Exploring the Role of Language in Two Systems for Categorization

Kayleigh Ryherd, PhD
University of Connecticut, 2019

Lorem ipsum dolor sit amet, consectetuer adipiscing elit. Etiam lobortis facilisis sem. Nullam nec mi et neque pharetra sollicitudin. Praesent imperdiet mi nec ante. Donec ullamcorper, felis non sodales commodo, lectus velit ultrices augue, a dignissim nibh lectus placerat pede. Vivamus nunc nunc, molestie ut, ultricies vel, semper in, velit. Ut porttitor. Praesent in sapien. Lorem ipsum dolor sit amet, consectetuer adipiscing elit. Duis fringilla tristique neque. Sed interdum libero ut metus. Pellentesque placerat. Nam rutrum augue a leo. Morbi sed elit sit amet ante lobortis sollicitudin. Praesent blandit blandit mauris. Praesent lectus tellus, aliquet aliquam, luctus a, egestas a, turpis. Mauris lacinia lorem sit amet ipsum. Nunc quis urna dictum turpis accumsan semper.

Exploring the Role of Language in Two Systems for Categorization

Kayleigh Ryherd

B.A., The George Washington University, 2014
M.S., University of Connecticut, 2016

A Dissertation
Submitted in Partial Fulfillment of the
Requirements for the Degree of
Doctor of Philosophy
at the
University of Connecticut
2019

APPROVAL PAGE

Doctor of Philosophy Dissertation

Exploring the Role of Language in Two Systems for Categorization
Kayleigh Ryherd, M.S.

Major Advisor: ______________________________
Nicole Landi

Associate Advisor: ______________________________
Letitia Naigles

Associate Advisor: ______________________________
James Magnuson

Associate Advisor: ______________________________
Eiling Yee

Associate Advisor: ______________________________
Clinton Johns

2019

Acknowledgments

Lorem ipsum dolor sit amet, consectetuer adipiscing elit. Etiam lobortis facilisis sem. Nullam nec mi et neque pharetra sollicitudin. Praesent imperdiet mi nec ante. Donec ullamcorper, felis non sodales commodo, lectus velit ultrices augue, a dignissim nibh lectus placerat pede. Vivamus nunc nunc, molestie ut, ultricies vel, semper in, velit. Ut porttitor. Praesent in sapien. Lorem ipsum dolor sit amet, consectetuer adipiscing elit. Duis fringilla tristique neque. Sed interdum libero ut metus. Pellentesque placerat. Nam rutrum augue a leo. Morbi sed elit sit amet ante lobortis sollicitudin. Praesent blandit blandit mauris. Praesent lectus tellus, aliquet aliquam, luctus a, egestas a, turpis. Mauris lacinia lorem sit amet ipsum. Nunc quis urna dictum turpis accumsan semper.

Contents

List of Figures

List of Tables

1 General Introduction

Categories help us organize the world. They help us predict and hypothesize about category members, helping us quickly select the most appropriate response for each situation. Language plays a key role in categorization and category learning. It provides structure in the form of category labels and affects how we think about and even perceive the categories themselves. As Lupyan (2012) puts it, language augments our thought. Thus, any thorough investigation of how we learn categories must consider the role of language.

Indeed, many theoretical frameworks of category learning consider the role of language. For example, COVIS, a key theory in perceptual category learning, stands for ”competition between verbal and implicit systems” (Ashby et al., 1998). Similarly, a theory put forth by Minda and colleagues is called ”A theory of verbal and nonverbal category learning” (Minda & Miles, 2010). However, most theories of category learning that consider language primarily focus on whether or not language has influences category learning, rather than exploring what kind of influences language might have .

Thus, the current work seeks to both define a theory of category learning and explore the role language has in this theory. In this review I will synthesize multiple approaches to category learning. Following the synthesis, I will review relevant literature that provides suggestions as to how language might be involved in category learning. Through these efforts, I will provide a theoretical framework and hypotheses for this dissertation.

1.1 Dual-systems model for category learning

Multiple theories converge on the idea that there are two systems for category learning. In this section, I will first describe a generalized dual-systems model that pulls threads from all of these theories and then go on to describe how each theory fits into the overarching framework.

1.1.1 Proposed model

The proposed model involves two systems for category learning. The first, which I title the **associative system**, uses associative mechanisms in an iterative manner to learn distributions of features. This system is best suited for learning multidimensional *similarity-based* categories such as natural kinds, where it is difficult to describe necessary and sufficient rules for inclusion. Similarity-based categories have features that are correlated and probabilistic, such that a given category instance may not have all of the category-relevant features but does tend to have some distribution of them. For example, although Manx cats do not have tails, a typical feature of cats, they are still undeniably members of the category *cat*. Thus, the associative system must be able to extract the most frequent pattern of features over many instances in order to learn a category.

In contrast, the **hypothesis-testing system** uses a more explicit learning method to test and adjust hypotheses about category boundaries. This method relies on selection of relevant features rather than representation of a distribution of feature probabilities. As such, it is most suited for learning rule-based categories, which typically have one or a few easily verbalizable rules for inclusion. For example, the *ad hoc* category *things to be sold at the garage sale* has a simple rule for inclusion that perfectly separates members from non-members.

Thus, we have two systems for category learning, each one ideal for learning a different type of category (similarity-based vs. rule-based). In the upcoming sections, I will describe theoretical and empirical evidence for a dual-systems model from five different approaches to category learning. I will show how each approach informs the current theoretical framework.

1.1.2 COVIS

COVIS stands for COmpetition between Verbal and Implicit Systems. First proposed by Ashby et al. in 1998, it is a prominent theoretical framework for perceptual category learning. This framework provides a dual-systems model that is grounded in neuropsychological data, allowing it to suggest neurobiological underpinnings for the two systems. It is important to note that this framework is mostly concerned with perceptual categories, which are defined as ”a collection of similar objects belonging to the same group”

(Ashby & Maddox, 2005, p. 151). This is in contrast to concepts, which Ashby and colleagues define as groups of related ideas. Thus, this approach focuses on categorizing objects that can be encountered and perceived in the real world.

As can be inferred from the title, the two category learning systems in COVIS are the **verbal** and **implicit** systems. The verbal system is COVIS' answer to our hypothesis-testing system. It is a declarative learning system that uses a hypothesis-testing method to learn category rules, typically for rule-based stimuli. Under COVIS, rule-based stimuli must have inclusion rules that are easy to describe verbally. Typically, rule-based stimuli used by Ashby and colleagues have a single rule for inclusion or two rules combined by "and" or "or." When a rule-based category involves multiple dimensions, decisions about each dimension are made separately, and these decisions are used to evaluate the logical operators. In other words, each dimension is considered on its own before their combination. These guidelines for rule-based categories ensure that an explicit hypothesis-testing method can be used to learn them. When learning a new category, the verbal system holds potential category inclusion rules in working memory that are then tested as stimuli are encountered. Over time, hypotheses are tested and switched until they reflect the optimal strategy for categorization. Individual differences in rule-based category learning have been shown to be related to an individual's cognitive flexibility (Reetzke et al., 2016), suggesting that the verbal system relies at least partially on executive function.

The implicit system from COVIS is most similar to our associative system. Like the associative system, it uses incremental learning to find category boundaries. It is most ideal for learning information-integration categories, which are like similarity-based categories but have also some specific guidelines. Information-integration categories are defined by some combination of dimensions. However, while each dimension can be considered separately in rule-based categories, all dimensions must be considered simultaneously for information-integration categories. Information-integration category membership depends on both the values associated with each dimension as well as the relationship between these values. Information-integration category boundaries are difficult or impossible to describe verbally. The structure of information-integration categories require an iterative, associative learning method. COVIS suggests that the implicit system relies on an information stream that connects stimuli, motor responses, and feedback to learn category membership.

One of the most substantial contributions of COVIS is its strong grounding in neurobiology. In the original paper, Ashby and colleagues proposed specific brain regions involved the verbal (hypothesis-testing) and implicit (associative) systems, supported by neuroimaging and patient studies. The verbal (hypothesis-testing) system relies on the prefrontal cortex (PFC), anterior cingulate cortex (ACC), striatum, hippocampus, and the head of the caudate nucleus. Information about the stimuli are processed in fronto-striatial loops, where potential category rules are generated. The PFC keeps these rules in working memory while the ACC and the head of the caudate nucleus mediate switching between rules based on feedback. Finally, the hippocampus stores longer-term memory of which rules have already been tested. The hippocampus is only involved when the task is complex enough that previously tested rules cannot all be stored in working memory (Ashby & Maddox, 2005, 2011).

Patient data shows that individuals with frontal damage as well as individuals with Parkison's disease, which affects the basal ganglia including the caudate nucleus, show difficulty in rule-based tasks such as the Wisconsin Cart Sorting Test (Robinson et al., 1980) and an experimental rule-based category learning task (Ashby et al., 2003). This suggests that both frontal regions and the basal ganglia are involved in rule-based categorization. More recent neuroimaging work, however, is still mixed as to the involvement of different areas specifically for rule-based categorization. Soto et al. (2013) found that two separate rule-based tasks could be differentiated based on activation in ventro-lateral PFC, suggesting that specific rules are stored in that region. Nomura et al. (2007) found activation in the medial temporal lobe (MTL), which contains the hippocampus, specifically for rule-based categorization. However, a later study failed to find any activation that was specifically greater for rule-based categorization (Carpenter et al., 2016). Thus, the neural underpinnings of the verbal (hypothesis-testing) system are still under debate.

The implicit (associative) system from COVIS has a different neurobiological pathway for category learning. It uses incremental learning rather than hypothesis testing to learn information-integration (similarity-based) categories. The main structure involved in this procedural learning system is the striatum, which is involved in reinforcement learning with dopamine as the reinforcement signal. From the striatum, information about the category is sent to the thalamus and the globus pallidus, which is within the basal ganglia.

Information then runs to motor and premotor cortex. This system links stimuli, motor responses during categorization, and feedback to allow the participant to learn categories. Neuroimaging studies using the implicit system again are mixed, with some finding activation in the caudate body while others fail to find that activation, instead seeing activity in parahippocampal regions (Carpenter et al., 2016; Nomura et al., 2007). A separate study also found a role for the putamen in similarity-based category learning (Waldschmidt & Ashby, 2011). As with the verbal system, the neural basis of the implicit system requires more study.

COVIS provides us with a few key insights. First, it is one of the most studied dual-systems theories of categorization. While Ashby and colleagues generally use visual stimuli for their tasks, this paradigm has been extended to other perceptual domains such as hearing/speech (Chandrasekaran et al., 2014, 2016). As such, research on the current theoretical framework (associative/hypothesis-testing systems) has much COVIS literature which we can compare it to. It also makes clear claims about the neurobiological basis of the two systems of category learning. While the specifics of these claims are still under debate in the literature, they at least provide regions of interest for researchers who want to conduct neuroimaging research on a dual-systems model of category learning. Finally, this approach is one of the only ones to consider how the two systems interact.

1.1.3 Dimensionality

The dimensionality approach, led by Lupyan and colleagues, considers categories in terms of the dimensions on which they cohere. Low-dimensional categories are the same on one or a small number of dimensions (e.g., color) and allow other dimensions to vary. Low-dimensional categories are similar to rule-based categories, as they can be described using relatively simple rules (e.g., *things that are red*). In fact, some of Lupyan's papers define low-dimensional categories as those that have a single dimension that can distinguish category members from non-members (Lupyan & Mirman, 2013). Examples of low-dimensional categories from this study include *things made of wood* and *things with handles*.

In contrast, high-dimensional categories are those that cohere on multiple dimensions, often so many that category rules are difficult to describe. Examples of high-dimensional categories from the previously-mentioned study include *birds*, *tools*, *things that fly*, and *objects that hold water*. Most natural kinds and artifacts are high-dimensional, as well as some *ad hoc* categories. Like similarity-based categories, high-dimensional categories require their members to be the same on most (but not all) relevant dimensions.

The core prediction tested using this approach is that low-dimensional categorization relies more heavily on language than high-dimensional categorization. The dimensionality approach postulates that language helps an individual select features, which is a process only helpful for low-dimensional categorization. High-dimensional categorization relies on creating associations across multiple features, which does not require language.

To explore this prediction, Lupyan and colleagues interfered with language ability in multiple ways across studies. In each study, they found that a reduction in language ability was associated with poorer performance on low- but not high-dimensional categorization. Lupyan & Mirman (2013) measured categorization ability in individuals with aphasia for both low- and high-dimensional categories. They found that the individuals with aphasia performed similarly to unimpaired controls on the high-dimensional categories, but showed significantly lower accuracy on the low-dimensional categories. Lupyan (2009) used a concurrent verbal load to reduce the verbal resources available during a categorization task. He found that individuals showed significantly poorer categorization with a verbal load as compared to a visuospatial load specifically for category judgments based on a single dimension (color or size) but not for those based on multiple dimensions (theme). Other studies manipulated language ability by using transcranial direct current stimulation (tDCS). One study found that reducing excitability in a language-critical region (left inferior frontal gyrus) led to poorer performance on low-dimensional but not high-dimensional categorization (Lupyan et al., 2012). Another study used stimuli that could either be categorized using a uni-dimenional or a bi-dimensional strategy. Reducing excitability over Wernicke's area made participants more likely to chose the bi-dimensional strategy, indicating that interfering with language functioning resulted in participants using higher-dimensional categorization (Perry & Lupyan, 2014).

The dimensionality approach to category learning and the studies done to test it provide multi-method evidence for the role of language in low-dimensional categorization. Unlike COVIS, where the verbal system largely uses language to describe and rehearse candidate category rules, the dimensionality approach suggests that language is used to select relevant features for a category. This idea has highly influenced this paper's dual-systems model, in which the hypothesis-testing system does select category-relevant features. However, the evidence for this approach is largely unable to speak for the system underlying high-dimensional categorization, as most of the effects for this system are null. Thus, it is not clear from this approach whether the hypothesized broad inter-item association building is in fact how individuals learn high-dimensional categories. In addition, while the authors claim that poorer low-dimensional categorization performance reflects difficulty in selecting category-relevant dimensions, the studies mentioned above do not directly test how interfering with language ability affects selection or inhibition ability.

1.1.4 Statistical Density

The statistical density framework focuses on the structure of categories defined by the relationships among members and non-members. Pioneered by Sloutsky, it proposes two category learning systems that are each used to extract different types of regularities from a stream of information, allowing for flexibility in the data collected (Sloutsky, 2010). Sloutsky's main metric for describing categories is called *statistical density*. In this section, I will describe statistical density in a broad sense; for more detailed information on how to calculate it, see Appendix A (p. 24).

The statistical density of a category is related to the ratio between the amount of entropy within a target category and the entropy between the target category and other categories in the set. In this context, entropy refers to variation within features. Consider a set of shapes. These shapes can vary in shape, size, and color. The within-category entropy for squares is all of the different sizes and colors for a single shape: squares. The between-category entropy includes all of the variation in size, color, and shape for the items in the set. **Sparse** categories have lots of within-category entropy; the items in the category cohere on only one or a few dimensions. All other dimensions are allowed to vary freely. In our shape example, a sparse square category would have squares of all color and sizes, such that color and size was not related to shape. Thus, to find the category *square*, an individual would have to isolate the "shape" feature.

In contrast, **dense** categories have little within-category entropy; their members have multiple intercorrelated features that together are predictive of category membership. There are few irrelevant features in dense categories. Within our set of shapes, the square category would be considered dense if all squares shared the same color and size. The distribution of these other features are what determine the statistical density of a category. If irrelevant features (here, color and size) are correlated with the relevant feature(s), the category is dense. If they vary independently of the relevant features, the category is sparse. Thus, statistical density expresses the relationships between features within a category as well as within an entire set of items. A particularly interesting feature of this metric is that statistical density is a continuous spectrum: categories can be very dense, very sparse, or anywhere in between.

This framework also outlines two systems used to learn categories with different densities. Dense categories are best learned by the compression-based system, which takes input and reduces it by representing some but not all features. With more instances, relevant features for a given category will be represented more frequently and survive the compression. In contrast, features that appear infrequently will be mostly filtered out. The compression-based system does not use conscious selection to determine which features are represented. Instead, redundant and probable features are more likely to continue on. The many correlated features of a dense category are easily extracted using this system, which is quite similar to our associative system.

The second learning system is called the selection-based system. This system directs attention towards relevant features, sampling those features for later representation. This system learns by aiming to reduce error. As feedback is encountered, the system shifts attention from those dimensions that create categorization errors to those that do not. The selection-based system relies heavily on multiple aspects of executive function, including inhibition and selection. It is best for learning sparse categories. While over time the compression-based system could be able to learn sparse categories, as the freely varying irrelevant features would eventually be less frequent than relevant features, this process would be much more inefficient than selecting and testing individual features. The selection-based system is Sloutsky's version of our hypothesis-testing system. Some research shows that sparse categorization is correlated with performance on a flanker task, which is often used to measure selection and inhibition (Perry & Lupyan, 2016). This suggests that at least some executive functions are related to sparse category learning.

The statistical density framework also discusses the development of these two systems. It suggests that children have access to the compression-based (associative) system early in development, as its mechanisms involve brain structures that develop relatively early, such as inferiotemporal cortex (Rodman, 1994). In contrast, the selection-based (hypothesis-testing) system involves more frontal regions that develop later, such as dorsolateral prefrontal cortex and anterior cingulate cortex (Eshel et al., 2007; Lewis, 1997; Segalowitz & Davies, 2004). Thus, this framework posits that the compression-based system develops before the selection-based system. Sloutsky and others have done some studies on different age groups testing the two systems with categories of different densities to verify this claim.

Kloos & Sloutsky (2008) tested both of these systems in children and adults. They engaged the two systems separately by modifying task demands. Some participants learned novel categories by being taught the rules for inclusion (e.g., "Ziblets have a short tail."). This activated the selection-based (hypothesis-testing) system. Other participants learned these categories by viewing multiple members, engaging the compression-based (associative) system. Thus, the authors could test how well individuals could learn novel categories of different densities depending on whether the category density matched the system being engaged. For both children and adults, learning performance was high when the category density and task instructions matched. However, while the adults were able to adapt and learn the categories in mismatch conditions, children were specifically unable to learn sparse categories just by viewing multiple instances. This suggests that children are not able to use the selection-based system without direct guidance from task instructions.

Other evidence for the developmental course of the two systems comes from a study of infants and adults. This study used a switching paradigm to investigate whether individuals were selecting specific features (using the selection-based/hypothesis-testing system) or processing the entire stimulus holistically (using the compression-based/associative system). They found that when viewing sparse categories, adults showed a significant switch cost as the category-relevant feature was changed, while infants did not. This indicates that adults viewing sparse categories were focusing on a specific feature, while infants were processing the entire stimulus. Eye movement data also suggested that even with holistic processing, infants were able to learn the categories (Best et al., 2013). Thus, this study found that adults and infants used different systems for learning the same categories.

A similar finding comes from a study of change detection. One study found that while both children and adults showed high accuracy when detecting change in a cued stimulus, children were better than adults at detecting change in task-irrelevant stimuli. Similar results were found when children and adults were asked to perform familiarity judgments on items seen during a visual search task: high performance for both groups when probing relevant features, and higher performance for children than adults when probing irrelevant features (Plebanek & Sloutsky, 2017). The results from both of these experiments suggest that children attend to a stimulus in a diffuse manner, even when task demands suggest a selective strategy. This is consistent with a later-developing selection-based system, as children may be processing these features using the compression-based system, which preserves even category-irrelevant features.

Thus, the statistical density approach to category learning provides two major points for consideration. First, the statistical density metric itself emphasizes the idea that there aren't two distinct types of categories (e.g., rule-based and similarity-based). Instead, categories exist on a spectrum ranging between these extremes. It is still unknown how a dual-system model would deal with stimuli that lie directly in the middle of this spectrum, however. Second, this framework is one of only a few that describes a developmental trajectory for a dual-systems framework of category learning.

1.1.5 Verbal/nonverbal

Like some of approaches discussed above, the verbal/nonverbal approach is a dual-systems model of category learning. While other approaches discuss the role of language in category learning, none make it as central as this approach by Minda & Miles (2010). The two systems in this approach are called the **verbal** and **nonverbal** systems. These systems align well both with the framework outlined in this paper as well as with other approaches. The verbal system uses hypothesis testing to determine the verbal rules best suited to characterize a category. In contrast, the nonverbal system uses associative mechanisms to learn categories, iteratively learning which features go together in predicting category membership.

A unique feature of this approach to category learning is its emphasis on traditional models of working memory and their role in the category learning process. Minda & Miles (2010) state that the verbal system relies heavily on working memory, especially the phonological loop and central executive, to rehearse and select potential rules (A. D. Baddeley & Hitch, 1974). The nonverbal system, meanwhile, uses the visuospatial sketchpad to store and rehearse visual information, but overall uses working memory to a lesser extent than the verbal system. Evidence for these hypotheses comes from a study showing that children exhibited adult-like performance when learning categories using the nonverbal system and reduced performance for categories that required use of the verbal system. This study also showed that adults showed more child-like performance when learning categories suited to the verbal system while under concurrent verbal load, suggesting that the verbal system indeed needs verbal working memory resources (Minda et al., 2008).

While the two systems described in Minda & Miles (2010) are quite similar to the systems hypothesized in this paper, there remains a core difference: the nonverbal system does not posit a role for language. This is likely due to the way Minda & Miles (2010) ground their dual-systems model in working memory. As will be discussed shortly, language can be very useful even for iterative, association-based learning, although perhaps not in the form of a verbal working memory resource. Thus, the verbal/nonverbal dual-systems model of category learning provides us with evidence that verbal working memory and executive resources support rule-based category learning but does not fully consider the ways in which language may influence category learning.

1.1.6 Taxonomic/thematic

As these previous frameworks have shown, when considering categories we must think carefully about how the items in a category relate to each other. The taxonomic/thematic framework is yet another way to consider relations within categories. **Taxonomically** related items are those we might think of as belonging to the same everyday category (e.g., animals, plants, tools, etc.). **Thematically** related items are those that go together in everyday life but are not necessarily part of the same category (e.g., needle and thread, apple and worm).

Similar to and perhaps even more so than the statistical density approach, the taxonomic/thematic framework has been able to provide many valuable insights about the developmental trajectory of categorization. The typical task in this line of research is a grouping task, where individuals are given a set of items and asked to group the ones that are ”alike” or ”the same.” Early research on this topic suggests that children primarily categorize items using thematic relations in kindergarten and switching to taxonomic relations later in childhood, although even this early work indicated that young children are able to learn taxonomic relations if necessary (Piaget et al., 1964; Vygotsky, 1962). Smiley & Brown (1979) found that the preference for taxonomic versus thematic relations switches between first and fifth grade as well as between college and old age, such that the very young and the elderly both show a preference for thematic relations. However, in another study, college-aged adult participants chose a thematically-related item more frequently in a triad task across ten different experiments, including one with the same stimuli used in Smiley and Brown's paper (Lin & Murphy, 2001).

Rather than being tied directly to age or ability, the preference for thematic or taxonomic classification may depend on an individual's goals. Markman & Hutchinson (1984) had children between the ages of 2 and 4 complete a triad task. The children were shown a target picture (e.g. a tennis shoe) as well as two options: one that was taxonomically related (e.g., a high-heeled shoe) and one that was thematically related (e.g., a foot). The children were then asked to ”find the one that is the same.” With these directions, the children chose the thematically-related object about half of the time. However, when a novel label was applied to the task (e.g., This is a *dax*. Can you find another *dax*?), the children were more likely to choose the taxonomically related item. Thus, having a category label focused the task and directed attention towards taxonomic category structure rather than thematic relations. Further research in children between the ages of 2 and 4 manipulated many parts of the typical triad task, including experimenter instructions and medium of presentation (pictures vs. physical objects). They found that the thematic preference seen in Smiley & Brown (1979) seemed to be strongly affected by task instructions and age (Waxman & Namy, 1997). Some research suggests that what is developing in young childhood is not a sensitivity to different types of relations but instead the ability to flexibly switch between thematic and taxonomic relations according to task demands (Blaye & Bonthoux, 2001).

Taxonomic and thematic categories and processing share many similarities with the approaches discussed above. Taxonomic categories are like similarity-based categories. Both are what a typical individual would consider to be a ”category;” they include natural kinds and artifacts. In contrast, thematic categories are more similar to rule-based categories. Both can be defined using a rule like ”usually found in a kitchen” or ”used for sewing.” Thinking about rule-based categories in terms of thematic relations brings a new aspect to these categories: situational similarity. Often, rule-based categories are *ad hoc*, or created for and bound to a certain situation (e.g., ”things to be sold at the garage sale”). Thus, when we think about how we learn and process rule-based categories using the hypothesis-testing system, we should keep in mind how we use our knowledge of situations or episodes in categorization.

1.2 Vocabulary/labels and category learning

Much theory and research has considered how having a single word for a category or concept affects how an individual learns and processes that category. In this document, we will consider the word form associated with a given category (either spoken or written) to be the **category label**. Thus, a category has two potential pieces. First, there is the category's meaning, or the way in which members belong to a category. As discussed previously, this can be a set of defined rules (e.g., anything you plan to sell is a part of the category *things to sell at the garage sale*) or an implicit set of fuzzy category boundaries (e.g., the ways in which you judge whether an item is a *chair*). The second piece of the category is its label. Individuals learning new categories often learn both the meaning and the label.

There have been multiple viewpoints on just how labels interact with the category or concept they describe and refer to. One line of thought postulates that labels are attached to concepts that can be formed in their absence (Gillette et al., 1999; Snedeker & Gleitman, 2004). This framework tends to focus on early-acquired object concepts, which are thought to be built nonverbally in the infant before language is acquired. Experiments done under this framework reveal interesting and important findings about the information that best supports a mapping between a category meaning and its label (e.g., having a syntactic frame for a category label leads to much quicker learning than just observing the use of the label in multiple situations). However, this viewpoint places little importance on the interplay between the label and the meaning; at best, the label is an additional way to access the meaning but does not seem to differ from any other feature.

Other researchers suggest that labels dynamically interact with meanings, and that having a single word for a meaning fundamentally changes how individuals think about and even perceive a category. In the words of Waxman & Markow (1995), words (labels) are ”invitations to form categories”. When a child encounters a novel word form applied to an object, they are initially biased to interpret that word form as a label for a category rather than the name of that singular object. Indeed, receiving a label for a category helps 12-month-old infants focus on common features more than just directive speech (Althaus & Mareschal, 2014). In adults, labels promote category learning even when they are redundant, and they do so even more than additional nonverbal features (Lupyan et al., 2007). Even more interestingly, having a label can change perceptual processing across development. Infants shown a certain set of objects without an accompanying label will sort these objects into multiple categories using visual features. However, if a single label is applied to the same set of objects, the infants will create only one category (Plunkett et al., 2008). In adults, hearing category labels affects visual perception. Participants asked to find 2s or 5s in a visual display showed better accuracy and shorter reaction time when hearing “two” or “five” immediately before the display appeared (Lupyan & Spivey, 2010).

The evidence cited above suggests that labels are special in some way—they are not simply additional features of fully-formed concepts. This may be because labels encourage individuals to focus on features that are more diagnostic (i.e., more often associated with members of a category) rather than features that are specific to a given instance. A number of studies from Lupyan and colleagues support this idea. For example, Edmiston & Lupyan (2015) found that adults tended to look at more typical instances of a category when hearing a label. Thus, when hearing the word ”bird,” participants were more likely to look at a robin (a more typical bird) than a penguin (a less typical bird). They also found that when listening to sounds associated with a category (e.g., bird chirp), participants tended to look at more likely sources of the sound (e.g., images of birds with their mouths open). This suggests that labels activate a typical, abstracted representation of a category while other sounds activate a more specific instance of that category that is congruent with the sound itself.

Similar findings come from a study looking at the formal category triangles. Triangles are by definition figures with three sides – any figure with three sides can be labeled a triangle. However, Lupyan (2017) found that typicality effects for triangles in multiple tasks were introduced when the word “triangle” was used. When asked to draw a triangle, participants most often drew isosceles or equilateral triangles with their base parallel to the horizontal (i.e., more canonical triangles). However, when instructed to draw a three-sided figure, participants drew a variety of triangles. The same typicality-related pattern of results was found for multiple other tasks, including typicality judgments, speeded recognition, and shape judgment. Another study found that pairing category instances with labels increased fixations on category-relevant features, as compared to pairing them with random words or silence, even for sparse categories (Barnhart et al., 2018). This study used an associative learning environment, where participants viewed many instances, were not asked to make category judgments, and were not provided any feedback on categorization. Thus, when the associative system is engaged, labels draw attention towards the most category-relevant features available.

This phenomenon is related to other research showing that other seemingly rule-based categories (e.g., grandmothers, odd numbers) show typicality effects (Armstrong et al., 1983; Lupyan, 2013). Armstrong and colleagues suggest that typicality effects are seen in what might be considered rule-based categories because these categories are defined both by rules for inclusion (e.g., having a grandchild) as well features that are used in identification (e.g., gray hair, tendency to bake cookies). This line of reasoning implies a continuum between rule-based and similarity-based categories, where categories with definite and verbalizable rules for inclusion are subject to processing most often associated with similarity-based categories. Thus, having a label for a category changes how individuals process that category, even when it has clearly-defined rules for inclusion.

Insight into why this might be the case comes from the Attentional Learning Account (ALA; Smith et al., 2002; Yoshida & Smith, 2005). The ALA posits that infants and young children extract statistical regularities from their environment and then use that knowledge to direct their attention towards future learning. For example, early-acquired words in English often refer to objects that are grouped based on their shape (e.g., ball). This regularity teaches the child to direct their attention towards shape when they learn a novel word. Children who are taught this regularity specifically in the laboratory also show greater vocabulary growth than untrained peers (Smith et al., 2002).

When thinking about the ALA, it is important to discuss the use of the word ”attention.” Attention can be driven either by the individual (endogenous) or by the environment (exogenous). In the endogenous case, the individual expends effort to focus on specific aspects of the stimulus (Engle & Kane, 2004). Alternatively, the environment can direct an individual’s attention to these different aspects. This exogenous case is more similar to the way attention is described in the ALA. As the individual learns that certain features tend to co-occur in a given stimulus (e.g., the name and shape of an object), an instance of one of those features draws attention towards the other. Since the label of a category is perhaps its most frequent feature, it co-occurs most often with other frequent (i.e., typical) features of that category. Thus, the typicality effects seen specifically for category labels may be the result of individuals learning statistical regularities between labels and features.

This type of iterative learning where feature distributions are learned over time closely matches the associative system. In contrast, the hypothesis-testing system is much more focused on selecting one or a few relevant features and discarding those that do not characterize category membership. In fact, many of the categories best learned by the hypothesis-testing system (e.g., *ad hoc* categories) do not have a single-word category label. Thus, a core hypothesis of this dissertation is that category labels affect learning in the associative system but not in the hypothesis-testing system. In the next section, I will discuss how language might play a role in the hypothesis-testing system.

1.3 Executive function and category learning

The hypothesis-testing system involves many executive functions (e.g., selecting and maintaining relevant category rules, inhibiting irrelevant rules). Both inhibitory control and working memory have been shown to be related to rule-based category learning (Rabi & Minda, 2014). In addition, interfering with language resources specifically affects the low-dimensional, rule-based categorization that is best processed by the hypothesis-testing system (Lupyan, 2009; Minda et al., 2008). This suggests that language is important for this system. Thus, it is possible that language and executive functions work together in the hypothesis testing system.

Indeed, language ability and executive function have been shown to be related in multiple studies to varying degrees. For example, Figueras et al. (2008) found significant positive correlations between language measures such as vocabulary and receptive grammar and a wide variety of executive function tasks for school-age children. Berninger et al. (2017) found that performance on inhibition and verbal fluency sub-tests of the D-KEFS, a standardized measure of executive function, was correlated with language outcomes in children between the ages of 9 and 15. Children with specific language impairment have also been shown to have some executive function deficits, specifically in updating and inhibition (Im-Bolter et al., 2006). Findings have been more mixed for the nature of the causal relationship between these skills. One study found a strong concurrent relationship between language and executive function longitudinally for children between ages 4 and 9, but no cross-lagged effects, suggesting that language and executive function are not directly influencing each other (Gooch et al., 2016). However, another study found that language ability at 2-3 years predicts executive function at 4 years (L. J. Kuhn et al., 2014). Thus, it is possible that the relationship between executive function and language ability changes over development. Regardless, language and executive function at least develop concurrently.

More evidence for the relationship between executive function and language comes from research on adults showing that interfering with verbal resources, usually through articulatory suppression, can negatively impact task switching (A. Baddeley et al., 2001; Emerson & Miyake, 2003). In a task-switching paradigm, performance typically decreases when an individual has to switch between tasks as compared to when they can perform the same task repeatedly. This decrease in performance is known as the switch cost. Articulatory suppression provides verbal interference by having the participant use language-related resources to repeat a nonsense string (e.g. “the the the”). In 6- and 9-year-old children, articulatory suppression has been shown to impair performance during task-switching but not during a flanker (inhibition) task (Fatzer & Roebers, 2012).

Interestingly, the negative effect of articulatory suppression on task switching is specific to instances where the individual must represent the task rules internally. For example, if participants must switch between different arithmetic functions such as addition and subtraction, verbal interference does not have an effect when the plus, minus, and equal signs are printed on the page A. Baddeley et al. (2001). A similar effect is found in a task-switching paradigm where participants must pay attention to different features of a stimulus. When the cue is the whole word (e.g., shape, color, etc.), articulatory suppression has no effect on switch cost. However, when the cue is just one letter (e.g., S, C, etc.), articulatory suppression increases the switch cost (Miyake et al., 2004). This effect suggests that task switching in these instances require a participant to use language to represent and formulate task rules (Cragg & Nation, 2010). These results indicate that language is important for representing and selecting rules, which may be similar to how the hypothesis testing system learns rule-based categories.

1.4 Interaction between category learning systems

While much research has focused on outlining dual-systems models of categorization, very little research has investigated how these two systems might interact. Both COVIS and the verbal/nonverbal approach suggest that the two systems in a dual-systems model operate in parallel. Stimuli are processed by both systems, but category decisions are made using the faster system or the system with the strongest evidence. However, some research suggests that the hypothesis-testing system may be the default. Behavioral studies encouraging participants to switch between hypothesis-testing and associative strategies in a perceptual category learning task show that unless participants are cued towards which type of strategy to use on a given trial, they tend to use hypothesis-testing strategies for all trials (Ashby & Crossley, 2010; Erickson, 2008). This suggests that the hypothesis-testing system can overpower the associative system when both are equally activated. Still, this line of research requires much more empirical evidence before definitive claims can be made.

1.5 Overview of the current study

This document consists of two experiments, aiming to answer three questions. First, **Experiment 1** investigates the relationship between the associative and hypothesis-testing systems. Almost all category learning studies using a dual-systems model utilize a between-subjects design to avoid transfer effects in learning. Thus, it is still unclear how an individual switches between the systems in response to task demands and stimulus characteristics. To investigate this switching, we test three order effects in Experiment 1 using a task adapted from Kloos & Sloutsky (2008). The first order effect compares switching between systems in matching blocks, where both the stimulus and the task demands push participants towards either the hypothesis-testing or associative system. The second order effect compares switching between the two systems for just dense stimuli, and the third for just sparse stimuli. We use task demands to encourage use of either the hypothesis-testing or associative system. We also test whether these order effects vary according to an individual's language ability. Thus, this experiment can both contribute to how we understand an individual's ability to switch between category-learning systems and assist in within-subjects experimental design with a dual-systems paradigm.

We use **Experiment 2** to answer the two remaining questions. This experiment is largely designed to test the core hypothesis that the associative system relies on labels while the hypothesis-testing system relies on executive function ability. We test individuals' vocabulary, executive function, and category learning in both systems. In addition, we are the first to directly compare category learning approaches within subjects. We use three different category learning tasks (from the COVIS, statistical density, and taxonomic-thematic approaches) to measure category learning ability. In this experiment, we directly compare performance on these three tasks for each participant to see if these category learning tasks, which have a similar theoretical basis, are indeed comparable.

2 Experiment 1

2.1 Method

2.1.1 Participants

Data was collected from 236 undergraduate psychology students at the University of Connecticut (161 Female, 67 Male, mean age = 18.94). Data for the category learning task was lost for 7 subjects due to technical errors. Thus, the final sample size was 229. Each subject was placed into one of six groups. Each group completed two blocks of the category learning task in a specific order. For more details, see Table 3. Unequal group sizes result from lost data due to technical errors.

Table 1. Group sizes for each order

Effect	Group	*N*
1	1	40
	2	38
2	3	39
	4	39
3	5	36
	6	37

2.1.2 Category Learning Task

This task measures learning of dense and sparse categories and is based off of a paradigm from previous research (Kloos & Sloutsky, 2008). Participants learn novel categories of items in four possible conditions in a 2 x 2 design. The first manipulation is learning type (supervised vs. unsupervised). In *supervised* learning, participants learn the categories by being instructed on the relevant features (e.g., “All friendly aliens have big noses.”). Images of the relevant features are provided along with the descriptions. In *unsupervised* learning, participants learn the categories by viewing sixteen instances of the category.

The second manipulation is category type (sparse vs. dense). Category type is measured by statistical density, which ranges from zero (where all features vary freely) to one (where all features co-occur perfectly). It is based on a comparison between within- and between-category entropy (Sloutsky, 2010). All categories in this experiment have seven dimensions. The *sparse* categories cohere on a single dimension, while the other dimensions vary freely (density = .25). In contrast, the *dense* categories cohere on six of the seven dimensions (density = .75). The seventh dimension is allowed to vary freely. For more details on how density was calculated, see Appendix A. Stimuli for each of the four blocks are different. See Fig. 1 for examples of the experimental manipulations. Table 2 summarizes how each experimental manipulation corresponds to the theorized category learning systems.

Table 2. Relationship between learning systems and experimental manipulations.

Experimental feature	Hypothesis-testing	Associative
Learning type	Supervised	Unsupervised
Stimulus type	Sparse	Dense

This task is within-subjects. Based on the group they were placed into, participants completed two of the four possible learning-category type combinations. In this experiment, I tested three main order effects. First, I tested order effects for the matching conditions (unsupervised-dense and supervised-sparse).The second order effect used unsupervised-dense and supervised-dense blocks. Finally, the third order effect tested the same sparse stimuli, testing unsupervised-dense and supervised-sparse blocks. This design led to six possible order groups that each participant could be placed into. See Table 3 for a summary.

In each block, participants were introduced to the task through a short cover story. They were told to learn which items go with a certain property (e.g., which aliens are friendly). Crucially, no labels were attached to the categories (e.g., some aliens are Ziblets). Then, participants completed a training block

Table 3. Block orders for statistical density task

Effect	Group	Block 1	Block 2
1	1	Unsupervised-dense	Supervised-sparse
	2	Supervised-sparse	Unsupervised-dense
2	3	Unsupervised-dense	Supervised-dense
	4	Supervised-dense	Unsupervised-dense
3	5	Unsupervised-sparse	Supervised-sparse
	6	Supervised-sparse	Unsupervised-sparse

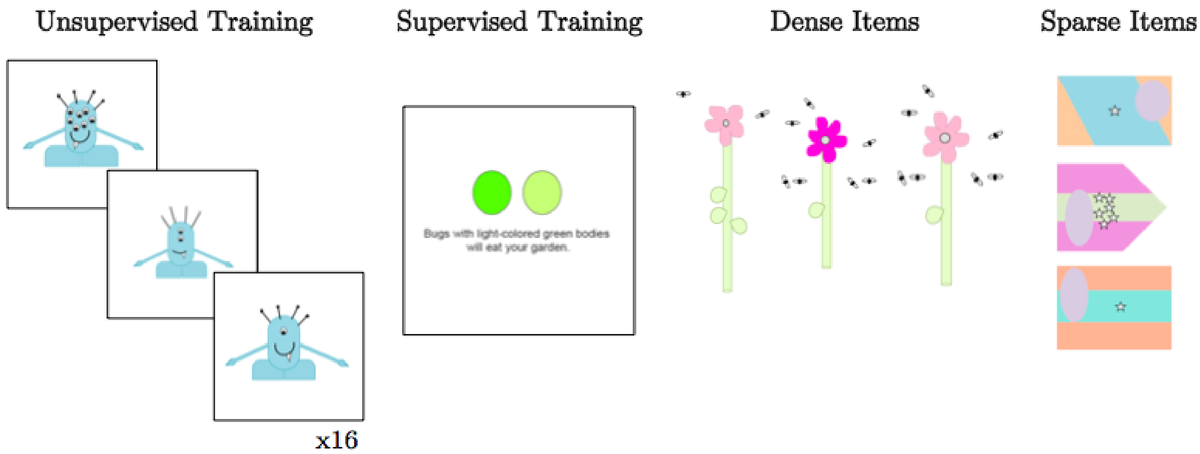

Figure 1. Examples of learning type and category type manipulations for category learning experiment.

(either supervised or unsupervised). After training, participants completed 40 test trials (16 target, 16 distractor, 8 catch), following the design of Kloos & Sloutsky (2008) . In each trial, participants saw a single item and used the keyboard to indicate whether the item matched the category they had just learned (e.g., if the alien is friendly). Catch items looked significantly different than both the target and competing categories, so participants should have always rejected them as members of the learned category. This experiment was presented using PsychoPy v.1.84.2 (Peirce, 2007).

2.1.3 Behavioral Measures

I used multiple assessments to test participants' language ability. The choice of assessments was based on the epiSLI criteria for language impairment (Tomblin et al., 1996), which includes comprehension, expression, vocabulary, grammar, and narrative. I adapted these requirements from a kindergarten population to a college-aged population. The epiSLI criteria have been shown to be robust for diagnosis of specific language impairment (SLI). In addition, other studies of language impairment more broadly have adapted a similar multidimensional approach to measuring language ability, sometimes including measures of phonological skills (Catts et al., 2006). Thus, using assessments that the many domains of language outlined in epiSLI criteria will allow me to get a fuller picture of individual differences in language ability. See Table 4 for a summary of the assessments and which domains of the epiSLI criteria they cover. The specific tests used in this experiment are detailed below.

Test of word reading efficiency (TOWRE) phonemic decoding subtest. TOWRE is a test of nonword fluency (Torgesen et al., 1992). This test is a part of the comprehension aspect of epiSLI, since the comprehension measure is reading-based. In the TOWRE, individuals have 45 seconds to read as many nonwords as possible. The nonwords become longer and more difficult as the list goes on. The raw score from the TOWRE is calculated by counting the number of words correctly pronounced before the time limit. These raw scores are then converted to standard scores using age-based norms. The standard scores are based on a distribution with a mean of 100 and a standard deviation of 15. In the current age range, a perfect raw score (63) on the TOWRE returns a standard score of ">120." For the purposes of this study, scores of ">120" will be trimmed to simply 120.

Woodcock Johnson-III word attack (WA) subtest. This task measures nonword decoding ability (Woodcock et al., 2001). Like the TOWRE, it is helpful for measuring the comprehension aspect of epiSLI. However, while the TOWRE measures word fluency, this task measures decoding accuracy. Participants read a list of nonwords out loud at their own pace. Raw scores are calculated by counting the number of words the participant said correctly. Raw scores are converted to standard scores using age-based norms. The standard score distribution has a mean of 100 and a standard deviation of 15.

Computerized reading comprehension. This test covers the comprehension and narrative aspects of epiSLI. This computerized reading comprehension (CRC) test is based on the Kaufman Test of Educational Achievement (KTEA) reading comprehension subtest (Kaufman & Kaufman, 2004). To create this test, I copied the passages and questions contained in the KTEA reading comprehension subtest into E-Prime (Schneider et al., 2002) for presentation on a computer. Then, I created multiple choice answers for the KTEA questions that did not already have them. In this task, participants read short expository and narrative texts and answer multiple-choice comprehension questions about them. Some questions are literal, while others require participants to make an inference. Participants completed as many questions as they could in 10 minutes. Once 10 minutes had elapsed, the participant was allowed to answer the question currently on the screen and then the assessment closed. Because this task is a modified version of the KTEA, I use raw scores in analysis rather than standardized scores based on the KTEA norms. Raw scores are calculated by counting the number of correctly answered questions for each participant.

Nelson-Denny vocabulary subtest. The Nelson-Denny vocabulary sub-test is a written assessment of vocabulary (Brown et al., 1981). This test covers the vocabulary aspect of epiSLI. This test has been used in multiple studies of college-aged adults and provides sufficient variability for individual difference investigations in this population (e.g., Boudewyn et al. 2015; Stafura & Perfetti 2014). In this test, participants are asked to choose the word closest to a target vocabulary word. The test has a total of 80 items. Participants were allowed unlimited time to complete all items. Raw scores were generated by counting the total number of correctly answered items. The raw scores were then converted to standard scores based upon a norming sample including students in 10th, 11th, and 12th grade as well as two- and four-year college students. The standard scores for this assessment have a mean of 200 and a standard deviation of 25.

Clinical Evaluation of Language Fundamentals recalling sentences subtest. I will use the recalling sentences subtest from the Clinical Evaluation of Language Fundamentals - Fourth Edition (CELF; Semel et al. 2006). This test covers the grammar and expression aspects of epiSLI. In this subtest, participants hear sentences and are asked to repeat them. Scoring is based on how many errors the participant makes in their repetition. Raw scores are calculated by adding up the number of points achieved for each item. These are then converted to standard scores using age-based norms. The standard scores are based on a distribution with a mean of 10 and a standard deviation of 3.

Raven's Advanced Matrices. Finally, I used Set II of Raven's Advanced Matrices (RAM) to measure nonverbal IQ (Raven, 1998). In this task, participants see a grid containing eight images and an empty space. The images are arranged in the grid according to some rule or rules. Participants must choose one of eight additional images that fits in the empty space. Due to time constraints, we restricted participants to 10 minutes in this task. Since this administration is different than the standard administration, we do not use standard scores. Raw scores are calculated by counting the number of correct answers given within 10 minutes.

Table 4. Assessments of language and their corresponding epiSLI domains.

Test	epiSLI Criteria
TOWRE WA	Comprehension (decoding aspect)
CRC	Comprehension, narrative
ND Vocab	Vocabulary
CELF RS	Grammar, expression

2.2 Procedure

Each participant completed the category learning task as well as all of the behavioral measures. TOWRE, WA, and CELF were audio-recorded to allow for offline scoring. To allow multiple subjects to be run in a single timeslot, some participants received tasks they could complete on their own (category learning, ND, computerized reading comprehension, Raven's) first while others completed tasks with the experimenter first (WA, CELF, TOWRE). All together, the seven tasks took approximately one hour.

2.3 Results

For all analyses shown below, accuracy was converted to *d'* values (Macmillan & Creelman, 2004) using the R package **neuropsychology** (Makowski, 2016). Correction for extreme values was done following (Hautus, 1995). Following prior research, all blocks where 5 or fewer catch items were correctly rejected were dropped from analysis. This resulted in 22 total missing blocks (out of 458 total), including both blocks from a single subject in group 5. For basic descriptive statistics on the category learning task, see Table 5. For reaction time, all incorrect trials were dropped, as well as any trials with a response faster than 250ms.

Table 5. Descriptive statistics for the category learning task.

Order Effect	Group	Block	Mean (SD) Accuracy	Mean (SD) RT (ms)
1	1	Unsupervised-dense	0.91 (0.19)	1074 (697)
		Supervised-sparse	0.93 (0.14)	809 (549)
	2	Supervised-sparse	0.72 (0.34)	864 (608)
		Unsupervised-dense	0.90 (0.18)	834 (563)
2	3	Unsupervised-dense	0.91 (0.18)	1063 (674)
		Supervised-dense	0.91 (0.23)	946 (636)
	4	Supervised-dense	0.90 (0.21)	960 (679)
		Unsupervised-dense	0.92 (0.18)	861 (561)
3	5	Unsupervised-sparse	0.57 (0.35)	1275 (761)
		Supervised-sparse	0.93 (0.14)	812 (543)
	6	Supervised-sparse	0.93 (0.12)	866 (592)
		Unsupervised-sparse	0.53 (0.38)	1003 (633)

2.3.1 Behavioral Assessments

For basic descriptive statistics on the behavioral measures, see Table 6. Before performing any statistical analyses using the individual difference measures, I checked their normality using the D'Agostino normality test from the R package **fBasics** (Wuertz et al., 2017). Four measures (CRC, ND Vocab, CELF RS, RAM) were significantly skewed. These measures were centered, scaled, and transformed using Yeo-Johnson transformations from R package **caret** (M. Kuhn, 2017). The remaining measures (TOWRE, WA) were not skewed and thus were simply scaled and centered.

Table 6. Descriptive Statistics for Behavioral Measures

Assessment	Mean	SD	Range
CELF Recalling Sentences SS	10.7	1.86	3-14
Computerized Reading Comprehension	21.7	5.12	7-48
Nelson-Denny Vocabulary SS	229	14.0	175-255
TOWRE SS	96.2	9.86	59-120
Word Attack SS	99.7	9.04	75-120
Raven's Advanced Matrices	15.1	4.58	0-26

Since my goal was to create a composite measure of language ability, I investigated the relationship between the behavioral measures. First, I constructed a correlation matrix between all of the behavioral measures (see Table 7). All pairs of measures had a significant positve correlation with the exception of CELF RS and RAM. To further test whether the behavioral measures could be combined into a single composite, I ran a principal components analysis (PCA) on the 5 assessments related to epiSLI (i.e., all assessments except RAM). The Kaiser-Meyer-Olkin overall measure of sampling adequacy was 0.69, above the commonly accepted threshold of 0.6. Bartlett's test of sphericity was also significant $\chi^2(10)$ = 236.16, *p* $<$0.001. These suggest that the 5 behavioral assessments were suitable for a PCA.

The first component in the PCA accounted for 47.74% of the variance and had an eigenvalue of 2.38. All of the factor loadings for this component were quite similar, ranging from -0.41 to -0.51. The second factor accounted for an additional 20.5% of the variance and had an eigenvalue of 1.02. This factor separated the two measures involved in decoding (TOWRE and WA) from the other measures (CRC, ND Vocab, and

CELF RS). The remaining components had eigenvalues below 1. Thus, of the two significant components, the first component explained almost half of the variance and had an eigenvalue more than double the second component, which largely represented decoding ability. Since the first component indicated that most of the measures loaded similarly, I decided to take a simple means approach to creating a language composite measure.

The language composite measure was created by averaging the 5 scaled, centered, and/or transformed measures. For participants with missing behavioral measures, the composite was created by averaging the remaining available measures. No subject was missing more than 1 measure. This composite measure was then scaled but not centered. This language composite measure and the centered, scaled, and transformed RAM measure are used in the analyses investigating order effects reported below.

Table 7. Correlations between behavioral measures.

	1	2	3	4	5	6
1. Computerized Reading Comprehension	-					
2. Nelson-Denny Vocabulary	0.57***	-				
3. CELF Recalling Sentences	0.31***	0.40***	-			
4. Raven's Advanced Matrices	0.31***	0.34***	0.09	-		
5. TOWRE	0.22**	0.28***	0.26***	0.16***	-	
6. Word Attack	0.22**	0.38***	0.29***	0.22***	0.53***	-

*$p < 0.05$, **$p < 0.001$, ***$p < 0.0001$

2.3.2 Order Effect 1: Matching Conditions

The first analysis investigated order effects for blocks in which the learning type (supervised vs. unsupervised) and category type (sparse vs. dense) both engaged the same category learning system (hypothesis testing vs. associative). Participants completed supervised-sparse and unsupervised-dense blocks.

Accuracy. I used linear mixed-effects models to examine the effects of block and order on accuracy at test. Accuracy in these models was measured by *d'* values for each subject by block. The base model included random intercepts for subject. Adding block and order as fixed effects significantly increased model fit, $\chi^2(2)$ = 13.21, p = 0.001. Adding the interaction between block and order further improved model fit, $\chi^2(1)$ = 6.03, p = 0.014. Thus, the final model including only experimental conditions had fixed effects of block, order, and the interaction between block and order as well as random intercepts for subject.

This model revealed two significant effects. First, there was a significant main effect of order, $F(1,75)$ = 8.60, p = 0.004. There was also a significant interaction between block and order, $F(1,74)$ = 6.10, p = 0.02. There was not a significant main effect of block, $F(1,76)$ = 2.61, p = 0.11. The interaction was broken down by conducting two separate models for each of the orders (unsupervised-dense first and supervised-sparse first). These analyses showed that when the associative system was engaged first (unsupervised-dense first), there was no significant main effect of block, $F(1,36)$ = 0.014, p = 0.91. When the hypothesis testing system was used first (supervised-sparse first), there was a significant effect of block, $F(1,37)$ = 7.52, p = 0.009. This shows that when participants complete engage the hypothesis-testing system first, performance on the supervised-sparse (hypothesis-testing) block is lower than in the unsupervised-dense (associative) block (see Table 5 Figure 2).

To investigate the effect of individual differences in language ability on the order effect, I used the final model above which included main effects for block and order as well as their interaction. I then added the language composite measure as a fixed effect. I also added RAM to control for nonverbal IQ. This model revealed no significant effects for RAM or the language composite; there remained a significant interaction between block and order.

Reaction time. Again, I used linear-mixed effects models to look at the effects of block and order on reaction time at test. While the accuracy measure was at the block level, reaction time here is modeled at the item level. The base model included random intercepts for subject and for block nested within subject. Adding the fixed effects of block and order increased the model fit, $\chi^2(1)$ = 21,02, $p < 0.001$. Further, adding the interaction between block and order improved model fit, $\chi^2(1)$ = 29.6, $p < 0.001$.

This model showed three significant effects. There was a significant main effect of block, $F(1,72)$ = 52.42, $p < 0.001$. There was also a significant main effect of order, $F(1,77)$ = 4.67, p = 0.03. Finally, there was a significant interaction between block and order, $F(1,72) = 35.0$, $p < 0.001$. To break down this interaction, I ran follow-up models for each of the two orders. This showed that when the associative system was engaged first (unsupervised-dense first), there was a significant main effect of block, $F(1,37) = 53.6$, $p < 0.001$. When the hypothesis testing system was used first (supervised-sparse first), there was no significant effect of block, $F(1,35) = 0.30$, $p = 0.59$. This result is the opposite of what was found in accuracy. When the associative system is engaged first, we see a difference in RT between blocks, but when the hypothesis-testing system in engaged first, there is no difference in RT.

Similar to the accuracy analysis, I added RAM and language ability as fixed effects to the final model from above. Neither one had any effect on RT. The main effects and interactions from above stayed significant.

Summary. While the findings from accuracy and reaction time seem to be opposing, they may in fact tell the same story. Group 1 engaged the associative system first. This group showed similar accuracy for both blocks but slower reaction time in their first block (unsupervised-dense/associative). Group 2 engaged the hypothesis-testing system first. They showed similar reaction times for both blocks, but lower accuracy in their first block (supervised-sparse/hypothesis-testing). Thus, both groups showed reduced performance (reflected in either reaction time or accuracy) on their first block, regardless of which system it engaged, perhaps reflecting a general learning effect across the task as a whole. Importantly, this learning effect is not modulated by language ability.

2.3.3 Order Effect 2: Dense Stimuli

The second order effect analysis compared groups 3 and 4. All participants learned only dense categories, with the order of learning types differing between groups.

Accuracy. Again, I used linear-mixed effects models to investigate the effects of block and order on accuracy at test. The base model included random intercepts for subject. Adding the fixed effects to the model did not significantly improve fit $\chi^2(2) = 0.07$, $p = 0.97$. Indeed, neither block, $F(1,145) = 0.053$, $p = 0.82$, nor order, $F(1,145) = 0.016$, $p = 0.90$, were significant predictors of accuracy. Thus, accuracy at test on dense categories was similar regardless of training type or block order. Next, I conducted the individual differences analysis. Since the goal of this investigation was to see whether the relationship between order and accuracy in each block changed as a function of language ability, I created a model with fixed effects for block, order, and language ability as well as RAM. The model showed no significant effects of any of the predictors.

Reaction time. I used the same linear mixed-effects model as above, with random intercepts for subject and for block nested within subject in the base model. Adding fixed effects of order and block did not significantly improve model fit, $\chi^2(2) = 2.54$, $p = 0.28$. Block, $F(1,72) = 0.05$, $p = 0.83$, and order, $F(1,77) = 2.49$, $p = 0.12$, did not have any effect on reaction time. Adding language ability and RAM to the model also did not improve fit, $\chi^2(2) = 2.46$, $p = 0.12$. These measures were not significant predictors of reaction time for dense stimuli.

Summary. There were no significant effects of block, order, or language ability found for dense stimuli. This may suggest that learning dense stimuli engages a single system regardless of the instructions. Alternatively, it may be that learning dense stimuli is overall an easy task, evidenced by the high accuracy values seen in these blocks.

2.3.4 Order Effect 3: Sparse Stimuli

The third order effect investigated differences in learning sparse categories based on learning type order, using data from groups 5 and 6.

Accuracy. I used the same type of linear mixed-effect models as the prior two order effects, with random intercepts for subject. Adding block and order significantly increased model fit, $\chi^2(2) = 57.5$, $p < 0.001$. However, adding the interaction between block and order did not increase model fit, $\chi^2(2) = 0.33$, $p = 0.56$. Thus, the final model included fixed effects for order and block but not their interaction. This model revealed a significant main effect of block, $F(1,67) = 75.69$, $p < 0.0001$, but no significant main effect of order, $F(1,67) = 0.0008$, $p = 0.98$. Participants showed significantly higher accuracy in supervised-sparse blocks than in unsupervised-sparse blocks (see Table 5).

As in the two previous analyses, I added RAM and language ability to the final model above. Adding the language composite improved model fit even after adding RAM, $\chi^2(2)$ = 5.34, p = 0.02. However, adding the interactions between block and language and order and language did not improve model fit, $\chi^2(2)$ = 1.94, p = 0.38. The final model, which included no interactions, showed the same main effect of block seen above as well as a significant main effect of language ability, F(1,63) = 5.21, p = 0.03. The effect of language ability was associated with a positive coefficient, b = 0.19, suggesting that accuracy and language ability were positively related. There was no main effect of RAM.

Reaction time. As above, I used a linear mixed-effect model with random intercepts for subject and block nested within subject as the base model. Adding the fixed effects of order and block significantly improved fit, $\chi^2(2)$ = 50.0, p <0.0001. In addition, adding the interaction between block and order improved fit, $\chi^2(2)$ = 25.4, p <0.0001. The final model showed significant main effects for block, F(1,67) = 39.46, p <0.0001, order, F(1,70) = 4.56, p = 0.04, as well as a significant interaction between block and order, F(1,67) = 29.57, p <0.0001. Follow-up models showed that for each order, there was a significant difference in reaction time by block. However, the difference between mean reaction time of the two blocks for group 5 (unsupervised-sparse first) was 463 ms, while the difference for group 6 (supervised-sparse first) was 137 ms. This suggests that the interaction represents a greater difference between blocks for participants who received the unsupervised-sparse block first.

For the individual differences analysis, I added RAM and the language composite to the final model from above. Adding the language composite did not improve the model fit. There was no effect of language ability on reaction time for sparse stimuli.

Summary. In terms of accuracy, participants showed higher accuracy on the supervised-sparse block than on the unsupervised-sparse block, regardless of order. In addition, accuracy on all blocks was positively related to language ability. This relationship did not vary by block or order. For reaction time, we saw and interaction between block and order, but no effect of language ability. The unsupervised-sparse block was by far the most difficult block for all participants who received it. Thus, this interaction may reflect this block difference crossed with learning effects. Participants who received the unsupervised-sparse block second were perhaps more comfortable with the task overall than participants who received the unsupervised-sparse block first, which lead to shorter reaction times for those receiving unsupervised-sparse second.

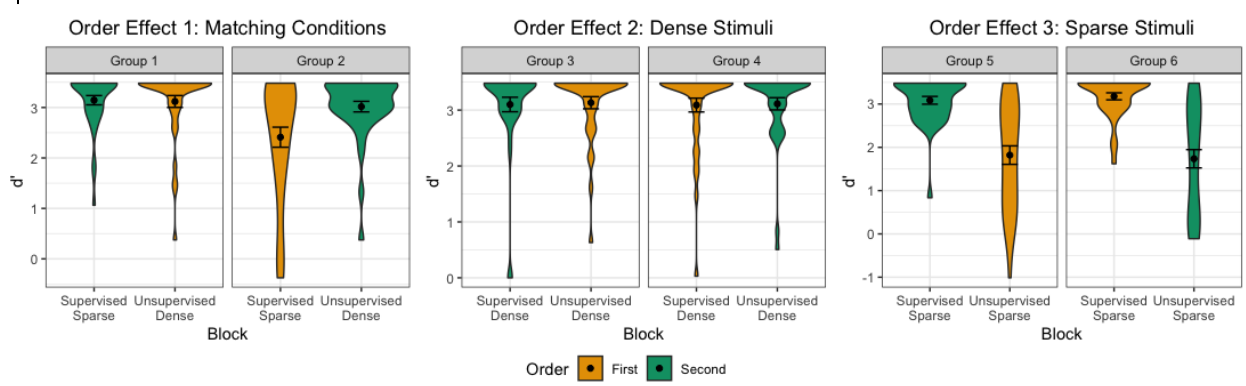

Figure 2. Accuracy (d') for each block completed by each group for all order effects. Colors indicate which block was encountered first by each group. Points indicate means with error bars reflecting standard error. Shaded portions represent the distribution of accuracy values; wider portions indicate more subjects with that accuracy value.

Figure 3. Reaction time (s) for each block completed by each group for all order effects. Colors indicate which block was encountered first by each group. Points indicate means with error bars reflecting standard error. Shaded portions represent the distribution of reaction times; wider portions indicate more trials with that reaction time.

2.4 Discussion

In this experiment, I tested three different order effects to see whether the order in which an individual engages the two category learning systems affects their learning using that system. The two manipulations in the statistical density task encouraged participants to use a particular system in two ways (learning type and stimulus type; see Table 2 for a summary). The analysis investigated whether block order affected performance in blocks where both the learning type and the stimulus type engaged the same system. The second analysis tested the effect of block order on performance when all stimuli were dense, and the third analysis did the same for only sparse stimuli.

2.4.1 Order Effects

The three analyses revealed what appears to be a general learning effect. It is most apparent in the first analysis, which showed that when both learning type and stimulus type engage the same category learning system, performance is better on the second block than on the first. Group 1 (unsupervised-dense first) showed slower reaction times in their first block, while Group 2 (supervised-sparse first) showed poorer accuracy in their first block, even though the first block for each of these groups was different. This result was also seen in a block by order interaction in the third analysis, where the difference between blocks in reaction times attenuated when the more difficult block (unsupervised-dense) was encountered second. Finally, while there were no significant effects in the second analysis, the mean reaction times are numerically higher for first blocks than for second.

The core hypothesis for this experiment was that engaging the hypothesis-testing system before the associative system would lead to reduced performance during associative blocks, but the reverse effect would not appear. This hypothesis was based on previous research that showed that when participants were required to switch between categories built using different category rules, they tended to rely more on executive function even if they were not actually switching between rule types (Erickson, 2008). Other research also has shown that when participants are asked to learn a hybrid category that combines different rule types, they end up using only a simple rule-based strategy (Ashby & Crossley, 2010). Thus, when individuals are bombarded with cues towards different systems on a trial-to-trial basis, they default to the more explicit strategies, reflecting reliance on the hypothesis-testing system. However, this type of result was not found in the current study. Instead of defaulting to the hypothesis-testing system and thus showing reduced performance on unsupervised or dense blocks that occurred second, better performance was almost always seen in second blocks. This may reflect a broad learning effect that was not seen in prior studies.

Differences in experimental paradigm may at least partially explain why this study shows learning effects while other studies show reliance on a single system. In the studies mentioned above, stimulus characteristics encouraged participants to switch between systems on a trial-by-trial basis. In the current study, trials were blocked and a single system was engaged for that block. Participants had short transition periods between blocks were new instructions and examples or rules were presented. The results presented here suggest that these transition periods were sufficient for participants to switch to a new system as stimulus and task demands changed.

2.4.2 Individual Differences

3 Experiment 2

3.1 Method

3.1.1 Participants

XX participants were recruited from the psychology undergraduate participant pool at the University of Connecticut (X Female, X Male, mean age = X).

3.1.2 Category Learning Tasks

This experiment used three different category learning tasks, each based on a different approach to category learning. We used these three tasks to investigate whether the paradigms used in different approaches engage category learning systems in a similar way. The order of category learning tasks was counterbalanced across participants. All category learning tasks were presented using PsychoPy v.1.84.2 (Peirce, 2007).

Sloustky statistical density task. This task used the same procedure and stimuli as the task described in Experiment 1. However, instead of completing only two blocks, participants completed all four blocks. Because the previous experiment showed few significant order effects, the order of the four blocks was randomly generated for each participant.

Ashby perceptual category learning task. There were two versions to this task: Information-Integration (II) and Rule-Based (RB). Participants completed the II version and then the RB version. Prior research has shown that when participants are asked to switch between the declarative (hypothesis-testing) and implicit (associative) systems, they end up using rule-based strategies from the declarative system for all trials. Thus, by engaging the implicit system first, we aimed to reduce transfer effects between versions as much as possible.

In each version of the task, participants were told that they would be learning two categories and that perfect performance was possible. They were also told to be as quick and accurate as possible. In each trial, participants viewed a Gabor patch that belonged to one of the two categories. Each patch subtended 11°of visual angle. The stimuli were generated using category parameters from Maddox et al. (2003). The participant then had 5000ms to press a key, indicating which category they believed the stimulus belonged to. After a response, the participant received feedback (”Correct” or ”Incorrect”). Feedback was presented for 1000ms, and then the next trial began. If the participant took more than 5000ms to respond, they saw ”Too Slow” and proceeded to the next trial. Participants completed five runs of each version. Each run had 80 trials (40 from each category) presented in a random order. Thus, in total participants completed 400 II trials and 400 RB trials.

Taxonomic/thematic task. This task was adapted from Murphy (2001) and Kalénine et al. (2009). There were also two versions of this task: one taxonomic and one thematic. Version order was counterbalanced across subjects, with some participants getting the taxonomic version first and others the thematic version first. Most versions of this type of task allow participants to choose the item that is most ”semantically related,” and thus do not ask participants to make either taxonomic or thematic choices on any given trial. As such, little research has looked at switching between taxonomic and thematic semantic judgments. Thus, counterbalancing was applied to control for order effects.

The stimuli were images taken from Konkle et al. (2010). We chose to use images in order to avoid automatic language processing. While participants likely did engage linguistic resources during the task, this should be due to how language relates to categorization rather than the features of the stimuli themselves. In each trial, four images were presented: a target, a taxonomically-related item, a thematically-related item, and an unrelated item. Taxonomically- and thematically-related items were chosen based on norms from Landrigan & Mirman (2016) where available. The Landrigan & Mirman (2016) norms were based on word stimuli rather than the images available from Konkle et al. (2010); as such, not all of the available images were normed. For images without norming information, we used our best judgment to pick items for each type of relation.

For each version, participants were told that they would be categorizing objects. They were told to pick the option that ”goes best with” (thematic) or is ”most similar to” (taxonomic) the target item. We chose these instructions based on previous research showing that slight differences in task instructions affect taxonomic and thematic judgments (Lin & Murphy, 2001). After instructions, participants got five practice trials. In each trial, the images were shown for 5000ms and participants had unlimited time to make a response. The practice trials were identical for the taxonomic and thematic versions of the task. After each response, participants received feedback ("Correct!" or "Oops!") for 1000ms. Once the practice trials were completed, participants received 24 test trials. While some images were seen in multiple trials, the 4-image combination for each trial was unique across the taxonomic and thematic versions of the task.

3.1.3 Executive Function Tasks

To measure executive function, we used three different tasks taken from the Psychology Experiment Building Language (PEBL) test battery (Mueller & Piper, 2014). We chose three tasks to try and tap multiple aspects of executive function, including inhibition, planning, and task-switching. All three tasks were presented using the PEBL software.

Flanker task (inhibition). This task was an implementation of the Eriksen & Schultz (1979) flanker task, using a method similar to Stins et al. (2007). In each trial, participants viewed a set of five arrows and were asked to respond based on the direction in which the center arrow was pointing (left or right). In congruent trials, all arrows faced the same way. In incongruent trials, the four distractor arrows pointed in the opposite direction of the target (center) arrow. In neutral trials, the four distractor arrows were just horizontal lines without arrowheads. Participants completed 20 trials for each condition in a 2 (direction; left vs. right) x 3 (condition; congruent vs. incongruent vs. neutral) design, for a total of 120 trials. **how many empty trials??** Each trial began with a 500 ms fixation, followed by the stimulus which appeared for 800ms. Participants were only allowed to respond during the 800ms that the stimulus was on the screen. After a response, there was an inter-trial interval of 1000ms. Participants received 12 practice trials before the actual experiment to get used to the timing of each trial. During practice trials, each response was followed by feedback ("Correct", "Incorrect") as well as a number indicating RT for that trial. This feedback was not provided for the test trials.

Switcher task (task-switching). This task was taken from Anderson et al. (2012). In this task, participants are presented with an array of colored shapes. Each colored shape has a single letter inside. For each trial, a single shape was indicated to be the target shape. Based on instructions at the top of the screen, participants were told to select a shape that matched the target shape on one of three dimensions (color, shape, or letter). Research from Miyake et al. (2004) has shown that cueing a dimension using its entire name (e.g. "shape") does not require as many language resources as cueing a dimension using a single letter (e.g., "s"). Since one of the core hypotheses of this study was that language supports executive functions in the hypothesis-testing system, we used a version of the switcher task that cued dimension using just a single letter. We expect that this version of the task requires individuals to represent dimensions/selection rules internally, similar to how they might represent possible category rules when learning rule-based categories.

The task consisted of nine different arrays of ten shapes. For each array, participants made ten responses. In the first three arrays, participants switched between two of the three dimensions in a fixed order (e.g., C - S - C - S, etc.). The relevant dimensions were different for each array. For the second three arrays, participants switched between all three dimensions still in a fixed order (e.g., S - C - L - S - C - L, etc). The specific order was different for each array. Finally, in the last three arrays participants switched between all three dimensions in a random order. Unlike previous arrays, in the last three participants were unable to anticipate the upcoming relevant dimension.

Tower of London task (planning). This task was a computerized version of the one described in Shallice (1982). In this task, participants were shown a setup of colored disks in three stacks as well as a target setup. They were given a limited number of moves to make their setup match the target setup. Participants could only have one disk in their "hand" at a time, and they could only pick the top disk up off of any stack. The trials varied in the number of steps required to match the target setup from 2 to 5, with easier (2 step) trials at the beginning of the task and harder (5 step) trials at the end of the task. Participants were encouraged to take their time and plan out their moves before beginning each trial.

3.1.4 Behavioral Measures

Finally, we used four different behavioral assessments to measure vocabulary, syntax, and nonverbal IQ.. **Neson-Denny vocabulary subtest.** To measure vocabulary, we used the same Nelson-Denny vocabulary subtest described in experiment 1.

Clinical Evaluation of Language Fundamentals recalling sentences and formulated sentences subtests. We used the CELF here to measure individuals differences in syntax production and perception. The recalling sentences subtest allowed us to look at receptive grammar, while the formulated sentences subtest provided a measure of expressive grammar. In the formulated subtest, participants view a scene and are asked to make a sentence containing a target word about that scene. Often, the target word encourages certain syntactic structures (e.g., ”because”).

Raven’s Advanced Matrices. We used Raven’s Advanced matrices to measure nonverbal IQ, as described in Experiment 1.

3.2 Procedure

Each participant completed all of the category learning and executive function tasks, as well as all of the behavioral measures. CELF responses were audio-recorded to allow for offline scoring. To allow multiple subjects to be run in a single timeslot, some participants received tasks in a shuffled order. All together, the tasks and behavioral measures took about an hour and a half.

3.3 Results

3.4 Discussion

4 General Discussion

5 Appendix A: Statistical Density Calculations

5.1 Statistical Density Formulae

Statistical density is the method that Sloutsky and colleagues use to define categories (Sloutsky, 2010). Dense categories have multiple intercorrelated features, while sparse categories have few relevant features. Statistical density can vary between 0 and 1. Higher values (closer to 1) are dense, while lower values (closer to 0) are sparse. We calculate statistical density (*D*) with the following formula, where H_{within} is the entropy within the category and $H_{between}$ is the entropy between the category and contrasting categories.

$$D = 1 - \frac{H_{within}}{H_{between}}$$

To find total entropy(*H*), we sum entropy due to varying dimension and entropy due to varying relations among dimensions.

$$H = H^{dim} + H^{rel}$$

This equation is the same whether you are calculating within-category entropy or between-category entropy. To find entropy due to dimensions, you use the following formulas, where *M* is the total number of varying dimensions, w_i is the attentional weight of a particular dimension (assumed to be 1), and p_j is the probability of value *j* on dimension *i*.

$$H^{dim}_{within} = \sum_{i=1}^{M} w_i [\sum_{j=0,1} within(p_j log_2 p_j)]$$

$$H^{dim}_{between} = \sum_{i=1}^{M} w_i [\sum_{j=0,1} between(p_j log_2 p_j)]$$

To find entropy due to relations, you use a similar set of formulas, where *O* is the total number of possible dyadic relations among the varying dimensions, w_k is the attentional weight of a relation (assumed to be 0.5), and p_{mn} is the probability of the co-occurrence of values *m* and *n* on dimension *k*.

$$H^{rel}_{within} = -\sum_{k=1}^{O} w_k [\sum_{\substack{\text{m=0,1} \\ \text{n=0,1}}} within(p_{mn} log_2 p_{mn})]$$

$$H^{rel}_{between} = -\sum_{k=1}^{O} w_k [\sum_{\substack{\text{m=0,1} \\ \text{n=0,1}}} between(p_{mn} log_2 p_{mn})]$$

All categories have 7 dimensions. For dense categories, 6 of these dimensions are correlated. The seventh dimensions is allowed to vary randomly. For sparse categories, 6 of the dimensions vary randomly. The seventh dimension is category-relevant and defines the category. All dimensions have two levels (e.g., for hair shape in aliens – curly and straight).

5.2 Statistical Density Calculations – Sparse

First, we calculate the entropy due to dimensions. We have 7 dimensions, so *M* = 7. Between categories (i.e., across all categories), each level of each dimension has a 0.5 probability of being present.

$$H_{between}^{dim} = -7 * 1(2 * 0.5log_2 0.5)$$
$$H_{between}^{dim} = -7log_2 0.5$$
$$H_{between}^{dim} = 7$$

Within categories, the relevant dimension does not vary – thus it does not contribute to the entropy. Its value goes to zero, leading to the following calculations.

$$H_{within}^{dim} = -6 * 1(2 * 0.5log_2 0.5)$$
$$H_{within}^{dim} = -6log_2 0.5$$
$$H_{within}^{dim} = 6$$

To find the entropy due to relations, we start by calculating *O*.

$$O = \frac{M!}{(M-2)! * 2!}$$
$$O = 21$$

Between categories, all dyadic relations have the same probability of co-occurrence (0.25). For each relation between dimensions, there are 4 possible combinations of the levels of those dimensions. They're all equally probable. Recall that for relations, we use an attentional weight of 0.5. So, we end up with the following.

$$H_{between}^{rel} = -21 * 0.5(4 * 0.25log_2 0.25)$$
$$H_{between}^{rel} = -10.5log_2 0.25$$
$$H_{between}^{rel} = 21$$

Within the target category, 15 of the dyadic relationships don't include the relevant feature. Thus, their probability of co-occurrence is .25. For 6 of the dyadic relations (any including the relevant feature), there is perfect co-occurrence: probability is either 0 or 1. This makes these terms go to zero, because $log_2 1 = 0$, and anything multiplied by zero is zero.

$$H_{within}^{rel} = -15 * 0.5(4 * 0.25log_2 0.25)$$
$$H_{within}^{rel} = -7.5log_2 0.25$$
$$H_{within}^{rel} = 15$$

Now, we use these calculated values to find entropy between and within categories.

$$H_{within} = 6 + 15$$
$$H_{within} = 21$$
$$H_{between} = 7 + 21$$
$$H_{between} = 28$$

Finally, we use the within- and between-category entropy to calculate the density.

$$D = 1 - \frac{21}{28}$$
$$D = 0.25$$

5.3 Statistical Density Calculations – Dense

The between category entropy for dense categories is the same as for sparse categories. $H_{between} = 28$

Next, we will consider within-category entropy due to dimensions. Six of the seven dimensions do not vary, so they do not contribute to the entropy. Their value goes to zero.

$$H_{within}^{dim} = -1 * 1(2 * 0.5 log_2 0.5)$$
$$H_{within}^{dim} = -log_2 0.5$$
$$H_{within}^{dim} = 1$$

Entropy due to relations is similar. Within the target category, 6 of the dyadic relationships don't include the relevant feature. Thus, their probability of co-occurrence is .25. For 15 of the dyadic relations, there is perfect co-occurrence, so their values go to zero.

$$H_{between}^{rel} = -6 * 0.5(4 * 0.25 log_2 0.25)$$
$$H_{between}^{rel} = -3 log_2 0.25$$
$$H_{between}^{rel} = 6$$

Next, we calculate the within-category entropy.

$$H_{within} = 1 + 6$$
$$H_{within} = 7$$

Finally, we use the within- and between-category entropy to calculate the density.

$$D = 1 - \frac{7}{28}$$
$$D = 0.75$$

References

Althaus, N., & Mareschal, D. (2014). Labels Direct Infants' Attention to Commonalities during Novel Category Learning. *PLoS ONE*, *9*(7), e99670. doi: 10.1371/journal.pone.0099670

Anderson, K., Deane, K., Lindley, D., Loucks, B., & Veach, E. (2012). *The effects of time of day and practice on cognitive abilities: The PEBL Tower of London, Trail-making, and Switcher tasks* (Tech. Rep.). Michigan Technological University.

Armstrong, S. L., Gleitman, L. R., & Gleitman, H. (1983). What some concepts might not be. *Cognition*, *13*(3), 263–308. doi: 10.1016/0010-0277(83)90012-4

Ashby, F. G., Alfonso-Reese, L. A., Turken, U., & Waldron, E. M. (1998). A neuropsychological theory of multiple systems in category learning. *Psychological Review*, *105*(3), 442–481. doi: 10.1037/0033-295X.105.3.442

Ashby, F. G., & Crossley, M. J. (2010). Interactions between declarative and procedural-learning categorization systems. *Neurobiology of Learning and Memory*, *94*(1), 1–12. doi: 10.1016/j.nlm.2010.03.001

Ashby, F. G., & Maddox, W. T. (2005). Human Category Learning. *Annual Review of Psychology*, *56*(1), 149–178. doi: 10.1146/annurev.psych.56.091103.070217

Ashby, F. G., & Maddox, W. T. (2011). Human category learning 2.0. *Annals of the New York Academy of Sciences*, *1224*(1), 147–161. doi: 10.1111/j.1749-6632.2010.05874.x

Ashby, F. G., Noble, S., Filoteo, J. V., Waldron, E. M., & Ell, S. W. (2003). Category learning deficits in Parkinson's disease. *Neuropsychology*, *17*(1), 115–124. doi: 10.1037/0894-4105.17.1.115

Baddeley, A., Chincotta, D., & Adlam, A. (2001). Working memory and the control of action: Evidence from task switching. *Journal of Experimental Psychology: General*, *130*(4), 641–657. doi: 10.1037//0096-3445.130.4.641

Baddeley, A. D., & Hitch, G. (1974). Working Memory. In *The psychology of learning and motivation* (Vol. 8, pp. 47–89). doi: 10.1016/S0079-7421(08)60452-1

Barnhart, W. R., Rivera, S., & Robinson, C. W. (2018). Effects of Linguistic Labels on Visual Attention in Children and Young Adults. *Frontiers in Psychology*, *9*, 1–11. doi: 10.3389/fpsyg.2018.00358

Berninger, V., Abbott, R., Cook, C. R., & Nagy, W. (2017). Relationships of Attention and Executive Functions to Oral Language, Reading, and Writing Skills and Systems in Middle Childhood and Early Adolescence. *Journal of Learning Disabilities*, *50*(4), 434–449. doi: 10.1177/0022219415617167

Best, C. A., Yim, H., & Sloutsky, V. M. (2013). The cost of selective attention in category learning: Developmental differences between adults and infants. *Journal of Experimental Child Psychology*, *116*(2), 105–119. doi: 10.1016/j.jecp.2013.05.002

Blaye, A., & Bonthoux, F. (2001). Thematic and taxonomic relations in preschoolers: The development of flexibility in categorization choices. *British Journal of Developmental Psychology*, *19*(3), 395–412. doi: 10.1348/026151001166173

Boudewyn, M. A., Long, D. L., Traxler, M. J., Lesh, T. A., Dave, S., Mangun, G. R., ... Swaab, T. Y. (2015). Sensitivity to Referential Ambiguity in Discourse: The Role of Attention, Working Memory, and Verbal Ability. *Journal of Cognitive Neuroscience*, *27*(12), 2309–2323.

Brown, J. I., Bennett, J. M., & Hanna, G. (1981). *The Nelson-Denny Reading Test*. Chicago, IL: Riverside Publishing.

Carpenter, K. L., Wills, A. J., Benattayallah, A., & Milton, F. (2016). A Comparison of the neural correlates that underlie rule-based and information-integration category learning. *Human Brain Mapping*, *37*(10), 3557–3574. doi: 10.1002/hbm.23259

Catts, H. W., Adlof, S. M., & Weismer, S. E. (2006). Language deficits in poor comprehenders: a case for the simple view of reading. *Journal of Speech, Language, and Hearing Research*, *49*(2), 278–293. doi: 10.1044/1092-4388(2006/023)

Chandrasekaran, B., Yi, H.-G., & Maddox, W. T. (2014). Dual-learning systems during speech category learning. *Psychonomic bulletin & review*, *21*(2), 488–95. doi: 10.3758/s13423-013-0501-5

Chandrasekaran, B., Yi, H.-G., Smayda, K. E., & Maddox, W. T. (2016). Effect of explicit dimensional instruction on speech category learning. *Attention, Perception, & Psychophysics*, *78*(2), 566–582. doi: 10.3758/s13414-015-0999-x

Cragg, L., & Nation, K. (2010). Language and the Development of Cognitive Control. *Topics in Cognitive Science*, *2*(4), 631–642. doi: 10.1111/j.1756-8765.2009.01080.x

Edmiston, P., & Lupyan, G. (2015). What makes words special? Words as unmotivated cues. *Cognition*, *143*, 93–100. doi: 10.1016/j.cognition.2015.06.008

Emerson, M. J., & Miyake, A. (2003). The role of inner speech in task switching: A dual-task investigation. *Journal of Memory and Language*, *48*(1), 148–168. doi: 10.1016/S0749-596X(02)00511-9

Engle, R. W., & Kane, M. (2004). Executive Attention, Working Memory Capacity, and a Two-Factor Theory of Cognitive Control. In B. H. Ross (Ed.), *The psychology of learning and motivation: Advances in research and theory* (pp. 145–199). New York: Elseview Science.

Erickson, M. A. (2008). Executive attention and task switching in category learning: Evidence for stimulus-dependent representation. *Memory & Cognition*, *36*(4), 749–761. doi: 10.3758/MC.36.4.749

Eriksen, C. W., & Schultz, D. W. (1979). Information processing in visual search: A continuous flow conception and experimental results. *Perception & Psychophysics*, *25*(4), 249–263. doi: 10.3758/BF03198804

Eshel, N., Nelson, E. E., Blair, R. J., Pine, D. S., & Ernst, M. (2007). Neural substrates of choice selection in adults and adolescents: Development of the ventrolateral prefrontal and anterior cingulate cortices. *Neuropsychologia*, *45*(6), 1270–1279. doi: 10.1016/j.neuropsychologia.2006.10.004

Fatzer, S. T., & Roebers, C. M. (2012). Language and Executive Functions: The Effect of Articulatory Suppression on Executive Functioning in Children. *Journal of Cognition and Development*, *13*(4), 454–472. doi: 10.1080/15248372.2011.608322

Figueras, B., Edwards, L., & Langdon, D. (2008). Executive Function and Language in Deaf Children. *Journal of Deaf Studies and Deaf Education*, *13*(3), 362–377. doi: 10.1093/deafed/enm067

Gillette, J., Gleitman, H., Gleitman, L., & Lederer, A. (1999). Human simulations of vocabulary learning. *Cognition*, *73*(2), 135–176. doi: 10.1016/S0010-0277(99)00036-0

Gooch, D., Thompson, P., Nash, H. M., Snowling, M. J., & Hulme, C. (2016). The development of executive function and language skills in the early school years. *Journal of Child Psychology and Psychiatry*, *57*(2), 180–187. doi: 10.1111/jcpp.12458

Hautus, M. J. (1995). Corrections for extreme proportions and their biasing effects on estimated values of d'. *Behavior Research Methods, Instruments, & Computers*, *27*(1), 46–51. doi: 10.3758/BF03203619

Im-Bolter, N., Johnson, J., & Pascual-Leone, J. (2006). Processing Limitations in Children With Specific Language Impairment: The Role of Executive Function. *Child Development*, *77*(6), 1822–1841. doi: 10.1111/j.1467-8624.2006.00976.x

Kalénine, S., Peyrin, C., Pichat, C., Segebarth, C., Bonthoux, F., & Baciu, M. (2009). The sensory-motor specificity of taxonomic and thematic conceptual relations: A behavioral and fMRI study. *NeuroImage*, *44*(3), 1152–1162. doi: 10.1016/j.neuroimage.2008.09.043

Kaufman, A. S., & Kaufman, N. L. (2004). *Kaufman test of educational achievement - comprehensive form.* American Guidance Service.

Kloos, H., & Sloutsky, V. M. (2008). What's behind different kinds of kinds: effects of statistical density on learning and representation of categories. *Journal of Experimental Psychology: General*, *137*(1), 52–72. doi: 10.1037/0096-3445.137.1.52

Konkle, T., Brady, T. F., Alvarez, G. A., & Oliva, A. (2010). Conceptual distinctiveness supports detailed visual long-term memory for real-world objects. *Journal of Experimental Psychology: General*, *139*(3), 558–578. doi: 10.1037/a0019165

Kuhn, L. J., Willoughby, M. T., Wilbourn, M. P., Vernon-Feagans, L., & Blair, C. B. (2014). Early Communicative Gestures Prospectively Predict Language Development and Executive Function in Early Childhood. *Child Development*, *196*(5), 1898–1914. doi: 10.1111/cdev.12249

Kuhn, M. (2017). caret: Classification and regression training [Computer software manual]. Retrieved from `https://CRAN.R-project.org/package=caret` (R package version 6.0-76)

Landrigan, J.-F., & Mirman, D. (2016). Taxonomic and Thematic Relatedness Ratings for 659 Word Pairs. *Journal of Open Psychology Data*, *4*(1), 2. doi: 10.5334/jopd.24

Lewis, D. (1997). Development of the Prefrontal Cortex during Adolescence: Insights into Vulnerable Neural Circuits in Schizophrenia. *Neuropsychopharmacology*, *16*(6), 385–398. doi: 10.1016/S0893-133X(96)00277-1

Lin, E. L., & Murphy, G. L. (2001). Thematic relations in adults' concepts. *Journal of Experimental Psychology: General*, *130*(1), 3–28. doi: 10.1037/0096-3445.130.1.3

Lupyan, G. (2009). Extracommunicative functions of language: verbal interference causes selective categorization impairments. *Psychonomic bulletin & review*, *16*(4), 711–718. doi: 10.3758/PBR.16.5.986

Lupyan, G. (2012). What Do Words Do? Toward a Theory of Language-Augmented Thought. In B. Ross (Ed.), *Psychology of learning and motivation - advances in research and theory* (Vol. 57, pp. 255–297). Academic Press. doi: 10.1016/B978-0-12-394293-7.00007-8

Lupyan, G. (2013). The difficulties of executing simple algorithms: Why brains make mistakes computers don't. *Cognition*, *129*(3), 615–636. doi: 10.1016/j.cognition.2013.08.015

Lupyan, G. (2017). The paradox of the universal triangle: Concepts, language, and prototypes. *The Quarterly Journal of Experimental Psychology*, *70*(3), 389–412. doi: 10.1080/17470218.2015.1130730

Lupyan, G., & Mirman, D. (2013). Linking language and categorization: Evidence from aphasia. *Cortex*, *49*(5), 1187–1194. doi: 10.1016/j.cortex.2012.06.006

Lupyan, G., Mirman, D., Hamilton, R., & Thompson-Schill, S. L. (2012). Categorization is modulated by transcranial direct current stimulation over left prefrontal cortex. *Cognition*, *124*(1), 36–49. doi: 10.1016/j.cognition.2012.04.002

Lupyan, G., Rakison, D. H., & McClelland, J. L. (2007). Language is not Just for Talking: Redundant Labels Facilitate Learning of Novel Categories. *Psychological Science*, *18*(12), 1077–1083. doi: 10.1111/j.1467-9280.2007.02028.x

Lupyan, G., & Spivey, M. J. (2010). Making the Invisible Visible: Verbal but Not Visual Cues Enhance Visual Detection. *PLoS ONE*, *5*(7), e11452. doi: 10.1371/journal.pone.0011452

Macmillan, N. A., & Creelman, C. D. (2004). *Detection theory: A user's guide*. Psychology press.

Maddox, W. T., Ashby, F. G., & Bohil, C. J. (2003). Delayed feedback effects on rule-based and information-integration category learning. *Journal of Experimental Psychology: Learning, Memory, and Cognition*, *29*(4), 650–662. doi: 10.1037/0278-7393.29.4.650

Makowski, D. (2016). Package 'neuropsychology': An r toolbox for psychologists, neuropsychologists and neuroscientists [Computer software manual]. Retrieved from `https://github.com/neuropsychology/neuropsychology.R` (0.5.0)

Markman, E. M., & Hutchinson, J. E. (1984). Children's sensitivity to constraints on word meaning: Taxonomic versus thematic relations. *Cognitive Psychology*, *16*(1), 1–27. doi: 10.1016/0010-0285(84)90002-1

Minda, J. P., Desroches, A. S., & Church, B. A. (2008). Learning rule-described and non-rule-described categories: A comparison of children and adults. *Journal of Experimental Psychology: Learning, Memory, and Cognition*, *34*(6), 1518–1533. doi: 10.1037/a0013355

Minda, J. P., & Miles, S. J. (2010). The influence of verbal and nonverbal processing on category learning. In B. H. Ross (Ed.), *The psychology of learning and motivation* (Vol. 52, pp. 117–162). Burlington: Academic Press. doi: 10.1016/S0079-7421(10)52003-6

Miyake, A., Emerson, M. J., Padilla, F., & Ahn, J.-c. (2004). Inner speech as a retrieval aid for task goals: the effects of cue type and articulatory suppression in the random task cuing paradigm. *Acta Psychologica*, *115*(2-3), 123–142. doi: 10.1016/j.actpsy.2003.12.004

Mueller, S. T., & Piper, B. J. (2014). The Psychology Experiment Building Language (PEBL) and PEBL Test Battery. *Journal of Neuroscience Methods*, *222*(1), 250–259. doi: 10.1016/j.jneumeth.2013.10.024

Murphy, G. L. (2001). Causes of taxonomic sorting by adults: A test of the thematic-to-taxonomic shift. *Psychonomic Bulletin & Review*, *8*(4), 834–839. doi: 10.3758/BF03196225

Nomura, E. M., Maddox, W. T., Filoteo, J. V., Ing, A. D., Gitelman, D. R., Parrish, T. B., ... Reber, P. J. (2007). Neural correlates of rule-based and information-integration visual category learning. *Cerebral Cortex*, *17*(1), 37–43. doi: 10.1093/cercor/bhj122

Peirce, J. W. (2007). PsychoPy—Psychophysics software in Python. *Journal of Neuroscience Methods*, *162*(1-2), 8–13. doi: 10.1016/j.jneumeth.2006.11.017

Perry, L. K., & Lupyan, G. (2014). The role of language in multi-dimensional categorization: Evidence from transcranial direct current stimulation and exposure to verbal labels. *Brain and Language*, *135*, 66–72. doi: 10.1016/j.bandl.2014.05.005

Perry, L. K., & Lupyan, G. (2016). Recognising a zebra from its stripes and the stripes from "zebra": the role of verbal labels in selecting category relevant information. *Language, Cognition and Neuroscience*, *3798*, 1–19. doi: 10.1080/23273798.2016.1154974

Piaget, J., Inhelder, B., & Lunzer, E. A. (1964). *The early growth of logic in the child: Classification and seriation*. Routledge & Kegan Paul.

Plebanek, D. J., & Sloutsky, V. M. (2017). Costs of Selective Attention: When Children Notice What Adults Miss. *Psychological Science*, *28*(6), 723–732. doi: 10.1177/0956797617693005

Plunkett, K., Hu, J.-F., & Cohen, L. B. (2008). Labels can override perceptual categories in early infancy. *Cognition*, *106*(2), 665–681. Retrieved from `http://linkinghub.elsevier.com/retrieve/pii/S0010027707001084` doi: 10.1016/j.cognition.2007.04.003

Rabi, R., & Minda, J. P. (2014). Rule-Based Category Learning in Children: The Role of Age and Executive Functioning. *PLoS ONE*, *9*(1), e85316. doi: 10.1371/journal.pone.0085316

Raven, J. C. (1998). *Raven's progressive matrices*. Oxford: Oxford Psychologists Press.

Reetzke, R., Maddox, W. T., & Chandrasekaran, B. (2016). The role of age and executive function in auditory category learning. *Journal of Experimental Child Psychology*, *142*(3), 48–65. doi: 10.1016/j.jecp.2015.09.018

Robinson, A. L., Heaton, R. K., Lehman, R. A., & Stilson, D. W. (1980). The utility of the Wisconsin Card Sorting Test in detecting and localizing frontal lobe lesions. *Journal of Consulting and Clinical Psychology*, *48*(5), 605–614. doi: 10.1037//0022-006X.48.5.605

Rodman, H. R. (1994). Development of Inferior Temporal Cortex in the Monkey. *Cerebral Cortex*, *4*(5), 484–498. doi: 10.1093/cercor/4.5.484

Schneider, W., Eschman, A., & Zuccolotto, A. (2002). *E-prime: User's guide*. Psychology Software Incorporated.

Segalowitz, S. J., & Davies, P. L. (2004). Charting the maturation of the frontal lobe: An electrophysiological strategy. *Brain and Cognition*, *55*(1), 116–133. doi: 10.1016/S0278-2626(03)00283-5

Semel, E. M., Wiig, E. H., & Secord, W. (2006). *CELF 4: clinical evaluation of language fundamentals*. Pearson: Psychological Corporation.

Shallice, T. (1982). Specific Impairments of Planning. *Philosophical Transactions of the Royal Society B: Biological Sciences*, *298*(1089), 199–209. doi: 10.1098/rstb.1982.0082

Sloutsky, V. M. (2010). From Perceptual Categories to Concepts: What Develops? *Cognitive Science*, *34*(7), 1244–1286. doi: 10.1111/j.1551-6709.2010.01129.x

Smiley, S. S., & Brown, A. L. (1979). Conceptual preference for thematic or taxonomic relations: A nonmonotonic age trend from preschool to old age. *Journal of Experimental Child Psychology*, *28*(2), 249–257. doi: 10.1016/0022-0965(79)90087-0

Smith, L. B., Jones, S. S., Landau, B., Gershkoff-Stowe, L., & Samuelson, L. (2002). Object Name Learning Provides On-the-Job Training for Attention. *Psychological Science*, *13*(1), 13–19. doi: 10.1111/1467-9280.00403

Snedeker, J., & Gleitman, L. (2004). Why It Is Hard To Label Our Concepts. In D. G. Hall & S. R. Waxman (Eds.), *Weaving a lexicon* (pp. 257–294). Cambridge, MA: MIT Press.

Soto, F. A., Waldschmidt, J. G., Helie, S., & Ashby, F. G. (2013). Brain activity across the development of automatic categorization: A comparison of categorization tasks using multi-voxel pattern analysis. *NeuroImage*, *71*, 284–297. doi: 10.1016/j.neuroimage.2013.01.008

Stafura, J. Z., & Perfetti, C. A. (2014). Word-to-text integration: Message level and lexical level influences in ERPs. *Neuropsychologia*, *64*(11), 41–53. doi: 10.1016/j.neuropsychologia.2014.09.012

Stins, J. F., Polderman, J. C. T., Boomsma, D. I., & de Geus, E. J. C. (2007). Conditional accuracy in response interference tasks: Evidence from the Eriksen flanker task and the spatial conflict task. *Advances in Cognitive Psychology*, *3*(3), 409–417. doi: 10.2478/v10053-008-0005-4

Tomblin, J. B., Records, N. L., & Zhang, X. (1996). A System for the Diagnosis of Specific Language Impairment in Kindergarten Children. *Journal of Speech Language and Hearing Research*, *39*(6), 1284–1294. doi: 10.1044/jshr.3906.1284

Torgesen, J. K., Wagner, R., & Rashotte, C. (1992). *TOWRE-2: test of word reading efficiency*. Austin, TX: Pro-Ed.

Vygotsky, L. S. (1962). Language and thought. *Massachusetts Institute of Technology Press, Ontario, Canada*.

Waldschmidt, J. G., & Ashby, F. G. (2011). Cortical and striatal contributions to automaticity in information-integration categorization. *NeuroImage*, *56*(3), 1791–1802. doi: 10.1016/j.neuroimage.2011.02.011

Waxman, S. R., & Markow, D. B. (1995). Words as Invitations to Form Categories: Evidence from 12- to 13-Month-Old Infants. *Cognitive Psychology*, *29*(3), 257–302. doi: 10.1006/cogp.1995.1016

Waxman, S. R., & Namy, L. L. (1997). Challenging the notion of a thematic preference in young children. *Developmental Psychology*, *33*(3), 555–567. doi: 10.1037/0012-1649.33.3.555

Woodcock, R. W., McGrew, K. S., Mather, N., & Schrank, F. (2001). *Woodcock-Johnson III NU tests of achievement*. Rolling Meadows, IL: Riverside Publishing.

Wuertz, D., Setz, T., & Chalabi, Y. (2017). fbasics: Rmetrics - markets and basic statistics [Computer software manual]. Retrieved from `https://CRAN.R-project.org/package=fBasics` (R package version 3042.89)

Yoshida, H., & Smith, L. B. (2005). Linguistic Cues Enhance the Learning of Perceptual Cues. *Psychological Science*, *16*(2), 90–95. doi: 10.1111/j.0956-7976.2005.00787.x