DISSERTATION

CUE COMPETITION AND FEATURE REPRESENTATION IN A CATEGORY LEARNING TASK: AN fMRI STUDY

Submitted by

Kade Jentink

Department of Psychology

In partial fulfillment of the requirements

For the Degree of Doctor of Philosophy

Colorado State University

Fort Collins, Colorado

Summer 2023

Doctoral Committee:

Advisor: Carol Seger

Agnieszka Burzynska
Don Rojas
Michael Thomas

ABSTRACT

CUE COMPETITION AND FEATURE REPRESENTATION IN A CATEGORY LEARNING TASK: AN fMRI STUDY

During learning, attention is limited, and therefore selecting what feature(s) to attend to in the environment is important. Sometimes, attention is captured by a cue or feature in such a way that other cues or features are not attended to, known as overshadowing. This process is not entirely understood in category learning, with some studies suggesting that it *enhances* learning of other features (Murphy et al., 2017), while others suggest that it *inhibits* (Lau et al., 2020). Furthermore, the location and organization of the neural representations that develop for category features during overshadowing has not been previously examined in this context. The present experiment used representational similarity analyses (RSA), a method for interrogating representational structure (Kriegeskorte et al., 2008), in order to examine where and how features were represented during overshadowing in a category learning task. Participants completed a category learning task in which categories were defined based on two informative features, one binary and one continuous. The binary feature was easier to learn (i.e., more salient), and it was hypothesized that it would overshadow learning of the more difficult continuous feature. This was demonstrated behaviorally: participants learned to categorize when the binary feature was present, then performed at chance when it was removed in a transfer task. Three different hypothetical models were fit to the neural data to determine underlying representational structure: a binary category model, an effector-specific motor model, and a model representing the degree of perceptual similarity for the continuous feature. During initial learning when the primary binary feature was present, the category model fit data from both early visual and object-specific areas

of visual cortex, while the motor model fit data from motor-related regions including primary somatomotor cortex and the cerebellum. The perceptual similarity model for the continuous feature did not fit any task data during either Training or Transfer. However, there was a trend for the category model to fit activity in the basal ganglia and lateral occipital complex (LOC) during the Transfer task when the only information available for categorization was the continuous feature. Taken together, these results suggest that, although overshadowing inhibits use of the overshadowed continuous feature as the basis of categorization behavior, it might still contribute to activation of neural representations of category membership.

TABLE OF CONTENTS

Chapter 1 – Introduction

Associative learning and cue competition

In early models of classical conditioning, a stimulus (conditioned stimulus), paired closely in time with an outcome (unconditioned stimulus), led to learning (Hull, 1943; Spence, 1956). However, mere temporal adjacency is not sufficient. Cue competition refers to the inhibition or prevention of acquisition of the relationship between some cue or feature of a stimulus and an outcome in a multi cue environment.

Cue competition can manifest in different ways. When pre-exposure to a cue prevents later learning of a compound cue (featuring the pre-exposed element and one additional cue), this is known as blocking. A failure to learn a compound cue because one element is more salient than the other is known as overshadowing (Kamin, 1965, 1967, 1969; Pavlov, 1927; Rescorla & Wagner, 1972). Early work speculated that cue competition might be the result of selective attention (Mackintosh, 1965; Sutherland, 1964), in which learning only occurs when attention, which has a finite capacity, is directed towards a stimulus (and away from others). Others, notably Kamin (1969; although see others e.g., Mackintosh, 1976) explained blocking in informational terms. They argued that after training with some initial stimulus, the inclusion of a second provided no new information because the unconditioned stimulus was already perfectly predicted by the initially presented conditioned stimulus. It was necessary for the organism to experience the outcome as at least partly unexplained, or in other words experience surprise, in order for learning to occur. This idea of tying the “surprisingness” of a

stimulus to learning became an important part of the Rescorla-Wagner (1972) model, although they characterized it as a “violation of expectations”.

Representations

Philosophy: what is a representation?

A representation is, in a general sense, a stand-in (Bechtel, 1998); it is an abstract concept: a symbol for something else, or a unit of information. It can be defined as the sum of the individual elements that comprise an object, more holistically, as a discrete unit, or even by its relationship to other representations. For example, a stray cat can represent both a cat that you own, or cats as a whole, and even more abstract ideas like “feline” character traits. A cat can also be defined by the elements that it is comprised of (i.e., its lines, shape, color, etc.), the sum of its parts (its “cat-ness”), or its distance in perceptual space from other cats compared to something like a dog. A conversation considering the full range of philosophical questions and approaches to defining the concept is outside the scope of the present study, which focuses on the nature of a representation as it is constructed neurally. Even still, the neuroimaging of representations is inherently complicated – it mixes difficult philosophical and scientific questions. For example, is a representation “formed” the moment there is activity related to object perception? Is it only a usable representation once a certain level of complexity is reached? What representational information is communicated from neural region to region? The present study focuses on “What can we measure regarding representations?” while only briefly considering “What does what we measure *truly* represent?”.

Measuring neural representations with neuroimaging

On a small scale, a neural representation could be thought of in terms of population coding. This refers to clusters of neurons, of varying sizes and at varying distances, which operate together, and can vary in firing rate/timing, as well as other factors (Panzeri et al., 2015). Information is commonly represented in such a way within the nervous system, and population coding accounts for the ways in which a myriad of different functions, from memory to motor movements, are represented. One benefit to this system is that it is robust: individual neurons within a population can be damaged without seriously harming the overall representation (Pouget et al., 2000). Furthermore, the scope of population coding can range from a single, finely-tuned neuron (e.g., "grandmother cell") to a vast array of more broadly useful neurons spread throughout an entire structure. When a small, localized set of functionally similar neurons support a representation, it is known as sparse coding, and when the opposite is true, and a large number of neurons are responsible for many different representations, this is known as distributed coding (Wixted et al., 2014; Young & Yamane, 1992).

A representation in this sense is a fundamental unit of the brain, and, as an internal state, acts as both storage and communication, capable of transmitting information that can be later decoded. Interestingly, it is even suggested that an organism extracts information from the activity of a population via an organic corrollary of the same type of machine learning algorithm we use to decode information from neurons in neuroimaging (Deneve et al., 1999).

An apposite upward scaling from population coding concerns representations that can be observed via neuroimaging methods such as functional magnetic resonance imaging (fMRI).

In fMRI, the activity of populations of neurons can be observed via blood flow (which is used as a proxy for said activity) in arbitrary regions of space known as voxels (three dimensional pixels)(Kwong et al., 1992). However, when using fMRI, a single voxel is typically not very informative. Instead, it is more interesting to consider where active voxels cluster together, since clusters of voxel activity can denote areas of representational and functional significance. Representations at this scale (multiple voxels) can be observed by comparing the activity of voxels across some number of conditions using what is known as univariate analysis (Friston et al., 1994). A group of voxels in a region, more active to faces than a baseline, might be said to represent faces. A region with voxels differentially active for large versus small objects (of any category) might generally code for size. Univariate analyses are great for attempting to identify the location of a stimulus in representational space, although this location is typically relative, as these analyses necessitate comparisons with other conditions to generate results (Davis & Poldrack, 2013).

One downside to univariate analyses is that the activity of neighboring voxels often gets averaged together in a process known as smoothing, which blurs single-voxel information in favor of increased ability to detect an effect (Nichols & Hayasaka, 2003; although see: Op de Beeck, 2010 for debate). Univariate analyses would find that the hypothetical region of interest (ROI) in Figure 1 cannot differentiate individual members of the category – the unique activity for each stimulus gets smoothed into one homogenous piece of information (color). However, there are methods which can preserve the unique pattern of information present in the ROI, allowing one to distinguish individual members of the category. These include multivoxel (or multivariate) pattern analysis (MVPA).

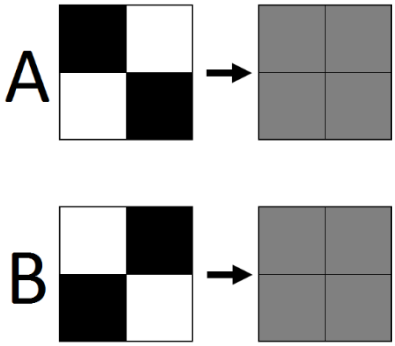

Figure 1: A simple example of the effect that averaging can have on the unique patterns of activity inherent in neuroimaging data.

MVPA capitalizes on the information contained in multiple voxels *simultaneously*. If univariate analyses are thought of as searching for *activation*, multivariate analyses can be thought of as searching for *information* (Kriegeskorte & Bandettini, 2007). These "pattern-based" spatial analyses are, in part, guided by the principles of population coding: that representations are distributed across functional clusters of neurons spread throughout the brain (Pouget et al., 2000). Furthermore, these distributed clusters, each connected to some stimulus or condition, are assumed to be linked to a discrete pattern of activation which can be translated into high-dimensional representational space. Although fMRI may not have direct access to the neuronal activity (it measures blood flow – a proxy), a difference in voxel patterns measured using MVPA can suggest a difference in neuronal patterns (Kriegeskorte & Bandettini, 2007). MVPA takes advantage of that information, and in turn, it can more directly measure said representational space than univariate analyses (Davis & Poldrack, 2013).

Types of MVPA Analyses

Decoding (Haxby et al., 2001) and representational similarity analysis (RSA; Kriegeskorte et al., 2008), are the two most popular MVPA analyses (Normal et al., 2006; Weaverdyck et al., 2020). They are useful for determining not just what a region represents, but how the

information is “encoded and organized” (Haxby et al., 2014; Haynes, 2015; Mahmoudi et al., 2012). Decoding is good for separating out the specifics of what a region represents. It accomplishes this in two steps: 1) a machine-learning pattern classifier is given labeled fMRI data, and “trained” to recognize which pieces of data belong to which label, then 2) it is given unlabeled data and tasked with seeing if what it has learned can be used to accurately assign labels to the trials. If a trained classifier is able to determine which trials belong to which trial type, the voxels which were sampled are said to contain information related to the representation of those trial types.

Decoding has been used to uncover representational information present in neural data which univariate analyses were previously unable to access. Haxby et al. (2001), in a landmark study, utilized MVPA to search for neural patterns of information that would let them classify visual images into several different categories. They were not only able to find a distinct pattern of response for each category, but also found that the information was sometimes overlapping: regions which responded strongly to one specific category (e.g., faces in the fusiform face area; FFA) also contained information which allowed for the decoding of other categories. In other words, it appeared that representational information related to multiple categories was present in areas previously found to only contain information for one type (e.g., Kanwisher et al., 1997).

Decoding has been used to help understand a wide variety of representational spaces, including: task type (Poldrack et al., 2009), vocalized emotion (Ethofer et al., 2009), behavioral performance (Raizada et al., 2010), and unconscious information (Haynes & Rees, 2005). Decoding has even allowed for the indirect observation of the representational space occupied

by dreams (Horikawa et al., 2013). However, describing the direct mappings from a task to the representation of its components (e.g., how is categoricity represented in the lateral occipital complex?) does not tell the whole story. It is also beneficial to understand the degree of representational similarity *between* representational spaces, or in other words, their second-order isomorphism (Figure 2)(Edelman, 1998; Kriegeskorte et al., 2008a; Shepard & Chipman, 1970).

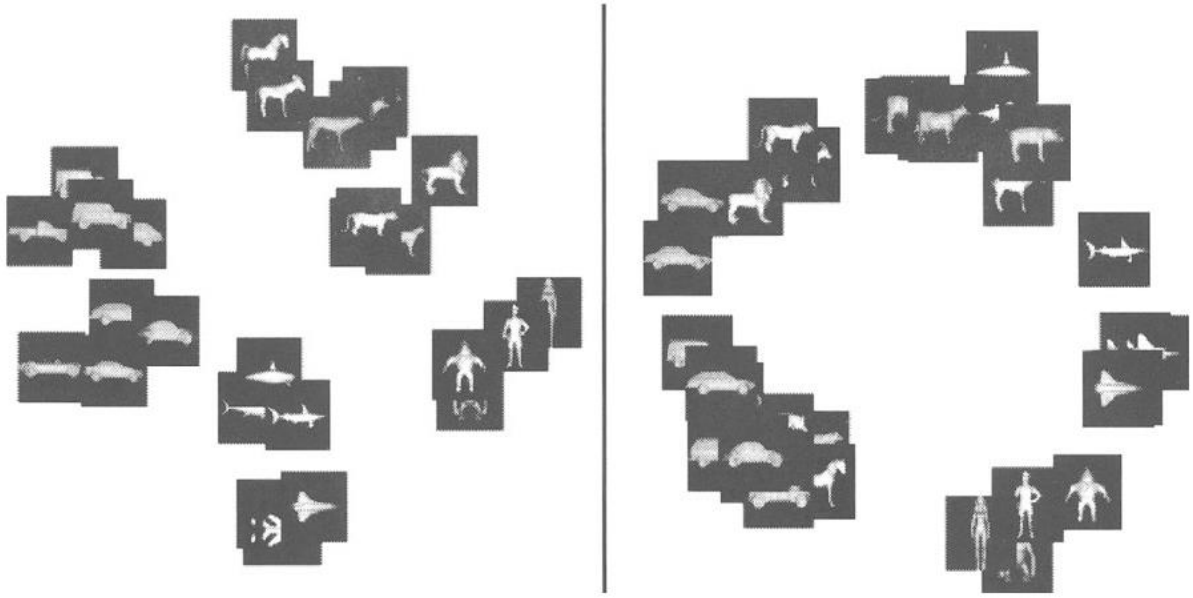

Figure 2: An example of two representational spaces. Left: Participant-reported perceptual similarity. Right: Similarity as determined by neuroimaging data. How similar are the two spaces?

Representational similarity analysis (RSA; Kriegeskorte et al., 2008a) is a method designed to examine these second-order isomorphisms. In a simple sense, it concerns the representational space of a stimulus as it relates to other stimuli – a representation of similarities – sometimes referred to as its "representational geometry" (Edelman, 1998; Kriegeskorte & Kievit, 2013). It is the difference between "What regions can differentiate faces from houses?" and "How different is the representation of a face from that of a house?". RSA

can even examine the relationships between stimuli within a single category (e.g., how different is one face stimulus from another). Like many multivariate methods, this is accomplished by accessing high-dimensional representational space. Whereas decoding seeks to draw a plane which separates points in that space into classes, RSA in contrast seeks to measure the distance between the points. It does this simply: multiple voxels are sampled for a stimulus, their activity is correlated with that of another stimulus, and the correlation value is assigned to a cell in a grid. This comparison is repeated for every stimulus combination, and eventually, a grid is assembled representing all possible combinations of correlations (Figure 3; left).

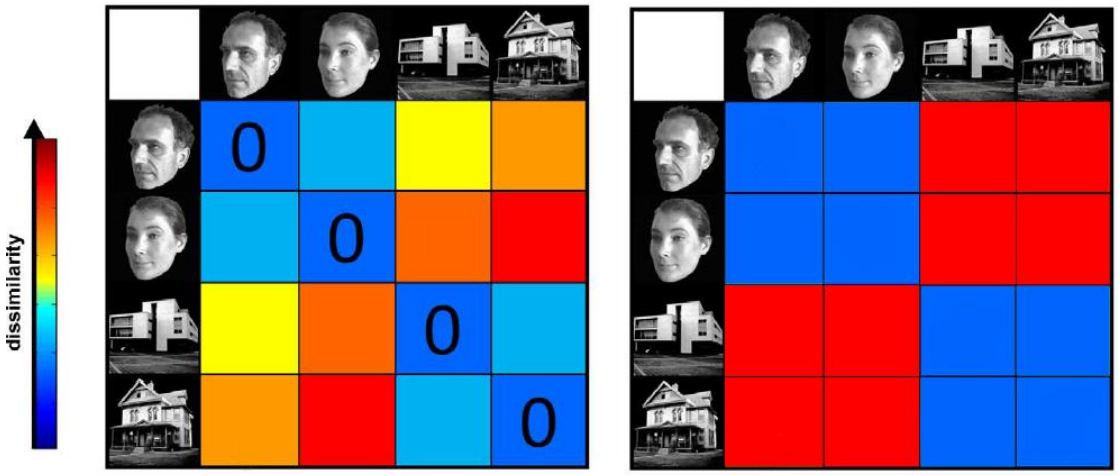

Figure 3: Two different representational dissimilarity matrices (RDMs). Left: A neural RDM created by comparing the patterns of voxel activity between two different individual stimuli. The diagonal is zero because a stimulus compared to itself is perfectly identical. Right: A hypothetical model of how a given neural region may represent a stimulus space. The color represents the hypothetical degree of correlation between two stimuli, with redder colors representing more dissimilar relationships.

Composing this representational dissimilarity matrix (RDM) for the neural data is usually only the first step. The RDM is then typically compared to something else; what it gets compared to depends on the representational question it is designed to answer. One study, for example, compared behavioral ratings of perceptual similarity of animals to the neural RDM

generated from viewing said animals in order to determine if the semantic distinctions identified by participants were reflected in representational space (Connolly et al., 2012). Not only can you use RDMs to compare representational spaces across ROIs (or compare the same ROI across different groups of people), but you can also compare an RDM to a model to see if the geometry fits your prediction for the representational space (Figure 3; right). Interestingly, because RDMs are abstracted from actual neural space, you can even compare them across *species* (Kriegeskorte et al., 2008b).

Overall, despite all the claims of "mind-reading" in journals and popular press (Normal et al., 2006), MVPA still has interpretational limitations on the representational claims it can make. The most important is that, for example, even if activity in a region can be used to successfully distinguish between cats and dogs that does not necessarily mean that the region represents every feature or property of cats and dogs. Neuroimaging data are inherently correlational (Poldrack, 2011), therefore the only conclusion that can logically be made is that the ROI contains information which allows that classifier to make a successful distinction between the two categories. In other words, the *where* question has been answered, but not the *what* (i.e., what is the nature of the representation) (Poldrack, 2011; Popov et al., 2018; Ritchie et al., 2019).

An anecdote illustrating this limitation comes from a competition that challenged competitors to develop advances in "brain reading and mind reading". Challengers were tasked with decoding specific elements from data in which participants watched a humorous TV show (e.g., actors, objects, spatial locations, humorous situations). It turned out that the best way to decode "humor" from the data was to tune a classifier on data from the ventricles, which

"…tended to jiggle whenever the participant felt an urge to laugh" (Tong & Pratte, 2012). Clearly, the ventricles do not code for humor, but they did contain information which allowed a classifier to make the distinction, an important point to remember when considering what representational information MVPA data might contain.

Overall, fMRI methods allow for the examination of representational spaces from many different perspectives, in more depth, and with a greater variety of methodological options, than I have space to describe. Moving forward, I will now focus on one particular representational space: novel categories defined by variation in separate features. I will introduce category learning and the way in which category representations are acquired, describe how features are learned, and lastly, introduce cue competition in the context of category learning.

Category learning

When presented with an unfamiliar stimulus, the brain must determine to what category it belongs in (e.g., food or predator) order to know how to act towards it (e.g., approach or avoid)(Seger & Miller, 2010). It is through the process of our interactions with category members that a representation is formed (Markman & Ross, 2003). However, the nature of the representational space, in terms of attention, features, and representations, has been approached in many different ways. I will discuss some theoretical models from a general perspective, the role of attention in the development of feature representations, and how these factors interact with task type to emphasize different representational spaces. Lastly, I will introduce how these factors can combine to produce competition effects similar to those discussed with associative learning paradigms before introducing the current project.

Multiple systems in category learning

Most early theories of category learning focused on the acquisition of category representations via a single system (e.g., Shepard et al., 1961). Briefly, some of these include prototype theory, which specified a single category representation constructed as an average of that class that all new stimuli were compared to (Rosch, 1973), and exemplar theory, in which new stimuli were compared to the representations of many previously seen exemplars to find a similar match (Nosofsky, 1986). As the field of memory research began to favor multiple systems, so did the field of category learning (Ashby & O'Brien, 2005; Ashby & Maddox, 2005; Ashby & Valentin, 2017).

Current theories paint a diverse picture of multiple category learning systems and representations, with many models implementing combinations of previous theories with components that load onto different memory systems (e.g., Erickson & Kruschke, 1998), or incorporate context specific mechanisms (i.e., how task demands affect category representations). One such model, COVIS (COmpetition between Verbal and Implicit Systems; Ashby et al., 1998), proposes that different representations are generated during category learning tasks based upon their memory system demands, such that tasks with an "easily verbalizable" rule for categorization compared to tasks without would load onto either declarative or implicit memory systems, respectively. SUSTAIN (Supervised and Unsupervised STratified Adaptive Incremental Network; Love et al., 2004) is another similarly diversified model which describes a representational landscape, dotted with clusters of specifically tuned category prototypes, that can flexibly create new clusters as required by task demands, irrespective of modality.

Category types, Task types, and within- or between-category representations

In order to categorize a stimulus, characteristics which define its membership must be identified. These characteristics, known as features, can be broadly defined in many different ways. In the simplest sense, they are low-level, easily separable visual properties such as size or shape. They can also be more complex, taking the form of integrated, less separable information such as color saturation and brightness (Cheng & Pachella, 1984). Some features of a category may not be perceptually available without classification training to increase their perceptual separability (Soto & Ashby, 2015). Features can furthermore be functionally defined in any number of arbitrary ways, and even created as needed, based on factors like conceptual knowledge about the stimuli (e.g., father/mother) or task demands (Davis, et al., 2017; Schyns et al., 1998). In short, features can range from perceptual to conceptual, and represent both similarities and differences within and between categories.

In naturalistic settings, feature dimensions typically vary continuously (e.g., height/color/etc.), however, in the lab, they often fluctuate around a small number of discrete values. For example, a participant presented with simple drawings of imaginary animals may find that each feature (e.g., the limbs) has only a binary set of values. The set of features and values that are shared amongst category members, as well as their relational nature, are known as their structure (Medin, 1989), and there are many different types. Rule-based structures can be learned via explicit reasoning processes, and often are defined by a unidimensional rule which can be easily verbally described, whereas information-integration structures require combining information from multiple different feature dimensions in a stage of cognition prior to conscious influence (Ashby & Ell, 2001).

The features that are focused on during learning are influenced by the context (i.e., task) in which they are presented (Goldstone, 1996; Yamauchi & Markman, 1998). Supervised tasks are those in which participants are given guidance about which category a stimulus belongs to. The most common supervised learning task structure used in category learning studies is trial-and-error learning with feedback, in which the participant indicates what they think the category membership is, and then receives corrective feedback. Both rule-based and information-integration category structures can be learned via trial and error task structures, but are thought to rely on different learning mechanisms.

Participants learning by trial and error typically begin by taking an explicit approach to learning: they try to search out rules, and test hypotheses, in order to learn the task (Nosofsky et al., 1994). Rule based tasks are well suited for learning in this manner and participants over time learn explicit rules sufficient to support high levels of performance. In rule based tasks participants perform best when they can verbalize the rules which constrain category membership (Ashby et al., 1998). Human participants can also learn to categorize in information integration tasks via trial and error learning with feedback. In these tasks the trial and error structure recruits reinforcement learning mechanisms (feedback is processed similarly to reward in reinforcement learning), and categories are learned by mapping the stimuli onto the category responses. In addition to supervised learning tasks, unsupervised tasks (such as passive viewing or self-organized sorting) can support category learning. Unsupervised tasks typically train the category learning system implicitly – learning goals may not be explicit, feedback is not provided, and categories can spontaneously develop (Love, 2002). Learning is often incidental rather than intentional (although not always: see, Love,

2002; Love, 2003), and stimuli are usually processed either by similarity (Wattenmaker, 1991), or simple unidimensional rules (Ashby et al., 1998). The present study used a supervised rule-based category learning task.

The "shape" of the representational geometry underlying different category structures can differ based on context. For example, some tasks lead participants to focus attention towards features which separate members *within* a category whereas others focus attention on features which define separation *between* categories (Bott et al., 2007; Hoffman & Rehder, 2003; Love et al., 2000; Wattenmaker, 1991). For example, unsupervised and supervised category learning formats generally lead to the learning of within- and between-category features, respectively (Chin-Parker & Ross, 2002, 2004; Goldstone, 1996; Markman & Ross, 2003; Yamauchi et al., 2002; Yamauchi & Markman, 1998). In inference tasks, a type of unsupervised task, participants are presented with an incomplete stimulus, and based on available information (e.g., other features or a category label), they must infer the missing feature. Since it is important to learn the relationships between category features which define their category membership, this type of task encourages participants to attend to all of the features – their within-category similarities – and participants tend to learn nondiagnostic as well as diagnostic features as they develop a category representation (Chin-Parker & Ross, 2002; Yamauchi & Markman, 1998).

In contrast, classification tasks, a type of supervised learning task, require that participants assign stimuli to arbitrary groups; they then receive feedback for their choice. Classification tends to support the learning of between-category differences. Attention is typically deployed to the feature or features which lead to the best performance, usually those

that are both perceptually salient and highly diagnostic, in spite of the fact that categories often contain many informative, correlated features (Shepard, 1961). If a diagnostic feature is suddenly changed, participants typically struggle since they usually have not learned any of the other potentially informative features (e.g., learned inattention: Deng & Sloutsky, 2016; Plebanek & Sloutsky, 2017). Partially diagnostic features are sometimes learned, but only if they are sufficiently easy, and the more highly (or perfectly) diagnostic features are harder to learn (Thomas et al., 2021). The present study examined the representation of both attended and unattended diagnostic features.

Although category learning in naturalistic settings often proceeds in an unsupervised manner (Gureckis & Love, 2003), most research has focused on supervised models (Nosofsky et al., 1994; Shepard et al., 1961), perhaps owing to the computational complexity of completely unconstrained unsupervised learning (Pothos & Chater, 2005). However, category learning models derived from supervised learning tasks have trouble extending across a wide variety of situations (Levering & Kurtz, 2015; Love, 2001; Schank et al., 1986). Even if it may seem more beneficial for people to develop rich and generalizable category representations by focusing on within-category differences, if the goal of a system is simply accurate classification, focusing on between-category differences generally results in better performance (Nosofsky et al., 1994). Therefore, it may be more efficient for a system to learn a minimal category representation focused on boundary conditions, despite the benefits (e.g., generalizability) that may stem from within-category representations (Vapnik, 1998). This can be observed by the tendency participants have to classify categories containing several related features (i.e., rich within-category similarity) by using a single dimension (e.g., Lassaline & Murphy, 1996).

Supervised and unsupervised learning contexts can interact with category structure in complicated ways, which can affect the type of representation which develops (Love, 2003). As an example, family resemblance categories have features which are correlated with one another, but there is no one feature which deterministically defines membership (Rosch & Mervis, 1975). One study found that, in both a supervised and unsupervised paradigm, not only did adding features, which should complicate learning, *increase* the total number of features learned, but it did so even when there was no benefit to accuracy (Hoffman & Murphy, 2006). Another similar study found equal, and in some cases, greater learning for the more complex (i.e., more features) structure (Minda & Smith, 2001).

The way in which attention is allocated to category features during different tasks has parallels with machine learning. For example, a discriminative classifier is concerned with the decision bound between two classes. Accuracy in assigning class labels is the most important factor, so classifiers tend to focus only on whatever features are necessary to complete the task with a high degree of accuracy. Generative classifiers, on the other hand, are more concerned with the class space, as a whole. They learn the distribution of an individual class, and therefore, acquire more information about the relationships which define it, relative to discriminative classifiers. Similar to category learning, discriminative classifiers tend to outperform generative classifiers when measured solely in terms of accuracy (Ng & Jordan, 2002).

Attentional constraints and category representation

Attention plays a prominent role in category learning models. Factors such as where and on what features we focus our attention are important for determining the type of category

representation that develops. Attending to a spatial location or a specific object enhances its processing while diminishing the processing of unattended locations/objects (Gazzaley et al., 2005; Moran & Desimone, 1985). The same pattern also applies to attention directed at individual features of an object and can allow for better separation of the values of the attended feature (Braunlich & Love, 2019; Polk et al., 2008). For example, attending to color over shape enhances processing of color, diminishes processing (and reduces learning) of shape, and makes it easier to distinguish the different color values present in the feature. This suggests that certain features (such as those in rule-based tasks) require explicit engagement in order to be processed.

Given a finite amount of attention, the more numerous the feature dimensions, the less attentional weighting each can receive, and attention will be focused on whichever is currently most relevant (Kruschke & Johansen, 1999; Nosofsky, 1986). This suggests that as the number of features increases, a lower percentage of the features will be learned as directing attention to the learned features overshadows the other features (but see Hoffman & Murphy, 2006). In a rule-based task, more features means more hypotheses to test regarding the classification rule (Markman & Ross, 2003). However, if the amount of attention required to learn the association between a feature and a category is low, then we would predict less overshadowing and that there may still be attention available for additional features (e.g., Hoffman & Murphy, 2006). For example, if attentional weights must sum to 1.0, and each feature requires a weight of .10, a category with four features will allow an organism to easily divide its attention to all of the features. Even if one or two more features are added, equal distribution of attention still allows for learning of all of the features. However, natural categories typically have some

features that can be more helpful for identification than others. For example, all extant species of bird have wings; identifying whether or not something has wings is the fastest way to tell if it might be a bird or not. In this case, it is not beneficial to divide attention equally. In a situation like this, in which one feature receives a disproportionate amount of attention, other features may receive little to no attentional weight, and may not be learned (Kruschke, 1992). Recognizing a bird is easy, but recognizing a specific species takes dedicated effort to learning its unique features. In other words, when one feature is exceptionally diagnostic of category membership, and accuracy is highly important, most, if not all attention, will be focused on the diagnostic feature; other features may not be learned.

Studies of cue competition in category learning

Error driven category learning models (e.g., Ashby et al., 1998; Gluck & Bower, 1988; Kruschke, 1992; Nosofsky et al., 1994) share similarities with error-driven associative learning models (e.g., Pearce & Bouton, 2001; Rescorla & Wagner, 1972). For example, once the relationship between a highly predictive feature and a category label is learned, adding additional features should have no effect on the distribution of attentional weights. This applies even if the additional features also predict category membership, so long as failing to learn them has no effect on performance (Hoffman & Murphy, 2006; Rehder & Murphy, 2003). In other words, both associative learning and category learning models suggest that learning should be guided towards the most relevant predictor for classification (Gluck & Bower, 1988; Kruschke, 1993; Love et al., 2004; Pearce & Bouton, 2001).

However, category learning may be fundamentally different from associative learning. For example, participants might be motivated to learn more features than strictly necessary in

order to generalize more efficiently, because they are interested, or especially in the case that the outcome of a categorization decision could have strong positive or negative outcomes (Blair et al., 2009; Markman, 1989; Murphy, 2002; Lau et al., 2020; Smith & Medin, 1981).

Athough there is some debate (Maes et al., 2016; Urcelay, 2017) cue competition in associative learning is generally well-supported, especially in animal models (e.g., Acebes et al., 2009; Blaser et al., 2004; Couvillon & Bitterman, 1989; Dickinson et al., 1984; Jennings & Kirkpatrick, 2006; Kamin, 1969; Lattal & Nakajima, 1998; Melchers et al., 2005; Pavlov, 1927; Prados et al., 2013; Wagner et al., 1968). However, the literature regarding cue competition in category learning is comparatively sparse. There are few publications, the ones that are available made very different methodological choices, and overall, the results are mixed (Bott et al., 2007; Lau et al., 2020; Murphy et al., 2017). *Neuroimaging studies* related to cue competition in a category learning context are seemingly nonexistent.

In one behavioral study investigating cue competition in category learning, the researchers were interested in whether or not aversive diagnostic features would draw attention to or away from other features (Murphy et al., 2017). For example, consider the sting of a yellow jacket: does the salience of the sting enhance or diminish what else you learn about the creature? It is possible that the sting draws so much of your attention that you fail to notice, for example, the striped pattern of its exoskeleton. Alternatively, you may be so motivated to learn to avoid the creature that your attention is drawn to its other features – it becomes crucial that you learn exactly what it looks like. Murphy et al. (2017) hypothesized, contrary to popular learning theories, that a highly salient feature would *increase* learning of other features rather than decrease it. They conducted several different category learning

experiments using one-away family resemblance categories in which a given feature value is associated with a category most of the time (except for its presence in one stimulus from another category). Their primary design concerned the influence of a noxious, salient feature (a shock representing a bug's "sting") on the learning of other non-noxious features.

Pre-training with the salient feature led to higher rates of learning of the salient feature compared to other features, but without this prior training, all features were learned equally well. They did not test pre-training with the "non-salient" features, so it is not possible to say whether it was the pre-training alone or the salience that caused the overshadowing. Interestingly, when the salient feature was perfectly diagnostic, as opposed to being diagnostic for only five of the six stimuli for each category, the learning of "non-critical" features was enhanced.

Lau and colleagues hypothesized that cue competition effects should be present in category learning, but that previous studies may not have been sensitive enough to measure them (Lau et al., 2020). Across three tasks, two groups of participants completed a category learning task. One group was presented with stimuli that contained only continuously varied features, while the other group saw the same stimuli with the inclusion of a binary feature. After a fixed amount of training, the first group merely continued the task without feedback, however, as the second group continued, they found that the binary feature had been removed. The researchers believed that the group which learned the category with the binary feature would have experienced overshadowing, which would result in inability to perform the task without the binary feature and the feedback. The results supported this prediction,

consistent with a single, salient feature (i.e., deterministic feature value), "overshadowing" learning of the other category features.

fMRI evidence for neural systems supporting category learning

There is no single neural region that supports the entire set of functions required for category learning (Seger & Miller, 2010). In fact, there are a myriad of distributed interactive systems involved in the different aspects of category learning, each differentially engaged depending on context (Reber et al., 2003; Zeithamova et al., 2008) and the category structures involved (Nomura & Reber, 2008), in such a way that multiple different neural systems may all be active simultaneously (Poldrack & Foerde, 2008; Rao et al., 1997; Roeder et al., 2017; Smith & Grossman, 2008). Key regions include the basal ganglia, visual cortex, parietal lobe, motor cortices, medial temporal lobe, and prefrontal cortex.

BASAL GANGLIA

The basal ganglia are a central hub for a wide variety of different processes (Gabrieli, 1998; Seger & Miller, 2010), such as motor control (Redgrave et al., 2010), and cognitive coordination (Stocco et al., 2010), and are recruited in many different types of category learning tasks (Nomura et al., 2007; Poldrack et al., 1999; Seger & Cincotta, 2005; Zeithamova et al., 2008), particularly those which utilize reinforcement learning (Villagrasa et al., 2018). There are several different functional subregions of the basal ganglia, the largest being the striatum, which can be functionally divided into the dorsal and ventral striatum, the former of which can be further subdivided into the caudate and putamen. The basal ganglia are active during the planning and execution of motor actions (Monchi et al., 2006), and help connect

stimuli with responses by learning the patterns and associations that define categories (Ashby et al., 1998; Seger & Cincotta, 2005).

Cortical regions communicate with the basal ganglia via corticostriatal loops: outputs from cortex travel to the striatum, and then back to the cortex through the thalamus (Alexander et al., 1986; Flaherty & Graybiel, 1991; Kemp & Powell, 1970). These recursive loops may be the basis for the construction of categories and abstractions; they all work together during learning (Seger, 2008; Seger & Cincotta, 2005; Seger & Miller, 2010). There are several functionally separate loops (although they somewhat overlap): executive, motivational, visual, and motor. During category learning, these loops help select category representations, and activate the appropriate responses (Lopez-Paniagua & Seger, 2011; Seger, 2008). The visual loop receives input from visual cortex, and communicates this information to executive and motor loops; through this mechanism, relevant responses can be chosen and implemented (Ashby et al., 1998). Feedback is processed by the executive loop, which also interacts with working memory during the selection of mental sets. Reward and feedback is further processed by the motivational loop (Seger & Miller, 2010). The executive and motivational loops are active early in category learning, when feedback is the most important (Williams & Eskander, 2006), and over time, as expertise develops, the motor loop takes over (Seger et al., 2010).

VISUAL CORTEX

Category learning in the visual modality is perhaps the best understood of all the sensory modalities. The ventral stream, a network of regions along the occipital and temporal lobes, is generally considered to encode object categories (DiCarlo et al., 2012; Gross, 1994). Category structure can be represented in regions as early as the primary visual and extrastriate

cortices, which can both represent simple category features like lines and shapes (Ashby & Ell, 2001; Reber 1998; Ullman et al., 2002). This is evidenced by changes in neural representations as categories are learned (Folstein et al., 2015; Reber et al., 1998). More complex categories, such as types of animals, have even been decoded from activity in low-level visual areas using MVPA (Cox & Savoy, 2003), however, it is difficult to draw conclusions about the category selectivity of these regions when there is no agreement regarding the information high-level areas in the ventral stream use to represent categories: complex category information, or low-level visual properties of objects (Andrews et al., 2015; Bracci et al., 2017; Coggan et al., 2016).

PARIETAL LOBE

Regions in the parietal lobe, such as the intraparietal sulcus and adjoining regions of the parietal lobules, are generally involved in visuospatial attention (Daniel et al., 2011; Little et al., 2004; Little & Thulborn, 2006), and function as part of a "multiple-demand network" (Assem et al., 2020; Duncan, 2010; Heekeren et al., 2008; LaBar et al., 1999) which is active during cognitively challenging tasks. Regarding category learning, specifically, parietal regions appear to link perceptual information (i.e., features) with potential motor responses (Aizenstein et al., 2000; Bracci et al., 2017; Grossman et al., 2002; Seger & Miller, 2010; Yi et al., 2016). For example, bilateral parietal activation was found in one study in which participants were required to analyze a set of category features in order to apply a rule for making an appropriate categorization decision (Seger et al., 2015).

MOTOR CORTICES

Many motor regions are active during categorization due to the response-dependent nature of most category learning tasks, including the primary, secondary, and premotor

cortices, as well as the motor components of the basal ganglia. During category learning tasks, a stimulus is associated with a category label, which is, in turn, associated with a motor response (Maddox et al., 2010; Sanders, 1971). The premotor and motor cortices receive input from regions like the prefrontal cortex, under guidance from the striatum, which all help guide the selection of the appropriate category response (Ashby et al., 1998; Ashby & Maddox, 2011; Braunlich & Seger, 2016). The premotor cortex has even been shown to contain feature representations along with response information (Wallis & Miller, 2003). Over time, as a task is learned more efficiently, these motor systems become more dominant while other initially active systems become less dominant (Seger & Miller, 2010). This shift in dominance towards motor areas typically signals an increased degree of automaticity (Helie et al., 2010; Muhammad et al., 2006)

MEDIAL TEMPORAL LOBE

The medial temporal lobe (MTL) plays several roles in category learning. It is involved in learning rules, as well as exceptions (Davis et al., 2012; Love et al., 2004), and can also aid category learning by acquiring the initial representations of individual exemplars (Seger & Cincotta, 2005; Seger et al., 2011). Although the MTL seems to primarily represent singular instances of category exemplars (Seger & Peterson, 2013), it can also help learn the relationships between stimuli (Giovanello et al., 2008; McCormick et al., 2010).

PREFRONTAL CORTEX

The striatum and MTL co-activate with other regions, notably the prefrontal cortex (PFC)(Poldrack & Rodriguez, 2004) – an area which, additionally, subserves category learning through the acquisition of new categories, primarily during early phases of learning (Asaad et

al., 1998; Seger & Miller, 2010). For example, the PFC has been shown to represent category boundaries, as well as individual features (Jiang et al., 2007), especially in the case of rule-based categories (Muhammad et al., 2006), such as those used in the present study. The PFC also appears to integrate perceptual information with executive processes, and has demonstrated greater activity during the feedback phase of category learning (Monchi et al., 2001; Toni et al., 2001). It is also active during hypothesis testing (Milton & Pothos, 2011).

RULE-BASED TASKS AND THE PRESENT STUDY

Different types of category learning tasks can load on different neural regions and systems. There are commonalities across tasks: visual category learning requires that the visual systems be active, and anything which requires a response will recruit motor systems. Past that, there can be substantial differences regarding the active systems. For example, rule-based and information-integration classification tasks appear to primarily load onto MTL and striatal systems, respectively (Ashby et al., 1998, 2020; Nomura et al., 2006; Waldron & Ashby, 2001; Xing & Sun, 2017).

Based on traditional models, rule-based tasks, such as the one used in the present study, should primarily recruit category learning regions such as the dorsolateral prefrontal cortex, anterior cingulate, caudate head, and medial temporal lobe (Nomura et al., 2007; Rao et al., 1997; Seger & Cincotta, 2006; Spiering & Ashby, 2008), regions important to the executive loop and working memory function. However, other regions generally useful during category learning should be active as well, including parts of the ventral stream and the multiple demand network. Lastly, relevant motor regions should be active as well.

Present study

Cue competition, and overshadowing, specifically, is present during category learning under specific conditions, as suggested by my pilot work presented below, and by others (e.g., Lau et al., 2020; Murphy et al., 2017). However, open questions remain regarding the neural correlates of feature representation. To my knowledge no previous neuroimaging research has examined the processing of features under cue competition during a category learning task.

The present study was concerned with the nature of the neural representations which develop during cue competition and overshadowing in a rule based, supervised category learning task. Participants were trained using the same paradigm and stimuli as in Lau et al. (2020) while simultaneously having fMRI data recorded. The stimuli contained two diagnostic features, one highly salient (i.e., easily to perceptually separate the values for each category) binary feature, and one more difficult, continuous feature. After a fixed number of training trials, participants completed a transfer phase in which the binary feature was removed, as well as the feedback. I predicted that the salient, binary feature would overshadow learning of the continuous feature, resulting in impaired performance during Transfer.

The primary questions of interest concern the nature of the neural representations that develop during learning and transfer. The primary analysis was a whole-brain searchlight RSA which attempted to fit three different hypothetical model matrices to the neural data. The first model was related to the binary feature and looked for areas that represented stimuli within each category equally which would suggest a representation of the category and/or the binary feature was present. The second model was effector-specific; it focused on the individual hands used to make the motor response. The third model was related to the continuous feature and

looked for areas that represented stimuli based on their perceptual similarity which would suggest a representation of the continuous feature was present. I used this RSA to test whether all features were represented equally and whether the more difficult continuous feature would be processed at all in the presence of a easy, highly salient one. Of interest were the regions which might code for discrete categories, motor responses, and/or features. I predicted that early perceptual regions would represent all features regardless of phase of learning as evidenced by a fit to the continuous feature model, but that higher order perceptual regions and frontoparietal regions would show overshadowing of the less salient, continuous feature as evidenced by lack of fit to the continous feature model. I further predicted that frontoparietal regions would code for discrete category membership as evidenced by a fit to the binary feature model, whereas motor regions would code for the motor response used to indicate category membership as evidenced by a fit to the effector-specific motor model.

To that end, several studies were conducted: a behavioral pilot study to verify the presence of overshadowing effects in our lab, an fMRI pilot study to assess feasability and examine whether or not the task recruited expected regions in a univariate analysis prerequisite to performing RSA analyses, and an RSA study, which was a full-scale version of the fMRI pilot study.

Chapter 2 – Behavioral Pilot

Description

The primary goal of this behavioral pilot was to produce an overshadowing effect in our lab. fMRI data collection is costly, both in time and money, and since overshadowing is required to answer my research question, it was important that the task I used consistently elicited it. Experiment 2 of the Lau et al. (2020) paper was chosen as a template for my design, and additionally, given that the aforementioned study had freely available materials, I used the actual stimuli from their experiment 2 (Figure 4), which elicited the strongest overshadowing effect of the two sets of stimuli used in the studies reported in their paper.

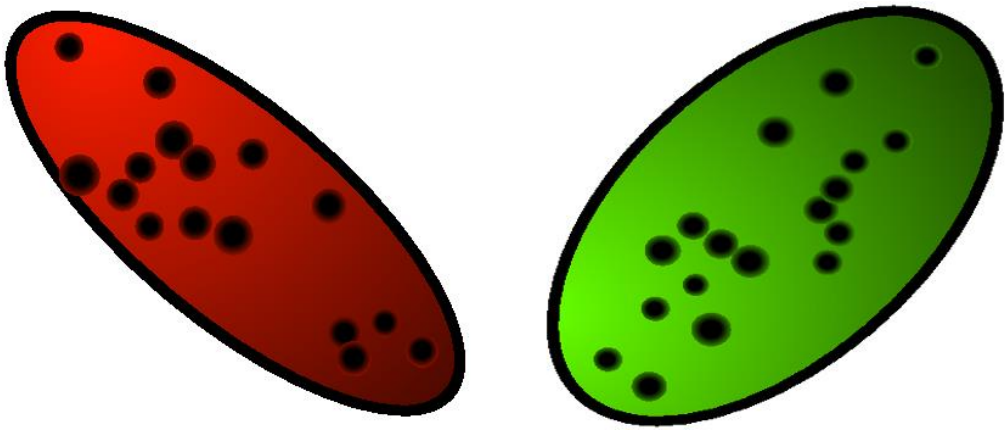

Figure 4: An example of two stimuli from the Lau et al. (2020) paper.

The stimuli in this pilot could be categorized based on either a salient, binary feature (rotation) or a continuous one (number of dots), or both. The stimuli also varied in two other dimensions unrelated to category membership: color and height. At some point in the task, the binary feature was removed, and performance became dependent on how well the continuous feature had been learned. My hypothesis was that a drastic reduction in performance would

occur in the second phase of the task when participants had to rely on the continuous diagnostic feature. This would replicate Experiment 2 in the Lau et al. (2020) paper, and would both demonstrate the robustness of the cue competition effect in this specific category learning task, as well as help fine-tune the parameters of the task to maximally elicit this effect for use in the scanner. Behavioral piloting was also important because some minor alterations were made to the stimuli (see Methods). A failure to replicate might suggest that either alternative factors beyond cue competition effects were responsible for the overshadowing observed in the Lau et al. paper, or the alterations to the stimuli were not as minor as assumed, and fundamentally altered the factors that lead to cue competition in the original study.

Methods

Participants

Participants (*n*=22) were recruited from students in the PSY100 and PSY250 research pool; students in those courses are each required to participate in research studies for class credit. In total, 22 students participated in two different groups which had minor procedural differences, reported separately as Group 1 and Group 2.

Stimuli

Experiment 2 in the Lau et al. (2020) study used oval-shaped, cell-like stimuli; they varied on the number of internal black dots, the color of the space between the dots, the height of the shape, and importantly, on the angle of rotation. The features for a given stimulus were, in general, defined by a mean parameter value and standard deviation. The mean values for the rotation feature (the binary feature) were 45° and -45° with a standard deviation of 10°; this made for a relatively simple task, perceptually ("Is it angled to the left or right?"). The height

and color parameters were also each sampled from unique Gaussian distributions, however, the mean and SD were identical for each category. The mean for the height feature was 220 pixels and the standard deviation was 30 pixels. The parameter value for the color feature for a given stimulus was programmed to randomly sample a value from the 360 degree HSB color wheel. Each of these three features had consistent parameter values across participants and categories. The number of dots (continuous feature) for each stimulus was chosen in a purely random order from a range of values unique to each category: 7-16 dots for one category (-45° rotation), and 18-27 dots for the other (45° rotation). The rotation and number of dots defined category membership, whereas the color and height were random across categories. A large bank of stimuli was generated using these parameters, and each participant had a randomly pre-defined set sampled for the task, with an equal number of stimuli from each category. For Group 1, stimuli were sampled in a purely random fashion from the stimulus bank, however, for Group 2, stimuli were sampled pseudo-randomly, so that an equal number of stimuli per dot bin were sampled.

Procedure

The task was presented to participants using PsychoPy (v2020.2.10; Peirce et al., 2019). Before the task began, participants were told about most of the aspects of the task including that they would be participating in a category learning task, and that accurate performance was the most important goal. Guided by feedback, they would learn through trial-and-error, and at some point, the feedback would be removed. Finally, no two stimuli would be identical. On the task computer, instruction screens presented prior to the task showed an example of two stimuli, one from each category. Once instructions were completed, the task began. First,

participants completed the Training phase. On each trial, they saw a stimulus, made a response, and received feedback. A blank screen was presented between trials, and between the stimulus and the feedback (see Figure 5 for timings). After a set number of trials, the Transfer phase began. At the beginning of the Transfer phase participants were informed that, not only would feedback be removed, but orientation would also be removed by presenting all stimuli horizontally across both categories; anything else they learned about how to categorize the stimuli would still apply.

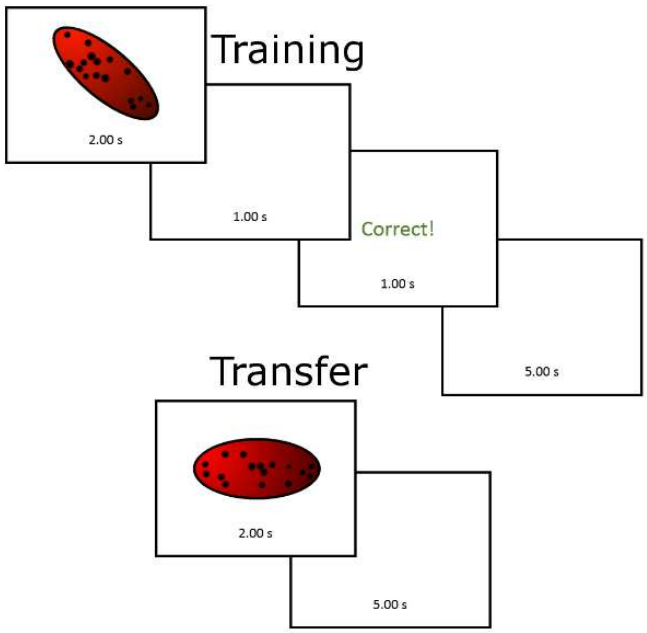

Figure 5: The task design used for the pilot studies. The Training phase is presented on top, and the Transfer phase on the bottom.

For Group 1, the stimuli were presented for three seconds. However, when presented for this relatively long duration, overshadowing was only evidenced in about half of the participants. Therefore, the stimulus duration was reduced to two seconds for Group 2, at which point all but two participants exhibited the overshadowing effect. Additionally, for Group 1, there were 150 Training trials, and 50 Transfer trials. However, given that all but one of the

participants had effectively mastered the task by 100 trials, for Group 2, the number of trials was adjusted: each phase was assigned an equal number (n=100). This was done so that 1) there would be more trials to use for analysis in the Transfer phase, and because 2) there was no reason to have such a long Training phase when participants were learning the task quickly.

Results

All data were analyzed in Python (3.7.10; Python Software Foundation) using these toolkits: Pandas (1.3.0; Reback et al., 2021), NumPy (1.20.3; Harris et al., 2020) and SciPy (1.6.2.; Virtanen et al., 2020) During Training, average accuracy in the first block for Group 1 began at 87.5%, and increased to 96.5% in the last block. For Group 2, accuracy began at 92.9%, and increased to 99.7%. During the Transfer phase, overall accuracy for Group 1 and Group 2 was 69.2% and 59.3% respectively. However, excluding those who did not demonstrate overshadowing, the accuracy during Transfer for Group 1 and Group 2 was 53.8% and 46.8% respectively. The overall average accuracy in each phase, for all participants (i.e., both those who did and did not demonstrate overshadowing), is plotted in Figure 6.

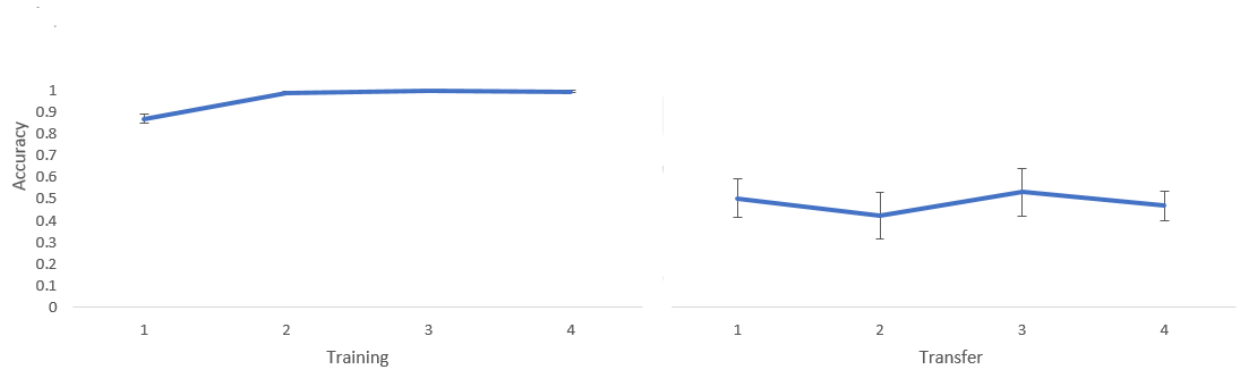

Figure 6: Behavioral data from the pilot task. The Training phase is on the left, and the Transfer phase is on the right.

Summary

This pilot task elicited a strong overshadowing effect, as demonstrated by a drop to chance levels after the removal of the binary feature. The results replicated Lau et al. (2020), and suggested that both 1) an overshadowing effect could be elicited in our lab, and 2) the alterations to the stimuli had minimal effect.

Regarding the alterations, although the number of stimuli for each category was equal, each participant had different numbers of stimuli for each dot bin (e.g., two stimuli with six dots, five stimuli with seven dots, etc.). This was rectified for the fMRI pilot, and each participant was instead assigned an equal number of stimuli from each dot bin (e.g., two stimuli with six dots, two stimuli with seven dots, etc.).

Chapter 3 - fMRI Pilot

Description

In this study, participants performed a two-part category learning task identical in structure to the behavioral pilot previously described. To reiterate: in the first phase (Training), participants were presented with oval-shaped stimuli (Figure 4) and asked to respond as to which of two categories each belonged. These stimuli had one binary, salient feature, one continuous feature, and two random features. At a certain point, the binary feature was removed (Transfer phase), and participants were asked to continue to categorize the stimuli, albeit without feedback. The goal of this study was to elucidate neural factors related to how features are represented by utilizing a task which had previously demonstrated overshadowing. My behavioral hypothesis was identical: having the binary feature present would overshadow learning of the continuous feature. Therefore, in the Transfer phase, removal of the binary feature would cause a reduction in accuracy relative to the Training phase due to dependence on the binary feature for accurate categorization.

Importantly, participants were scanned using fMRI during the behavioral task. For this measure, it was hypothesized that, given the basic category learning structure of the task, a univariate analysis would unveil regions common to category learning tasks (Seger & Miller, 2010) during the Training phase, such as those described in the Introduction (e.g., visual cortex, basal ganglia, etc.).

Methods

Participants

Participants were recruited from an ad posted to the Psychology and MCIN graduate student listservs, as well as from word-of-mouth. In total, data from 9 participants was included in the analyses. Data from one additional participant was excluded due to interference from a permanent wire retainer.

Stimuli

Stimuli were again presented using PsychoPy, and were identical to those used for Group 2 in the behavioral pilot. The defined list of stimuli was again randomly generated for each participant, with stimuli from each dot bin represented equally.

Procedure

Participants were first given verbal instructions, which were effectively identical to those in the behavioral pilot (including an additional request that participants remain still while in the scanner).

They were then given a practice task outside of the scanner. They were told that the purpose of the practice task was to get them used to the format of the task they would complete in the scanner (e.g., timing of stimuli, what feedback looks like, etc.), but that the stimuli would be different. Written instructions presented on the computer before the practice task began included a brief recap of the verbal instructions. Twenty stimuli were presented: the stimuli consisted of a blue star and an orange square.

In the scanner, two tasks were completed in a counterbalanced order: the present task, and a separate unrelated experiment (not reported here). The flow and timings were identical

to those used for Group 2 of the behavioral pilot (see Figure 5 for specific task-flow and timings). Because having several, shorter sets of data can be beneficial for MVPA analyses, participants were shown 20 trials at a time (a "run"). Four identical runs of Training data were collected in this manner, and each run contained one stimulus for each dot bin, presented in a random order. Afterwards, the Transfer phase began. During this phase, another four runs of 20 trials with identical timings (except for the absence of feedback) were collected. After four runs of the Transfer phase, the task, and scanning, was complete.

Image Acquisition

Brain images were acquired at Colorado State University using a 3 T Siemens Magnetom Skyra scanner with a 64-channel head coil. A single high-resolution T1-weighted anatomical scan was acquired at the beginning of each scanning session with these parameters: 0.9 mm isotropic voxels; TE = 2.32 ms; TR = 2400 ms; field of vision [FOV] = 230 mm; flip angle = 8°. Each functional scan was acquired using an accelerated Echo-Planar imaging (EPI) sequence with these parameters: TE = 38 ms, TR = 859 ms; 2.5 mm isotropic voxels; 60 transverse slices with 2.5 mm thickness; flip angle = 52°; FOV = 210 mm, 84 X 84 matrix.

fMRI Preprocessing

The results of this study come from preprocessing performed using a Docker image of fMRIPrep 20.2.1 (Esteban et al., 2019; Esteban et al., 2020), which is based on Nipype 1.5.1 (Gorgolewski et al., 2011; Gorgolewski et al., 2018). Many internal operations of fMRIPrep additionally use Nilearn 0.6.2 (Abraham et al., 2014), mostly within the functional processing workflow.

ANATOMICAL DATA PREPROCESSING

The T1-weighted (T1w) images were corrected for intensity non-uniformity (INU) with N4BiasFieldCorrection (Tustison et al., 2010), distributed with ANTs 2.3.3 (Avf et al., 2008), and used as T1w-references throughout the workflow. The T1w-references were then skull-stripped with a Nipype implementation of the antsBrainExtraction.sh workflow (ANTs), using OASIS30ANTs as target template. Brain tissue segmentation of cerebrospinal fluid (CSF), white-matter (WM) and gray-matter (GM) was performed on the brain-extracted T1w using fast (FSL 5.0.9; Zhang et al., 2001). Volume-based spatial normalization to one standard space (MNI152NLin2009cAsym) was performed through nonlinear registration with antsRegistration (ANTs), using brain-extracted versions of both T1w reference and the T1w template. The following template was selected for spatial normalization: ICBM 152 Nonlinear Asymmetrical template version 2009c (Fonov et al., 2009),

FUNCTIONAL DATA PREPROCESSING

For each of the 8 BOLD runs per participant, the following preprocessing was performed. First, a reference volume and its skull-stripped version were generated using a custom methodology of fMRIPrep. The BOLD reference was then co-registered to the T1w reference using flirt (FSL 5.0.9; Jenkinson & Smith, 2001) with the boundary-based registration (Greve & Fischl, 2009) cost-function. Co-registration was configured with nine degrees of freedom to account for distortions remaining in the BOLD reference. Head-motion parameters with respect to the BOLD reference (transformation matrices, and six corresponding rotation and translation parameters) were estimated before any spatiotemporal filtering using mcflirt (FSL 5.0.9; Jenkinson et al., 2002). BOLD runs were slice-time corrected using 3dTshift from

AFNI 20160207 (Cox & Hyde, 1997). The BOLD time-series were resampled onto each participant's original, native space by applying the transforms to correct for head-motion. These resampled BOLD time-series will be referred to as preprocessed BOLD in original space, or just preprocessed BOLD.

The BOLD time-series were resampled into standard space, generating a preprocessed BOLD run in MNI152Nlin2009cAsym space. First, a reference volume and its skull-stripped version were generated using a custom methodology of fMRIPrep. A confound time series derived from head motion estimates was calculated using the preprocessed BOLD and used, later, during the GLM phase. All resamplings were performed with a single interpolation step by composing all the pertinent transformations (i.e., head-motion transform matrices and co-registrations to anatomical and output spaces). Gridded (volumetric) resamplings were performed using antsApplyTransforms (ANTs), configured with Lanczos interpolation to minimize the smoothing effects of other kernels (Lanczos, 1964). Spatial smoothing was performed using the "Smooth" function of SPM12 (version: 7771; Friston et al., 2014) in MATLAB (version: 9.6.0.1135713 (R2019a) Update 3; The MathWorks Inc.) with an isotropic 8 mm full-width at half-maximum Gaussian kernel.

Univariate analyses

Smoothed data were subjected to a univariate analysis using the general linear model (GLM). All stimuli were included as a single regressor, and voxels were identified with activity which correlated with the presentation of each stimulus. One additional regressor was included for feedback, and as well as six for head motion estimates (rotation and translation along the x-, y-, and z-axis). The duration for both stimuli and feedback was calculated as the time that each

item was physically present on the screen. This was identical for the feedback, but due to an issue with PsychoPy coding, the duration of the stimuli was slightly variable, fluctuating around 1.68 s. An implicit baseline was used which consisted of unmodeled timepoints.

All analyses were cluster corrected for multiple comparisons using the family-wise error rate method with an initial (uncorrected) threshold of $p<.001$ and an FWE corrected threshold of $p<.05$. Cluster-extent correction is suggested for studies with low power as opposed to voxel-level corrections, which can be overly stringent in such cases (Woo et al., 2014). The specific *p*-values I chose have been demonstrated to be appropriate for parametric analyses (Bansal & Peterson, 2018; Eklund et al., 2016). Although the spatial specificity is lower than voxel-wise methods (Yeung, 2018), the goal of this analysis was to simply identify general category learning networks.

Results

Behavioral

Data were analyzed using the same software as the behavioral pilot: Pandas, NumPy, and SciPy (same versions). During Training, accuracy increased from 78.9% in the first half (runs 1 and 2) to the 97.8% in the last (runs 3 and 4). During the Transfer phase, accuracy dropped to 46.8%, with most participants experiencing an average drop in accuracy of 54% from the last block of Training to the first block of Transfer. This difference in accuracy between the Training and Transfer blocks was statistically significant ($t(8)=19.3$, $p<.001$, $d=0.99$) using a paired-samples *t*-test. This is consistent with the binary feature overshadowing the other category-relevant feature.

One participant may have reversed the response buttons in the Transfer phase, as their accuracy was well below chance (21.2%). This would also suggest that they were resilient to overshadowing. However, even if their accuracy was reversed and included in the average for the Transfer phase, the overall accuracy for the Transfer condition remained centered around chance (53.1%). Without their data, accuracy was exactly 50.0% in the Transfer phase.

fMRI

Univariate fMRI analysis was performed using SPM12 and MATLAB (same versions as preprocessing). The intent, here, was to verify that basic visual category learning networks were active during the Training phase. To do this, stimuli from the Training phase were compared to an implicit baseline (Figure 7).

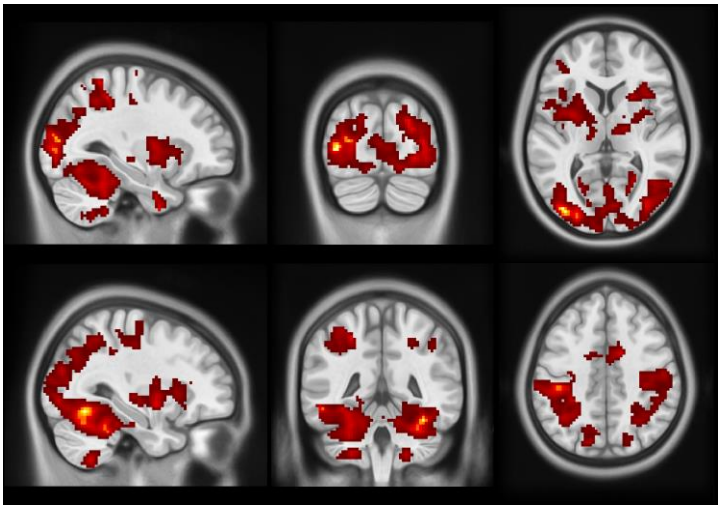

Figure 7: Neuroimaging data from the univariate analysis of the pilot fMRI data. Brighter colors indicate greater *t*-values. All data were initially thresholded at $p<.001$ (uncorrected), and FEW cluster corrected at $p<.05$.

Regions of activity included primary and extrastriate visual areas, portions of the basal ganglia, including the caudate nucleus and putamen, bilateral motor and premotor areas, left ventral temporal cortex, posterior cingulate cortex, right dorsolateral prefrontal cortex, right

fusiform gyrus, left inferior temporal gyrus, right precentral gyrus, lingual gyrus, left insular cortex, and the cerebellum. These locations match previously identified regions of category learning/visuomotor task activity (Poldrack et al., 1998; Seger & Miller, 2010; Seger et al., 2010).

Summary

Overall the pilot fMRI study was successful: participants showed strong and consistent behavioral overshadowing effects, and a univariate analyses found that the task recruited areas typical for rule-based category learning. These were necessary foundations for performing RSA analyses in the following study as they demonstrated that the task was tapping into category learning systems.

The process of running this pilot study suggested several methodological refinements to incorporate into the following study. An open question from the pilot data concerned the processing of the stimulus features; there was a chance that the activity observed during the training phase was in response to rotation, alone. That is, the results may not have been related to feature processing, in general, as much as they were related to perceptual processing specific to rotation. This was addressed by the use of two different task groups, each using a unique, perceptually salient set of features (see Methods - Stimuli).

There may have been an issue with interference from motor response activation, as well. During the task, the location of the motor response for a category never changed, so participants always used the same hand to respond to each category. Therefore, category-specific information and motor-response activation were confounded. Any area which appeared to contain feature information may instead have been reflecting motor response

information. For the following study, categories were assigned to motor responses on a trial-by-trial basis via category labels on the right and left sides of the screen such that category and motor response were orthogonal to each other (see Methods – Stimuli).

Additionally, the timing of the feedback may have been too temporally adjacent to the presentation of the stimulus. Therefore, the interval between stimulus and feedback was increased in the following study. Although SPM does pre-whiten the data to reduce autocorrelation, shorter TRs can lead to more significant autocorrelation (Purdon & Weisskoff, 1998); no software package can entirely remove it, although SPM is still relatively effective (Olszowy et al., 2019). The single-trial modeling used in the following study does help account for collinearity (Mumford et al., 2012), but increasing the time between the stimulus and feedback seemed beneficial (Friston et al., 1995; Woolrich et al., 2001). As this is a rule-based task, the timing of the feedback is less important to learning, anyway (Ashby et al., 2002).

Lastly, between trials participants viewed a totally blank, white screen. One minor change was that, between trials, the white screen was modified to include a fixation cross. Although it may have little to no impact compared to a blank screen beyond increasing bloodflow to visual areas (Gusnard & Raichle, 2001; Patriat et al., 2013), the intention was to help provide standardization to the implicit baseline, and additionally, including a fixation cross has been shown to help reduce drowsiness (Agcaoglu et al., 2019).

Chapter–4 - RSA Experiment

Description

The overall goal was to examine the development of neural representations during overshadowing in a category learning context. In overshadowing, multiple informative features are present, but one captures attention (and dominates learning) due to its salience. As previously discussed, there is disagreement in the literature regarding whether or not overshadowing is beneficial (Murphy et al., 2017) or detrimental (Lau et al., 2020) to feature learning during category learning tasks. Additionally, there is not currently any neuroimaging literature which addresses how feature representations develop during overshadowing. Therefore, this experiment examined how representations develop for informative features during a supervised, rule-based category learning task. This included whether or not representations developed for the overshadowed feature, and if so, how they were organized.

In order to accomplish this, participants first completed a Training phase in which they were presented with one stimulus at a time in a supervised manner and tasked with categorizing it. There were two informative features present – one was salient by virtue of its difficulty (it was very easy to learn), effectively having only two possible values, while the other was more difficult, and instead, the feature values were on a spectrum. After the Training phase, participants began a Transfer phase in which the salient feature was removed. They still had to categorize the stimuli, however, there was no longer any feedback present, leaving them without the salient feature, as well as any way to identify whether their responses were accurate or not. In addition to the category learning task, a 1-back task was also administered

to the participants. It was intended that this task would function as a feature localizer due to the nature of how stimuli must be processed holistically in order to succeed on the task.

I hypothesized that, similar to the results of the behavioral pilot study, participants would demonstrate an overshadowing effect in which they quickly learned *only* the salient discrete feature during Training and would therefore perform poorly during Transfer when it was removed. For the RSA, I hypothesized that early perceptual areas would represent all perceptual features during both phases of the task as evidenced by a fit to the continuous perceptual distance model, but that higher order perceptual and frontoparietal regions would show overshadowing of the continuous feature as evidenced a lack of fit to the continuous perceptual distance model. I further hypothesized that frontoparietal regions would represent discrete category membership as evidenced by a fit to the binary feature model, and that motor regions would represent motor responses as evidenced by a fit to the motor model.

Methods

Participants

Twenty-four participants were recruited from South China Normal University. This sample size is in line with relatively recent estimates for the median sample size in top neuroimaging journals (Poldrack et al., 2017; Szucs & Ioannidis, 2017, 2020). One participant was excluded for not performing the task correctly, leaving twenty-three participants for the analyses.

Stimuli

Stimuli were presented using PsychoPy and were similar to those used in the pilot studies in that they were oval-shaped and varied on parameters related to height, color, rotation, and the number of dots. Each task used a unique set of stimuli.

For the 1-back task, one set of stimuli was used for all participants: every feature varied continuously and had a randomly sampled feature value. Stimuli could have one of a 1-360 degree range of rotational values, one of a range of 7-27 dots, one of a range of heights of 145-295 pixels, and one of an HSB color wheel value ranging from 1-360 degrees.

For the category learning task, in order to control for perceptual effects, participants were randomly assigned to counter-balanced groups differing in regard to which feature was binary and which was continuous. One group had categories defined by the same features as the pilot studies, rotation (binary) and number of dots (continuous), and the other had categories defined by color (binary) and height (continuous). These groups were labeled the "rotation" and "color" groups, respectively.

For the rotation group, the rotational angles of each category were identical to the pilot studies (mean: -45° and 45°; SD: 10°), as were the number of dots (7-16 or 18-27). The height and color feature values, however, were instead completely random (i.e., there was no longer a mean and standard deviation used to select these features). Height was chosen randomly from between 145-295 pixels, and color was chosen randomly from the full range of HSB color wheel values (1-360 degrees) for both categories.

For the color group, the HSB color wheel defined two categories using 0° (red) and 180° (blue) as the mean values and a standard deviation of 8° for both categories. Height values for

each category were between either 183-219 pixels or 221-257 pixels (and only occurred in 4-pixel increments; e.g., 183, 187, 191, etc.). Rotational angle and number of dots were chosen randomly. Angle varied from 1-360 degrees, while the number of dots ranged from 7-27 for both categories.

Additionally, for both the 1-back and the category learning task, the response buttons for each task choice alternated pseudorandomly throughout the task. On each trial, two labels were presented: one on the bottom corner of each side of screen. An arrow pointing to the side of the screen also accompanied each label (e.g., for the label on the bottom-right corner of the screen, an adjacent arrow pointed to the right). The side of the screen and accompanying arrow signified which of two response boxes participants should use to respond to which category. An arrow/label on the left meant that the left-hand response box should be used if the label belongs to the on-screen stimulus. During the 1-back task, the words "yes" and "no" were used for each label. For the category learning task, the imaginary category names, "Goopie" and "Plunkett", were displayed as the two labels. Each stimulus was presented twice – once for each response button layout. Although the stimulus selection was pseudorandom (i.e., dot bin and height values were controlled), the presentation order for each participant was randomized.

Procedure

Outside of the scanner, participants completed the same Training task as in the pilot study in order to familiarize themselves with the basic format of the category learning task. Once in the scanner, they first completed a 1-back task. They were instructed to respond "yes" or "no" depending on whether the currently presented stimulus matched the previously

presented one, and to ignore responding to the first stimulus (which has no prior stimulus for comparison). Each stimulus was presented for 2 seconds, followed by a blank screen lasting 3 seconds. There were four small blocks of 20 stimuli, separated by 30 second breaks, with a pseudorandom sampling, counterbalancing the number of "yes" and "no" responses, as well as the times each hand responded to each answer (i.e., the left hand responded "yes" 40 times, and "no" 40 times, same as the right hand). The 1-back task was included in order to have the option of defining participant-specific functional ROIs for visual features. However, this analysis was ultimately unnecessary, therefore the 1-back task will not be further discussed. There was a small break once the task finished while the next task loaded that lasted at least two minutes.

Participants then completed a category learning task with a structure identical to the pilot fMRI study. First, there was a Training phase (with feedback), including the salient, binary feature, followed by a Transfer phase in which that feature (as well as the feedback) was removed. Both the Training and the Transfer phase consisted of four blocks of 20 trials each.

Image acquisition

Brain images were acquired at South China Normal University using a 3 T Siemens Magnetom Prisma Fit scanner with a 64-channel head coil. A single high-resolution T1-weighted anatomical scan was acquired at the end of each scanning session with these parameters: 0.8 mm isotropic voxels; TE = 2.07 ms; TR = 1800 ms; field of vision [FOV] = 256 mm; flip angle = 9°. Each functional scan was acquired using an accelerated Echo-planar imaging (EPI) sequence with these parameters: TE = 31 ms; TR = 1000 ms; 2.4 X 2.4 X 3.5 mm voxel size; 39 transverse slices with 3.5 mm thickness; flip angle = 70°; FOV = 211 mm, 88 X 88 matrix.

Image preprocessing

The steps and software used for preprocessing were identical to those used in the pilot study. In short, anatomical scans were corrected for non-uniformity and skull-stripped. Functional data were motion corrected, slice-time corrected, and coregistered to the participant's anatomical space. RSA was performed in native participant space on unsmoothed data.

Representational similarity analysis

The RSA was performed on single-trial, *t*-statistic images (Misaki et al., 2010), generated using the Least-Squares Separate procedure (LSS; Mumford et al., 2012; Mumford et al., 2014) in SPM. An LSS model takes as regressors the trial to be modeled, all other trials of that type, and then finally all other trials (in addition to any other regressors e.g., motion parameters). The parameter estimate for this first trial is retained and brought forward into subsequent analyses. This method can reduce collinearity when creating single-trial models for fast event-related fMRI designs (Mumford et al., 2012). These single-trial *t*-statistic images were used as the basis for a whole-brain searchlight as well as a more targeted ROI approach.

MODELS

There were three primary models used for comparison in the RSA, hereafter referred to as "category", "motor", and "perceptual distance" (see Figure 8). The category matrix focused on strict binary categoricity; all stimuli from one category were defined as equally similar to each other, and equally dissimilar from all stimuli in the other category. The motor matrix was arranged by motor response; all responses from the left hand were defined as equally similar to one another, and equally dissimilar from all right-handed responses. In the perceptual distance

matrix, stimuli were arranged by the perceptual similarity of the continuous category feature; each stimulus was more similar to its immediate neighbor in perceptual space than to any other stimulus. As an example, for the group which had dots as the continuous category characteristic, stimuli were arranged from least to most dots, and stimuli with the least dots shared the greatest dissimilarity with stimuli that had the most dots.

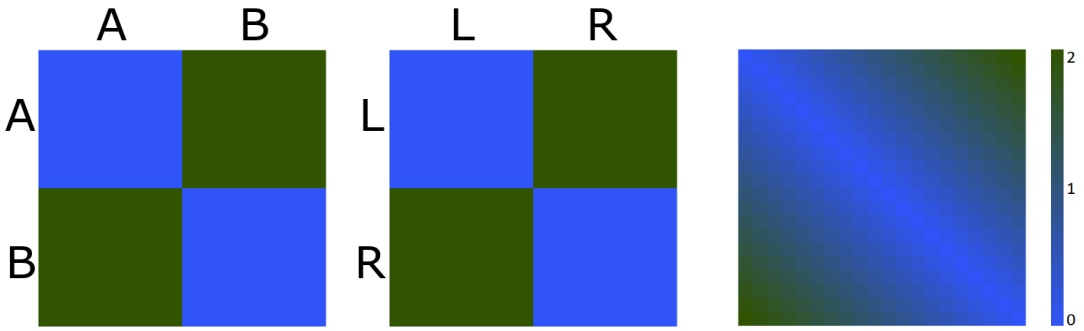

Figure 8: Different hypothetical model dissimilarity matrices are displayed. Blue represents similarity and green represents dissimilarity. The grids are composed of task stimuli, with each grid modeling different relationships between the stimuli; the stimuli along are identical in each matrix, varying only in their arrangement. *Left*: This matrix represents binary categoricity. All stimuli from the same category are equally similar, and all stimuli from the opposite category are equally dissimilar. *Middle*: This matrix represents a binary motor response. All stimuli that received a left-hand response are equally similar to each other and equally dissimilar from stimuli which received a right-hand response. *Right*: This matrix represents perceptual distance for the continuous value dimensions (e.g., dot number and height), with stimuli arrayed along the top and side axes in order of similarity. There is very little perceptual difference between a stimulus and its neighbor on the grid in perceptual space. Dissimilarity gradually increases as stimuli progress from left-to-right/top-to-bottom.

WHOLE-BRAIN ANALYSIS

The data selected for each whole-brain searchlight RSA came from either the Training or the Transfer phase. Data from the Training phase used an average of the second, third, and fourth runs, while data from the Transfer phase used an average of all four runs. The first run of the Training phase was excluded due to the fact that performance had not stabilized yet and the present study was interested in representations that develop during skilled performance. The *t*-statistic images for each trial were used to create neural RDMs (see Figure 8). The

configuration of the RDM varied depending on the planned model comparison for the RSA, but the same 20 trials (i.e., all trials in each run) were always used. For the category model, stimuli were separated by category. For the motor model, stimuli were separated by the hand used to make the response, and only included if the trial was correct. For the perceptual distance model, the trials in the neural RDM were arranged by the perceptual similarity of the continuous category feature (i.e., neighboring trials in the matrix were more similar to one another, perceptually, compared to those further away).

To generate RSA scores for each voxel, a whole-brain searchlight (Kriegeskorte et al., 2006) was performed, one participant at a time, using the "Brain Imaging Analysis Kit" (BrainIAK 0.11; Kumar et al., 2020a; Kumar et al., 2020b). The searchlight utilized a "ball" shape with a radius of 4 voxels (approximately 10-mm; 257 voxels). Although there is not one "best fit" size for a searchlight radius (Etzel et al., 2013), this radius has been generally shown to work well for studies focused on category selective regions (e.g., Chadwick et al., 2016; de Gardelle et al., 2013; He et al., 2020), although ranges from 3-mm to 10-mm have also been shown to be reasonably effective in a similar context (Clarke & Tyler, 2014).

The voxels selected for the searchlight were chosen from a brain mask generated using the compute_background_mask function of NiLearn (0.8.0; Abraham et al., 2014; Pedregosa et al., 2011) which chooses, as a mask, any voxel containing a non-zero value, effectively selecting every voxel with data (and excluding voxels outside the brain). The voxels which contained data had been previously masked during the single-trial GLM using a functional brain mask output by fMRIPrep.

Center voxels in a searchlight were not assigned an RSA score if the searchlight cluster contained fewer voxels than at least half of the total possible searchlight size. This selection process tends to exclude voxels along the edge of the brain, however, this returns more stable RDM calculations (Dimsdale-Zucker & Ranganath, 2018)

For each searchlight cluster taken from the masked data, the *t*-statistics in each voxel, for each trial, were Pearson correlated (corrcoef; NumPy) and 1 – *r* transformed such that high values represented greater dissimilarity. This resulted in a 20x20 neural RDM in which each cell contained a value representing the correlation between trials for all voxels in the searchlight cluster. Of primary interest were the comparisons between these RDMs and different feature processing models. In order to conduct these analyses, the RDM for a given searchlight was Spearman correlated (stats.spearmanr; SciPy) with a model matrix (using only the upper half of each and excluding the diagonal). This resulted in a single correlation value which was then assigned to the searchlight center.

The output of comparing the neural RDM to these model matrices was a map of Spearman correlation values in participant space. These were Fisher z-transformed (Dimsdale-Zucker & Ranganath, 2018), and then normalized to a standardized template (MNI152NLin2009cAsym) using antsApplyTransforms (ANTs 2.3.5; Avants et al., 2014) with a standardized anatomical scan in the template space as the reference.

The normalized Spearman correlation maps were subjected to a one-sample *t*-test in conjunction with maximum statistic permutation testing using randomise (FSL 6.0; Winkler et al., 2014) with 4096 permutations corrected for multiple comparisons using threshold-free cluster enhancement (Smith & Nichols, 2009).

ROI ANALYSIS

Following a priori predictions for possible locations of representational information during the task, several ROIs were used for a representational similarity analysis similar to what was performed during the whole-brain searchlight. That is, an RDM for each ROI was calculated and compared to a hypothetical model matrix. The targeted locations included the middle frontal gyrus (mFG), basal ganglia, lateral occipital complex (LOC), and precentral gyrus (see Figure 9). The mask for the mFG was borrowed with permission from a study on neural representations of category membership (Mok & Love, 2020), while the masks for the precentral gyrus and basal ganglia were generated using a version of the Harvard-Oxford atlas. The mask for the LOC was generated by using Neurosynth.org (Yarkoni, 2023) which was used to compile a functional association map from 226 studies using the term “lateral occipital”. The peak coordinate was extracted from one hemisphere and mirrored, then from those starting points, a 5 mm spherical mask was generated. All masks began in MNI space. Prior to analysis, masks were transformed from MNI space to each participant’s unique participant space using antsApplyTransforms (ANTs 2.3.5), using the participant’s anatomical scan as a reference.

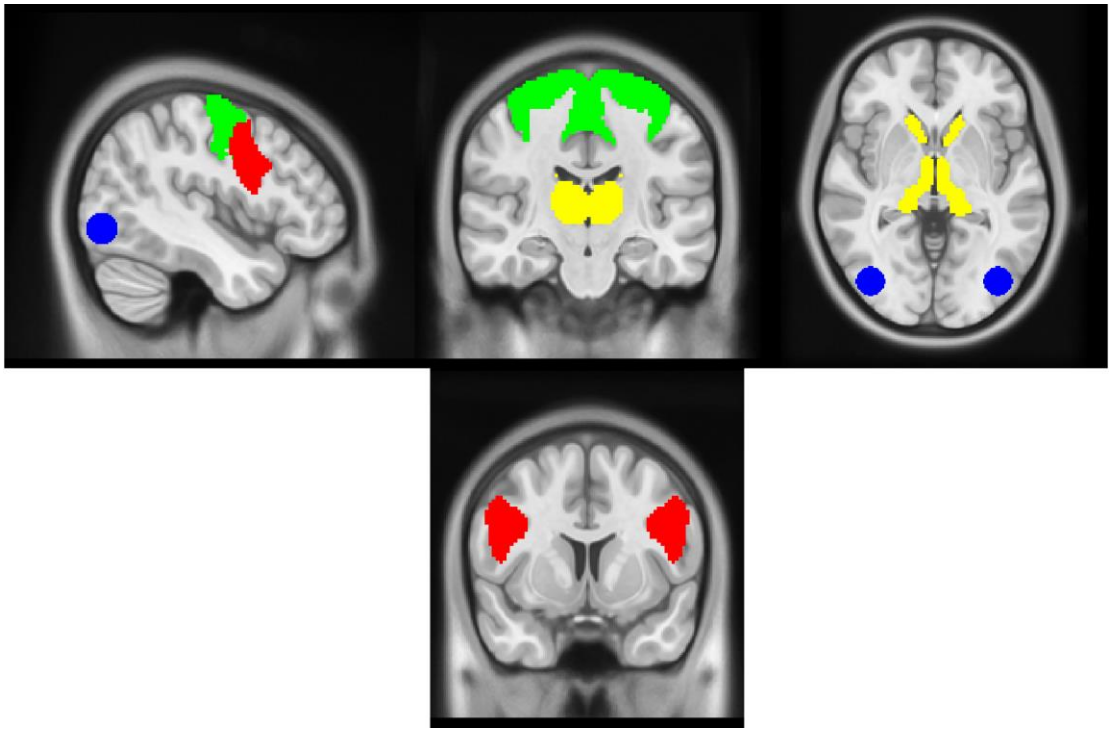

Figure 9: Each color represents a different anatomical location used for the ROI analyses. Blue = lateral occipital complex; green = precentral gyrus; red = middle frontal gyrus; yellow = basal ganglia. The locations shown are in MNI space, however, each mask was transformed to participant space for the ROI analysis.

Afterwards, the ROI voxels were treated like a searchlight cluster in the aforementioned whole-brain searchlight RSA: the *t*-statistics in each voxel, for each trial, were Pearson correlated, 1 – *r* transformed, correlated with the particular hypothetical model matrix (Figure 8), then Fisher *z*-transformed. The final output was a single RSA score per ROI per model per participant. These scores were analyzed using one_sample_permutation (nltools 0.4.7; Chang et al., 2022), which conducted a similar one-sample permutation-based *t*-test (5,000 permutations) as the one used on the whole-brain searchlight data.

Results

Behavioral

One participant was excluded for responding randomly throughout the task, leaving 24 participants in the analysis. A 2 x 4 (task phase x run) repeated measures ANOVA found a main effect for task phase ($F(1,24) = 177.01$, $p<.001$, $\eta^2 = .741$), demonstrating that accuracy during the Training phase (M=89.7%, SD=10.3%) was significantly higher than during the Transfer phase (M=48.6%, SD=13.8%; see Figure 10). The effect size (η^2) for the main effect of task run indicated a large effect. There was also a main effect for scan run ($F(3,72) = 8.00$, $p<.001$, $\eta^2 = .02$), as well as an interaction effect between the variables ($F(3,72) = 6.95$, $p<.001$, $\eta^2 = .018$). The effect sizes (η^2) for the main effect for scan run and the interaction effect indicated a small effect. Looking *within* each individual phase of the task, post-hoc tests found that the *only* significant difference across runs occurred during the Training phase in which performance for run 1 was significantly lower than the subsequent three runs. All other run comparisons within each phase were nonsignificant.

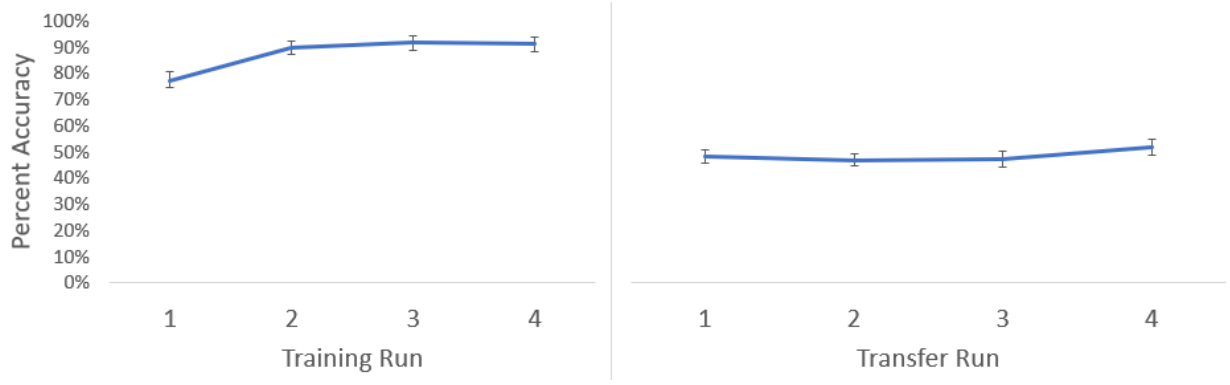

Figure 10: Percent accuracy across the task, separated by phase/run. Each run consisted of 20 trials.

Only one participant failed to demonstrate an overshadowing effect during the Transfer phase, performing at 85% accuracy across the four runs. Otherwise, all other participants performed roughly at chance (50%).

RSA

SEARCHLIGHT

There were three models used for the whole-brain searchlight RSA: category, motor, and perceptual distance (see Figure 8). Each RSA compared these models with a particular segment of task data from either the Training or Transfer phase. Because the present study is concerned with the nature of representations that develop during learning and transfer, data from different phases were used to answer different questions. The category and motor models were compared to data from runs 2-4 of the Training phase, and the perceptual distance model was compared to data from all four runs of the Transfer phase. Training blocks were used for the category and motor models because this is considered to be a learning phase in which category/motor representations develop. The Transfer phase was used for the perceptual distance model with the goal of searching for neural representations of the continuous overshadowed feature. Figures 11-13 display the resulting RSA maps for the searchlight.

CATEGORY

In the rotation condition, the category model identified regions in runs 2-4 that were primarily localized to bilateral medial portions of the superior, middle, and inferior occipital gyri, illustrated in Figure 11. For the color condition, the category model did not identify any statistically significant regions.

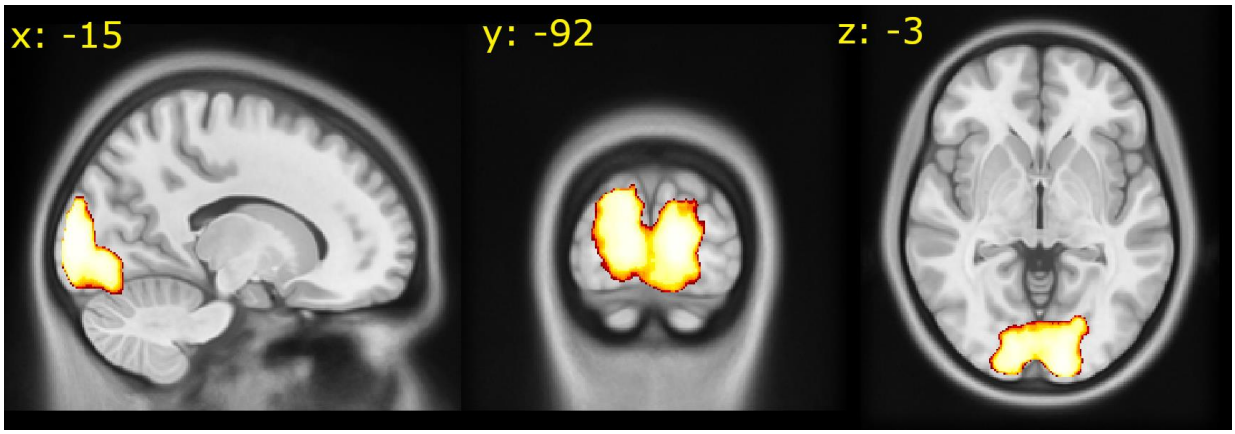

Figure 11: Regions of significance for the RSA utilizing the category model matrix in the rotation task group.

MOTOR

The searchlight analysis of the Training data found relatively consistent results across both task groups, illustrated in Figures 12 and 13. Locations where neural RDMs for both groups fit the motor model were primarily located across the somatomotor cortex in the precentral and postcentral gyrus. The motor model additionally fit neural data in the lingual gyrus, as well as a superior and medial location in the cerebellum that extended slightly into the left hemisphere.

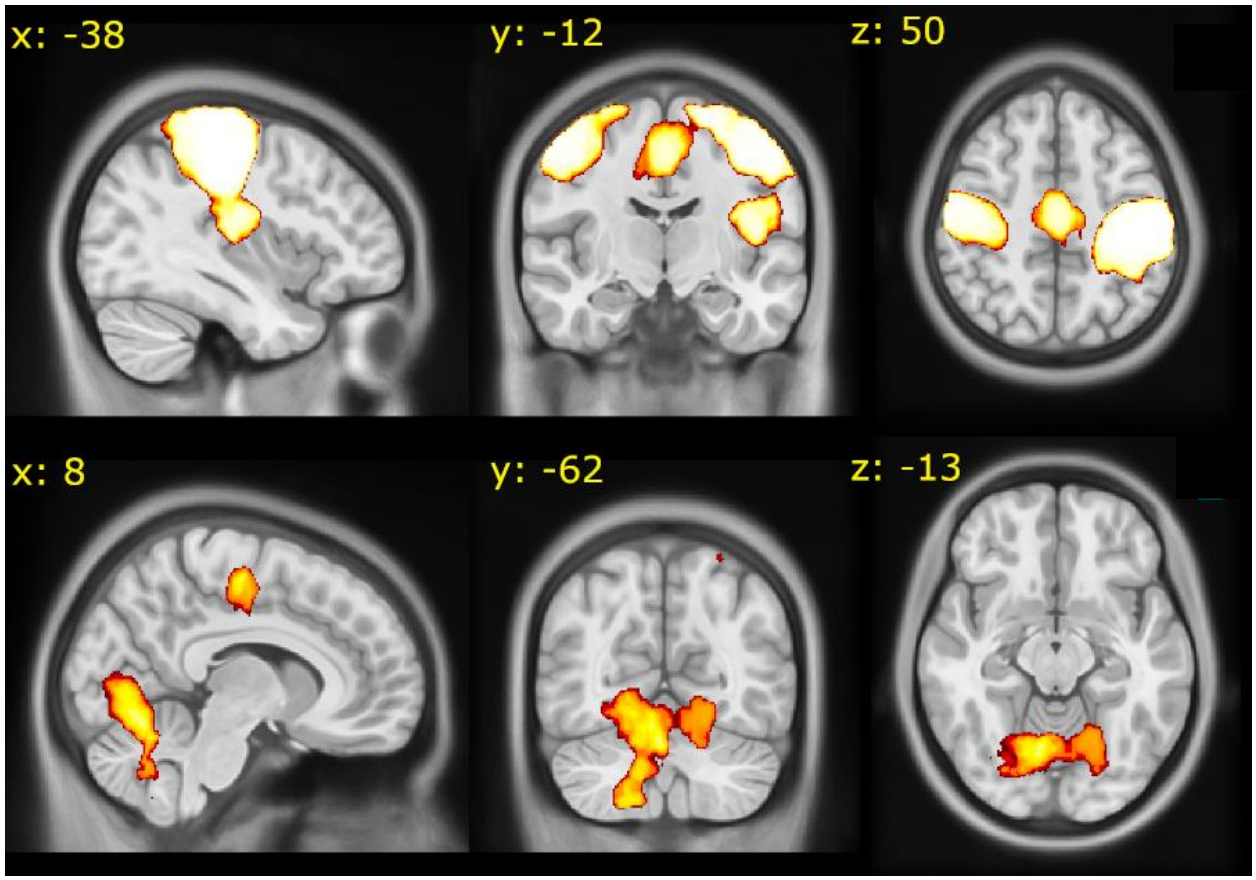

Figure 12: Both sets of images demonstrate regions of significance for the RSA utilizing the motor model matrix in the rotation task group.

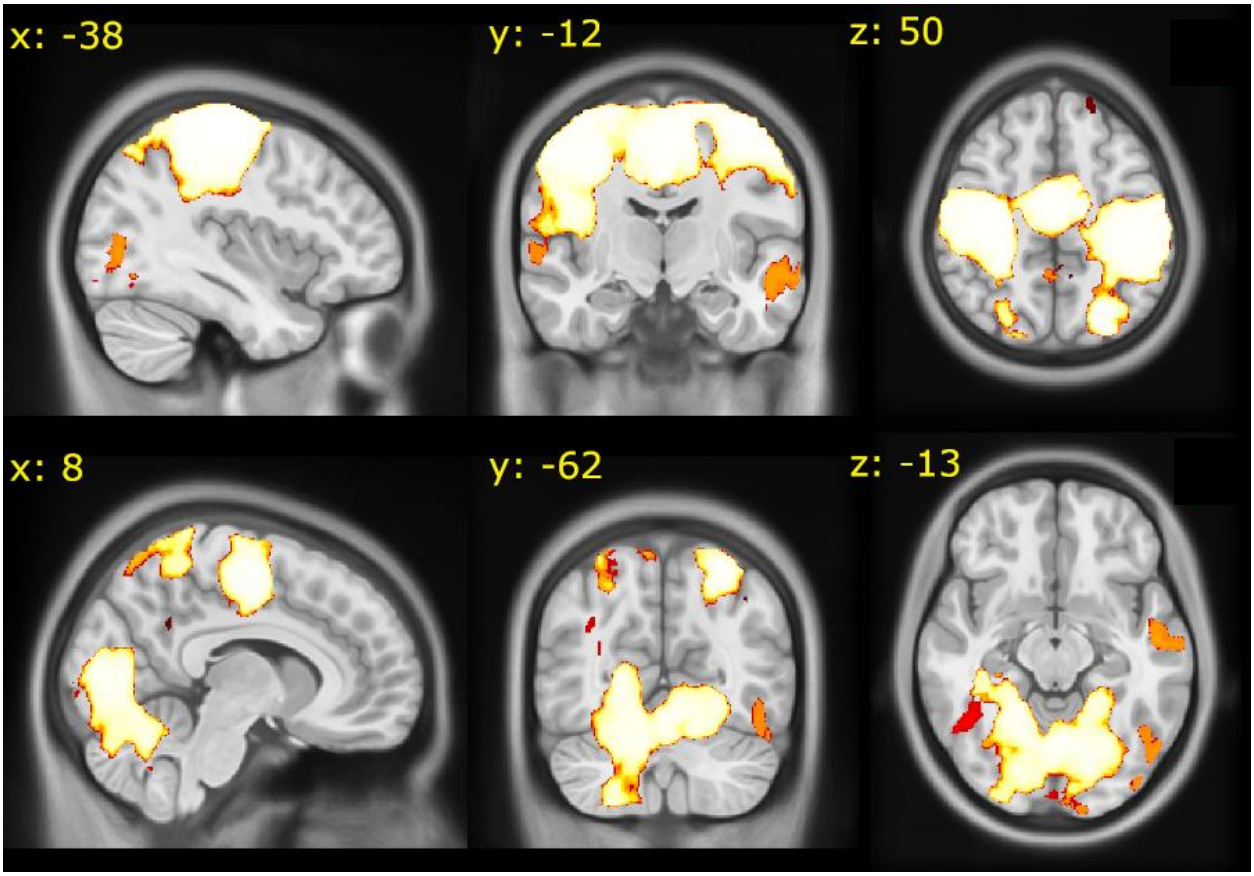

Figure 13: Both sets of images demonstrate regions of significance for the RSA utilizing the motor model matrix in the color task group.

PERCEPTUAL DISTANCE

There were no areas in which the perceptual distance model fit the neural data in either of the task groups during the searchlight analysis. This extended to exploratory analyses utilizing more lenient multiple comparison corrections and *p*-value thresholds.

ROI

A region of interest RSA was performed using ROIs in the following regions: a portion of the middle frontal gyrus (mFG), bilateral portions of the lateral occipital complex (LOC), the basal ganglia, and the precentral gyrus. The mFG, basal ganglia, and LOC were selected based on previous work (Mok & Love, 2020; Ritchie & Op de Beeck, 2019; Seger et al., 2015) as

regions sensitive to category representations, while the precentral gyrus was selected as a confirmatory region from the whole-brain searchlight analysis (Etzel, 2013). See Methods and Figure 9 for more information about the ROI definitions.

Table 1 contains the correlations between each ROI and the model matrix used for the RSA. Some of the significant ROIs were also regions which were significant in the whole-brain searchlight. Most notably, the neural RDM for the precentral gyrus was significantly correlated with the motor model for both task groups during the Training phase. There were also significant results unique to the ROI analysis; these were primarily in the LOC. During the Training phase, the motor model fit neural data for the color group in the LOC, and during the Transfer phase, neural data from the rotation group fit the category model in the LOC. Lastly, there were a few trending ROI results. During the Training phase, the motor model was trending with the neural RDM from the LOC for the rotation group. Additionally, the category model was trending with the RDM from the basal ganglia for the category group, and the perceptual distance model was trending with the RDM from the mFG for the color group. Lastly, the category model was trending with the neural RDM from the precentral gyrus for the color group during the Transfer phase.

Table 1

RSA scores for the training and transfer phases from each individual ROI

ROI	Model	Rotation		Color		Rotation		Color	
		rho	*p*	rho	*p*	rho	*p*	rho	*p*
		Training				Transfer			
Precentral gyrus	Category	.00	.940	.00	.872	.03	.208	-.02	.068
	Perceptual	.02	.442	.00	.862	-.01	.635	-.01	.724
	Motor	.21	< .01	.24	< .001				
LOC	Category	.01	.385	.00	.913	.03	< .05	.00	.780
	Perceptual	.03	.249	.01	.551	.04	.159	.00	.876
	Motor	.04	.078	.03	< .05				
Basal Ganglia	Category	-.01	.404	.00	.982	.02	.079	-.03	.050
	Perceptual	.02	.426	.01	.807	-.01	.696	.02	.317
	Motor	.00	.841	.01	.644				
mFG	Category	-.02	.190	.00	.872	.04	.124	-.01	.137
	Perceptual	.01	.783	.00	.988	-.02	.551	-.03	.074
	Motor	.01	.314	.02	.231				

Spearman's rho indicates the degree of correlation between the RDM and the model matrix

Discussion

Although cue competition is well understood in associative learning (Kamin, 1965; Pearce & Bouton, 2001; Prados et al., 2013; Rescorla & Wagner, 1972), the mechanisms of action are less clear in category learning paradigms. In fact, there is conflicting research regarding how and why cue competition functions in category learning (Bott et al., 2007; Lau et al., 2020; Murphy et al., 2017), and no research regarding the neural mechanisms of feature representation in this context. The present study was designed to investigate whether or not overshadowing was present in a rule-based category learning task, as well as how the category features were represented neurally during overshadowing. The task utilized one highly salient (i.e., easy to categorize) feature as well as a more difficult, but still informative feature. It was hypothesized that early perceptual regions would represent all of the features as evidenced by a fit to the perceptual distance model, but that higher order perceptual and frontoparietal

regions would show an overshadowing effect as evidenced by a lack of fit to the perceptual distance model. It was further hypothesized that frontoparietal regions would code for discrete category membership as evidenced by a fit to the category model, and that motor regions would represent the motor response as evidenced by a fit to the motor model.

In order to address these hypotheses, representational similarity analysis was used in conjunction with a searchlight analysis to try and identify regions containing information for each informative feature.

Behavioral overshadowing

The specific behavioral hypothesis in the present study was that participants would quickly learn the highly salient (i.e., easily categorizable) binary feature during Training, but would fail to learn the more difficult but still diagnostic continuous feature. Using the binary salient feature to categorize should result in highly accurate performance during Training but would result in overshadowing of the continuous feature that should lead to poor performance during Transfer. Indeed, this is the pattern of behavior that was demonstrated. Only one participant who performed the task failed to demonstrate an overshadowing effect; their accuracy after the removal of the salient feature was better than chance (about 85%). Overall, this suggests that, although it is possible to learn the more difficult continuous feature, attention and task demands may simply lead too strongly towards the more salient binary feature during learning.

Research suggests that, in tasks similar to the present study (i.e., classification tasks) where accuracy is important, participants will often fixate on the easiest, most diagnostic feature and ignore the rest (Deng & Sloutsky, 2016; Kruschke. 1992; Plebanek & Sloutsky, 2021;

Shepard, 1961), even if other features are diagnostic of category membership (Hoffman & Murphy, 2006; Rehder & Murphy, 2003). Although other research has suggested that, during cue competition, a highly salient feature may *enhance* learning of less salient ones (Murphy et al., 2017), the results of the present study suggest that cue competition in category learning typically is consistent with predictions from associative learning models in which competing cues typically do not share attention equally.

Neural representations

The primary analysis for the present study consisted of a representational similarity analysis (Kriegeskorte et al., 2008). Contrary to analyses focused on activation, this type of analysis searches for information. The multivariate patterns of activity elicited for a set of stimuli are examined in order to determine what the representational space looks like and how the information within it might be organized. For example, does a region represent faces and houses using similar neuronal patterns of activity (are the representations similar), or are the patterns unique enough to be considered separate representations? Furthermore, in the same example, it can help determine if the region codes for individual faces or houses, or the category more generally. This is accomplished by comparing neural data to models of each hypothesized type of representational structure in order to find which representations fit the data. Three different models were used in the present study. The first defined a representational space for the two discrete categories that was strictly binary. The second was a motor model which separated representations based on the hand used to make a response. The third was a model representing perceptual distance in which the representational space was structured around unique representations for individual stimuli.

In the present study, RSA was used in conjunction with a category learning task to examine where and how features are represented. The primary analysis, a whole-brain searchlight RSA, found that early visual areas represented categories (based on fit to the category model), whereas motor areas represented the hand used to make a response (based on fit to the motor model). There were no regions identified as representing the overshadowed feature (based on fit to the perceptual distance model). ROI analyses found patterns of model fit similar to those found with the whole-brain searchlight, including activity in the precentral gyrus being fit by the motor model. Additional model fits in the ROIs beyond those identified in the searchlight analysis were also observed. Notably, these included a significant fit to the category model in the lateral occipital complex (LOC) and a trending fit to the category model in the basal ganglia.

CATEGORY

The results of the whole-brain searchlight RSA using the category model suggest that early visual areas maintain information related to category membership during training when rotation is the binary salient feature. Furthermore, for the same task group, the ROI analysis found that the category model fit patterns of neural activity in the lateral occipital complex during transfer. The ROI analysis failed to find any other model fits for the category model for the color group except for trending model fits for both task groups in the basal ganglia during the Transfer phase. The category model represented categories as discrete entities: that is, all stimuli from the same category were considered to be equally similar and all stimuli from the other category were equally dissimilar. This means that, for example, although there were degrees of rotation within each category for the rotation group (i.e., each category did not use

a single rotational value), the discrete category model cannot speak to whether early visual areas distinguished between the different orientation value of each stimulus. What it can potentially speak to is discrete category structure.

The visual cortex is concerned with visual information, and category structure has been identified in regions as early as the primary visual and extrastriate cortices (Ashby & Ell, 2001; Reber, 1998), where it is hypothesized to be related to simple features like lines and shapes. The representations that develop in these early visual areas can also be modulated by top-down attentional processes related to task goals (e.g., supervisory feedback), resulting in rapidly developed category representations (Hochstein & Ahissar, 2002), especially when task demands are minimal (Hammer & Sloutsky, 2016). Early perceptual regions are also involved in selective attention, as well as task-dependent working memory (e.g., holding stimuli in working memory such as during the delay between stimulus presentation and feedback) (Ashby & Ell, 2001; Postle, 2006). Given the supervised nature of the classification task used in the present study, as well as the minimal task demands, it was hypothesized that early perceptual regions would code for discrete category membership. The results of the present study support this hypothesis, however, not all hypotheses were supported; several other frontoparietal regions involved in category learning were expected to contain information regarding category membership.

Parietal regions, such as the intraparietal sulcus, are involved in visuospatial attention (Daniel et al., 2011). These regions can link information such as category features to potential motor responses (Seger & Miller, 2010; Yi et al., 2016), and have been shown to be active in similar rule-based categorization tasks (Seger et al., 2015; Seger et al., 2000). It is thought that

the role of these category-sensitive regions is to accumulate perceptual information (Sestieri et al., 2014).

During categorization, the prefrontal cortex is thought to be involved in processes such as hypothesis testing (Milton & Pothos, 2011) and the implementation of rules (Antzoulatos & Miller, 2011). It is active during early phases of learning (Seger & Miller, 2010), and in rule-based tasks similar to the present study, it has been shown to represent category boundaries as well as individual features (Jiang et al., 2007).

During the searchlight RSA, the neural RDMs were able to be fit to the category model for only one of the two task groups: the rotation group. This creates an interpretative issue. Part of the reason the sample was split into two task groups was to try and avoid the issue of confounding category with perceptual feature in the analyses. In other words, if just one set of perceptual features were used, any areas where the neural RDM fit the model could be said to contain information for one of two reasons: either the area contained information for *category features*, in general, or the area contained information for that specific *perceptual feature* (e.g., rotation, color). Having two feature sets would have meant that, if both task groups recruited the same neural area, it would be more likely that the area represented the category instead of just one specific perceptual feature (e.g., rotation). The lack of significant results in early visual cortex for the color group means that the significant results for the rotation group could have occurred either because the region contains information related to category membership, or alternatively, the region contains information related to perceptual details like the rotation of an object independent of category representation.

From early single-cell work in cats (Hubel & Wiesel, 1962), to more modern work utilizing multivariate decoding methods (Carlson, 2014; Kamitani & Tong, 2005; Ragni et al., 2020), it is well documented that regions of early visual cortex are sensitive to the orientation of visual stimuli. The stimuli in the rotation group were oriented diagonally to the left or the right (at an average tilt of -45° and 45°, respectively), which may have created a perceptually distinct dichotomy in terms of overall orientation. The perceptual difference in rotation between each category may have also been distinct enough that early visual areas sensitive to orientation registered each category (i.e., rotational direction) separately. This would explain the model fit between the category model and the neural RDMs acquired from early visual regions; the visual areas were representing orientation as a dichotomy rather than a spectrum.

For the ROI analysis, this would, in part, explain the lack of any model fits in the lateral occipital complex during the Training phase. The LOC is part of the ventral stream and is strongly associated with object recognition (Grill-Spector et al., 2001; Malach et al., 1995) responding to images of objects more strongly than non-objects (e.g., scrambled images/textures). Decoding studies focusing on this region have discovered that object representations cluster by category (Cichy et al., 2014; Contini et al., 2017), and that this clustering occurs quite rapidly, with one study finding that coarse category distinctions were decodable within one second (Carlson et al., 2013). Although there were no a priori hypotheses for this region, it was included in the analysis as an area sensitive to category representations (Connolly et al., 2012; Folstein et al., 2013; Hammer & Sloutsky, 2016). Still, despite research suggesting that this area is sensitive to category distinctions, there were no significant results for the category model in this region during the Training phase.

It is unclear as to why the category model did not predict activity in early visual areas sensitive to color in the task group that had color as the salient feature. There were only two colors used for the stimuli in this condition: red and blue, which differ significantly in neural representation (Kim et al., 2020; Xiao et al., 2007). There was also very little within-category variation in the base color, so participants would have been able to represent orientation as a dichotomy rather than a spectrum, as was the case for the rotation feature. Given that participants in this task group demonstrated an overshadowing effect (high accuracy in training, chance performance in transfer), it is likely that they were paying attention to the color of the stimuli during Training. Despite the lens of attention being focused on color, there were no neural areas that fit the category model.

One possible explanation for the lack of a fit to the model has to do with how early visual areas code for color and form (e.g., rotation/orientation). Research suggests that these features are coded separately in early visual areas (Livingstone & Hubel, 1984; Zeki, 1978), and other studies focusing on these areas using RSA have, indeed, found regions containing information related to separate colors using similar category-like models to what was used in the present study (e.g., Sun et al., 2021). However, these studies often use simple colored circles as their stimuli. For the present study, stimuli in the color task group included rotation and number of dots as random features, and each category contained an equal sample of both features (i.e., each category had an equivalent set of rotation and dot values depending on the task group). The regions that represent these additional features in the brain may have caused interference. Some studies suggest that there are many small, intermingled regions sensitive to color and orientation in early visual areas (Seymour et al., 2010), so it is possible that each

searchlight could have included regions representing both types of features. Even if there were actually unique representations for each color, the information related to rotation would have been equal for both groups. This lack of a difference between conditions (for rotation) may have reduced the overall ability of the searchlight to detect the information available for each color. Perhaps there is more information available for rotation than for color, which would explain why the rotation group showed significant areas of information – this group also included color, but as a random feature. However, the inclusion of random color in the rotation group does not appear to have caused an issue detecting an effect.

This explanation, however, fails to account for the full picture. fMRI has long demonstrated the ability to use multivariate analyses to decode information related to features such as rotation, despite the sparse distribution of other noisy regions sensitive to different features (e.g., Kamitani & Tong, 2005). The analysis, in principle, should be capable of parsing information related to features such as color despite informational interference from surrounding areas. These aforementioned issues, then, may simply be related to the size of the searchlight; a smaller searchlight radius may have been better able to detect color information in the brain from the small regions which represent it. Many of these studies that decoded information such as rotation utilized voxel resolutions smaller than the size of the searchlights used in the present study (e.g., 3 mm isometric voxels in other studies compared to a 4 mm *radius* for each searchlight). This may impact the overall sensitivity of the analysis by including too much additional information to disentangle.

There was a trend for the category model to fit the data from the basal ganglia during the Transfer phase, but not the training phase. One possible explanation has to do with the

nature of the task. The COVIS model (Ashby et al., 1998) proposes that there are two different functional networks active during category learning tasks: one utilizes declarative systems (e.g., working memory, executive attention) such as the medial temporal lobe, and one utilizes procedural (e.g., skill-based), striatal-based systems (Ashby & Maddox, 2011). These systems can cooperate, but sometimes, they are in competition (Wang & Voss, 2014). During early learning in the type of rule-based, category learning task utilized in the present study, this model suggests declarative systems should be active *over* striatal ones (Freedberg et al., 2020); it is possible that striatal systems were simply not engaged during this phase of the task due to competition from declarative systems.

Once feedback was removed in the Transfer phase, however, the declarative, hypothesis-testing system may have been no longer active. Instead, participants would have had to rely on a "gut" feeling as to whether or not they were accurately categorizing. This type of implicit categorization fits with the role of the basal ganglia during procedural learning (Ashby et al., 1998; Smith et al., 2015). Under these circumstances, it is possible that category representations were activated, however, participants did not appear to have access to this information as shown by at-chance accuracy. This fits with the hypothesis that category knowledge learned implicitly is inaccessible by declarative consciousness (Ashby & Maddox, 2011).

Another trend for the category model fit was in the LOC, a region involved in representing and perceiving objects (Grill-Spector et al., 2001). This region could be responding, in part, towards the visual characteristics of the stimuli. For example, it could be responding to the density, or even the texture of the dots used to indicate category membership (Chen et al.,

2020). The number of dots varied with category (effectively, each category had either a “low” or “high” number of dots), and clusters of these dots may have acted as emergent structures in the stimuli, varying in their density by category (with stimuli in the “low” category having fewer clusters than those in the “high” category). The LOC, then, would be organizing the stimuli by the number of clusters of darkly colored circles. This model fit could represent a subconscious, but fundamental part of the visual ventral processing stream in which abstract features of objects are sorted into categories (Pietrini et al., 2004). This explanation, however, only applies when the feature was the number of dots – there was no apparent model fit for the color group (i.e., the height of the stimuli), and it is less clear as to why this perceptual characteristic of the stimuli was not similarly processed by the LOC. Perhaps stimulus height is just not salient enough.

Both of these trends for category model fits suggest the possibility that some type of category learning of the continuous feature (i.e., the number of dots) may have been occurring during the transfer phase. However, if so, the information was not cognitively accessible. Still, this presents the exciting possibility that categorical representations can develop for unattended features in an unsupervised context. That the category model, and not the perceptual distance model, fit the data also suggests that these may be true category representations and not just representations of perceptual variability.

It is unclear as to why areas outside early visual regions, such as the intraparietal sulcus, that can code for features such as rotation (Bettencourt & Xu, 2015), failed to demonstrate a significant association with the category model. This may be reflective of a greater overall issue with the RSA models (see Limitations and Future Directions), or perhaps there was no need to

maintain full category representations in frontoparietal regions because the task was very simple: participants only had to identify whether the stimulus was oriented to the left or the right or whether it was red or blue. Previous research has found that simple rules for categorization are learned quickly and then require very little activity to maintain (Poldrack et al., 2001; Zeithamova et al., 2008).

MOTOR

A whole-brain searchlight RSA comparing the motor model to data from the Training phase found regions of model fit in areas such as the precentral gyrus and motor cortex, as well as portions of the cerebellum. Parts of the postcentral gyrus and lingual gyrus also fit the motor model. In the ROI analysis, the motor model was able to be fit to data from the precentral gyrus (both task groups) and LOC (color task group only); there were no other regions that fit the model. The motor model used for this RSA was structured such that all trials in which a response was made using the left hand were defined as being equally similar, and equally dissimilar to all responses made using the right hand. This suggests that the regions that fit the model contained information representing what hand was used to make a *correct* response. Incorrect responses were not included in the model.

The way that the task was designed also means that the information is independent of the category membership of the stimulus, and independent of what side of the screen the category label was presented on. This is because, during the task, the correct responses for both categories alternated – the category labels could be on either side of the screen, which required the participants to monitor the category label location and appropriately adjust which hand they used to respond to which category. The labels additionally alternated in such a way

that if a participant got every trial correct, they would have used each hand to respond to each category on each side of the screen an equal number of times. For these reasons, if the motor model fit, it provided strong evidence that the information was related to the hand used to make a response and less likely related to a perceptual characteristic of the stimuli.

This presents a bit of a mystery, then, regarding the model fit of the motor model during the ROI analysis of the LOC. This region, as previously discussed, is largely related to representing differences between categories. Given the level of control for category features across the left- and right-hand responses (i.e., equal presentation of category labels for each category for each hand), there should be no difference in available perceptual information.

The lingual gyrus is more closely linked to visual processing than motor functioning (Palejwala et al., 2022), however, some previous research has found lingual gyrus involvement in abnormal motor function. For example, one study identified the lingual gyrus as being linked to gait disturbances in patients with idiopathic normal pressure hydrocephalus (Suzuki et al., 2022), while another identified lower gray matter volume in this region in patients with tardive dyskinesia (Yu et al., 2018). Given the lack of literature on the role of the lingual gyrus in typical motor functioning, it is hard to speculate as to why it contained information related to the appropriate motor response in the present study. Alternatively, apparent activity in this area could be due to a normalization error. The lingual gyrus is adjacent to the cerebellum, and the significant cluster which incorporated the lingual gyrus was part of the cluster in the cerebellum. It could be that the normalization of the participant space data caused a small amount of “bleed” from the cerebellum into the lingual gyrus.

Activity in the precentral and postcentral gyri is consistent with research suggesting that these regions contain information representing which hand will be chosen to perform a motor action. Hirose et al. (2018) had participants freely tap a finger of their own choosing from their left or right hand while fMRI data was recorded. The researchers then attempted to take the data and decode which hand they were going to use to perform the tap. What they found was that regions of motor cortex, similar to those identified in the present study, contained sufficient information to predict which hand was going to be used to make the response.

Although the cerebellum has more recently been implicated in a host of other cognitive and associative functions (e.g., vision; van Es et al., 2019), it is still primarily linked to motor function. Research suggests that this region, in addition to others such as the striatum (Hoover & Strick, 1999), may be active as a sort of "monitor" during motor actions, providing feedback on the current location of limbs (Shadmehr et al., 2010), and correcting movements as necessary (Robinson, 1975; Shadmehr & Krakauer, 2008), regardless of the limb used (Alahmadi et al., 2015). Considering the role of the cerebellum, then, it is not surprising that information related to the limb used to make a response was found, as the cerebellum was likely monitoring the motor response of each participant during the task to help ensure that the appropriate motor program was selected.

One question that remains is why other regions involved in motor processes were not fit by the motor model in either the searchlight or ROI RSA during the training phase. For example, the basal ganglia are involved in several aspects of motor functioning such as inhibiting competing movements (Mink, 1996) and selecting the appropriate motor response during category learning tasks (Gordon et al., 2023; Seger, 2008). However, as previously discussed, it

is possible that striatal systems were not engaged due to preferential control of motor resources by declarative systems (Ashby & Valentin, 2017). Interestingly, this hypothesis is not without challenge, with some studies suggesting that both systems can learn simultaneously (Crossley & Ashby, 2015; Foerde et al., 2006). This would suggest that there should have been a motor model fit to the basal ganglia during Training, which is not what the present study found. As there is currently no consensus in the literature regarding whether these category learning systems function cooperatively or competitively, it is hard to suggest what the reason might be for these results.

One possible explanation is related to the multi-purpose role of basal ganglia. This region is composed of several functionally distinct sub-regions (e.g., putamen, caudate, nucleus accumbens), and each can play a different part in the categorization process. For example, the body/tail of the caudate communicates with visual cortex in order to assign stimuli to a category, while the putamen, which is connected to motor planning areas, helps determine the appropriate motor response based on category membership (Seger & Miller, 2010). Neural activity in these regions would likely fit with the category model (caudate) and motor model (putamen) if processed individually. However, the mask used for the basal ganglia in the ROI analysis included these sub-regions together, which may have impacted the ability of the RSA to find an appropriate fit for each of the individual models.

It is worth mentioning that it was not possible to appropriately analyze data from the Transfer phase using the motor model. The motor model was defined such that only correct trials were included in the RDM (trials in which the intended motor response was appropriate). In the Transfer phase, where accuracy was roughly at chance, it was difficult to find participants

with a sufficient number of trials per run to fit this model. In addition, left- and right-hand responses were not evenly distributed across transfer phase blocks. However, there were no unique hypotheses specified for the motor model in the Transfer phase; it was not expected that motor representations would change between Training and Transfer.

PERCEPTUAL DISTANCE

One unanticipated outcome was that the perceptual distance model did not fit to any neural region during the whole-brain searchlight RSA or the ROI analyses for both task groups. It was hypothesized that the overshadowed feature would at the very least be represented in early visual regions due to bottom-up processing of the feature. It was further hypothesized that higher level perceptual regions and frontoparietal regions might also represent the overshadowed feature even in the absence of attention to the feature, and a goal of this study was to identify how far along the visual and category learning pathway the overshadowed feature could be detected. However, despite examining two different sets of perceptual features (i.e., the two different task groups), neither of the two continuous features (number of dots and stimulus height) represented in the perceptual distance model fit activity in any neural region. This model assumed that stimuli were represented on a continuum, with stimuli close to each other in perceptual space having greater similarity than those further along the perceptual spectrum. Stimuli with seven dots would be more similar to stimuli with eight dots compared to stimuli with twenty-seven dots. The lack of a model fit suggests that either there is no representational information available, or that this model does not accurately fit the representational information that is available.

It is unclear what this overall lack of significant results might mean given the lack of previous research and the somewhat exploratory nature of the analysis. Perhaps behavioral overshadowing precludes development of neural representations for the unattended feature, despite the possibility that, at some point, the continuous feature may have been the focus of attention. Additionally, it should be noted that this model was fit to the data during the Transfer phase, in which no feedback was provided. It is possible that previously learned representations may not be elicited, or new representations learned in the absence of feedback. Although it is not necessary for learning in all rule-based classification tasks (Ashby et al., 2002; Ashby & O'Brien, 2005), behaviorally, the lack of feedback appeared to impede further learning during the Transfer phase: participants remained at chance levels of performance across all four blocks.

Alternatively, although behavioral data suggest the continuous feature was overshadowed by the binary feature, it is possible that information related to the continuous feature was incorporated into the representation formed in the Training phase. This would prevent the model from appropriately fitting to the representation. Instead of representing the feature on a spectrum, which was assumed in the perceptual distance model, the feature representations may have been structured in discrete categories similar to how the binary feature may have been represented (at least in the rotation group as there was no significant model fit for the color group). This would partially explain the lack of a model fit to neural activity in the Transfer phase – as soon as the binary feature vanished, the representational structure came apart. This possibility, however, is difficult to test using the present task design. An RDM organized in the same way as the one used for the category model fits would be

representationally equivalent to one sorted by the binary feature; this precludes any sort of unique examination of the continuous feature during Training using the category model (i.e., the RDM used to compare against the category model is the same no matter whether it's organized by the binary or continuous feature).

As a note, exploratory analyses (i.e., analyses at reduced *p*-values) did show a trending model fit to the perceptual distance model in early visual areas, however, interpretating this is difficult given the reduction in statistical rigor required to achieve the results.

Limitations and Future Directions

A failure to fit a hypothetical model matrix to a neural RDM literally means that the data sampled for the RDM did not fit the geometry of the model matrix. This could be related to issues with how the single trial information used for the RDM was extracted, or alternatively, it could be that the multivariate data simply does not contain information that fits the hypothetical structure of the model matrix (e.g., a region might maintain separate representations for faces and houses, or it may not differentiate between the two). This does not mean that the RDM does not contain any representational information, rather, the model just needs to be in another shape.

RSA is used to examine the similarities between representations (Kriegeskorte et al., 2008). In this capacity, it is able to suggest the underlying representational structure, or the "representational geometry", of a region in the brain (Kriegeskorte & Kievit, 2013). In other words, it can help identify how representations are related to one another. It suggests areas of *information* rather than *activation*, but the nature of this information is not always clear. A perfect fit between a neural RDM containing face and house stimuli and a binary category

model could suggest that the region sampled for the RDM contains high-level information for faces and houses, separately, but it could also suggest that there is information related to some other shared low-level feature between the stimuli (e.g., faces are *circular,* and houses are *rectangles*). Furthermore, finding a model that fits your neural RDM only means that you have found a likely candidate for the representational model – there can always be a better fitting or more accurate model for the representational space (Popov et al., 2018).

If there had been only one set of perceptual features used to indicate category membership in the present study, significant RSA results would have had an unclear interpretation – the information could have been related to category representation (i.e., separate information for each category), however, it could have also just been perceptual information related to characteristics of the salient feature ("left" or "right" orientation). The present study attempted to control for this confound between category and perceptual feature by using two different feature sets to indicate category membership. If both feature sets fit with the same category model in the same area, it would be more likely that the information available to the RSA was related to high-level category representations rather than simple perceptual information. However, the only model that fit both task groups was the motor model; the other two models were either impossible to fit to any neural RDMs (perceptual distance) or could only be fit for one task group (category).

In exchange for attempting to orthogonalize category and perceptual characteristics in the RSA by using two task groups, the sample was divided in half. Although the overall number of participants for the task was in line with current recommendations for detecting medium to large effects in fMRI studies (Poldrack et al., 2017; Szucs & Ioannidis, 2020), the effective

sample size for each task group was relatively low. It is possible that the present study was concerned with phenomena which have real but small effects, in which case, there may not have been sufficient power to detect an effect given the sample size utilized. In hindsight, given the difficulty in finding an effect for the perceptual distance model, it may have been more beneficial to accept the interpretational issue in exchange for greater power to find an effect.

An additional methodological consideration that is worth acknowledging relates to the process by which raw task data is broken down into isolated single trial segments for use in multivariate forms of analysis such as RSA. The goal in this process is to take the global signal and extract data specifically related to whatever the researcher wants to analyze (e.g., stimuli/feedback/etc.). In blocked or slow event-related designs it is relatively easy to isolate the signal of interest as it is often separated by large enough periods of time that it is not contaminated by signal from other trials or types of events. However, in the rapid event-related design used for the present study, relatively short inter-stimulus intervals made it possible for signal from nearby trials to overlap in time, becoming collinear (i.e., highly correlated) (Visser et al., 2016). If this issue is left unaddressed, it can lead to inflated rates of false positives in pattern similarity analyses such as those used in the present study (Mumford et al., 2014).

Although there are several possible collinearity issues in the present study (e.g., rapid event-related design/relatively short ISI), steps were taken to try and mitigate this potential problem. The first step was taken during the design stage when specifying the ISI. Early simulation research suggested that 4 s was an appropriate amount of time between stimuli to reduce collinearity (Zarahn et al., 1997), and other studies have similarly suggested that 2-4 s is appropriate depending on how the single-trial data is modeled (Mumford et al., 2014).

Therefore, the present study utilized a 5 s ISI which allowed for the presentation of the required number of stimuli within the scan time limit while also giving the longest ISI possible.

Furthermore, each participant was presented with stimuli in a unique random order, another step which can help reduce problems from collinearity (Mumford et al., 2014). During the single-trial modeling phase, stimuli were also modeled following the least squares-separate procedure, which was demonstrated to be, in one particular study, the best method out of several tested modeling procedures for mitigating collinearity issues in single-trial data (Mumford et al., 2012). These steps, along with others described in the Methods, such as the use of updated autocorrelation modeling techniques (Olszowy et al., 2019), have all been chosen to help address the issue of collinearity.

However, despite multiple precautions, it is not possible to know if sufficient procedures have been implemented to satisfactorily reduce collinearity. Longer ISIs do seem to benefit pattern similarity analyses (Visser et al., 2016), so it is possible that fewer runs for each session with longer ISIs may have been better capable of capturing the single-trial patterns, while also meeting task demands (albeit with fewer trials). However, ISIs such as those used in the present study have still been demonstrated to be capable of detecting differences in pattern estimations under conditions such as those used in the present study (Zeithamova et al., 2017).

In addition to the category learning task, a 1-back task was also administered to the participants. It was intended that this task would function as a feature localizer due to the nature of how stimuli must be processed holistically in order to succeed on the task. To perform accurately, you must identify whether the currently presented stimulus is identical to the previously presented one; each feature of the stimulus must be processed, in turn. If there had

been significant effects for the perceptual distance model RSA, this localizer could have been used to define a region of interest for a follow-up analysis to confirm that information related to the continuous feature was maintained at the category level. However, because the perceptual distance model failed to correlate with any of the neural RDMs during the searchlight analysis, there was no follow-up needed.

Future studies should attempt to untangle feature confounds in the data. These include the confounding of the binary feature and category, as well as the binary feature and the continuous feature. It might be beneficial to include more realistic features, as well. In a research task such as the one used in the present study, participants may not engage with stimuli the same way they would in a naturalistic setting. Learning that an orange is orange (more or less) does not prevent you from learning that it is round. It may be beneficial to include a wider variety of stimuli, as well. This has been shown to enhance learning of the features that define a category's internal structure (Roads et al., 2018), and might result in less overshadowing, as well as the ability to generalize the internal category structure to new, unseen stimuli (Ell et al., 2017).

Conclusions

Stimuli in the environment contain innumerable features and selective attention is limited, therefore, selecting what feature(s) to attend to becomes an important decision. The selected feature is often the one which stands out the most – this could be because of its physical salience, or alternatively, because it provides a large amount of useful information. The present experiment examined the impact of a highly salient stimulus on category learning by monitoring behavioral responses and neural representations.

The presence of a highly informative feature behaviorally overshadowed learning of a less informative feature during learning, which led to a decline in performance when the informative feature was removed. Representational pattern similarity was analyzed using RSA, an fMRI analysis method. The primary analyses found that a binary category model fit data from both early visual and object-specific areas, while an effector-specific motor program model fit data from motor regions.

The present study has contributed to the body of literature providing behavioral evidence in favor of overshadowing as a preventative (e.g., Lau et al., 2020) rather than an enhancing (e.g., Murphy et al., 2017) factor in category learning. Furthermore, the present study characterized how information might be represented when overshadowing occurs during category learning tasks. Importantly, it suggests for the first time that an overshadowed feature might still be represented in category-specific regions of the brain, despite an absence of cognitive access to this information that is expressed in categorization behavior.

References

Abraham, A., Pedregosa, F., Eickenberg, M., Gervais, P., Mueller, A., Kossaifi, J., Gramfort ,A., Thirion, B., & Varoquaux, G. (2014). Machine learning for neuroimaging with scikit-learn. *Front. Neuroinform., 8(14).*

Acebes, F., Solar, P., Carnero, S., & Loy I. (2009). Blocking of conditioning of tentacle lowering in the snail (Helix aspersa). *The Quartely Journal of Experimental Psychology, 62(7), 1315-1327.*

Agcaoglu, O., Wilson, T. W., Wang, Y.-P., Stephen, J., & Calhoun, V. D. (2019). Resting state connectivity differences in eyes open versus eyes closed conditions. *Hum Brain Mapp, 40(8), 2488-2498.*

Aizenstein, H. J., MacDonald, A. W., Stenger, A., Nebes, R. D., Larson, J. K., Ursu, S., & Carter, C. S. (2000). Complementary category learning systems identified using event-related functional MRI. *Journal of Cognitive Neuroscience, 12(6), 977-987.*

Alahmadi, A. A. S., Pardini, M., Samson, R. S., D'Angelo, E., Friston, K. J., Toosy, A. T., & Wheeler-Kingshott, C. A. M. G. (2015). Differential involvement of cortical and cerebellar areas using dominant and nondominant hands: An fMRI study. *Human Brain Mapping, 36, 5079-5100.*

Alexander, G. E., DeLong, M. R., & Strick, P. L. (1986). Parallel organization of functionally segregated circuits linking basal ganglia and cortex. *Annu. Rev. Neurosci., 9, 357-381.*

Alexandre, A., Pedregosa, F., Eickenberg, M., Gervais, P., Mueller, A., Kossaifi, J., Gramfort, A., Thirion, B., & Varoquaux, G. (2014). Machine learning for neuroimaging with Scikit-Learn. *Frontiers in Neuroinformatics, 8(14).*

Alink, A., Krugliak, A., Walther, A., & Kriegeskorte, N. (2013). fMRI orientation decoding in V1 does not require global maps or globally coherent orientation stimuli. *Front. Psychol., 12*.

Amedi, A., Jacobson, G., Hendler, T., Malach, R., & Zohary, E. (2002). Convergence of visual and tactile shape processing in the human lateral occipital complex. *Cerebral Cortex, 12(11), 1202-1212.*

Andrews, T. J., Watson, D. M., Rice, G. E., & Hartley, T. (2015). Low-level properties of natural images predict topographic patterns of neural response in the ventral visual pathway. *Journal of Vision, 15(7), 1-12.*

Antzoulatos, E. G., & Miller, E. K. (2011). Differences between neural activity in prefrontal cortex and striatum during learning of novel abstract categories. *Neuron, 71(2), 243-249.*

Asaad, W. F., Rainer. G., & Miller, E. K. (1998). Neural activity in the primate prefrontal cortex during associative learning. *Neuron, 21, 1399-1407.*

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, 442-481.*

Ashby, F. G., & Ell, S. W. (2001). The neurobiology of human category learning. *Trends in Cognitive Sciences, 5, 204-210.*

Ashby, F. G., & Gott, R. E. (1988). Decision rules in the perception and categorization of multidimensional stimuli. *Journal of Experimental Psychology: Learning, Memory, & Cognition, 14, 33-53.*

Ashby, F. G., & Maddox, W. T. (1993). Relations between prototype, exemplar, and decision bound models of categorization. *Journal of Mathematical Psychology, 37(3), 372-400.*

Ashby, F. G., & Maddox, W. T. (2005). Human category learning. *Annu. Rev. Psychol., 56, 149-178.*

Ashby, F. G., & Maddox, W. T. (2011). Human category learning 2.0. *Ann. N. Y. Acad. Sci., 1224, 147-161.*

Ashby, F. G., Maddox, W. T., & Bohil, C. J. (2002). Observational versus feedback training in rule-based and information-integration category learning. *Memory & Cognition, 30(5), 666-677.*

Ashby, F. G., & O'Brien, J. B. (2005). Category learning and multiple memory systems. *TRENDS in Cognitive Sciences, 9(2), 83-89.*

Ashby, F. G., Smith, J. D., & Rosedahl, L. A. (2020). Dissociations between rule-based and information-integration categorization are not caused by differences in task difficulty. *Memory & Cognition, 48, 541-552.*

Ashby, F. G., & Valentin, V. V. (2017). Multiple systems of perceptual category learning: Theory and cognitive tests. In H. Cohen & C. Lefebvre (Eds.), *Handbook of categorization in cognitive science* (pp. 157–188). Elsevier Academic Press.

Assem, M., Glasser, M. F., Van Essen, D. C., & Duncan, J. (2020). A domain-general cognitive core defined in multimodally parcellated human cortex. *Cerebral Cortex, 30(8), 4361-4380.*

Avants, B.B., Epstein, C.L., Grossman, M., & Gee, J.C. (2008). Symmetric diffeomorphic image registration with cross-correlation: Evaluating automated labeling of elderly and neurodegenerative brain. *Medical Image Analysis, 12(1), 16-41.*

Avants, B. B., Tustison, N., & Johnson, H. (2009). Advanced normalization tools. *Insight J, 2, 1-35.*

Bansal, R., & Peterson, B. S. (2018). Cluster-level statistical inference in fMRI datasets: The unexpected behavior of random fields in high dimensions. *Magnetic Resonance Imaging, 49, 101-115.*

Bechtel, W. (1998). Representations and cognitive explanations: Assessing the Dynamicist's challenge in cognitive science. *Cognitive Science, 22(3), 295-318.*

Becker, S., & Wojtowicz, J. M. (2007). A model of hippocampal neurogenesis in memory and mood disorders. *Trends Cogn. Sci., 11, 70-76.*

Behzadi, Y., Restom, K., Liau, J., & Liu, T.T. (2007). A component based noise correction method (CompCor) for BOLD and perfusion based fMRI. *NeuroImage, 37(1), 90-101.*

Berns, G. S., Cohen, J. D., & Mintun, M. A. (1997). Brain regions responsive to novelty in the absence of awareness. *Science, 276, 1272-1275.*

Bettencourt, K. C., & Xu, Y. (2015). Decoding the content of visual short-term memory under distraction in occipital and parietal areas. *Nature Neuroscience, 19, 150-157.*

Blair, M. R., Watson, M. R., & Meier, K. M. (2009). Errors, efficiency, and the interplay between attention and category learning. *Cognition, 112(2), 330-336.*

Blaser, R. E., Couvillon, P. A., & Bitterman, M. E. (2004). Backward blocking in honeybees. *The Quarterly Journal of Experimental Psychology, 57B(4), 349-360.*

Bott, L., Hoffman, A.B., & Murphy, G.L. (2007). Blocking in category learning. *J Exp Psychol Gen., 136(4), 685-699.*

Bracci, S., Daniels, N., & Op de Beeck, H. (2017). Task context overrules object- and category-related representational content in the human parietal cortex. *Cerebral Cortex, 27(1), 310-321.*

Bracci, S., Ritchie, J. B., & Op de Beeck, H. (2017). On the partnership between neural representations of object categories and visual features in the ventral visual pathway. *Neuropsychologia, 105, 153-164.*

Braunlich, K., & Love, B. C. (2019). Occipitotemporal representations reflect individual differences in conceptual knowledge. *Journal of Experimental Psychology: General, 148(7), 1192-1203.*

Braunlich, K., & Seger, C. A. (2016). Categorical evidence, confidence, and urgency during probabilistic categorization. *NeuroImage, 125(15), 941-952.*

Bruner, J. S., Goodnow, J. J., Austin, G. A. (1956). *A study of thinking*. Wiley; Oxford.

Bundesen, C. (1998). A computational theory of visual attention. *Philos Trans R Soc Lond B Biol Sci., 353(1373), 1271-1281.*

Carlson, T. A. (2014). Orientation decoding in human visual cortex: New insights from an unbiased perspective. *J Neurosci., 34(24), 8373-8383.*

Carlson, T. A., Tovar, D. A., Alink, A., & Kriegeskorte, N. (2013). Representational dynamics of object vision: The first 1000 ms. *Journal of Vision, 13(1)*.

Carter, C. S., Braver, T. S., Barch, D. M., Botvinick, M. M., Noll, D., & Cohen, J. D. (1998). Anterior cingulate cortex, error detection, and the online monitoring of performance. *Science, 280, 747-749.*

Chadwick, M. J., Anjum, R. S., Kumaran, D., Schacter, D. L., Spiers, H. J., & Hassabis, D. (2016). Semantic representations in the temporal pole predict false memories. *PNAS, 113(36), 10180-10185.*

Chang, L., Sam., Jolly, E., Cheong, J. H., Burnashev, A., Chen, A., Clark, M., Frey, S., & Fitzpatrick, P. (2022). cosanlab/nltools: 0.4.7 (v0.4.7). *Zenodo*. https://doi.org/10.5281/zenodo.7015135

Chen, S., Weidner, R., Zeng, H., Fink, G. R., Muller, H. J., & Conci, M. (2020). Tracking the completion of parts into the whole objects: Retinotopic activation in response to illusory figures in the lateral occipital complex. *NeuroImage, 207, 116426.*

Cheng, P. W., & Pachella, R. G. (1984). A psychophysical approach to dimensional separability. *Cognitive Psychology, 16, 279-304.*

Cichy, R. M., Pantazis, D., & Oliva, A. (2014). Resolving human object recognition in space and time. *Nature Neuroscience, 17, 455-462.*

Chin-Parker, S., & Ross, B. H. (2002). The effect of category learning on sensitivity to within-category correlations. *Memory & Cognition, 30, 353-362.*

Chin-Parker, S., & Ross, B. H. (2004). Diagnosticity and prototypicality in category learning: A comparison of inference learning and classification learning. *Journal of Experimental Psychology: Learning, Memory, and Cognition, 30(1), 216-226.*

Cincotta, C. M., & Seger, C. A. (2007). Dissociation between striatal regions while learning to categorize during feedback and via observation. *J. Cogn. Neurosci., 19, 249-265.*

Clarke, A., & Tyler, L. K. (2014). Object-specific semantic coding in human perirhinal cortex. *The Journal of Neuroscience, 34(14), 4766-4775.*

Coggan, D. D., Liu, W., Baker, D. H., & Andrews, T. J. (2016). Category-selective patterns of neural response in the ventral visual pathway in the absence of categorical information. *NeuroImage, 135, 107-114.*

Connolly, A. C., Guntupalli, J. S., Gors, J., Hanke, M., Halchenko, Y. O., Wu, Y.-C., Abdi, H., & Haxby, J. V. (2012). The representation of biological classes in the human brain. *The Journal of Neuroscience, 32(8), 2608-2618.*

Contini, E. W., Wardle, S. G., & Carlson, T. A. (2017). Decoding the time-course of object recognition in the human brain: From visual features to categorical decisions. *Neuropsychologia, 105, 165-176.*

Couvillon, P. A., & Bitterman, M. E. (1989). Reciprocal overshadowing in the discrimination of color-odor compounds by honeybees: Further tests of a continuity model. *Animal Learning & Behavior, 17, 213-222.*

Cox, D. D., & Savoy, R. L. (2003). Functional magnetic resonance imaging (fMRI) “brain reading”: Detecting and classifying distributed patterns of fMRI activity in human visual cortex. *NeuroImage, 19, 261-270.*

Cox, R.W., & Hyde, J.S. (1997). Software tools for analysis and visualization of fMRI data. *NMR in Biomedicine, 10(4-5), 171-178.*

Crossley, M. J., & Ashby, F. G. (2015). Procedural learning during declarative control. *Journal of Experimental Psychology: Learning, Memory, & Cognition, 41(5), 1388-1403.*

Daniel, R., Wagner, G., Koch, K., Reichenbach, J. R., Sauer, H., & Schlösser, R. G. M. (2011). Assessing the neural basis of uncertainty in perceptual category learning through varying levels of distortion. *Journal of Cognitive Neuroscience, 23(7), 1781-1793.*

Davis, T., Goldwater, M., & Giron, J. (2017). From concrete examples to abstract relations: The rostrolateral prefrontal cortex integrates novel examples into relational categories. *Cerebral Cortex, 27, 2652-2670.*

Davis, T., Love, B. C., & Preston, A. R. (2012). Learning the exception to the rule: Model-based fMRI reveals specialized representations for surprising category members. *Cerebral Cortex, 22, 260-273.*

de Gardelle, V., Stokes, M., Johnen, V. M., Wyart, V., & Summerfield, C. (2013). Overlapping multivoxel patterns for two levels of visual expectation. *Front. Hum. Neurosci., 7, 158.*

Deneve, S., Latham, P. E., & Pouget, A. (1999). Reading population codes: A neural implementation of ideal observers. *Nature Neuroscience, 2(8), 740-745.*

Deng, W., & Sloutsky, V. M. (2016). Selective attention, diffused attention, and the development of categorization. *Cognitive Psychology, 91, 24-62.*

Desimone, R., Albright, T. D., Gross, C. G., & Bruce, C. (1984). Stimulus-selective properties of inferior temporal neurons in the macaque. *J. Neurosci., 4, 2051-2062.*

DiCarlo, J. J., Zoccolan, D., & Rust, N. C. (2012). How does the brain solve visual object recognition. *Neuron, 73(3), 415-434.*

Dickinson, A., Shanks, D., & Evenden, J. (1984). Judgement of Act-Outcome Contingency: The Role of Selective Attribution. *The Quarterly Journal of Experimental Psychology Section A, 36(1), 29-50.*

Duncan, J. (2010). The multiple-demand (MD) system of the primate brain: Mental programs for intelligent behavior. *Trends in Cognitive Sciences, 14(4), 172-179.*

Edelman, S. (1998). Representation is representation of similarities. *Behavioral and Brain Sciences, 21, 449-498.*

Edelman, S., Grill-Spector, K., Kushnir, T., & Malach, R. (1998). Toward direct visualization of the internal shape representation space by fMRI. *Psychobiology, 26(4), 309-321.*

Edgell, S. E., Castellan, N. J., Roe, R. M., Barnes, J. M., Ng, P. C., Bright, R. D., & Ford, L. A. (1996). Irrelevant information n probabilistic categorization. *Journal of Experimental Psychology: Learning, Memory, and Cognition, 22, 1463-1481.*

Egger, M. C., & Miller, N. E. (1962). Secondary reinforcement in rats as a function of information value and reliability of the stimulus. *Journal of Experimental Psychology, 64(2), 97-104.*

Eklund, A., Nichols, T. E., & Knutsson, H. (2016). *PNAS, 113(28), 7900-7905.*

Elison, J. M., Leggit, V. L., Thomson, M., Oyoyo, U., & Wycliffe, N. D. (2008). Influence of common orthodontic appliances on the diagnostic quality of cranial magnetic resonance images. *American Journal of Orthodontics and Dentofacial Orthopedics, 134(4), 563-572.*

Erickson, M. A., & Kruschke, J. K. (1998). Rules and exemplars in category learning. *Journal of Experimental Psychology: General, 127, 107-140.*

Esteban, O., Markiewicz, C.J., Blair, R.W., Moodie, C.A., Isik, I., Erramuzpe, A., Kent, J.D., Concalves, M., DuPre, E., Snyder, M., Oya, H., Ghosh, S.S., Wright, J., Durnez, J., Poldrack, R.A., & Gorgolewski, K.J. (2019). fMRIPrep: A robust preprocessing pipeline for functional MRI. *Nature Methods, 16, 111-116.*

Esteban, O., Markiewicz, C.J., Burns, C., Goncalves, M., Dorota, J., Ziegler, E., Soshana, B., Gage, D.E., Pinsard, B., Madison, C., Waskom, M., Notter, M.P., Clark, D., Manhães-Savio, A., Clark, D., Jordan, K., Dayan, M., Halchenko, Y.O., Loney, F., Salo, T., … Ghosh, S. (2020). nipy/nipype: 1.5.1 (Version 1.5.1). *Zenodo.* https://doi.org/10.5281/zenodo.4035081

Esteban, O., Markiewicz, C.J., Goncalves, M., DuPre, E., James, K.D., Salo, T., Rastko, C., Basile, P., Ross, B.W., Poldrack, R.A., & Gorgolewski, K.J. (2020). fMRIPrep: A robust preprocessing pipeline for functional MRI (Version 20.2.1). *Zenodo*. https://doi.org/10.5281/zenodo.4252786

Ethofer, T., Van De Ville, D., Scherer, K., & Vuilleumier, P. (2009). Decoding of emotional information in voice-sensitive cortices. *Current Biology, 19(12), 1028-1033.*

Etzel, J. A., Zacks, J. M., & Braver, T. S. (2013). Searchlight analysis: Promise, pitfall, and potential. *NeuroImage, 78, 261-269.*

Flaherty, A. W., & Graybiel, A. M. (1991). Corticostriatal transformations in the primate somatosensory system: Projections from physiologically mapped body-part representations. *J. Neurophysiol., 66, 1249-1263.*

Folstein, J. R., Palmeri, T. J., & Gauthier, I. (2013). Category learning increases discriminability of relevant object dimensions in visual cortex. *Cerebral Cortex, 23, 814-823.*

Folstein, J. R., Palmeri, T. J., Van Gulick, A. E., Gauthier, I. (2015). Category learning stretches neural representations in visual cortex. *Curr Dir Psychol Sci., 24(1), 17-23.*

Fonov, V.S., Evans, A.C., McKinstry, R.C., Almli, C.R., & Collins, D.L. (2009). Unbiased nonlinear average age-appropriate brain templates from birth to adulthood. *NeuroImage, 47, Supplement 1: S102.*

Freedberg, M., Toader, A. C., Wassermann, E. M., & Voss, J. L. (2020). Competitive and cooperative interactions between medial temporal and striatal learning systems. *Neuropsychologia, 136, 107257.*

Freedman, D. J., Riesenhuber, M., Poggio, T., & Miller, E. K. (2003). A comparison of primate prefrontal and inferior temporal cortices during visual categorization. *J. Neurosci., 23, 5235-5246.*

Friston, K. J., Ashburner, J., Barnes, G., Flandin, G., Gitelman, D., Glauche, V., Hutton, C., Litvak, V., Moran, R., Oostenveld, R., Penny, W., Phillips, C., Pinotsis, D., Ridgway, G., Seghier, M., Stephan, K. E., Andersson, J., Brett, M., Buechel, C., Chen, C. C., Chumbley, J., Daunizeau, J., Harrison, L., Heather, J., Henson, R., Holmes, A., Rosa, M., Kiebel, S., Kilner, J., Mattout, J., Nichols, T., Poline, J. B., Worsley, K. J. (2014). "Statistical Parametric Mapping." The Wellcome Trust Centre for Neuroimaging, URL http://www.fil.ion.ucl.ac.uk/spm/.

Friston, K. J., Holmes, A. P., Poline, J.-B., Grasby, P. J., Williams, S. C. R., Frackowiak, R. S. J., & Turner. R. (1995). Analysis of fMRI time-series revisited. *NeuroImage, 2, 45-53.*

Friston, K. J., Frith, C. D., Passingham, R. E., Liddle, P. F., & Frackowiak, R. S. J. (1992). Motor practice and neurophysiological adaptation in the cerebellum: A positron tomography study. *Proc. R. Soc. London, 248, 223-228.*

Friston, K. J., Jezzard, P., & Turner, R. (1994). Analysis of functional MRI time-series. *Human Brain Mapping, 1, 153-171.*

Gabrieli, J. D. E. (1998). Cognitive neuroscience of human memory. *Annual Review of Psychology, 49, 87-115.*

Gazzaley, A., Cooney, J. W., McEvoy, K., Knight, R. T., & D'Esposito, M. (2005). Top-down enhancement and suppression of the magnitude and speed of neural activity. *J Cogn Neurosci., 17(3), 507-517.*

Giovanello, K. S., Schnyer, D., & Verfaellie, M. (2008). Distinct hippocampal regions make unique contributions to relational memory. *Hippocampus, 19(2), 111-117.*

Gluck, M. A., & Bower, G. H. (1988). From conditioning to category learning: An adaptive network model. *Journal of Experimental Psychology: General, 117, 227-247.*

Gluck, M. A., Meter, M., & Myers, C. E. (2003). Computational models of the hippocampal region: Linking incremental learning and episodic memory. *Trends Cogn. Sci., 7, 269-276.*

Goldstone, R. L. (1996). Isolated and interrelated concepts. *Memory & Cognition, 24, 608-628*.

Gordon, E. M., Chauvin, R. J., Van, A. N., Rajesh, A., Nielsen, A., Newbold, D. J., Lynch, C. J., Krimmel, S. R., Scheidter, K. M., Monk, J., Miller, R. L., Metoki, A., Montez, D. F., Zheng, A., Elbau, I., Madison, T., Nishino, T., Myers, M. J., … Dosenbach, U. F. (2023). A somato-cognitive action network alternates with effector regions in motor cortex. *Nature, 617, 351-359.*

Gorgolewski, K.J., Burns, C.D., Madison, C., Clark, D., Halchenko, Y.O., Waskom, M.L., & Ghosh, S. (2011). Nipype: A flexible, lightweight and extensible neuroimaging data processing framework in Python. *Frontiers in Neuroinformatics, 5(13).*

Greve, D.N., & Fischl, B. (2009). Accurate and robust brain image alignment using boundary-based registration. *NeuroImage, 48(1), 63-72.*

Grill-Spector, K., Kourtzi, Z., & Kanwisher, N. (2001). The lateral occipital complex and its role in object recognition. *Vision Research, 41, 1409-1422.*

Gross, C. G. (1994). How inferior temporal cortex became an object area. *Cerebral Cortex, 5, 455-469.*

Grossman, M., Smith, E. E., Koenig, P., Glosser, G., DeVita, C., Moore, P., & McMillan, C. (2002). The neural basis for categorization in semantic memory. *NeuroImage, 17(3), 1549-1561.*

Gureckis, T. M., & Love, B. C. (2003). Towards a unified account of supervised and unsupervised category learning. *J. Expt. Theor. Artif. Intell., 15, 1-24.*

Gusnard, D. A., & Raichle, M. E. (2001). Searching for a baseline: Functional imaging and the resting human brain. *Nature Reviews Neuroscience, 2, 685-694.*

Hammer, R. (2015). Impact of feature saliency on visual category learning. *Front. Psychol., 6.*

Hammer, R., & Sloutsky, V. (2016). Visual category learning results in rapid changes in brain activation reflecting sensitivity to the category relations beween perceived objects and to decision correctness. *Journal of Cognitive Neuroscience, 28(11), 1804-1819.*

Harris, C. R., Millman, K. J., van der Walt, S. J., Gommers, R., Virtanen, P., Cournapeau, D., Wieser, E., Taylor, J., Berg, S., Smith, N. J., Kern, R., Picus, M., Hoyer, S., van Kerkwijk, M.

H., Brett, M., Haldane, A., del Río, J. F., Wiebe, M., Peterson, P., … Oliphant, T. E. (2020). Array programming with NumPy. *Nature, 585, 357-362.*

Harris, S. A., & Tong, F. (2009). Decoding reveals the contents of visual working memory in early visual areas. *Nature, 458(7238), 632-635.*

Hasanin, M., Kaplan, S. E. F., Hohlen, B., Lai, C., Nagshabandi, R., Zhu, X., & Al-Jewair, T. (2019). Effects of orthodontic appliances on the diagnostic capability of magnetic resonance imaging in the head and neck region: A systematic review. *International Orthodontics, 17(3), 43-414.*

Haxby, J. V., Connolly, A. C., & Guntupalli, J. S. (2014). Decoding neural representational spaces using multivariate pattern analysis. *Annu. Rev. Neurosci., 37, 435-456.*

Haxby, J. V., Gobbini, M. I., Furey, M. L., Ishai, A., Schouten, J. L., & Pietrini, P. (2001). Distributed and overlapping representations of faces and objects in ventral temporal cortex. *Science, 293, 2425-2430.*

Haynes, J.-D. (2015). A primer on pattern-based approached to fMRI: Principles, pitfalls, and perspectives. *Neuron, 87, 257-270.*

Haynes, J.-D., & Rees, G. (2005). Predicting the orientation of invisible stimuli from activity in human primary visual cortex. *Nature Neuroscience, 8(5), 686-691.*

He, C., Hung, S.-C., & Cheung, O. S. (2020). Roles of category, shape, and spatial frequency in shaping animal and tool selectivity in the occipitotemporal cortex. *The Journal of Neuroscience, 40(29), 5644-5657.*

Heekeren, H. R., Marrett, S., & Ungerleider, L. G. (2008). The neural systems that mediate human perceptual decision making. *Nature Reviews Neuroscience, 9, 467-479.*

Helie, S., & Ashby, G.F. (2012). Learning and transfer of category knowledge in an indirect categorization task. *Psychological Research, 76, 292-303.*

Helie, S., Waldschmidt, J. G., & Ashby, F. G. (2010). Automaticity in rule-based and information-integration categorization. *Attent, Percept Psychophys, 72, 1013-1031.*

Hirose, S., Nambu, I., & Naito, E. (2018). Cortical activation associated with motor preparation can be used to predict the freely chosen effector of an upcoming movement and reflects response time: An fMRI decoding study. *NeuroImage, 183, 584-596.*

Hochstein, S., & Ahissar, M. (2002). View from the top: Hierarchies and reverse hierarchies in the visual system. *Neuron, 36(5), 791-804.*

Hoffman, A. B., & Murphy, G. L. (2006). Category dimensionality and feature knowledge: When more features are learned as easily as fewer. *J Exp Psychol Learn Mem Cog, 32(2), 301-315.*

Hoffman, A. B., & Rehder, B. (2010). The costs of supervised classification: The effect of learning task on conceptual flexibility. *Journal of Experimental Psychology: General, 139(2), 319-340.*

Hoover, J. E., & Strick, P. L. (1999). The organizatino of cerebellar and basal ganglia outputs to primary motor cortex as revealed by retrograde transneuronal transport of herpes simplex virus type 1. *The Journal of Neuroscience, 19(4), 1446-1463.*

Hubel, D. H., & Wiesel, T. N. (1962). Receptive fields, binocular interaction and functional architecture in the cat's visual cortex. *J Physiol., 160, 106-154*.

Hull, C. L. (1943). *Principles of behavior.* New York: Appleton-Century-Crofts.

Jenkins, D., & Kirkpatrick, K. (2006). Interval duration effects on blocking in appetitive conditioning. *Behavioural Processes, 2-3, 318-329.*

Jenkinson, M., Bannister, P., Brady, M., & Smith, S. (2002). Improved optimization for the robust and accurate linear registration and motion correction of brain images. *NeuroImage, 17(2), 825-841.*

Jenkinson, M., & Smith, S. (2001). A global optimisation method for robust affine registration of brain images. *Medical Image Analysis, 5(2), 143-156.*

Jiang, X., Bradley, E., Rini, R. A., Zeffiro, T., Vanmeter, J., Riesenhuber, M. (2007). Categorization training results in shape- and category-selective human neural plasticity. *Neuron, 53, 891-903*.

Kamin, L. J. (1965). Temporal and intensity characteristics of the conditioned stimulus. In W. F. Prokasy (Eds.), *Classical Conditioning* (pp. 118-147). New York: Appleton-Century-Crofts.

Kamin, L. J. (1967). "Attention-like" processes in classical conditioning. In M. R. Jones (Eds.), *Miami symposium on the prediction of behavior*, *1967: Aversive stimulation* (pp. 9-31). Coral Gables, Fl. University of Miami Press.

Kamin, L. J. (1969). Predictability, surprise, attention and conditioning. In B. A. Campbell & R. M. Church (Eds.), *Punishment and Aversive Behavior* (pp. 279-296). New York: Appleton-Century-Crofts.

Kamitani, Y., & Tong, F. (2005). Decoding the visual and subjective contents of the human brain. *Nature Neuroscience, 8(5), 679-685.*

Kanwisher, N., McDermott, J., & Chun, M. M. (1997). The Fusiform Face Area: A module in human extrastriate cortex specialized for face perception. *The Journal of Neuroscience, 17(11), 4302-4311.*

Kanwisher, N., Woods, R. P., Iacoboni, M., & Mazziota, J. (1997). A locus in human extrastriate cortex for visual shape analysis. *Journal of Cognitive Neuroscience, 9, 133-142.*

Kemp, J. M., & Powell, T. P. (1970). The cortico-striate projection in the monkey. *Brain, 93, 525-546.*

Kim, I., Hong, S. W., Shevell, S. K., & Shim, W. M. (2020). Neural representations of perceptual color experience in the human ventral visual pathway. *PNAS, 117(23), 13145-13150.*

Kim, J.-K., & Zatorre, R. J. (2011). Tactile-auditory shape learning engages the lateral occipital complex. *The Journal of Neuroscience, 31(21), 7848-7856.*

Kowler, E., Anderson, E., Dosher, B., & Blaser, E. (1995). The role of attention in the programming of saccades. *Vision Res., 35(13), 1897-1916.*

Kriegeskorte, N., & Bandettini, P. (2007). Analyzing for information, not activation, to exploit high-resolution fMRI. *NeuroImage, 38(4), 649-662.*

Kriegeskorte, N., Goebel, R., & Bandettini, P. (2006). Information-based functional brain mapping. *PNAS, 103(10), 3863-3868.*

Kriegeskorte, N., & Kievit, R. A. (2013). Representational geometry: Integrating cognition, computation, and the brain. *Trends in Cognitive Sciences, 17(8), 401-412.*

Kriegeskorte, N., Mur, M., & Bandettini, P. A. (2008). Representational similarity analysis – connecting the branches of systems neuroscience. *Frontiers in Systems Neuroscience, 2(4), 1-28.*

Kriegeskorte, N., Mur, M., Ruff, D. A., Kiani, R., Bodurka, J., Esteky, H., Tanaka, K., & Bandettini, P. A. (2008). Matching categorical object representations in inferior temporal cortex of man and monkey. *Neuron, 60, 1126-1141.*

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

Kruschke, J. K. (1993). Three principles for models of category learning. In Nakamura, G. V., Taraban, R., & Medin, D. L. (Eds), *The psychology of learning and motivation (vol. 29): Categorization by humans and machines* (pp. 57-90). Academic Press.

Kruschke, J. K., & Johansen, M. K. (1999). A model of probabilistic category learning. *Journal of Experimental Psychology: Learning, Memory, and Cognition, 25(5), 1083-1119.*

Kumar, M., Anderson, M. J., Antony, J. W., Baldassano, C., Brooks, P. P., Cai, M. B., Chen, P.-H., Ellis, C., Henselman-Petrusek, G., Huberdeau, D., Hutchinson, J. B., Li, Y. P., Lu, Q., Manning, J. R., Mennen, A. C., Nastase, S. A., Richard, H., Schapiro, A. C., Schuck, N. W., … Norman, K. A. (2020). BrainIAK: The Brain Imaging Analysis Kit. OSF. https://doi.org/10.31219/osf.io/db2ev.

Kumar, M., Ellis, C. T., Lu, Q., Zhang, H., Capotă, M., Willke, T. L., Ramadge, P. J., Turk-Browne, N. B., & Norman, K.A. (2020). BrainIAK tutorials: User-friendly learning materials for advanced fMRI analysis. *PLoS Computational Biology, 16(1), e1007549*.

Kwong, K. K., Belliveau, J. W., Chesler, D. A., Goldberg, I. A., Weisskoff, R. M., Poncelet, B. P., Kennedy, D. N., Hoppel, B. E., Cohen, M. S., Turner. R., Cheng., H.-M., Brady, T. J., & Rosen, B. R. (1992). Dynamic magnetic resonance imaging of human brain activity during primary sensory stimulation. *Proc. Natl. Acad. Sci. USA, 89, 5675-5679.*

LaBar, K. S., Gitelman, D. R., Parrish, T. B., & Mesulam, M.-M. (1999). Neuroanatomic overlap of working memory and spatial attention notworks: A functional MRI comparison within subjects. *NeuroImage, 10, 695-704.*

Lanczos, C. (1964). Evaluation of noisy data. *Journal of the Society for Industrial and Applied Mathematics Series B Numerical Analysis 1(1), 76-85.*

Lassaline, M. E., & Murphy, G. L. (1996). Induction and category coherence. *Psychonomic Bulletin & Review, 3(1), 95-99.*

Lattal, K. M., & Nakajima, S. (1998). Overexpectation in appetitive Pavlovian and instrumental conditioning. *Animal Learning & Behavior, 26, 351-360.*

Lau, J. S.-H., Casale, M.B., & Pashler, H. (2020). Mitigating cue competition effects in human category learning. *Quarterly Journal of Experimental Psychology, 73(7), 983-1003.*

Levering, K. R., & Kurtz, K. J. (2015). Observation versus classification in supervised category learning. *Mem Cogn, 43, 266-282.*

Little, D. M., Klein, R., Shobat, D. M., McClure, E. D., & Thulborn, K. R. (2004). Changing patterns of brain activation during category learning revealed by functional MRI. *Cognitive Brain Research, 22(1), 84-93.*

Little, D. M., & Thulborn, K. R. (2006). Prototype-distortion category learning: A two-phase learning process across a distributed network. *Brain and Cognition, 60(3), 233-243.*

Livingstone, M. S., & Hubel, D. H. (1984). Anatomy and physiology of a color system in the primate visual cortex. *J Neurosci., 4(1), 309-356.*

Logan, G. D. (2004). Cumulative progress in formal theories of attention. *Annual Review of Psychology, 55, 207-234.*

Lopez-Paniagua, D., & Seger, C. A. (2011). Interactions within and between corticostriatal loops during component processes of category learning. *Journal of Cognitive Neuroscience, 23(10), 3068-3083.*

Love, B. C. (2001). Three deadly sins of category learning modelers. *Behavioral & Brain Sciences, 24, 687-688.*

Love, B. C. (2002). Comparing supervised and unsupervised category learning. *Psychonomic Bulletin & Review, 9(4), 829-835.*

Love, B. C. (2003). The multifaceted nature of unsupervised category learning. *Psychonomic Bulletin & Review, 10(1), 190-197.*

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

Mackintosh, N. J. (1965). Selective attention in animal discrimination learning. *Psychological Bulletin, 64, 124-150.*

Mackintosh, N. J. (1976). Overshadowing and stimulus intensity. *Animal Learning and Behavior, 4(2), 186-192.*

Maddox, W. T., Glass, B. D., O'Brien, J. B., Filoteo, J. V., & Ashby, F. G. (2010). Category label and response location shifts in category learning. *Psychological Research, 74, 219-236.*

Maes, E., Boddez, Y., Joaquín, M.A., Krypotos, A.-M., D'Hooge, R., De Houwer, J., & Beckers, T. (2016). The elusive nature of the blocking effect: 15 failures to replicate. *J Exp Psychol Gen., 145(9), e49-71.*

Mahmoudi, A., Takerkart, S., Regragui, F., Boussaoud, D., & Brovelli, A. (2012). Multivoxel pattern analysis for fMRI data: A review. *Computational and Mathematical Methods in Medicine, 2012.*

Malach, R. (2012). Targeting the functional properties of cortical neurons using fMR-adaptation. *NeuroImage, 62(2), 1163-1169.*

Malach, R., Reppas, J. B., Benson, R. R., Kwong, K. K., Jiang, J. Kennedy, W. A., Ledden, P. J., Brady, T. J., Rosen, B. R., & Tootell, R. B. H. (1995). Object-related activity revealed by functional magnetic resonance imaging in human occipital cortex. *PNAS, 92, 8135-8139.*

Markman, A. B. (1989). *Categorization and naming in children: Problems of induction.* MIT Press.

Markman, A. B., & Ross, B. H. (2003). Category use and category learning. *Psychol. Bull., 129(4), 592-613.*

MATLAB. (2019). version 9.6.0.1135713 (R2019a) Update 3. Natick, Massachusettes: The MathWorks Inc.

McCormick, C., Moscovitch, M., Protzner, A. B., Huber, C. G., & McAndrews, M. P. (2010). Hippocampal-neocortical networks differ during encoding and retrieval of relational memory: Functional and effective connectivity analysis. *Neuropsychologia, 48(11), 3272-3281.*

Medin, D. L. (1989). Concepts and conceptual structure. *American Psychologist, 44(12), 1469-1481.*

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

Melchers, K. G., Shanks, D. R., & Lachnit, H. (2008). Stimulus coding in human associative learning: Flexible representations of parts and wholes. *Behavioural Processes, 77, 413-427.*

Melchers, K. G., Üngör, M., & Lachnit, H. (2005). The experimental task influences cue competition in human causal learning. *Journal of Experimental Psychology: Animal Behavior Processes, 31(4), 477–483.*

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

Minda, J. P., & Ross, B. H. (2004). Learning categories by making predictions: An investigation of indirect category learning. *Memory & Cognition, 32(8), 1355-1368.*

Minda, J. P., & Smith, J. D. (2001). Prototypes in category learning: The effects of category size, category structure, and stimulus complexity. *Journal of Experimental Psychology: Learning, Memory, and Cognition, 27, 775-799.*

Mink, J. W. (1996). The basal ganglia: Focused selection and inhibition of competing motor programs. *Progress in Neurobiology, 50(4), 381-425.*

Misaki, M., Kim, Y., Bandettini, P. A., & Kriegeskorte, N. (2010). Comparison of multivariate classifiers and response normalizations for pattern-information fMRI. *NeuroImage, 53, 103-118.*

Mok, R. M., & Love, B. C. (2022). Abstract neural representations of category membership beyond information coding stimulus or response. *Journal of Cognitive Neuroscience, 34(10), 1719-1735.*

Monchi, O., Petrides, M., Petre, V., Worsley, K., & Dagher, A. (2001). Wisconsin Card Sorting revisited: Distinct neural circuits participating in different stages of the task identified by event-related functional magnetic resonance imaging. *Journal of Neuroscience, 21(19), 7733-7741.*

Moran, J., & Desimone, R. (1985). Selective attention gates visual processing in extrastriate cortex. *Science, 229, 782-784.*

Muhammad, R., Wallis, J. D., & Miller, E. K. (2006). A comparison of abstract rules in the prefrontal cortex, premotor cortex, inferior temporal cortex, and striatum. *Journal of Cognitive Neuroscience, 18(6), 974-989.*

Mumford, J. A., Davis, T., & Poldrack, R. A. (2014). The impact of study design on pattern estimation for single-trial multivariate pattern analysis. *NeuroImage, 103, 130-138.*

Mumford, J. A., Turner, B. O., Ashby, F. G., & Poldrack, R. A. (2012). Deconvolving BOLD activation in event-related designs for multivoxel pattern classification analyses. *NeuroImage, 59, 2636-2643.*

Murphy, G. L. (2002). *The big book of concepts.* MIT Press.

Murphy, G. L., & Dunsmoor, J. E. (2017). Do salient features overshadow learning of other features in category learning? *J Exp Psychol Anim Learn Cogn., 43(3), 219-230.*

Murphy, G. L., & Medin, D. L. (1985). The role of theories in conceptual coherence. *Psychological Review, 92, 289-316.*

Naselaris, T., Kay, K. N., Nishimoto, S., & Gallant, J. (2011). Encoding and decoding in fMRI. *NeuroImage, 56, 400-410.*

Ng, A. Y., & Jordan, M. I. (2002). On discriminative vs. generative classifiers: A comparison of logistic regression and naïve Bayes. *Advances in Neural Information Processing Systems, 14, 841-848.*

Nichols, T., & Hayasaka, S. (2003). Controlling the familywise error rate in functional neuroimaging: A comparative review. *Statistical Methods in Medical Research, 12, 419-446.*

Nomura, E. M., Maddox, W. T., Filoteo, J. V., Ing, A. D., Gitelman, D. R. Parrish, T. B., Mesulam, M.-M., & Reber, P. J. (2007). Neural correlates of rule-based and information-integration visual category learning. *Cereb. Cortex, 17(1), 37-43.*

Nomura, E. M., & Reber, P. J. (2008). A review of medial temporal lobe and caudate contributions to visual category learning. *Neuroscience & Biobehavioral Reviews, 32(2), 279-291.*

Norman, K. A., Polyn, S. M., Detre, G. J., & Haxby, J. V. (2006). Beyond mind-reading: multi-voxel pattern analysis of fMRI data. *Trends Cogn Sci., 10(9), 424-430.*

Nosofsky, R. M. (1984). Choice, similarity, and the context theory of classification. *Journal of Experimental Psychology: Learning, Memory, and Cognition, 10(1), 104-114.*

Nosofsky, R. M. (1986). Attention, similarity, and the identification-categorization relationship. *J Exp Psychol Gen, 115(1), 39-61.*

Nosofsky, R. M., Gluck, M. A., Palmeri, T. J., McKinley, S. C., & Glauthier, P. (1994). Comparing models of rule-based classification learning: A replication and extension of Shepard, Hovland, and Jenkins (1961). *Memory & Cognition, 22, 352-369*.

Nosofsky, R. M., & Johansen, M. K. (2000). Exemplar-based accounts of "multiple-system" phenomena in perceptual categorization. *Psychon Bull Rev, 7(3), 375-402.*

Nosofsky, R. M., Palmeri, T. J., & McKinley, S. C. (1994). Rule-plus exception model of classification learning. *Psychological Review, 101(1), 53-79.*

O'Craven, K. M., Downing, P. E., & Kanwisher, N. (1999). fMRI evidence for objects as the units of attentional selection. *Nature, 401, 584-587.*

Olszowy, W., Aston, J., Rua, C., & Williams, G. B. (2019). Accurate autocorrelation modeling substantially improves fMRI reliability. *Nature Communications, 10(1220), 1-11.*

Oosterhof, N., Tipper, s., & Downing, P. (2012). Visuo-motor imagery of specific manual actions: A multi-variate pattern analysis fMRI study. NeuroImage, 63, 262-271.

Op de Beeck, H. P. (2010). Against hyperacuity in brain reading: Spatial smoothing does not hurt multivariate fMRI analysis? *NeuroImage, 49(3), 1943-1948.*

Op de Beeck, H. P., Wagemans, J., & Vogels, R. (2008). The representation of perceived shape similarity and its role for category learning in monkeys: A modeling study. *Vision Research, 48(4), 598-610.*

Palejwala, A. H., Dadario, N. B., Young, I. M., O'Connor, K., Briggs, R. G., Conner, A. K., O'Donoghue, D. L., & Sughrue, M. E. (2021). Anatomy and white matter connections of the lingual gyrus and cuneus. *World Neurosurgery, 151, e426-e437.*

Palmeri, T. J., & Flanery, M. A. (2002). Memory systems and perceptual categorization. *Psychology of Learning and Motivation, 41, 141-189.*

Panzeri, S., Macke, J. H., Gross, J., & Kayser, C. (2015). Neural population coding: Combining insights from microscopic and mass signals. *Trends in Cognitive Sciences, 19(3), 162-172.*

Patriat, R., Molloy, E. K., Meier, T. B., Kirk, G. R., Nair, V. A., Meyerand, M. E., Prabhakaran, V., & Birn, R. M. (2013). The effect of resting condition on resting-state fMRI reliability and consistency: A comparison between resting with eyes open, closed, and fixated. *NeuroImage, 78, 463-473.*

Paus, T., Petrides, M., Evans, A. C., & Meyer, E. (1993). Role of the human anterior cingulate cortex in the control of oculomotor, manual, and speech responses: A positron emission tomography study. *J Neurophysiol, 70, 453-469.*

Pavlov, I. P. (1927). *Conditioned reflexes.* (G. V. Anrep., Trans.). London: Oxford University Press.

Pearce, J. M. & Bouton, M. E. (2001). Theories of associative learning in animals. *Annual Review of Psychology, 52, 111-139.*

Pedregosa, F., Varoquaux, G., Gramfort, A., Michel, V., Thirion, B., Grisel, O., Blondel, M., Prettenhofer, P., Weiss, R., Dubourg, V., Vanderplas, J., Passos, A., Cournapeau, D., Brucher, Mm., Perrot, M., & Duchesnay, E. (2011). Scikit-learn: Machine learning in Python. *Journal of Machine Learning Research, 12(85), 2825-2830.*

Peirce, J. W., Gray, J. R., Simpson, S., MacAskill, M. R., Höchenberger, R., Sogo, H., Kastman, E., & Lindeløv, J. (2019). PsychoPy2: Experiments in behavior made easy. *Behavior Research Methods, 51, 195-203.*

Pietrini, P., Furey, M. L., Ricciardi, E., Gobbini, M. I., Wu, W.-H. C., Cohen, L., Guazzelli, M., & Haxby, J. V. (2004). Beyond sensory images: Object-based representation in the human ventral pathway. *Biological Sciences, 101(15), 5658-5663.*

Plebanek, D. J., & Sloutsky, V. M. (2017). Costs of selective attention: When children notice what adults miss. *Psychological Science, 28, 723-732.*

Poldrack, R. A. (2011). Inferring mental states from neuroimaging data: From reverse inference to large-scale decoding. *Neuron, 72, 692-697.*

Poldrack, R. A., Baker, C. I., Durnez, J., Gorgolewski, K. J., Matthews, P. M., Munafò, M. R., Nichols, T. E., Poline, J.-B., Vul, E., & Yarkoni, T. (2017). Scanning the horizon: Towards transparent and reproducible neuroimaging research. *Nature Reviews Neuroscience, 18, 115-126.*

Poldrack, R. A., Clark, J., Pare-Blagoev, E. J., Shohamy, D., Moyano, J. C., Myers, C., & Gluck, M. A. (2001). Interactive memory systems in the human brain. *Nature, 414, 546-550.*

Poldrack, R. A., Desmond, J. E., Glover, G. H., & Gabrieli, J. D. E. (1998). The neural basis of visual skill learning: An fMRI study of mirror reading. *Cerebral Cortex, 8, 1-10.*

Poldrack, R. A., Halchenko, Y. O., & Hanson, S. J. (2009). Decoding the large-scale structure of brain function by classifying mental states across individuals. *Psychological Science, 20(11), 1364-1372.*

Poldrack, R. A., & Foerde, K. (2008). Category learning and the memory systems debate. *Neuroscience and Biobehavioral Reviews, 32(2), 197-205.*

Poldrack, R. A., Prabhakaran, V., Seger, C. A., & Gabrieli, J. D. E. (1999). Striatal activation during acquisition of a cognitive skill. *Neuropsychology, 13, 564-574.*

Poldrack, R. A., & Rodriguez, P. (2004). How do memory systems interact? Evidence from human classification learning. *Neurobiology of Learning and Memory, 82(3), 324-332.*

Polk, T. A., Drake, R. M., Jonides, J. J., Smith, M. R., & Smith, E. E. (2008). Attention enhances the neural processing of relevant features and suppresses the processing of irrelevant

features in humans: A functional magnetic resonance imaging study of the Stroop task. *The Journal of Neuroscience, 28(51), 13786-13792*.

Popov, V., Ostarek, M., & Tenison, C. (2018). Practices and pitfalls in inferring neural representations. *NeuroImage, 174, 340-351.*

Posner, M. I. (1980). Orienting of attention. *Quartely Journal of Experimental Psychology, 32, 3-25.*

Postle, B. R. (2006). Worknig memory as an emergent property of the mind and brain. *Neuroscience, 139(1), 23-38.*

Pothos, E. M., & Chater, N. (2005). Unsupervised categorization and category learning. *The Quartely Journal of Experimental Psychology, 58A(4), 733-752.*

Pouget, A., Dayan, P., & Zemel, R. (2000). Information processing with population codes. *Nature Reviews Neuroscience, 1, 125-132.*

Power, J.D., Mitra, A., Laumann, T.O., Snyder, A.Z., Schlaggar, B.L., & Petersen, S.E. (2014). Methods to detect, characterize, and remove motion artifact in resting state fMRI. *NeuroImage, 84 (Supplement C), 320-341.*

Prados, J., Alvarez, B., Acebes, F., Loy, I., Sansa, J., & Moreno-Fernández, M. M. (2013). Blocking in humans, rats and snails using a within-subjects design. *Behavioural Processes, 100, 23-31.*

Purdon, P. L., & Weisskoff, R. M. (1995). Effect of temporal autocorrelation due to physiological noise and stimulus paradigm on voxel-level false-positive rates in fMRI. *Human Brain Mapping, 6, 239-249.*

Python Software Foundation. Python Language Reference, version 3.7.10. Available at http://www.python.org.

Raizada, R. D. S., Tsao, F.-M., Liu, H.-M., & Kuhl, P. K. (2010). Quantifying the adequacy of neural representations for a cross-language phonetic discrimination task: Prediction of individual differences. *Cerebral Cortex, 20, 1-12.*

Reback, J., jbrockmendel, McKinney, W., Van den Bossche, J., Augspurger, T., Cloud, P., Hawkins, S., gfyoung, Sinhrks, Roeschke, M., Klein, A., Petersen, T., Tratner, J., She, C., Ayd, W., Hoefler, P., Naveh, S., Garcia, M., Schendel, J., … Dong, K. (2021). pandas-dev/pandas: Pandas 1.3.0 (v1.3.0). Zenodo. https://doi.org/10.5281/zenodo.5060318

Reber, A. S., & Millward, R. B. (1968). Event observation in probability learning. *Journal of Experimental Psychology, 77(2), 317-327.*

Reber, P. J., Stark, C. E. L., & Squire, L. R. (1998). Cortical areas supporting category learning identified using functional MRI. *PNAS, 95(2), 747-750.*

Reber, P, J., Gitelman, D. R., Parrish, T. B., & Mesulam, M. M. (2003). Dissociating explicit and implicit category knowledge with fMRI. *Journal of Cognitive Neuroscience, 15(4), 574-583.*

Redgrave, P., Rodriguez, M., Smith, Y., Rodriguez-Oroz, M. C., Lehericy, S., Bergman, H., Agid, Y., DeLong, M. R., and Obeso, J. A. (2010). Goal-directed and habitual control in the basal ganglia: implications for Parkinson's disease. *Nat. Rev. Neurosci. 11, 760–772*.

Rehder, B., & Hoffman, A. B. (2005). Eyetracking and selective attention in category learning. *Cognitive Psychology, 51, 1-41.*

Rehder, B., & Murphy, G. L. (2003). A knowledge-resonance (KRES) model of category learning. *Psychological Bulletin & Review, 10, 759-784.*

Rao, S. M., Bobholz, J. A., Hammeke, T. A., Rosen, A. C., Woodley, S. J., Cunningham, J. M., Cox, R. W., Elliot, A., & Binder, J. R. (1997). Functional MRI evidence for subcortical participation in conceptual reasoning skills. *NeuroReport, 8(8), 1987-1993.*

Rescorla, R. A., & Wagner, A. R. (1972). A theory of Pavlovian conditioning: Variations in the effectiveness of reinforcement and nonreinforcement. In A. H. Black & W. F. Prokasy (Eds.), *Classical Conditioning II: Current Research and Theory* (pp. 64-99). New York: Appleton Century Crofts.

Richler, J. J., & Palmeri, T. J. (2013). Visual category learning. *WIREs Cognitive Science, 5(1), 75-94.*

Ritchie, J. B., Kaplan, D. M., & Klein, C. (2019). Decoding the brain: Neural representation and limits of multivariate pattern analysis in cognitive neuroscience. *Brit. J. Phil. Sci., 70, 581-607.*

Roads, B. D., Xu, B., Robinson, J. K., & Tanaka, J. W. (2017). The easy-to-hard training advantage with real-world medical images. *Cognitive Research: Principles and Implications, 3(38).*

Robinson, D. A. (1975). Occulomotor control signals. In *Basic Mechanisms of Ocular Motility and Their Clinical Implications,* ed. P Bachyrita, G Lennerstrand, pp. 337-374. Oxford, UK, Pergamon.

Roeder, J. L., Maddox, W. T., & Filoteo, J. V. (2017). The neuropsychology of perceptual category learning. In H. Cohen & C. Lefebvre (Eds.), *Handbook of categorization in cognitive science* (pp. 189-225). Elsevier.

Roelfsema, P. R, Tolboom, M., & Khayat, P. S. (2007). Different processing phases for features, figures, and selective attention in the primary visual cortex. *Neuron, 56(5),* 785-792.

Rosch, E. H. (1973). Natural categories. *Cognitive Psychology, 4, 328-350.*

Rosch, E. H., & Mervis, C. B. (1975). Family resemblance: Studies in the integral structure of categories. *Cognitive Psychology, 7, 573-605.*

Sanders, B. (1971). Factors affecting reversal and nonreversal shifts in rats and children. *Journal of Comparative Physiological Psychology, 74(2), 192-202.*

Satterthwaite. T. D., Elliot, M.A., Gerraty, R.T., Ruparel, K., Loughead, J., Calkins, M.E., Eickhoff, S.B., Hakonarson, H., Gur, R.C., & Wolf, D.H. (2013). An improved framework for confound regression and filtering for control of motion artifact in the preprocessing of resting-state functional connectivity data. *NeuroImage, 64, 240-256.*

Schank, R. C., Collins, G. C., & Hunter, L. E. (1986). Transcending inductive category formation in learning. *Behavioral & Brain Sciences, 9, 639-686.*

Schyns, P. G., Goldstone, R. L., & Thibaut, J.-P. (1998). The development of features in object concepts. *Behavioral and Brain Sciences, 21, 1-54.*

Seger, C. A. (2008). How do the basal ganglia contribute to categorization? Their roles in generalization, response selection, and learning via feedback. *Neurosci. Biobehav. Rev., 32, 265-278.*

Seger, C. A., Braunlich, K., & Liu, Z. (2021). Protogamble. Manuscript in preparation.

Seger, C. A., Braunlich, K., Wehe, H. S., & Liu, Z. (2015). Generalization in category learning: The roles of representational and decisional uncertainty. *The Journal of Neuroscience, 35(23), 8802-8812.*

Seger, C. A., & Cincotta, C. M. (2005). The roles of the caudate nucleus in human classification learning. *J. Neurosci., 25, 2941-2951.*

Seger, C. A., & Cincotta, C. M. (2006). Dynamics of frontal, striatal, and hippocampal systems during rule learning. *Cerebral Cortex, 16(11), 1546-1555.*

Seger, C. A., Dennison, C. S., Lopez-Paniagua, D., Peterson, E. J., & Roark, A. A. (2011). Dissociating hippocampal and basal ganglia contributions to category learning using stimulus novelty and subjective judgments. *NeuroImage, 55, 1739-1753.*

Seger, C. A., & Miller, E. K. (2010). Category learning in the brain. *Annu. Rev. Neurosci., 33, 203-219.*

Seger, C. A., Peterson, E. J., Cincotta, C. M., Lopez-Paniagua, D., & Anderson, C. W. (2010). Dissociating the contributions of independent corticostriatal systems to visual categorization learning through the use of reinforcement learning modeling and Granger causality modeling. *NeuroImage, 50, 644-656.*

Seger, C. A., Poldrack, R. A., Prabhakaran, V., Zhao, M., Glover, G. H., & Gabrieli, J. D. E. (2000). Hemispheric asymmetries and individual differences in visual concept learning as measured by functional MRI. *Neuropsychologia, 38(9), 1316-1324.*

Sestieri, C., Tosoni, A., Mignogna, V., McAvoy, M. P., Shulman, G. L., Corbetta, M., & Romani, G. L. (2014). Memory accumulation mechanisms in human cortex are independent of motor intentions. *J Neurosci, 34, 6993-7006.*

Seymour, K., Clifford, C. W. G., Logothetis, N. K., & Bartels, A. (2010). Coding and binding of color and form in visual cortex. *Cerebral Cortex, 20, 1946-1954.*

Shadmehr, R., & Krakauer, J. W. (2008). A computational neuroanatomy for motor control. *Experimental Brain Research, 185, 359-381.*

Shadmehr, R., Smith, M. A., & Krrakauer, J. W. (2010). Error correction, sensory prediction, and adaptation in motor control. *Annual Review of Neuroscience, 33, 89-108.*

Shawn, W. E., Smith, D. B., Peralta, G., & Helie, S. (2017). The impact of category structure and training methodology on learning and generalizing within category representations. *Attention, Perception, & Psychophysics, 79, 1777-1794.*

Shepard, R. N. (1964). Attention and the metric structure of the stimulus space. *Journal of Mathematical Psychology, 1(1), 54-87.*

Shepard, R. N. (1980). Multidimensional scaling, tree-fitting, and clustering. *Science, 210, 390-397.*

Shepard, R. N., & Chipman, S. (1970). Second-order isomorphism of internal representations: Shape of states. *Cogn. Psychology, 1, 1-17.*

Shepard, R. N., Hovland, C. I., Jenkins, H. M. (1961). Learning and memorization of classifications. *Psychological Monographs, 75(517).*

Smith, E. E. (2008). The case for implicit category learning. *Cognitive, Affective, & Behavioral Neuroscience, 8(1), 3-16.*

Smith, E. E., & Grossman, M. (2008). Multiple systems of category learning. *Neuroscience & Biobehavioral Reviews, 32(2), 249-264.*

Smith, E. E., & Medin, D. L. (1981). *Categories and concepts.* Harvard University Press.

Smith, S. M., & Nichols, T. E. (2009). Threshold-free cluster enhancement: Addressing problems of smoothing, threshold dependence and localisation in cluster inference. *NeuroImage, 44(1), 83-98.*

Smith, J. D., Zakrzewski, A. C., Herberger, E. R., Boomer, J., Roeder, J. L., Ashby, F. G., & Church, B. A. (2015). The time course of explicit and implicit categorization. *Attention, Perception, & Psychophysics, 77, 2476-2490.*

Sobel, D.M., Tenenbaum, J.B., & Gopnik, A. (2004). Children's causal inferences from indirect evidence: Backwards blocking and Bayesian reasoning in preschoolers. *Cognitive Science, 22, 525-538.*

Soto, F. A., & Ashby, F. G. (2015). Categorization training increases the perceptual separability of novel dimensions. *Cognition, 139, 105-129*.

Spence, K. W. (1956). *Behavior theory and conditioning.* New Haven, CT: Yale University Press.

Spiering, B. J., & Ashby, F. G. (2008). Response process in information-integration category learning. *Neurobiology of Learning and Memory, 90(2), 330-338.*

Stocco, A., Lebiere, C., and Anderson, J. R. (2010). Conditional routing of information to the cortex: a model of the basal ganglia's role in cognitive coordination. *Psychol. Rev. 117, 541–574*.

Sun, M., Hu, L., Xin, X., & Zhang, X. (2021). Neural hierarchy of color categorization: From prototype encoding to boundary encoding. *Front. Neurosci., 15.*

Sutherland, N. S. (1964). Visual discrimination in animals. *British Medical Bulletin, 20, 54-59.*

Suzuki, Y., Iseki, C., Igari, R., Sato, H., Koyama, S., Kawahara, H., Itagaki, H., Sonoda, Y., & Ohta, Y. (2022). Reduced cerebral blood flow of lingual gyrus associated with both cognitive

impairment and gait disturbance in patients with idiopathic normal pressure hydrocephalus. *Journal of the Neurological Sciences, 437, 120266.*

Szucs, D., & Ioannidis, J. P. A. (2017). Empirical assessment of published effect sizes and power in the recent cognitive neuroscience and psychology literature. *PLOS Biology, 19(3), e3001151.*

Szucs, D., & Ioannidis, J. P. A. (2020). Sample size evolution in neuroimaging research: An evaluation of highly-cited studies (1990–2012) and of latest practices (2017–2018) in high-impact journals. *NeuroImage, 221, 117164.*

Tanaka, K. (1996). Inferotemporal cortex and object vision. *Annu. Rev. Neurosci., 19, 109-139.*

Thomas, S., Kapoor, A., & Srinivasan, N. (2021). Supervised category learning: When do participants use a partially diagnostic feature. *Proceedings of the Annual Meeting of the Cognitive Science Society, 43(43), 1313-1319.*

Tong, F., & Pratte, M. S. (2012). Decoding patterns of human brain activity. *Annu. Rev. Psychol., 63, 483-509.*

Toni, I., Ramnani, N., Josephs, O., Ashburner, J., & Passingham, R. E. (2001). Learning arbitrary visuomotor associations: Temporal dynamic of brain activity. *NeuroImage, 14(5), 1048-1057.n*

Tustison, N.J., Avants, B.B., Cook, P.A., Zheng, Y., Egan, A., Yushkevich, P.A., & Gee, J.C. (2010). N4ITK: Improved N3 bias correction. *IEEE Transactions on Medical Imaging, 29(6), 1310-1320.*

Tversky, A. (1977). Features of similarity. *Psychological Review, 84(4), 327-352.*

Ullman, S., Vidal-Naquet, M., & Sali, E. (2002). Visual features of intermediate complexity and their use in classification. *Nature Neuroscience, 5, 682-687.*

Urcelay, G. P. (2017). Competition and facilitation in compound conditioning. *Journal of Experimental Psychology: Animal Learning and Cognition, 43(4), 303–314.*

Urcelay, G. P., & Miller R. R. (2009). Potentiation and overshadowing in Pavlovian fear conditioning. *J Exp Psychol Anim Behav Process, 35(3), 340-356.*

van Es, D. M., Zwaag, W., Knapen, T. (2019). Topographic maps of visual space in the human cerebellum. *Current Biology, 29(10), 1689-1694.*

Vapnik, V. N. (1998). *Statistical learning theory.* John Wiley & Sons.

Villagrasa, F., Baladron, J., Vitay, J., Schroll, H., Antzoulatos, E. G., Miller, E. K., & Hamker, F. H. (2018). On the role of cortex-basal ganglia interactions for category learning: A neurocomputational approach. *J Neurosci, 38(44), 9551-9562.*

Virtanen, P., Gommers, R., Oliphant, T. E., Haberland, M., Reddy, T., Cournapeau, D., Burovski, E., Peterson, P., Weckesser, W., Bright, J., van der Walt, S. J., Brett, M., Wilson, J., Millman, J., Mayorov, N., Nelson, A. R. J., Jones, E., Kern, R., Larson, E., … van Mulbregt, P. (2020). SciPy 1.0: Fundamental algorithms for scientific computing in python. *Nature Methods, 17(3), 261-272.*

Visser, R. M., De Haan, M. I. C., Beemsterboer, T., Haver, P., Kindt, M., & Scholte, H. S. (2016). Quantifying learning-dependent changes in the brain: Singl-trail multivoxel pattern analysis requires slow event-related fMRI. *Psychophysiology, 53, 1117-1127.*

Vong, W. K., Hendrickson, A. T., Navarro, D. J., & Perfors, A. (2019). Do additional features help or hurt category learning? The curse of dimensionality in human learners. *Cognitive Science, 43, e12724.*

Wagner, A. R., Logan, F. A., & Haberlandt, K. (1968). Stimulus selection in animal discrimination learning. *Journal of Experimental Psychology, 76(2, Pt.1), 171–180.*

Waldmann, M.R., & Holyoak, K.J. (1992). Predictive and diagnostic learning within causal models: Asymmetries in cue competition. *Journal of Experimental Psychology: General, 121, 222-236.*

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

Wallis, J. D., & Miller, E. K. (2003). From rule to response: Neuronal processes in the premotor and prefrontal cortex. *Journal of Neuorphysiology, 90(3), 1790-1806.*

Wang, J. X., & Voss, J. L. (2014). Brain networks for exploration decisions utilizing distinct modeled information types during contextual learning. *Neuron, 82, 1171-1182.*

Wardle, S. G., Ritchie, J. B., Seymour, K., & Carlson, T. A. (2017). Edge-related activity is not necessary to explain orientation decoding in human visual cortex. *The Journal of Neuroscience, 37(5), 1187-1196.*

Wattenmaker, W. D. (1991). Learning modes, feature correlations, and memory-based categorization. *Journal of Experimental Psychology: Learning, Memory, & Cognition, 17, 908-923.*

Weaverdyck, M. E., Lieberman, M. D., & Parkinson, C. (2020). Multivoxel pattern analysis in fMRI: A practical introduction for social and affective neuroscientists. *Social Cognitive and Affective Neuroscience, 15(4), 487-509.*

Williams, D. A., Sagness, K. E., & McPhee, J. E. (1994). Configural and elemental strategies in predictive learning. *Journal of Experimental Psychology: Learning, Memory, and Cognition, 20, 694-709.*

Williams, Z. M., & Eskander, E. N. (2006). Selective enhancement of associative learning by microstimulation of the anterior caudate. *Nat. Neurosci., 9, 1790-1806.*

Winkler, A. M., Ridgway, G. R., Webster, M. A., Smith, S. M., & Nichols, T. E. (2014). Permutation inference for the general linear model. *NeuroImage, 92, 381-397.*

Wixted, J. T., Squire, L. R., Jang, Y., Papesh, M. H., Goldinger, S. D., Kuhn, J. R., Smith, K. A., Treiman, D. M., & Steinmetz, P. N. (2014). Sparse and distributed coding of episodic memory in neurons of the human hippocampus. *PNAS, 111(26), 9621-9626.*

Woo, C.-W., Krishnan, A., & Wager, T. D. (2014). Cluster-extent based thresholding in fMRI analysis: Pitfalls and recommendations. *NeuroImage, 1(91), 412-419.*

Woolrich, M. W., Ripley, B. D., Brady, M., & Smith, S. M. (2001). Temporal autocorrelation in univariate linear modeling of fMRI data. *NeuroImage, 14, 1370-1386.*

Xiao, Y., Casti, A., Xiao, J., & Kaplan, E. (2007). Hue maps in primate striate cortex. *NeuroImage, 35(2), 771-786.*

Xing, Q., & Sun, H. (2017). Differential impact of visuospatial working memory on rule-based and information-integration category learning. *Front. Psychol., 8(530), 1-13.*

Yamauchi, T., Love, B. C., & Markman, A. B. (2002). Learning nonlinearly separable categories by inference and classification. *Journal of Experimental Psychology: Learning, Memory, and Cognition, 28, 585-593.*

Yamauchi, T., & Markman, A. B. (1998). Category learning by inference and categorization. *Journal of Memory and Language, 39, 124-148.*

Yamauchi, T., & Markman, A. B. (2000). Inference using categories. *Journal of Experimental Psychology: Learning, Memory, and Cognition, 26, 776-795.*

Yarkoni, T. (2023). *Neurosynth*. https://www.neurosynth.org.

Yi, H.-G., Maddox, W. T., Mumford, J. A., & Chandrasekaren, B. (2016). The role of corticostriatal systems in speech category learning. *Cerebral Cortex, 26(4), 1409-1420.*

Yizhar, O., Tal, Z., & Amedi, A. (2023). Loss of action-related function and connectivity in the blind extrastriate body area. *Front Neurosci., 17, 973525.*

Young, M. P., & Yamane, S. (1992). Sparse population coding of faces in the inferotemporal cortex. *Science, 256(5061), 1327-1331.*

Yu, T., Fan, F., Cao, H., Luo, X., Tan, S., Yang, F., Zhang, X., Shugart, Y. Y., Hong, L. E., Li, C.-S. R., & Tan, Y. (2018). Decreased gray matter volume of cuneus and lingual gyrus in schizophrenia patients with tardike dyskinesia is asociated with abnormal involuntary movement. *Scientific Reports, 8(12884).*

Zarahn, E., Aguirre, G., & D'Esposito, M. (1997). A trial-based experimental design for fMRI. *NeuroImage, 6, 122-138.*

Zeithamova, D., Maddox, W. T., & Schnyer, D. M. (2008). Dissociable prototype learning systems: Evidence from brain imaging and behavior. *The Journal of Neuroscience, 28(49), 13194-13201.*

Zeithamova, D., Maria-Alejandra, A. S., & Adke, A. (2017). Trial timing and pattern-information analyses of fMRI data. *NeuroImage, 153, 221-231.*

Zeki, S. M. (1978). Functional specialization in the visual cortex of the monkey. *Nature, 274, 423-428.*

Zhang, Y., Brady, M., & Smith, S. (2001). Segmentation of brain MR images through a hidden M arkov Random Field Model and the Expectation-Maximization Algorithm. *IEEE Transactions on Medical Imaging, 20(1), 45-57.*