A Neural Network Model of the Effect of Prior Experience with Regularities on Subsequent Category Learning

Casey L. Roark[1,2] **(croark@pitt.edu)**

David C. Plaut[1,2] **(plaut@cmu.edu)**

Lori L. Holt[1,2,3] **(loriholt@cmu.edu)**

[1]Department of Psychology, Carnegie Mellon University, Pittsburgh, PA 15213 USA
[2]Center for the Neural Basis of Cognition, Pittsburgh, PA 15213 USA
[3]Neuroscience Institute, Carnegie Mellon University, Pittsburgh, PA 15213 USA

Abstract

A popular dual systems theory of category learning argues that the structure of categories in perceptual space determines the mechanisms that drive learning. However, less attention has been paid to the nature of the perceptual dimensions defining the categories. Researchers typically assume that there is a direct, linear relationship between experimenter-defined physical input dimensions and learners' psychological dimensions, but this assumption is not always warranted. Through a set of simulations, we demonstrate that, based on the nature of prior experience, the psychological representations of experimenter-defined dimensions can place drastic constraints on category learning. We compare the model's behavior to several human studies 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 is a critical aspect to understanding the mechanisms that drive category learning.

Keywords: neural network; perception; category learning; statistical regularities

Introduction

Forming perceptual categories is thought to be at the heart of many complex cognitive processes, such as object recognition (Richler & Palmeri, 2014) and speech perception (Holt & Lotto, 2010). A popular dual systems model of category learning (Ashby et al., 1998) suggests that the mechanisms supporting category learning are engaged differently with different types of category structure. Many studies demonstrate differences in learning rule-based categories—requiring selective attention to individual input dimensions—and information-integration categories—requiring pre-decisional integration across dimensions (Ashby, Alfonso-Reese, Turken, & Waldron, 1998; Ashby & Maddox, 2011; Yi & Chandrasekaran, 2016). However, claims of fundamental differences between these two learning problems have been questioned (Carpenter, Wills, Benattayallah, & Milton, 2016; Edmunds, Milton, & Wills, 2015; Milton & Pothos, 2011; Wills, Suret, & McLaren, 2004).

Apart from the debate of dual versus single category learning systems, a critical assumption of many theories of category learning is that experimenter-defined input dimensions align with participants' internal psychological dimensions.

For some of the most well-studied pairs of dimensions in the visual domain, such as line length and orientation or spatial frequency and orientation of lines in a Gabor patch, this assumption is likely true. The underlying psychological and neural representations of simple visual input dimensions are well understood. However, more recent applications of these theories in the auditory domain make clear that the assumption may be problematic (Roark & Holt, 2019; Scharinger, Henry, & Obleser, 2013). In the auditory domain, the specific coding of many dimensions of acoustic signals is unclear. Beyond that, it is likely also the case that the representations of many of these dimensions interact in some way, potentially because the statistical structure of the dimensions is such that the neural coding for the dimensions is not independent (Garner, 1974; Wang, 2007).

One approach to avoiding an implicit assumption that physical input dimensions align with psychological dimensions is to estimate psychological representation of perceptual spaces using multidimensional scaling based on similarity ratings (Nosofsky, 1992; Shepard, 1980). However, this approach is somewhat limited in the conclusions that can be made about how existing representations influence learning. By systematically varying the relationship among perceptual dimensions, we are able to investigate the interaction between perception and cognition during category learning.

In the case of the dual-systems model of category learning, an argument is made about how so-called 'rule-based' category structures, which require selective attention to individual input dimensions, are optimally learned by an entirely different learning system than so-called 'information-integration' category structures, which are said to require 'pre-attentive' integration across more than one input dimension. However, there is little consideration as to what these dimensions are, and whether dimensions that are manipulated independently by the experimenter are actually independent in the psychological and neural representations of perceivers. That is, the underlying psychological dimensions may place strong constraints on the interpretation of what is 'rule-based' and what is 'information-integration' in terms of the actual problem being solved by the mind and brain.

Multiple cognitive science literatures demonstrate that underlying psychological dimensions are not necessarily homologous with input dimensions. Prior experience, including category learning, influences perception of dimensions (Goldstone, 1998) and prior knowledge and concepts can influence attention to input dimensions and category learning outcomes (Kaplan & Murphy, 2000; Kaplan & Murphy, 1999). Nonetheless, these literatures do not directly address the question of how category learning driven by distinct systems, as in a dual systems account of category learning, would be impacted when apparently orthogonal input dimensions are not psychologically independent. In the current work, we unite these perspectives to probe the influence of prior knowledge on learning with a dual systems approach that specifies that unique mechanisms drive category learning depending on category structure.

In the current investigation, we present a neural network model that demonstrates how underlying psychological representations that are formed by structured experience in the sensory environment may place strong constraints on novel category learning, depending on the structure of the categories in the physical space. We first describe the model architecture. We then describe the training and evaluation procedure for the model's behavior. Finally, we compare the model's behavior to human behavior from several perceptual category learning studies in the literature.

Model Architecture

There are two components to the model architecture (see Figure 1): the lower-level part supports **representation learning**, in which perceptual representations are gradually shaped through extensive experience prior to the experiment; the higher-level part supports **category learning**, in which the evoked representations of different stimuli are relatively rapidly associated with particular behavioral responses within an experimental context.

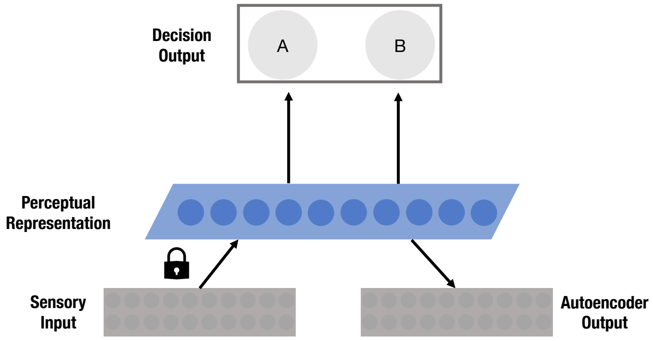

Figure 1: Model architecture.

Representation Learning

In the model, representational learning over two physical input dimensions *x* and *y* is implemented by an autoencoder. That is, the model receives structured sensory input that it must recreate in the autoencoder output layer via a smaller, "bottleneck" layer. A 20 unit 'sensory input' layer connects to a ten-unit hidden layer which connects to a 20 unit 'autoencoder output' layer. Ten of these 20 units represent the physical x-dimension value and ten represent the physical y-dimension value. For each dimension, a particular value was represented as an unnormalized 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 1,0). This encoding allows for graded input, which reflects population encoding of information in sensory cortex. Activations in the sensory input layer also have 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 is to recreate the input in the output layer. As a result of the training experience, the network will learn perceptual representations over the intermediate (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 (Figure 2)—a positive relationship between two dimensions (Positive), a negative relationship (Negative), the x-dimension is represented in more detail than the y-dimension (X-Dimension), the y-dimension is represented in more detail than the x-dimension (Y-Dimension), and where there is no correlation or relationship between the two dimensions (Independent). These environments are not meant to capture any particular natural signal statistics, but rather to reflect clear alternative scenarios to demonstrate how these simple relationships might be encoded in the perceptual system.

We trained the network for 50,000 epochs on batched learning across the 289 stimuli within each training distribution with a learning rate of 0.0001 and a bound of 1.0 on the length of the weight change vector. These learning parameters are intentionally conservative and were chosen solely to ensure that representation learning was stable and effective.

Category Learning

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). A two-unit category decision output layer was then connected to the hidden layer, reflective of decision responses in two-category problem.

For each of the five representation environments, the model was separately trained on four category learning problems (Figure 3). Each of the category learning problems was identical in terms of statistical structure (category variance and overlap between categories). The key difference is the rotation of the categories in physical space, such that

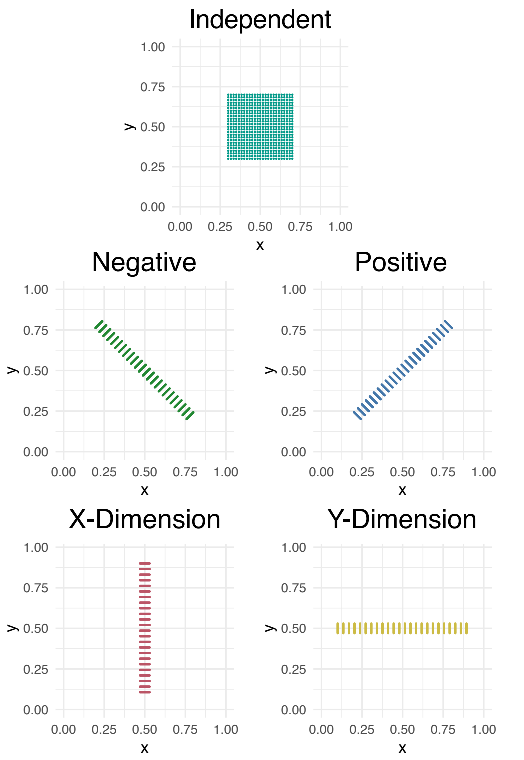

Figure 2: Representation learning distributions.

the category distinction requires different usage of the physical input dimensions (representative of the experimenter-defined dimensions). These category environments were designed to reflect two rule-based (RB) problems (RB-X dimension, RB-Y dimension) and two information-integration (II) problems (II-Positive boundary and II-Negative boundary). Critically, as a consequence of the representation learning phase, the physical dimensions (i.e., the experimenter-defined dimensions) *do not necessarily align* with the model's internal perceptual representations.

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. The network was trained separately on each of the category learning problems and exemplars were presented in random order without replacement. The model was trained and tested on one presentation of each of 200 stimuli from each category learning environment (100/category) and the same stimuli were used for training and testing. Because generalization of category learning was not a main component of our experiment and because we tested a conservative approach as a proof-of-concept demonstration, we trained and tested the model on the same stimuli. This is consistent in some ways with many studies of category learning that often do not directly test generalization. Future work should explore the extent of the model to generalize learned knowledge to novel exemplars.

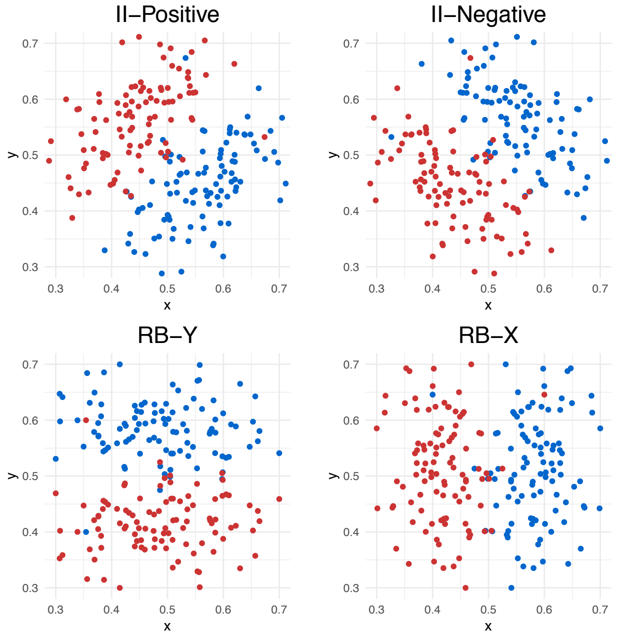

Figure 3: Category learning distributions.

During category learning, the network was trained with steepest descent using a learning rate of 0.5. Ten simulated subjects were run for each category learning type. For each simulated subject, after a single sweep through all 200 exemplars with the model updating its weights after each exemplar, the model was tested on the same stimulus set while keeping the weights stable (i.e. providing no feedback to the model).

Results

We examined the model's categorization performance after training with all 200 exemplars. This reflects the situation in which a human has encountered all category exemplars (with feedback) and is then tested on them without feedback.

Categorization accuracy

We quantified accuracy as the percent of category exemplars for which the model met a target activation criterion of 0.5. The simulated subjects' accuracies for each of the representation environments and category problems are shown in Figure 4.

The performance of the model on these categorization problems greatly depended on the nature of the pre-trained representations. Each representation environment (Independent, Positive, Negative, X-Dimension, Y-Dimension) makes specific predictions about the pattern of accuracy for the four different category types.

Training with the Independent distribution led to very high accuracy among the four category problems, with no significant differences across the four types ($F(3,36) = 1.51$, $p = .23$, $\eta_p^2 = .11$). However, there appears to be numerical differences in the means, such that the two RB problems have higher accuracies than the two II problems.

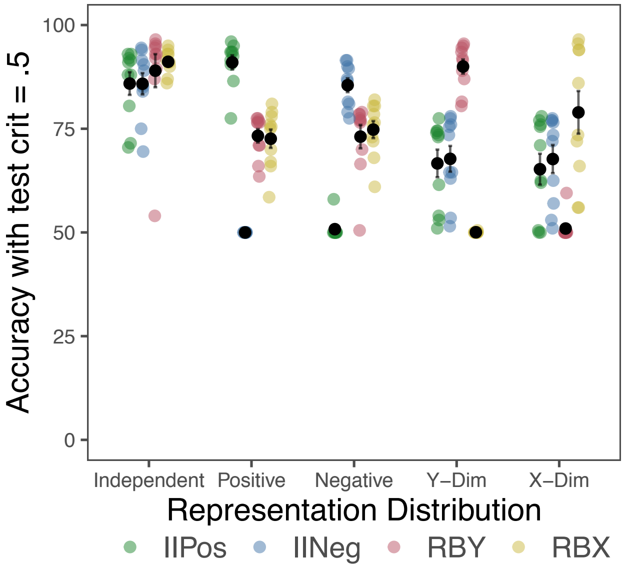

Figure 4: Model simulation accuracies in the four category learning environments with different perceptual representations.

The Negative and Positive correlation distributions had complementary results. For the Negative training distribution, accuracy was highest for the II-Positive problem and lowest for the II-Negative problem, with the two RB problems at intermediate levels ($F(3, 36) = 106.79$, $p < 0.0005$, $\eta_p^2 = .899$). For the Positive training distribution, accuracy was highest for the II-Negative category and lowest for the II-Positive category, with the two RB categories at intermediate levels ($F(3,36) = 54.13$, $p < 0.0005$, $\eta_p^2 = .92$).

The Y-Dimension and X-Dimension distributions, for which one dimension was represented more faithfully, and in more detail, than the other dimension had complementary results. For the Y-Dimension training distribution, accuracy was highest for the RB-Y problem, for which correct categorization required a distinction along the y-dimension and lowest for the RB-X problem, requiring a distinction along the x-dimension ($F(3,36) = 45.3$, $p < .0005$, $\eta_p^2 = .79$). The opposite pattern was observed for the X-Dimension distribution, with the RB-X problem having the highest accuracy and the RB-Y problem having the lowest ($F(3,36) = 10.03$, $p < .0005$, $\eta_p^2 = .46$). The two II problems were at intermediate accuracy levels for both the Y-Dimension an X-Dimension distributions.

These results demonstrate the potential for existing perceptual representations to have a major impact on the outcomes of category learning, especially when the physical dimensions or experimenter-defined dimensions do not align with the dimensions of representations.

Comparison with Human Behavior

In this section, we compare the model's behavior to human behavior on the four category learning types. When we can observe the pattern of accuracy in humans across several category learning types in the same physical 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 is not well understood, as with complex auditory dimensions.

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). 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).

Roark and Holt (2019) found that the category problems with the highest accuracy were the II-Positive and RB-MF, with RB-CF learned at more moderate levels, and II-Negative learned at the lowest levels (Figure 5). This overall pattern most closely aligns with the model's behavior for the Positive distribution, indicating that these acoustic dimensions may have a representation that reflects a long-term positive relationship between CF and MF.

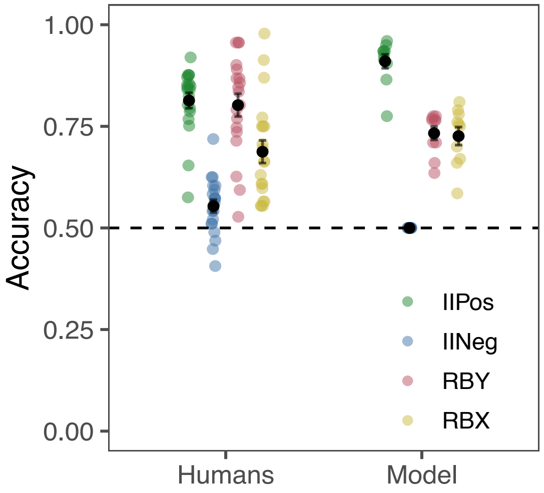

Figure 5. Human categorization performance in the generalization test in Roark & Holt (2019) compared with model performance with Positive distribution.

Ell, Ashby and Hutchinson (2012): Visual dimensions

In Ell, Ashby and Hutchinson (2012), Experiment 2, participants learned categories based on the visual dimensions of saturation and brightness. As in the 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-Negative but was not significantly different from II-Positive. 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. The nature of the current model simulations is that the training is extreme to demonstrate a first-pass confirmation that the nature of the representations can have drastic impacts on learning outcomes. However, it is likely the case that to match human behavior and representations more closely, the training representations will need to be less extreme.

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.

The purpose of these experiments was not to compare accuracies of the two RB and two II tasks. As such, Smith et al. (2014) do not compare accuracy across the four tasks. Although they trained on all four training types, they report the average accuracy for the two RB tasks to the average accuracy for the two II tasks. This comparison stems from their investigation into the differences between RB and II category learning but distorts the 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, which aligns with the model's behavior with the Independent distributions, reflecting a situation where the two sensory dimensions are encoded independently. The explanation makes sense in that cross-modal dimensions are likely to be encoded by 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. This exact pattern is not represented directly in the model's behavior. However, it is still mostly aligned with the Independent representations, with some combination for which one dimension is represented slightly more faithfully than the other dimension, resulting in disproportionate perceptual salience across the two sensory dimensions. Though there are some limitations in our ability to directly compare the effects to the model behavior, it seems reasonable that one of these dimensions may be more salient than the other, which may have influenced learning outcomes.

Conclusions

The current set of simulations demonstrates that the nature of experience in a sensory environment can shape the representations of input dimensions in a way that can drastically impact category learning behavior. Depending on the nature of the representations, which are shaped by experience, some category learning problems are easily learnable, whereas others are completely unlearnable. This model demonstrates that consideration of perceptual processing and acknowledgement of the constraints that the perceptual system and existing representations place on learning are critical to understanding the mechanisms at play during perceptual category learning. The nature of the learning problem may differ substantially depending on the perceptual representations across the very same input dimensions.

The current model used relatively simple and somewhat extreme training spaces that are clearly much more abstract than the way sensory information is presented 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 expansion of this model should include a simulation of the kind of variability and noise that exists in real-world sensory environments. Additionally, different kinds of relationships in the input should also be tested to make clear predictions about how the many different kinds of relationships (rather than just independence or a perfect correlation) can be represented by the model to affect behavior. Finally, this model is restricted to a two-dimensional space. The world beyond simple experiments has many more dimensions, some of which are relevant, others irrelevant, some present and varying, some rarely present and stable. A future iteration 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.

An important next step would be to train the model and humans on identical distributions and to compare visual, auditory, and cross-modal dimensions on the same category distributions. However, the current investigation provides valuable insight into the nature of the representations of these dimension comparisons and provides a proof-of-concept investigation that indicates that rule-based and information-integration distinctions 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 learning challenge.

The current model used an autoencoder for training the representations to reflect sensory regularities. Of course, an autoencoder does not fully reflect the complexities of human learning or sensory experience. Future work might expand the current demonstration to include more nuanced training methods, including feedback-based learning. We suspect that training with feedback may impact subsequent category learning behavior even more strongly.

The goal of the current neural network model simulation experiment was to illuminate the influence of perceptual representations on category learning. This formulation of the model allows us to make predictions about how perceptual representations of sensory information influence category learning mechanisms. What is clear from this investigation is that the nature of learning problem can vary dramatically based on the network's existing hidden unit representations. The same is likely true with human learning. While much of the human perceptual category learning research has used simple, verbalizable dimensions that are likely represented independently both neurally and in mental representations, it is a much more difficult and interesting problem to understand what happens when perception is not so straightforward. The current model demonstrates scenarios in which the influence of perception on cognition can be drastic.

This approach challenges the typical assumption made in 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. Therefore, labeling problems as rule-based or information-integration based on experimenter-defined dimensions does not capture the true complexity of the problem or the nature of the problem for the human perceptual system.

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). The influence of the psychological representation of dimensions was also a focus of earlier work in the domain of perception and learning (Garner, 1974; Kemler & Smith, 1979; Kemler Nelson, 1993; Melara & Marks, 1990). However, the *perceptual* side of perceptual category learning has drifted out of focus of current theories of learning. The current set of simulations demonstrates that the psychological representation of information, shaped by experience, can have strong influences on the nature of the learning problem.

Acknowledgements

The project described was supported by Award Number T32GM081760 from the National Institute of General Medical Sciences.

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.

Ashby, F. G., & Maddox, W. T. (2011). Human category learning 2.0. *Annals of the New York Academy of Sciences*, *1224*, 147–161.

Best, C. T. (1995). A Direct Realist View of Cross-Language Speech Perception. In *Speech Perception and Linguistic Experience: Issues in Cross-Language Research*.

Carpenter, K. L., Wills, A. J., Benattayallah, A., & Milton, F. (2016). A Comparison of the neural correlates that underlie rule-based and information-integration category learning. *Human Brain Mapping*, *37*(10), 3557–3574.

Edmunds, C. E. R., Milton, F., & Wills, A. J. (2015). Feedback can be superior to observational training for both rule-based and information-integration category structures . *The Quarterly Journal of Experimental Psychology*, 37–41.

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.

Garner, W. R. (1974). *The Processing of Information and Structure*. Hillsdale, NJ: Erlbaum.

Goldstone, R. L. (1998). Perceptual learning. *Annual Review of Psychology*, *49*, 585–612.

Holt, L. L., & Lotto, A. J. (2010). Speech perception as categorization. *Attention, Perception, & Psychophysics*, *72*(5), 1218–1227.

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.

Kaplan, A. S., & Murphy, G. L. (2000). Category learning with minimal prior knowledge. *Journal of Experimental Psychology: Learning, Memory, and Cognition*, *26*(4), 829–846.

Kaplan, Audrey S., & Murphy, G. L. (1999). The acquisition of category structure in unsupervised learning. *Memory and Cognition*, *27*(4), 699–712.

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.

Kemler Nelson, D. G. (1993). Processing integral dimensions: the whole view. *Journal of Experimental Psychology: Human Perception and Performance*, *19*(5), 1105–1113.

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.

Milton, F., & Pothos, E. M. (2011). Category structure and the two learning systems of COVIS. *European Journal of Neuroscience*, *34*(8), 1326–1336.

Nosofsky, R. (1992). Similarity Scaling And Cognitive Process Models. *Annual Review of Psychology*, *43*(1), 25–53.

Richler, J. J., & Palmeri, T. J. (2014). Visual category learning. *Wiley Interdisciplinary Reviews: Cognitive Science*, *5*(1), 75–94.

Roark, C. L., & Holt, L. L. (2019). Perceptual dimensions influence auditory category learning. *Attention, Perception, and Psychophysics*, *81*(4), 912–926.

Scharinger, M., Henry, M. J., & Obleser, J. (2013). Prior experience with negative spectral correlations promotes information integration during auditory category learning. *Memory & Cognition*, *41*, 752–768.

Shepard, R. N. (1980). Multidimensional Scaling, Tree-Fitting, and Clustering. *Science*, *210*(24), 390–398.

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*, 1473–1484.

Wang, X. (2007). Neural coding strategies in auditory cortex. *Hearing Research*, *229*(1–2), 81–93.

Wills, A. J., Suret, M., & McLaren, I. P. L. (2004). Brief communication: the role of category structure in determining the effects of stimulus preexposure on categorization accuracy. *The Quarterly Journal of Experimental Psychology. B, Comparative and Physiological Psychology*, *57*(1), 79–88.

Yi, H. G., & Chandrasekaran, B. (2016). Auditory categories with separable decision boundaries are learned faster with full feedback than with minimal feedback. *The Journal of the Acoustical Society of America*, *140*(2), 1332–1335.