structures through a combination of unsupervised and supervised training within the same experimental session (Bröker, Love, & Dayan, 2021). The results of these combination studies support our current framework and demonstrate that understanding the challenge presented to the learner (whether unsupervised or supervised) requires understanding of the alignment of underlying representations and the task defined by the experimenter.

Here, the extensive representation training (50,000 epochs) was designed to approximate a lifetime of human experience that may or may not align with the requirements of a short-term environment. The influence of existing representations on learning is a major focus of the speech and language learning fields (Best, 1995; Iverson & Kuhl, 1995; Scharinger et al., 2013). Specifically, theories demonstrate that the extent of conflict between one's native language categories and novel second language categories determine how difficult those categories are to learn (Best, 1995). When there is little or no conflict (e.g., Zulu click categories for native English listeners), learning proceeds quickly and effortlessly (Best, McRoberts, & Sithole, 1988). When there is high conflict (e.g., English /r/−/l/ categories for native Japanese listeners), learning is difficult (Lotto, Sato, & Diehl, 2004). The current framework provides insight about why this difference exists – long-term experience with a native language enhances representation of dimensions that are relevant to that experience and diminishes representation of dimensions that are irrelevant. The resulting effect is that input categories that align with learners' existing representations, maximizing distinctions that need to be made, are readily learned and input categories that are orthogonal to those representations are difficult to learn.

The influence of the psychological representations of dimensions on perception and learning was also a focus of earlier work (Garner, 1974; Kemler Nelson, 1993; Kemler & Smith, 1979; Melara & Marks, 1990). While some dimensions are represented independently leading to enhanced selective attention to those dimensions (i.e., separable dimensions), others have interdependent representations making selective attention more difficult (i.e., integral dimensions). Because of these underlying differences, learning categories that require selective attention to the underlying dimensions proceeds easily with separable dimensions and is more difficult with integral dimensions, whereas learning categories that require integration across dimensions proceeds easily with integral, but not separable dimensions (Ell et al., 2012; Garner, 1976; Maddox & Dodd, 2003; Roark & Holt, 2019). This prior work demonstrates the utility of the current framework in understanding how long-term sensory experience may support specific psychological representations that reflect that experience and subsequently

influence category learning based on the physical input dimensions.

Despite its limitations, the current investigation provides valuable insight into the influence of learned perceptual representations on category learning and provides proof-of-concept evidence that the category structure in input space (i.e., rule-based or information-integration) may not be the key determiner in understanding categorization. Instead, these results demonstrate that it is imperative to understand the nature of the perceptual representations of the dimensions involved to understand the problem for the learner. While much of the research on human perceptual category learning has used simple, verbalizable dimensions that are likely represented independently both neurally and in cognitive representations, it is a much more difficult problem to understand what happens when perception is not so straightforward.

Our framework challenges a typical assumption made by experimenters testing theories of category learning – that experimenter-defined dimensions are aligned with participants' psychological representations. If there is a misalignment between these concepts of dimensions, then what may appear to the experimenter to be a 'rule-based' problem may not actually be 'rule-based' for the perceptual system. Our intention is not to explain or differentiate rule-based and information-integration learning problems or to directly contribute to the vast literature that attempts to explain how categories with different structures may be learned by a single system or separate systems (e.g., Ashby & Maddox, 2011; Newell et al., 2011). Instead, we argue that defining the categorization problem based on experimenter-defined dimensions does not capture the true complexity of the problem for the human perceptual system. As such, this framework has implications for understanding category learning more generally, beyond the distinction between rule-based and information-integration categories. The nature of psychological representations of dimensions developed across long-term experience has implications for a wide variety of theories of category learning. Understanding the nature of psychological dimensions has implications for interpreting which dimensions are attentionally weighted in exemplar models of categorization (e.g., Francis & Nusbaum, 2002; Nosofsky, 1986) or interpreting similarity in representations in clustering models of learning (e.g., Love, Medin, & Gureckis, 2004). It is important to note that other models like SUSTAIN (Love et al., 2004) capture the statistical structure of the input through other methods of recoding (e.g., cluster representations compared to continuous representations in the current model).

In general, the perceptual component of perceptual category learning has drifted out of focus of current theories of learning. The current set of simulations demonstrates that psychological representations of the sensory world, shaped by long-term experience, can strongly influence the nature of the problem for the learner.

Funding

This work was supported by the National Institutes of Health [T32GM081760, F32DC018979 to C. L. R.] and National Science Foundation [NSF BCS 1950054 to L. L. H.].

Declarations of interest

None.

Appendix B. Supplementary data

The data are available through the Open Science Framework repository at https://osf.io/w64nu/ (Roark et al., 2020). This archive includes the raw data (error and accuracy) for both batch and online learning for all distribution types (Independent, Negative, Positive, X-Dimension, Y-Dimension) and categories (II-Negative, II-Positive, RB-X, RB-Y). The archive also contains the error data for the uniform representation training distribution (Fig. 6) and the proportion category A response data for category learning (Fig. 7). Supplementary data to this article can be found online at [https://doi.org/10.1016/j.cognition.2021.104997].

References

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. https://doi.org/10.1037/0033-295x.105.3.442

Ashby, F. G., & Ell, S. W. (2001). The neurobiology of human category learning. *Trends in Cognitive Sciences, 5*(5), 204–210. https://doi.org/10.1016/s1364-6613(00)01624-7

Ashby, F. G., & Gott, R. E. (1988). Decision rules in the perception and categorization of multidimensional stimuli. *Journal of Experimental Psychology: Learning, Memory, and Cognition, 14*(1), 33–53. https://doi.org/10.1037//0278-7393.14.1.33

Ashby, F. G., & Maddox, W. T. (2011). Human category learning 2.0. *Annals of the New York Academy of Sciences, 1224*, 147–161. https://doi.org/10.1111/j.1749-6632.2010.05874.x

Ashby, F. G., & Soto, F. A. (2015). Multidimensional signal detection theory. In *The Oxford handbook of computational and mathematical psychology* (pp. 13–34).

Ashby, F. G., & Waldron, E. M. (1999). On the nature of implicit categorization. *Psychonomic Bulletin & Review, 6*(3), 363–378. https://doi.org/10.3758/bf03210826

Ashby, F. G., & Waldron, E. M. (2000). The neuropsychological bases of category learning. *Current Directions in Psychological Science, 9*(1), 10–14. https://doi.org/10.1111/1467-8721.00049

Aslin, R. N., & Newport, E. L. (2014). Distributional language learning: Mechanisms and models of category formation. *Language Learning, 64*(September), 86–105. https://doi.org/10.1111/lang.12074

Best, C. T. (1995). A direct realist view of cross-language speech perception. In *Speech perception and linguistic experience: Issues in cross-language research*.

Best, C. T., McRoberts, G. W., & Sithole, N. M. (1988). Examination of perceptual reorganization for nonnative speech contrasts: Zulu click discrimination by english-speaking adults and infants. *Journal of Experimental Psychology: Human Perception and Performance, 14*(3), 345–360. https://doi.org/10.1037/0096-1523.14.3.345

Bröker, F., Love, B. C., & Dayan, P. (2021). When unsupervised training benefits category learning. *PsyArXiv.*. https://doi.org/10.31234/osf.io/k5pzu

Burns, B., & Shepp, B. E. (1988). Dimensional interactions and the structure of psychological space: The representation of hue, saturation, and brightness. *Perception & Psychophysics, 43*, 494–507. https://doi.org/10.3758/bf03207885

Carvalho, P. F., Chen, C., & Yu, C. (2021). The distributional properties of exemplars affect category learning and generalization. *Scientific Reports, 11*(1), 11263. https://doi.org/10.1038/s41598-021-90743-0

Crossley, M. J., & Ashby, F. G. (2015). Procedural learning during declarative control. *Journal of Experimental Psychology: Learning, Memory, and Cognition, 41*(5), 1388–1403. https://doi.org/10.1037/a0038853

Drucker, D. M., Kerr, W. T., & Aguirre, G. K. (2009). Distinguishing conjoint and independent neural tuning for stimulus features with fMRI adaptation. *Journal of Neurophysiology, 101*(6), 3310–3324. https://doi.org/10.1152/jn.91306.2008

Eimas, P. D. (1975). Auditory and phonetic coding of the cues for speech: Discrimination of the [r-l] distinction by young infants. *Perception & Psychophysics, 18*(5), 341–347. https://doi.org/10.3758/bf03211210

Ell, S. W., Ashby, F. G., & Hutchinson, S. (2012). Unsupervised category learning with integral-dimension stimuli. *The Quarterly Journal of Experimental Psychology, 65*(8), 1537–1562. https://doi.org/10.1080/17470218.2012.658821

Elman, J. L., & Zipser, D. (1988). Learning the hidden structure of speech. *The Journal of the Acoustical Society of America, 83*(4), 1615–1626. https://doi.org/10.1121/1.395916

Ester, E. F., Sprague, T. C., & Serences, J. T. (2019). Categorical biases in human occipitoparietal cortex. *The Journal of Neuroscience, 40*(4), 917–931. https://doi.org/10.1523/jneurosci.2700-19.2019

Everson, R. M., Prashanth, A. K., Gabbay, M., Knight, B. W., Sirovich, L., & Kaplan, E. (1998). Representation of spatial frequency and orientation in the visual cortex. *Proceedings of the National Academy of Sciences, 95*(14), 8334–8338. https://doi.org/10.1073/pnas.95.14.8334

Feldman, J. (2021). Mutual information and categorical perception. *Psychological Science, 32*(8), 1298–1310. https://doi.org/10.1177/0956797621996663

Feldman, N. H., Griffiths, T. L., Goldwater, S., & Morgan, J. L. (2013). A role for the developing lexicon in phonetic category acquisition. *Psychological Review, 120*(4), 751–778. https://doi.org/10.1037/a0034245

Fleming, R. W., & Storrs, K. R. (2019). Learning to see stuff. *Current Opinion in Behavioral Sciences, 30*, 100–108. https://doi.org/10.1016/j.cobeha.2019.07.004

Foard, C. F., & Kemler Nelson, D. G. (1984). Holistic and analytic modes of processing: The multiple determinants of perceptual analysis. *Journal of Experimental Psychology. General, 113*(1), 94–111. https://doi.org/10.1037/0096-3445.113.1.94

Folstein, J. R., Gauthier, I., & Palmeri, T. J. (2010). Mere exposure alters category learning of novel objects. *Frontiers in Psychology, 1*(AUG), 1–6. https://doi.org/10.3389/fpsyg.2010.00040

Francis, A. L., & Nusbaum, H. C. (2002). Selective attention and the acquisition of new phonetic categories. *Journal of Experimental Psychology: Human Perception and Performance, 28*(2), 349–366. https://doi.org/10.1037//0096-1523.28.2.349

Frost, R., Armstrong, B. C., & Christiansen, M. H. (2019). Statistical learning research: A critical review and possible new directions. *Psychological Bulletin, 145*(12), 1128–1153. https://doi.org/10.1037/bul0000210

Garner, W. R. (1974). *The processing of information and structure*. Hillsdale, NJ: Erlbaum.

Garner, W. R. (1976). Interaction of stimulus dimensions in concept and choice processes. *Cognitive Psychology, 8*(1), 98–123. https://doi.org/10.1016/0010-0285(76)90006-2

Getz, L. M., Nordeen, E. R., Vrabic, S. C., & Toscano, J. C. (2017). Modeling the development of audiovisual cue integration in speech perception. *Brain Sciences, 7*(3), 32. https://doi.org/10.3390/brainsci7030032

Gifford, A. M., Cohen, Y. E., & Stocker, A. A. (2014). Characterizing the impact of category uncertainty on human auditory categorization behavior. *PLoS Computational Biology, 10*(7), Article e1003715. https://doi.org/10.1371/journal.pcbi.1003715

Goldstone, R. L. (1998). Perceptual learning. *Annual Review of Psychology, 49*, 585–612. https://doi.org/10.1146/annurev.psych.49.1.585

Goldstone, R. L., Lippa, Y., & Shiffrin, R. M. (2001). Altering object representations through category learning. *Cognition, 78*, 27–43.

Goudbeek, M., Cutler, A., & Smits, R. (2008). Supervised and unsupervised learning of multidimensionally varying non-native speech categories. *Speech Communication, 50*(2), 109–125. Retrieved from http://www.sciencedirect.com/science/article/pii/S0167639307001410.

Gureckis, T. M., & Goldstone, R. L. (2008). The effect of the internal structure of categories on perception. In *Proceedings of the 30th annual conference of the cognitive science society* (pp. 1876–1881).

Hinton, G. E. (1989). Connectionist learning procedures. *Artificial Intelligence, 40*, 185–234. https://doi.org/10.1016/0004-3702(89)90049-0

Holt, L. L., & Lotto, A. J. (2010). Speech perception as categorization. *Attention, Perception, & Psychophysics, 72*(5), 1218–1227. https://doi.org/10.3758/app.72.5.1218

Iverson, P., & Kuhl, P. K. (1995). Mapping the perceptual magnet effect for speech using signal detection theory and multidimensional scaling. *Journal of Acoustical Society of America, 97*(1), 553–562.

Johannesson, M. (2001). *The problem of combining integral and separable dimensions. Technical Report at the Department of Computer Science*. Sweden: University of Skövde, Sweden and Lund University Cognitive Studies, Lund University. https://citeseerx.ist.psu.edu/viewdoc/download?.

Kalish, M. L., Newell, B. R., & Dunn, J. C. (2017). More is generally better: Higher working memory capacity does not impair perceptual category learning. *Journal of Experimental Psychology: Learning, Memory, and Cognition, 43*(4), 503–514. https://doi.org/10.1037/xlm0000323

Kemler, D. G., & Smith, L. B. (1979). Accessing similarity and dimensional relations: Effects of integrality and separability on the discovery of complex concepts. *Journal of Experimental Psychology: General, 108*(2), 133–150. https://doi.org/10.1037/0096-3445.108.2.133

Kemler Nelson, D. G. (1993). Processing integral dimensions: The whole view. *Journal of Experimental Psychology: Human Perception and Performance, 19*(5), 1105–1113. https://doi.org/10.1037/0096-1523.19.5.1105

Lewandowsky, S., Yang, L.-X., Newell, B. R., & Kalish, M. L. (2012). Working memory does not dissociate between different perceptual categorization tasks. *Journal of Experimental Psychology: Learning, Memory, and Cognition, 38*(4), 881–904. https://doi.org/10.1037/a0027298

Lewicki, M. S. (2002). Efficient coding of natural sounds. *Nature Neuroscience, 5*(4), 356–363. https://doi.org/10.1038/nn831

Lim, S.-J., Fiez, J. A., & Holt, L. L. (2019). Role of the striatum in incidental learning of sound categories. *Proceedings of the National Academy of Sciences, 116*(10), 201811992. https://doi.org/10.1073/pnas.1811992116

Lisker, L. (1986). "Voicing" in English: A catalogue of acoustic features signaling /b/ versus /p/ in trochees. *Language and Speech, 29*(1), 3–11. https://doi.org/10.1177/002383098602900102

Lotto, A. J., Sato, M., & Diehl, R. L. (2004). *Mapping the task for the second language learner: The case of Japanese acquisition of /r/ and /l/. from sound to sense: 50+ years of discoveries in speech communication* (pp. 181–186).

Love, B. C., Medin, D. L., & Gureckis, T. M. (2004). SUSTAIN: A network model of category learning. *Psychological Review, 111*, 309–332. https://doi.org/10.1037/0033-295x.111.2.309

Maddox, W. T., & Dodd, J. L. (2003). Separating perceptual and decisional attention processes in the identification and categorization of integral-dimension stimuli. *Journal of Experimental Psychology: Learning, Memory, and Cognition, 29*(3), 467–480. https://doi.org/10.1037/0278-7393.29.3.467

Maye, J., Werker, J. F., & Gerken, L. (2002). Infant sensitivity to distributional information can affect phonetic discrimination. *Cognition, 82*(3), 101–111. https://doi.org/10.1016/s0010-0277(01)00157-3

McMurray, B., & Jongman, A. (2011). What information is necessary for speech categorization? Harnessing variability in the speech signal by integrating cues computed relative to expectations. *Psychological Review, 118*(2), 219–246. https://doi.org/10.1037/a0022325.what

Melara, R. D., & Marks, L. E. (1990). Hard and soft interacting dimensions: Differential effects of dual context on classification. *Perception & Psychophysics, 47*(4), 307–325. https://doi.org/10.3758/bf03210870

Newell, B. R., Dunn, J. C., & Kalish, M. (2011). Systems of category learning fact or fantasy? *Psychology of Learning and Motivation, 54*, 167–215. https://doi.org/10.1016/b978-0-12-385527-5.00006-1

Nixon, J. S. (2020). Of mice and men: Speech sound acquisition as discriminative learning from prediction error, not just statistical tracking. *Cognition, 197*, Article 104081. https://doi.org/10.1016/j.cognition.2019.104081

Nixon, J. S., & Tomaschek, F. (2021). Prediction and error in early infant speech learning: A speech acquisition model. *Cognition, 212*, Article 104697. https://doi.org/10.1016/j.cognition.2021.104697

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

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

Pierrehumbert, J. B. (2003). Phonetic diversity, statistical learning, and acquisition of phonology. *Language and Speech, 46*(2–3), 115–154. https://doi.org/10.1177/00238309030460020501

Richler, J. J., & Palmeri, T. J. (2014). Visual category learning. *Wiley Interdisciplinary Reviews: Cognitive Science, 5*(1), 75–94. https://doi.org/10.1002/wcs.1268

Roark, C. L., & Holt, L. L. (2019). Perceptual dimensions influence auditory category learning. *Attention, Perception, & Psychophysics, 81*(4), 912–926. https://doi.org/10.3758/s13414-019-01688-6

Roark, C. L., & Holt, L. L. (2020). Statistical learning does not overrule perceptual priors during category learning. *PsyArXiv.*. https://doi.org/10.31234/osf.io/sdf7y

Roark, C. L., Plaut, D. C., & Holt, L. L. (2020). *A neural network model of the effect of prior experience with regularities on subsequent category learning*. https://doi.org/10.17605/OSF.IO/W64NU. October 7.

Saffran, J. R., & Kirkham, N. Z. (2018). Infant statistical learning. *Annual Review of Psychology, 69*, 181–203. https://doi.org/10.1146/annurev-psych-122216-011805

Scharinger, M., Henry, M. J., & Obleser, J. (2013). Prior experience with negative spectral correlations promotes information integration during auditory category learning. *Memory & Cognition, 41*(5), 752–768. https://doi.org/10.3758/s13421-013-0294-9

Schwartz, O., & Simoncelli, E. P. (2001). Natural signal statistics and sensory gain control. *Nature Neuroscience, 4*(8), 819–825. https://doi.org/10.1038/90526

Schyns, P. G., Goldstone, R. L., & Thibaut, J. P. (1998). The development of features in object concepts. *Behavioral and Brain Sciences, 21*(1), 1–17. discussion 17-54 https://doi.org/10.1017/s0140525x98000107.

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

Smith, J. D., Johnston, J. J. R., Musgrave, R. D., Zakrzewski, A. C., Boomer, J., Church, B. A., & Ashby, F. G. (2014). Cross-modal information integration in category learning. *Attention, Perception, & Psychophysics, 76*(5), 1473–1484. https://doi.org/10.3758/s13414-014-0659-6

Smith, L. B., Jayaraman, S., Clerkin, E., & Yu, C. (2018). The developing infant creates a curriculum for statistical learning. *Trends in Cognitive Sciences, 22*(4), 325–336. https://doi.org/10.1016/j.tics.2018.02.004

Tijsseling, A. G., & Gluck, M. A. (2002). A connectionist approach to processing dimensional interaction. *Connection Science, 14*(1), 1–48. https://doi.org/10.1080/09540090210138859

Toscano, J. C., & McMurray, B. (2010). Cue integration with categories: Weighting acoustic cues in speech using unsupervised learning and distributional statistics. *Cognitive Science, 34*(3), 434–464. https://doi.org/10.1111/j.1551-6709.2009.01077.x

Venables, W. N., & Ripley, B. D. (2002). *Modern applied statistics with S* (4th ed.). Springer.

Wang, X. (2007). Neural coding strategies in auditory cortex. *Hearing Research, 229*(1–2), 81–93. https://doi.org/10.1016/j.heares.2007.01.019

Werker, J. F., Yeung, H. H., & Yoshida, K. A. (2012). How do infants become experts at native-speech perception? *Current Directions in Psychological Science, 21*(4), 221–226. https://doi.org/10.1177/0963721412449459