THESIS

AN INVESTIGATION INTO THE FORMATION OF REPRESENTATIONAL ASSOCIATIONS IN VISUAL CATEGORY LEARNING

Submitted by

Kade Garrett Jentink

Department of Psychology

In partial fulfillment of the requirements

For the Degree of Master of Science

Colorado State University

Fort Collins, Colorado

Spring 2017

Master's Committee:

Advisor: Carol Seger

Don Rojas
Agnieszka Burzynska

ABSTRACT

AN INVESTIGATION INTO THE FORMATION OF REPRESENTATIONAL ASSOCIATIONS IN VISUAL CATEGORY LEARNING

Category learning allows us to use previous information we have accumulated, and extend it to new situations. Multiple systems are proposed to underlie learning, including: an explicit, rule-based system, and an implicit, procedural system. Information integration tasks are thought to load heavily onto the latter. In these tasks, a high degree of accuracy is reached only if participants can integrate incommensurable dimensions, often without being able to verbally describe how they are categorizing each stimulus. Learning in this type of task is thought to occur as participants associate a given stimulus with a category label, and then that label to a motor response. The present study sought to examine whether there may be an additional associative stage in which a stimulus is first associated with a "category representation" – a representation of the critical characteristics of a given category – which is then associated with a category label. Two experiments were conducted which attempted to determine whether this form of category representation is learned in information integration tasks. Both experiments reversed the category representation – category label association for a subset of stimuli and tested if subjects would transfer this reversal to the remaining stimuli, as should happen if they learned to associate each label with a single abstract category representation. Experiment 1 trained subjects with two sets of labels, each of which was associated with the same abstract category representation, to see if reversing one set of labels would alter the other. Experiment 2 trained subjects with 1 set of labels and tested if learning to reverse half of the stimulus space would transfer to the remaining half. In addition, the consistency of category label and motor response associations were manipulated in Experiment 2, with the hypothesis that subjects learning under inconsistent mappings would be forced to learn category labels and be more likely form an abstract category representation, whereas subjects learning under consistent conditions might only learn basic stimulus – response associations. Subjects in Experiment 1 did not

transfer the reversal to the second set of category labels, inconsistent with the hypothesis that subjects would form an abstract category representation. However, over half the subjects in Experiment 2 did transfer reversed category label associations to untrained stimuli. Furthermore, a greater number of subjects transferred the reversals in the Inconsistent mapping condition. This is the first study to present evidence suggesting the existence of an abstract category representation and to provide a unique dissociation between consistent and inconsistent mappings for an information-integration task.

TABLE OF CONTENTS

CHAPTER 1: INTRODUCTION

Categorization is necessary for survival. In addition to helping us recognize threats, individuals, and locations, categorization is necessary for forming meaningful associations between similar items, and in the case of encountering something unknown, allowing us to extrapolate based on past experiences. We can even use previously acquired representations from memory to imagine what the future could be like. This is because our minds store commonalities in addition to specific representations. In general, categorization is something humans can do quickly and effectively (Seger & Miller, 2010). While research into how we learn to categorize stretches across several domains (e.g. somatosensory, auditory, emotion), visual categorization is the most studied area (Richler & Palmeri, 2014).

Categorization, decision making, and generalization may all be intertwined to a greater degree than is currently emphasized in the literature (Seger & Peterson, 2013). Research into the latter two areas, therefore, may benefit through a greater understanding of category learning. For example, understanding the mechanisms of how we learn to categorize stimuli might elucidate how we use this information to make a decision to act. Furthermore, category representations allow us to generalize and transfer our knowledge to new situations and new task demands. Therefore, by adding to the available knowledge on category learning, other applied research areas can benefit as well (e.g. neuroeconomics, top-down mediation of perception).

I begin by describing the types of tasks used in category learning studies, as much of this terminology will be used throughout the paper. Next, I discuss early theories of category learning, which form the basis for modern theories of category learning. Afterwards, I discuss the single- versus multiple-systems debate, presenting evidence from both sides on the nature of how category representations are formed and used. I will focus largely on behavioral dissociation studies, along with some evidence from neuroimaging, that are relevant for understanding the formation of associations between stimuli, categories, and motor responses. Finally, I describe two studies conducted to investigate whether

intermediate category representations between stimuli and category labels are formed during information integration category learning.

Categorization tasks

It is important to begin by introducing some of the terminology used in describing the different types of category learning tasks. While there are others, the tasks most relevant for the proposed study are rule-based (RB) and information-integration (II) tasks. (Ashby & Maddox, 2011). Rule-based tasks are ones in which the categories can be determined through logical reasoning. Generally, the rule by which stimuli can be most optimally categorized can be explicitly stated by participants, and it is often a one-dimensional rule (e.g. "if blue, category A, if green, category B"), although it does not necessarily have to be. One example is the Wisconsin Card Sorting Task (WCST), in which participants must learn rules to sort cards into groups (Maddox, Ashby, & Bohil, 2003). These types of tasks are thought to rely on declarative memory systems, including both working and episodic memory, along with executive function systems.

Information-integration tasks are thought to rely more on implicit, procedural memory systems. In information integration tasks, participants must integrate information from two, often incommensurable dimensions. A commonly used task, and one which is directly relevant to the proposed study, presents circular, sine-wave gradient stimuli which vary in bar rotation and bar width. The perceptual space the stimuli are sampled from is divided by a line moving through the space at a 45-degree angle. In this way, as one dimension moves along the x-axis, the characteristic of the y-axis increases as well. To succeed, participants cannot rely on a verbal rule (e.g. when is width greater than rotation?) but instead must learn to associate responses with regions of perceptual space (Ashby, Paul, & Maddox, 2011; refer to figure 4 and 8 in methods for illustration). The procedural system is commonly associated with skills learned through practice, and there is generally little conscious recollection of the associated memories. Learning in this domain also requires consistent feedback, and is slow and incremental. Accuracy in some information-integration tasks is often maximized when information from two or more stimulus dimensions is integrated at some predecisional stage, but in general, learning in

information-integration tasks takes place at an unconscious level. An example of a procedural learning task is the Serial Reaction Time (SRT) task, in which reaction-time performance to rapid button presses is reduced through repetition of extended patterns of stimulus presentation; this occurs even when participants are not aware of the repetition (Maddox, Ashby, Ing, & Pickering, 2004).

Early Theories: The Prototype Model

There have been several competing theories put forth over the decades in regards to the mechanisms through which category representations are learned and used. Each makes assumptions about how stimuli representations are created, what information is needed to recognize a category, and how a category decision is eventually made (Ashby & Maddox, 1993). One commonality across theories is an emphasis on how category learning utilizes fundamental memory systems. In the early years of category learning research theories typically assumed that learning relied on a single memory system. The question of which memory system that may be, and in what manner the memory system is recruited to facilitate category learning, was a common subject of debate (Ashby & O'Brien, 2005). An early theory of category learning suggested that all categories are learned through the acquisition of logical rules, which are determined through simple, explicit hypothesis testing. This was even suggested to apply to how animals might learn categories (Bourne, 1970). However, this explanation was argued as being too artificial (Richler & Palmeri, 2014), and although rule based category learning is still accepted as one form of category learning it is not thought to encompass all learning. Instead, it is thought that there may be several different forms of category learning.

An appreciation of the limitations of rule based theory led to the development of prototype theory, which argued that a prototypical representation is learned. Effectively, the prototype serves as an internal representation of a category, and it is constructed from features abstracted from exemplars of that category (Goldman & Homa, 1977). Posner and Keele (1968) proposed that, during learning, the commonalities within a group are abstracted from individual exemplars and stored in memory. To demonstrate this, they used a visual categorization task in which, during a training phase, participants learned highly distorted versions of different dot-pattern prototypes until they had correctly identified two

sets of stimuli with perfect accuracy. The dot-pattern stimuli presented during training were generated by first creating a "prototype" for each group (composed of 9 dots, arbitrarily arranged in a 30 x 30 point matrix), and varying the deviation of each of the starting points of the dots to different degrees based on certain statistical procedures. As the deviation of the dots from their original starting points increased, they were said to have greater perceptual variability from the prototype. In the test phase, participants were shown the trained stimuli as well as other related ones: new stimuli with equal perceptual variability, stimuli with greater perceptual variability, and the prototype stimuli they were all based on. They found that identification of the prototypes was less error prone, and elicited quicker responses, compared to other exemplars, even those which were presented during training. They suggested that during learning, participants learned the central tendency of the category (the prototype) through being exposed to stimuli which varied in their perceptual dimensions. Therefore, when presented with the prototype stimuli the trained stimuli were based on, they were quickly able to identify the most representative member of the categories they had learned.

Prototype theory accounted well for the results from several category learning studies. The quick and accurate identification of the prototype in Posner and Keele's study suggested that participants learned the prototypical features of each category even without direct observation of the prototype. Another study, which used similar methods but different stimuli, inserted a one-week delay between the training and testing phase and found that while training stimuli were mostly forgotten, prototypes and new patterns were easily identified. The forgetting of specific stimuli and their features suggested that category learning happens at a more abstract level and that the prototype representation is the important part of the category learning process (Goldman & Homa, 1977).

One criticism leveled towards the prototype theory was that, at the time, almost all category learning models predicted excellent classification of a prototypical stimulus (Reed, 1972). A prototype, by definition, has the greatest number of similarities to its exemplars, and additionally, it is highly unlikely that one prototype would be similar to exemplars from any other category. The next wave of category learning theories was exemplar driven. They suggested that, instead of learning a prototype through

abstraction, categories were learned by combining features only from exemplars which had been previously presented. This process did not require abstraction, but could still account for effects in the prototype theory literature. It also emphasized the importance of a participant's contextual knowledge (i.e. presented exemplars) in how categories are learned as well (Richler & Palmeri, 2014).

Early Theories: The Exemplar Model

One specific exemplar based model was called the context theory of classification (Medin & Schaffer, 1978). It proposed that when a stimulus is presented, its features should activate an associative network created by aggregating features from similar, previously presented stimuli. One important difference is that, rather than the categorization decision depending on a comparison utilizing an equally weighted average of known diagnostic features, as in prototype theory, the context model specified a multiplicative combination rule, wherein high similarity to some features is more important than average similarity overall. An example is that, while a mannequin may very closely resemble a human (e.g. anatomical similarity, clothing, etc.) the lack of animacy (i.e. one feature) is a far greater determinate of category membership than its average similarity (Medin & Schaffer, 1978). In a series of four experiments, Medin and Schaffer (1978) utilized stimuli with several different binary dimensions (e.g. geometric shapes; size: big or small). They found that their statistical models not only outperformed many other existing theories in accounting for their results, but they were also able to account for the results of studies originally interpreted as supporting other theories as well. They suggested that this was evidence of their model being more broadly applicable.

An extension to the context model came in the form of the generalized context model (GCM; Nosofsky, 1986). One advantage of the GCM included accommodating stimuli with multivalued, continuous features instead of just binary ones. This was important because while the context model was proposed with the idea that natural stimuli are more arbitrarily constrained, they only used binary-choice features. The GCM also proposed that the development of categories through exposure to exemplars was more complicated than initially proposed. It was suggested that as participants focus on certain salient features of stimuli, they may inappropriately infer that some continuous measure of a particular stimulus

feature reflects category membership. This would alter the psychological dimensions for category membership between participants, which they interpreted as meaning that each different subject could potentially maintain a different mental representation of each category. A small study Nosofsky (1986) conducted concurred with these hypotheses, and he cautioned interpreting results in other studies as suggesting "a direct reflection of the underlying similarity representation, or of attention and decision processes that operate on this representation" (p. 54).

Early Theories: The Decision-Bound Model

Another well-supported theory was known as the decision-bound model. This model had its basis in general recognition theory (GRT; Ashby & Townsend, 1986). General recognition theory assumes that participants learn categories by associating specific responses with a corresponding region in perceptual space. Because repeated presentations of a stimulus do not always necessarily generate the same perceptual event due to perceptual noise, the likelihood of whether a stimulus will be perceived as belonging to one category or another follows a normal distribution. This causes categories to be separated into groups by more than just their corresponding exemplars, but also "as a probability mixture of the individual exemplar distributions" (Ashby & Maddox, 1992, p.53). This model proposes that participants divide the representational stimulus space into response regions each associated with a category label, and this partition between response regions is referred to as the decision bound. The GRT assumes participants attempt to respond optimally, but encounter certain limitations: "perceptual noise, selection of a sub-optimal decision bound, variability in the memory of this bound, response bias, and variability in the memory of the response criterion" (Ashby & Maddox, 1993, p. 377). This model predicts that, eventually, categorization should become automatic once a category becomes associated with a particular region. This is in contrast to exemplar theory, in which it is proposed that a participant must make a comparison between the current exemplar and all other presented exemplars every time (Ashby & Maddox, 1990).

It was common for researchers to conduct a category learning study in which they would apply several different models to the data in an attempt to see which one best could account for the data. The

"best fitting" model would then be assumed to be correct. Eventually, however, some concerns arose concerning the efficacy of this approach (Chandrasekaran, Koslov, & Maddox, 2014). Ashby and Maddox (1993) demonstrated that, at the level of the data, the prototype, exemplar, and decision-bound models were mathematically equivalent. Additionally, some data emerged which demonstrated behavioral dissociations between category learning tasks which could not be explained easily by any single-system model (Ashby, Alfonso-Reese, Turken, & Waldron, 1998). Finally, neuroimaging evidence began to emerge suggesting that, based on certain category learning task demands, multiple neural regions associated with different learning and memory functions could be recruited simultaneously; sometimes in parallel (both increasing), sometimes competitively (one increasing while the other decreased). Commonly examined learning and memory regions included components of the medial temporal lobe (MTL; hippocampus) responsible for declarative memory, and the basal ganglia, which can be involved in skills and instrumental learning. This suggested that different learning systems were engaged in a context driven manner, and that category learning may actually rely on more than one representational system (Poldrack & Packard, 2003). The single- versus multiple-system debate is complex, and arguments from both sides will be considered.

Single- versus Multiple-systems views

Current models of category learning are much more complicated than the early models, and involve converging evidence from behavior, neurological, and computational modeling domains (Richler & Palmeri, 2014). Early theories of category learning never specified which memory system may be in use during the categorization process. However, as research from the memory literature began to suggest there may be multiple memory systems, inquiry into which system(s) may be responsible for which aspects of category learning began. This interest was supplemented by advances in imaging methods (e.g. functional magnetic resonance imaging; fMRI) which allowed for localization of activity in the brain. In the beginning, virtually all category learning theories assumed a single-system model. While the term "single-system" may have at first referred to a single *neural* system, single-system proponents now take the view that there is a single *representational* system which may be used differently depending on the

circumstances and is not necessarily confined to a single neural system. Single-system theorists suggest that one representation may be shared across different types of tasks and demands, while multiple-system theorists argue that separate and independent systems are recruited based on the task to be performed (Richler & Palmeri, 2014).

Single- versus Multiple-systems views: Multiple interpretations

On a broad, conceptual level, the main argument put forth by single-system theorists is one of parsimony: that the evidence put forth to suggest a multiple-systems model can be explained using computational models that do not involve multiple systems. The argument, then, is that the single-systems view is preferable (Poldrack & Foerde, 2008). To explain evidence from neuroscience demonstrating differential neural activity during certain tasks, single-systems theorists argue that the activity reflects the use of different neural systems for different computations, but that it still reflects the usage of one representation. Additionally, some systems may be recruited more than others for certain category learning tasks (Richler & Palmeri, 2014). There are even relatively modern category learning models that reflect this single-systems perspective. The attention learning covering map (ALCOVE; Kruschke, 1992) is an extension of the exemplar-based general context model, and incorporates "perceptual processing, perceptual memories, selective weighting based on diagnosticity, learned associations between exemplars and categories, and categorization decision mechanisms, all of which can be subject to top-down executive control" (Richler & Palmeri, 2014, p. 85). This model, however, has had a hard time accounting for data from recently conducted dissociation studies, which will be discussed later (Maddox et al., 2003). While some published studies claim that "many researchers now accept the strength of the evidence supporting multiple systems" (Ashby & Maddox, 2011), there is still some debate regarding the interpretation of this evidence.

One such example comes from an fMRI study which investigated differences in intentional (explicit) or incidental (implicit) learning tasks using dot-prototype stimuli (Reber, Gitelman, Parrish, & Mesulam, 2003). The participants in the intentional condition were told they would see a number of dots and that the configuration reflected category membership (for only one category); they were told that

should try to learn this relationship. In the incidental condition participants were told nothing regarding the categorical nature of the task or stimuli. Instead, they were told it was a mental imagery study, and they were instructed to imagine pointing at the dot in the center of the screen. During the testing phase, the categorical nature of the task was revealed, and they were instructed to respond as to whether each presented stimulus did or did not belong to the previously presented category. While accuracy between groups was equal, they found different patterns of brain activity between tasks. The intentional group showed greater activity in the hippocampus and some cortical regions; areas proposed to be associated with explicit category learning tasks, while the incidental group showed a reduced activation in the right middle occipital cortex (primary visual) for novel stimuli from the trained category. Repetition suppression (RS), a phenomena which describes reduced neural activation in response to repeated stimuli (Grill-Spector, Henson, & Martin, 2006), could have possibly accounted for the latter effect, however, the authors suggested that RS models do not suggest this effect can generalize to novel, related stimuli. Therefore, based on previous research, the authors interpreted these data as representing fluent category processing. Taking these results in combination, the researchers argued that their results demonstrated separate category representations, in support of the multiple-systems view (Reber et al., 2003).

A criticism leveled at this study had to do with the way the task was explained to participants; the separate "representations" could have been due to differences in stimulus-encoding processes. A study by Gureckis, James, and Nosofsky (2011) replicated and extended the Reber and colleagues (2003) paper to decouple factors which they claimed could have produced data in favor of multiple-systems. In addition to directly replicating the conditions used in the Reber et al. (2003) paper, they also added two additional conditions. In an additional intentional condition, participants were still told they would be learning categories, but that the most important diagnostic feature was the center dot and that they should imagine pointing to it. The additional incidental condition was a similar reversal in that, while they were still not told about the categorical nature of the task, they were asked to focus on the configuration of the dots. They found that their encoding instructions (attention to the center dot versus overall configuration) strongly influenced the observed patterns of activation, regardless of the explicit/implicit nature of the

task, and they suggested that the observed activity was better explained by how participants visually processed the stimuli rather than evidence for multiple-systems (Gureckis et al., 2011).

This example demonstrates the complexity of this debate, and how easily critics from both sides can interpret the evidence as favoring their views. Typically, criticisms originating from single-system views are often methodological in nature, focusing on critique of a single experiment supporting multiple systems. While evidence in favor of the multiple-systems view is large and still growing (Ashby & Maddox, 2005; Ashby & Maddox, 2011), critics suggest, in light of studies such as the aforementioned fMRI study, the current amount of evidence may not be enough to completely reject the single-system view. While they do not suggest that the multiple-systems view is entirely incorrect either, they place a high burden of proof on these complex, multiple-system models (Zaki & Kleinschmidt, 2014). While the single-system view has some valid criticisms, it cannot entirely account for evidence to the contrary, either. Additionally, proponents of this theory are at a loss to describe the broader set of results from neuroscience and animal data (Poldrack & Foerde, 2008).

Competition between verbal and implicit systems (COVIS); a multiple-systems model

The COVIS model is a well-supported multiple-systems model which the proposed study is based on. COVIS attempted to incorporate modern behavioral, neurological, and computational modeling data into one coherent format. As suggested by its name, this model proposed that category learning was a COmpetition between Verbal and Implicit Systems (COVIS; Ashby et al., 1998; Ashby et al., 2011). For the studies proposed here, the implicit system within the COVIS model will be the primary focus.

COVIS has been experimentally tested primarily through the use of rule-based and information-integration tasks, which were described earlier. Ashby and colleagues argued that each of these tasks recruited primarily one of the COVIS mechanisms, with rule-based involving the verbal system, and information integration the implicit system. They proposed that performance on rule-based tasks was governed by a verbal, explicit system that relied on semantic knowledge and was under conscious control. Information-integration performance, they suggested, was conversely governed by a non-verbal, implicit knowledge system which utilizes procedural learning (Ashby et al., 1998), although the latter aspect has

endured some criticisms (Smith, 2008), and the nature of their explanation for the implicit system was not entirely refined at that point. The authors suggested that during category learning, performance should initially be dominated by the verbal system. However, the implicit system should eventually take over and begin to "automate" performance. In general, which system ends up being the most dominant should depend on the dimensions of the stimuli to be categorized: a stimulus with separable dimensions in which only one dimension is relevant should be easy to categorize using a unidimensional rule, hence, the explicit system should dominate. However, if the dimensions are inseparable and more than one dimension is relevant, it should be difficult, if not impossible, to develop a rule, leaving the implicit system to dominate (Ashby et al., 1998).

Ashby and colleagues furthermore proposed neural loci for the explicit and implicit systems, and presented behavioral data on category learning tasks from special populations, such as those with Parkinson's and Huntington's disease, as evidence of dissociations in performance; possibly related to specific insults to the verbal or implicit category learning system. They suggested that the verbal system relies on a network connecting the prefrontal cortex, head of the caudate, and the anterior cingulate, which allows for switching from ineffective rules and selecting the appropriate rule respectively. The implicit system, they suggested, functioned via an associative learning mechanism in which the tail of the caudate receives projections from visual areas and projects to premotor cortex; synaptic plasticity within the tail of the caudate then can result in a learned association between each stimulus and a particular response. This would, in essence, form a direct stimulus-response relationship (Ashby et al., 1998).

Dissociation studies: Feedback differences

Although COVIS has been supported by a wide variety of evidence from behavioral, neurological, and computational studies, the focus here will be on the studies that not only support the existence of multiple mechanisms, but that further characterize the nature of the implicit and explicit systems.

The first area that will be covered is the relationship between feedback and task performance. The COVIS model predicted that, due to the conscious nature of the explicit system, in rule based tasks the

form of feedback and how it is processed should not constrain task performance. In contrast the implicit system was predicted to rely, in part, on neurally constrained reinforcement learning; implying that the nature of feedback presentation in information integration tasks should be much more critical for learning. An early study found that rule-based learning was not impeded by the lack of feedback, compared to a severe decrement in performance due to lack of feedback in an information-integration task. Ashby, Queller, and Berretty (1999) investigated the ability of participants to perform simple (rule-based) or complex (information-integration) tasks with or without feedback. Without feedback, participants in the rule-based task were still able to perform with almost perfect accuracy, while those in the information-integration performed sub-optimally. Ashby, Maddox, and Bohil (2002) trained participants using either observational training, in which a stimulus and its label are presented simultaneously and no response is collected, or feedback training, in which participants must respond with their best guess of category membership and are given feedback afterwards. On the basis of predictions made by COVIS, they proposed that these two conditions would have differing effects on rule-based and information-integration systems. Rule-based systems, are based on working memory and executive attention, thus the timing and nature of the category membership information in both conditions should not matter. However, implicit systems involved in information-integration learning rely on dopaminergic reward signaling after a stimulus and response; if the label is presented alongside the stimulus as in the observational condition, without a response, these systems will not be recruited. Therefore, they hypothesized a deficit in performance in the observational training condition for the information-integration task, but not in the rule-based task. They found that participants performing the information-integration task in the observational training condition performed less accurately than all other groups, and were also more likely to use sub-optimal rule-based strategies compared to their counterparts in the feedback training condition, suggesting the implicit system was not utilized (Ashby et al., 2002).

The finding that that feedback in information-integration tasks is more effective when presented after the response raises the question of what amount of delay between response and feedback is most optimal for learning. As COVIS suggests that the implicit learning system is dopamine mediated, there

should be an optimal period of time after dopamine is released but before it disappears during which learning is most effective. A study by Maddox and colleagues (2003) varied the time between response and feedback in rule based and information integration category learning tasks (conditions: Immediate, 2.5 s, 5 s, 10 s). For the rule-based task there was no difference in performance related to the timing of the feedback. For the information-integration task, they found that performance was best in the immediate feedback condition compared with the 2.5 s, 5 s, and 10 s delays. The authors suggested that since 2.5 s was enough time to see a decrease in synaptic efficacy and a weakening of the reward response, the optimal amount of time might lie between a 0 and 2.5 second delay (Maddox et al., 2003). A later study (Worthy, Markman, & Maddox, 2013), hypothesized, based on neuroscience studies (Lindskog, Kim, Wikström, Blackwell, & Kotaleski, 2006) published after the Maddox et al. (2003) paper, that "learning is best when calcium (mediated by glutamate) and dopamine levels peak simultaneously, and that this is likely to occur when feedback is given 500 ms after a response has been made" (p. 292). Worthy and colleagues (2013) found that, in the rule-based task, there was no effect for feedback timing. In the information-integration task, they found that the optimal feedback delay was 500 milliseconds.

These studies on feedback timing seem to support the predictions made by COVIS in regards to the different mechanisms between explicit and implicit category learning systems; as the explicit system is conscious, feedback can be processed at will, whereas implicit systems learn more automatically and rely on more biological constraints.

Dissociation studies: Dual-task performance

Another area of research examining dissociations between rule-based and information-integration tasks is in the domain of dual-task performance. Another prediction by COVIS had to do with the nature of how each system relies on cognitive resources such as executive functions. The implicit mechanism in COVIS is independent of these resources, whereas the explicit mechanism relies on them for learning. In a dual-task study, participants must keep track of and respond to two different tasks with different performance goals. If tasks loading onto the explicit or implicit system were given the same dual-task, one which required the use of an executive function system, it was predicted that the explicit task would

suffer greater performance deficits. This is because the explicit system was proposed to rely on areas responsible for executive function, whereas the implicit system was proposed to rely more on procedural learning systems. Therefore, a simultaneous executive function task should affect the explicit task more so than the implicit task due to the competition for the same system's resources.

Waldron and Ashby (2001) taught participants to categorize geometric shapes varying in a binary fashion across shape, background color, shape color, and numerosity, using either a unidimensional rule (rule-based) or a complex, three-dimensional rule (information-integration). During a second session, participants had to learn new, additional rules, and also were required to perform a numerical analog of the Stroop task as a dual task. In this concurrent Stroop task, a number was presented on either side of the screen during presentation of the stimuli to be categorized. The numbers varied in physical size as well as numerical value, and after the participant responded to the categorization aspect of the task, they were asked "value" or "size," and had to indicate which was larger for whichever option was presented. They were also instructed to prioritize performance on the Stroop task, and to think about the categorization task as a secondary priority. The researchers suggested that, as they had to hold these value and size aspects of the numbers in working memory while categorizing, it should add to the difficulty of the task. What they found was that the concurrent Stroop task produced severe decrements in performance on the arguably easier explicit categorization task, while performance on the more complicated implicit task was largely spared. These results are contrary to a single-system model, which would predict performance becoming worse with a concurrent task the more complex the category structure becomes (Waldron & Ashby, 2001).

A study attempting to determine the generalizability of Waldron and Ashby's work was conducted using almost identical procedures (including the concurrent numerical Stroop task), except instead of stimuli varying along binary dimensions, they used stimuli which varied on continuous ones. In the experiment, which similarly compared performance on a rule-based versus information-integration task, they found results which were identical: poorer performance in the rule-based compared to the information-integration task during concurrent administration of the numerical Stroop task (Zeithamova

& Maddox, 2006). These dissociation studies of dual-task loading seem to support COVIS as well: specifically, the idea that explicit and implicit systems function using separate neural systems. Executive function tasks interfered with the explicit, declarative system proposed to be in use during rule-based tasks, while the implicit, procedural system seemed not to suffer serious performance decrements. Additionally, that these systems can function independently during the same task is a direct prediction of COVIS (Ashby et al., 1998).

Dissociation studies: Motor responses

Another area of dissociation research, and a very important one in the context of the proposed study, involves the differential effects of motor response manipulations. COVIS postulates that the implicit system involves the procedural memory system (Ashby et al., 1998), suggesting a much closer link between information integration categories learned via procedural systems and a motor response.

One of the first studies that attempted to investigate this procedural memory aspect of COVIS was conducted by Ashby, Ell, and Waldron (2003). There were three different response conditions that participants completed while they performed either a rule-based or information-integration task. The study was broken up into a training phase and a transfer phase, and there were three different possible conditions which were identical across both types of tasks. In the control condition, participants used their left and right hands, positioned on the left- and right-hand side of a keyboard, to respond as to whether a stimulus belonged to categories A or B respectively; the transfer phase was identical. In the hand-switch condition, participants used their right hand on the left-hand side of the keyboard and their right hand on the left-side of the keyboard during training, and they uncrossed them during transfer (similar to the control condition). The button-switch condition had participants begin the training phase identically to the control condition, but in the transfer phase, instead of changing hand positions, the category label that each button referred to was reversed. In this way, during the hand-switch condition, only the motor responses were reversed, whereas in the button-switch condition, the response locations as well as the motor responses reversed.

In the rule-based task, there was no interference across all experimental manipulations. The experimenters suggested this was due to the fact that participants learned abstract category labels mediated by an explicit, rule-based system, and as the explicit system does not involve a motor component they were able to easily adjust their responses. The results for the information-integration task, however, were quite different. The hand-switch condition exhibited a slight decrease in accuracy which was eventually recovered from, but the button-switch condition produced significant interference which did not decrease with practice. The authors suggested this was due to the fact that in the hand-switch condition, while the motor response was changed, the actual response keys remained the same. However, in the button-switch condition, they suggested that the performance deficit likely occurred due to the fact that in information-integration tasks, participants learn to execute a specific response location (category A, left-hand side of keyboard) more so than a specific motor response (category A, left hand) (Ashby et al., 2003).

A follow-up to this study was conducted by Maddox, Bohil, and Ing (2004). In their study, they attempted to provide further evidence of the procedural aspect of the implicit learning system. Participants were assigned to one of two conditions in a rule-based or information-integration task. In what they referred to as the "A-B" condition, stimuli identical to those presented in the Ashby et al. (2003) study were presented along with the query "Is this an A or B?" Participants pressed one key for category A, and another key for category B. The other condition was referred to as the "yes-no" condition. In this case, participants were either asked "Is this an A?" or "Is this a B?" and were then supposed to respond with one key for "yes" or another key for "no" with the idea being that in the yes-no condition, the response locations were constantly changing. They hypothesized that if there is a procedural element to information-integration tasks, then an inconsistent set of response mappings should prevent stimulus-response associations from being formed as effectively compared to having response locations consistently mapped. They found a decrease in performance for the yes-no condition relative to the A-B condition for the information-integration task that was not present in the rule-based tasks. They suggested this was further evidence that the explicit system does not require a consistent response

mapping; the only thing that participants need to learn is an abstract representation of category labels. This is in stark contrast to the results from the information-integration task, providing further evidence that these may be multiple-representational systems (Maddox et al., 2004).

While the previously mentioned evidence seems to suggest that a consistent response location is crucial for effective category learning, one study provides evidence which appears to contradict that idea. Ashby and colleagues (2003) provided evidence that changing response locations *after* learning severely disrupted performance, and while Maddox and colleagues (2004) seemed to present evidence that inconsistent response locations during training can be problematic, the nature of the yes-no component of the task led to some criticisms as to what systems exactly were being recruited. Therefore, Spiering and Ashby (2008) wanted to further investigate the effect of inconsistent response locations during performance of an information-integration task. In their first experiment, participants completed an information-integration in which the category labels were represented by either two circles of different colors or the letters A and B. In one condition, the circles remained in the same location, whereas in the other two, the location of the circles and letters varied randomly. They found that while the random locations started with worse performance and took longer to learn, the asymptotic difference between the consistent and inconsistent group performance was non-significant. They suggested that, in regards to the Ashby et al. (2003) paper, it was the blocked nature of the task which caused the interference when participants suddenly switched to inconsistent mappings. This was because they had been able to rely on both a spatial and feature association, and they suggested that the spatial association may be the more effective of the two. The participants in the Spiering and Ashby (2008) study never had a spatial association to rely on, so they were forced to learn the weaker, feature association instead.

In the second experiment reported by Spiering and Ashby (2008), they wished to examine the yes-no aspect of the Maddox and colleagues (2004) paper. They performed a similar version of that study, using an information-integration task, however instead of the prompt being “Is this an A?” or “Is this a B?” each side of the screen just had the words yes or no, which remained stationary, and instead at the bottom of the screen one of the two colored circles would be presented randomly with a question mark.

They reasoned that it could be either the extra logical decision or the inconsistent response locations for the yes-no keys in the Maddox et al. (2004) paper which caused such poor performance. In their study, then, they wanted to test the logical decision aspect only. What they found was that performance in the yes-no experiment was worse than the random location condition from their first experiment. They suggested that the logical decision of the yes-no task must have been the reason for the performance deficit in the study by Maddox and colleagues, not the varying response locations. They further reasoned that the yes-no task might recruit executive processes in neural regions which are poorly connected to where the implicit learning takes place; this communication problem would cause the observed performance deficit. Recall, also, that previously reported studies found that dual-task performance affected rule-based tasks more than information-integration tasks. While this seems to be contrary to their results, they argued that the difference between their task and others is that the yes-no task does not load as heavily onto working memory as the other dual-tasks which reported opposite behavioral patterns. In other words, the amount of working memory required for the yes-no condition is relatively light, which does not challenge the capacity required for the basic, rule-based category decisions (Spiering & Ashby, 2008).

Data from the serial reaction time (SRT) task literature provide some supporting evidence in favor of effector flexibility. In these tasks, sequences of stimuli are presented, and participants must respond to each stimulus with a specified motor effector (e.g. stimuli at four locations of on the screen are mapped to four different response buttons, each of which is pressed, with a separate finger). Reaction time in these tasks decreases for repeated sequences in comparison with random sequences, but without conscious awareness of the sequence by participants. SRT tasks are thought to utilize the implicit, procedural system, (especially the basal ganglia; Curran 1995; Wächter et al., 2009) similarly to information-integration tasks, and there has been some research looking at effector specificity in this domain. One study found that participants trained to responded using 3 separate fingers on the same hand were able to transfer their sequence knowledge when using one finger to press all three buttons which required them to use the arm as the effector rather than individual fingers (Cohen, Ivry, & Keele, 1990,

Experiment 2). Two additional SRT tasks found that effector programs could be "mirrored" and maintain the response time (RT) reduction for trained patterns versus random ones (Verwey & Clegg, Experiment 1; Verwey & Wright, 2004), which the authors of both studies suggested as evidence of implicit motor programs which can exist independently of or dependent on a specific effector (but still be separate representations). It is possible in information-integration tasks that the procedural system may acquire motor programs which are both effector dependent (e.g. the index finger from either hand), and effector independent (e.g. respond "A" to category "A" stimuli regardless of which effector is required). Another study using monkeys found that effector specificity transfer was a function of the amount of time that effector was used to perform a specific pattern, with less time spent in practice resulting in easier transfer (Rand et al., 2000). Given that the length of time it took to interfere with effector transfer was several days of practice, it is not improbable that the procedural system can be malleable with regards to which effector it requires to perform procedural tasks well.

All together, these papers provide strong evidence that the implicit system, specifically in the context of information-integration tasks, has a procedural memory component (Ashby et al., 2003, Spiering & Ashby, 2008). At the time COVIS was proposed, and later while these papers were being published, the involvement of the procedural memory system in implicit category learning was not entirely known, however, it is currently much more widely accepted (Cantwell, Crossley, & Ashby, 2015). However, recently it has been suggested that the initially hypothesized direct mapping of a stimulus to its associated response could be more complicated. Instead, it has been proposed that there is an additional component that mediates this linkage in the form of a category label association that links a stimulus to a response.

A two-stage representational model

The classic COVIS framework postulates that the explicit and implicit systems form different direct relationships between stimuli and other representations. In the explicit system, stimuli become associated with abstract labels via rules. In the implicit system, small regions of perceptual space surrounding each stimulus become associated with a motor response. Many other models, including

exemplar-based models such as ALCOVE (Kruschke, 1992), theorize a similar stimulus-response relationship. Recently, evidence has emerged that suggests that implicit category learning is more complicated than just the stimulus-response association postulated by the COVIS framework. This evidence suggests that, in-between the stimulus and response representations, there could be a mediating category label representation (Kruschke, 1996; see Figure 1 for a visualization of the one- and two-stage

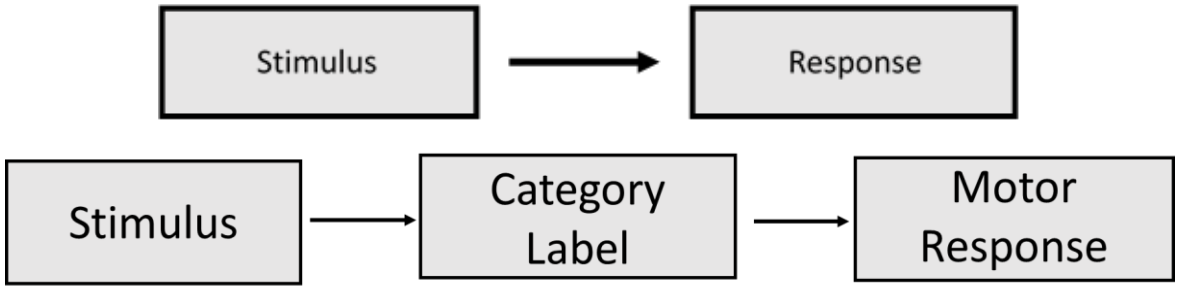

Figure 1: A diagram of the one-association, stimulus – motor response model, and the two-association, stimulus – category label – motor response model

association models). This means that instead of a direct stimulus-response relationship, an association is formed between a stimulus and a category label, and it is this category label that then becomes associated with the appropriate response.

An early theory including an intermediate category label was the AMBRY model (Kruschke, 1996), a variant of the ALCOVE model (Kruschke, 1992; AMBRY is not an acronym as ALCOVE is, instead, it is a play on words as an ambry is a special type of alcove). The AMBRY model postulated that exemplars in a category, rather than being individually mapped to a specific response, were instead first linked to a common category membership (Kruschke, 1996).

The possibility that a category label representation is formed when learning information integration tasks was investigated by Maddox, Glass, O'Brien, Filoteo, and Ashby (2010). They used a four-category task with three different conditions: a control condition, a category-switch condition, and a response-switch condition. Once participants had trained to a specific accuracy criterion, either the category labels changed or the responses used to indicate category membership changed, dependent on the assigned condition. Participants were told that the categories had changed, and were instructed to re-

learn the task using the trial-by-trial feedback. Maddox and colleagues hypothesized that if direct stimulus-response associations form the basis of information integration learning, then these two conditions should result in identical performance because the stimulus – response associations formed during training are broken in both conditions. They further hypothesized that if learning involves forming two associations, there might be differential effects on performance between the category label and response location conditions. They found that the category label group suffered a greater performance cost, but also experienced a faster recovery, than the response location group. These findings were consistent with the two-association hypothesis, but not with the classic COVIS model limited to a single stimulus-response association. Maddox and colleagues suggested that COVIS might be extended to account for these results, but that more neurological and behavioral evidence would be needed (Maddox et al., 2010).

Since this study, the two-stage associative model has become more accepted, and a formal model of COVIS incorporating these stages has been proposed. In this model, input from cortical visual association areas projects to the body and tail of the caudate nucleus, and then on to the pre-supplementary motor area (preSMA). Projections from the preSMA extend to the posterior putamen, which in turn projects to the supplementary motor area (SMA). Learning to associate a stimulus with the appropriate category label occurs on the path from cortical visual areas to the preSMA (via the tail of the caudate), and learning to associate a category label with the desired motor response takes place on the path from the preSMA to the SMA (via the putamen) (Cantwell et al., 2015). Learning occurs via synaptic plasticity driven by a dopamine mediated reinforcement signal at cortical-striatal synapses in the body and tail of the caudate (for the first stage), and in the posterior putamen (for the second stage) (Crossley, Ashby, & Maddox, 2014).

Three stage models

The two-stage model allows for stimuli to be assigned to a category with a common label, but does not account for the possible learning of a particular category structure, e.g., a prototype. Some early research studies posited that another mediating representational layer may exist between the stimulus and the category label (Kendler & Kendler, 1962; Sanders, 1971; see Figure 2). This "category

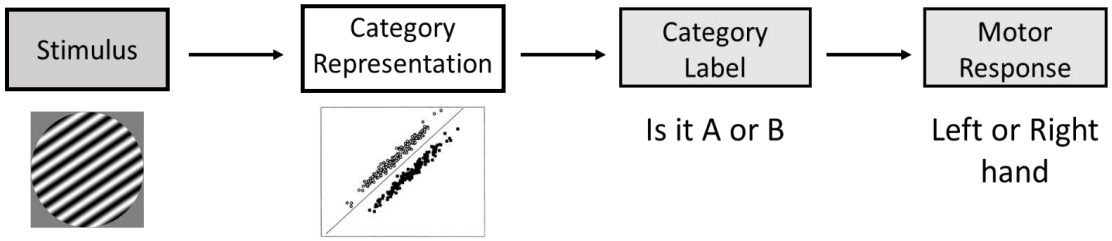

Figure 2: A diagram of the three-association model proposed by Wills

representation" could exist as an abstract conceptualization of the category structure as a whole; it would maintain the features that relate individual stimuli within categories.

A method was developed by Wills, Noury, Moberly, and Newport (2006) to test whether such a mediating category representation exists through a manipulation involving the category label – motor response association. They hypothesized that if a unique category representation is formed for each encountered categorization problem, and learning is manipulated for a subset of category members, such as reversing the category labels, then participants should later extend the learned manipulation to the remaining category members as well. This is because, in the context of this reversal example, reversal of the category labels will cause an alteration to the association between the category representation and its category label. However, the association between the stimuli and the category representation remains unchanged. When the stimuli which were not subjected to response reversal are presented later, they still will activate the same category representation, but now this representation has been linked to a new category label, and subsequently, to a different motor response.

Wills and colleagues (2006, experiment 2) tested this hypothesis by training participants in their study on two separate "family resemblance" categorization problems using a single set of category labels.

All the stimuli and features were unique to an individual category problem, but across the two problems the same labels for the alternative categories were used (A and B). Once they reached a certain accuracy criterion, one of the categorization problems was selected, and participants were trained to reverse the category labels for only a subset of stimuli from both categories. In a final testing phase, all stimuli from both categorization problems were presented. They found that the reversal of the label-response relationship for a subset of stimuli affected the label-response relationship for the entire categorization problem; participants applied the reversal to all of the stimuli from the categorization problem, even those stimuli they had not been trained to reverse. They interpreted this result as indicating that the relationship of the category representation to the category label had been altered in the reversal phase. When the remaining stimuli which had not been presented during the reversal phase appeared, they still activated the same category representation, however, now the representation was associated with the opposite category label, which caused a reversed response.

Overview of studies

The Wills et al. (2006) study provides evidence for an additional category representational layer between stimuli and category labels. However, they used a family resemblance category learning task, and it is unclear whether a similar mediating representational layer is formed when learning information integration tasks (Maddox et al., 2004). Although there is an array of supporting evidence for a two-stage association model for implicit information integration learning, it is unclear whether this two-stage model can fully account for learning, and whether an abstract category representation might be learned as well. I used the method developed by Wills et al. (2006) to test whether an abstract category representation is formed during information-integration category learning. In Study 1, participants learned one information integration category structure with two sets of labels for the individual categories. One set of labels was then trained in reverse, and a final testing phase investigated if they learned one category representation for both sets of labels by examining reversal behavior for the label set that was not trained in reverse. Study 2 trained participants on one information integration category structure with one set of labels. A

spatially defined subset of each category was then trained in reverse, and a final testing phase examined reversal behavior for all stimuli, including those that were not trained in reverse.

CHAPTER 2: STUDY 1

Introduction

The goal of this study was to investigate whether or not an intermediate category representation between stimuli and category labels is learned in information-integration tasks. The design was based on the one used in the Wills et al. (2006) paper in which participants learned two category learning tasks, learned to reverse a subset of stimuli from one category, and then completed a final task in the absence of feedback testing to examine whether or not they extended the reversal to untrained stimuli. Participants were told they would learn via trial-by-trial feedback to categorize stimuli for two different categorization problems using two sets of category labels. In reality, the two categorization problems shared the same stimuli, and there were merely two sets of labels assigned to the same stimulus distributions. After reaching the accuracy criterion on label set 1 and on label set 2, the entire set of stimuli belonging to label set 1 was reversed, and participants again trained to the same accuracy criterion. In the final phase, stimuli with either set of labels were presented intermixed with each other, and feedback was no longer given. My hypothesis was that, if mediating category representations are acquired in information integration learning, that participants would learn a single representation of the category and would learn to assign both label sets to this category representation. Reversing one label set should therefore lead to participants reversing the other label set, because the category representation-category label was changed. Alternatively, if participants did not show reversal of the second label set, that would suggest several alternative possibilities: that completely independent category representations and/or label associations were formed for each label set such that reversal of one set did not affect the other, or that no category representation was formed.

Methods

Participants

Participants were recruited from the PSY100 and PSY250 research pool. Each student in those courses is required to participate in research studies for class credit. In total, 98 students participated,

however, only 58 students completed the task. Initially, the task was scheduled for one hour, but a large proportion could not complete the task in that amount of time; extending the time to two hours reduced attrition to reasonable levels.

Stimuli

The task was presented to participants using Psychtoolbox (Brainard, 1997; Pelli, 1997; Kleiner et al, 2007). This set of programming tools functions as a Matlab extension. Stimuli were Gabor patches: circular sine-wave gradients which vary in terms of the bar rotation (orientation) and bar width (spatial frequency). These were generated by first defining a point in perceptual space to serve as the base for generating other stimuli (black dots on Figure 3; both clusters: mean Y = 225, SD Y = 20, SD X = 14; Category A: mean X = 260; Category B: mean X = 440) within an arbitrary 0:700 space (wherein the initial arbitrary values of the *x* and *y* axis had been transformed into orientation and frequency values respectively) which had been rotated 45º. A y-range of approximately 425 "units", split evenly around both sides of the black dot, was used to generate each set of category stimuli. Approximately 1000 stimuli were generated by sampling randomly from a bivariate normal distribution, which was constrained by the mean and y-range specified. The resulting stimulus distributions for each category are shown in Figure 3. On each trial, a randomly sampled stimulus was presented on the computer screen. Each stimulus was approximately 4 cm in diameter, and subtended a visual angle of 30º on average. We did not control visual angle for each participant by fixing head position, so this value likely differed between participants and changed across the experiment due to factors such as shifting posture and chair distance from the screen.

Procedure

Participants were told by the experimenter that they would be participating in a visual categorization task, and that the goal was to learn to categorize stimuli as accurately as possible. They

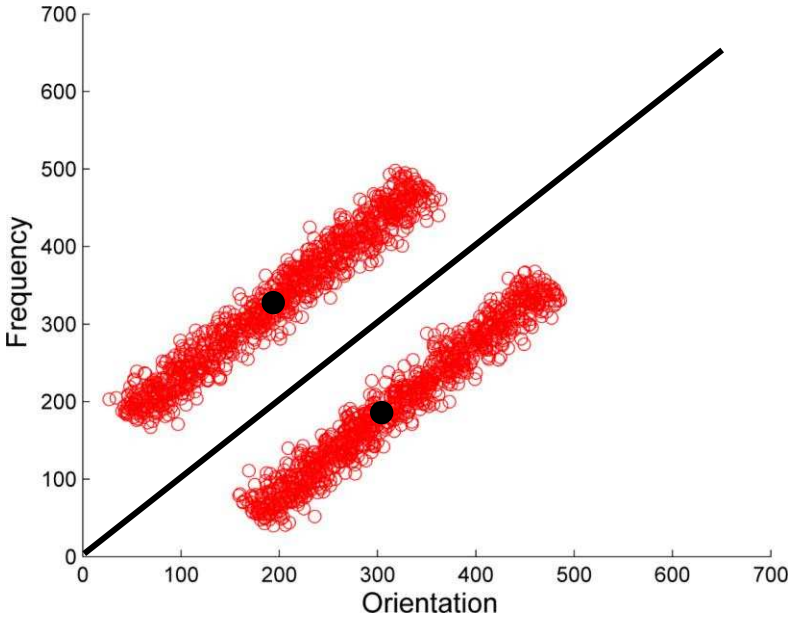

Figure 3: The perceptual space from which the stimuli were sampled. Black dots denote the center point from which stimuli were generated, black line represents decision bound.

were also told they would have to learn the task via trial and error, but that they would receive feedback on their responses. Participants were instructed to categorize stimuli into two categories; each category would have a category label (e.g., R and I). They were further told that at first the labels would always be on the same location on the screen and that they should select the desired category by pressing the button on the corresponding side of the screen with their left or right hand. After reaching a certain accuracy criterion for a number of blocks, the labels on the bottom of the screen indicating category membership would begin alternating locations; for example, for the category labels R and I, sometimes the R label would be in the lower left hand corner, and sometimes in the lower right hand corner. They would then have to continue to respond by pressing the button (right or left hand) corresponding with the side of the screen that the category label appeared on.

The task began with an instruction screen reiterating the verbal instructions, and then the training portion of the task began. Participants trained first with one set of category labels and then with the second set. The two category label sets were R-I and E-O and were assigned to the first or second learning task in a counter-balanced fashion. Training continued on each label set until they reached 80%

accuracy for five 30-trial blocks. Each stimulus was presented for two seconds or until the participant responded; this was followed by a half-second of a blank screen, then feedback was presented for another

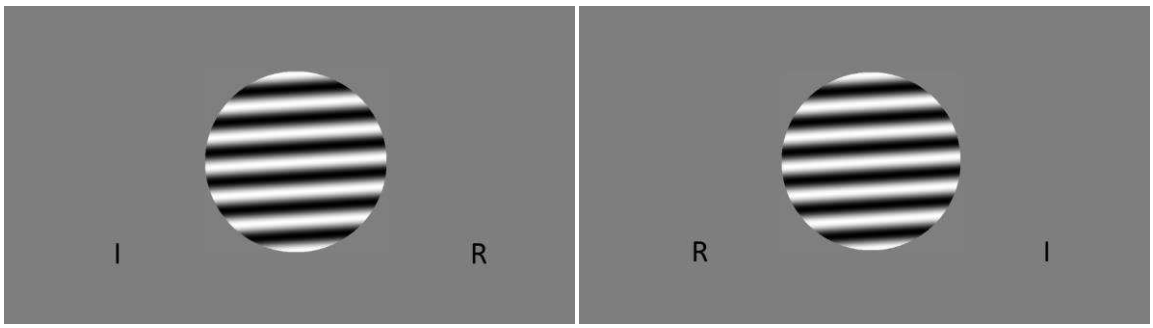

Figure 4: An example of alternating category label positions

half-second. The feedback indicated whether they were correct, incorrect, or had been too slow to respond. After reaching the accuracy criterion two times, the labels began alternating between lower left and lower right corners in a pseudorandom sequence and continued to do so until they completed the training block for that set of labels (See Figure 4). Afterwards, the other set of labels was presented, and the task progressed identically until they reached the accuracy criterion five more times.

Afterwards, the reversal phase began in which the first set of labels they had been presented with during training was now selected to be trained in reverse; the correct responses for each category were switched. First, an instruction screen was displayed which told them that they were entering a new portion of the task and that they should continue to categorize stimuli as accurately as possible. The nature of the change and the presence of the reversal was not disclosed. The reversal phase proceeded just like the training phase; participants had to complete five 30-trial blocks with 80% accuracy.

After completing the reversal phase, the final test portion of the task began. The procedure for this section was different from the training tasks. Instead of 30-trial blocks, 300 trials were presented back-to-back without feedback and without breaks. Both sets of category labels were used and trials with the different label sets were intermixed randomly. Labels from different sets were not mixed within a single trial; the options were still either R-I or E-O. At the end of this block, the task was completed, and instructions displayed telling participants they had finished.

Results

It took participants 36 blocks on average to complete the initial training (range: 16-83), and 10 blocks to complete the reversal training (range: 5-24), collapsed across both sets of labels; there were no significant differences in performance as a function of which category label set was reversed. For the transfer phase, the dependent variable was the percentage of stimuli categorized as belonging to category R or E (equivalent labels) within each of the originally trained categories R/E and I/O. This measure was chosen to avoid the ambiguity inherent in judging which categorization choice is correct (in accordance with original training, or the reversal training). For each subject, proportion of R/E responses was calculated separately for stimuli from originally trained categories R/E and I/O (factor 1; category), and for stimuli from labels in phase 2 that subjects trained to reverse [trained stimuli], and labels that were not trained to be reversed [transfer stimuli]; (factor 2; training). A 2x2x2 ANOVA was conducted on factors 1 and 2, with the RI and EO reversal conditions as a third, between-subjects factor (see Figures 5 and 6 respectively). There was an interaction effect for all factors($F(1,48) = 7.58$, $p < .01$), with trained stimuli receiving a significantly higher proportion of reversed responses than transfer stimuli in both label reversal conditions. Post-hoc tests (see Table 1), indicated that trained

Table 1

Pair-wise comparison, by condition, for category labels trained in reverse

Label Reversal	Mean Difference	SE	*p*-value	95% CI
RI	-24.71	8.38	0.005*	-41.56 | -7.87
EO	-50.27	8.38	0.0001*	-67.11 | -33.42

Pair-wise comparison, by condition, for transfer stimuli

Label Reversal	Mean Difference	SE	*p*-value	95% CI
RI	28.85	6.68	0.0001*	15.41 | 42.82
EO	48.15	6.68	0.0001*	34.72 | 61.58

Note: mean difference is calculated by subtracting the percent endorsement for categories R/E when presented with category R/E stimuli from the percent endorsement of category R/E for category I/O stimuli. Negative values indicate reversal, positive values indicate maintenance of original category

* Significant at 0.05 level

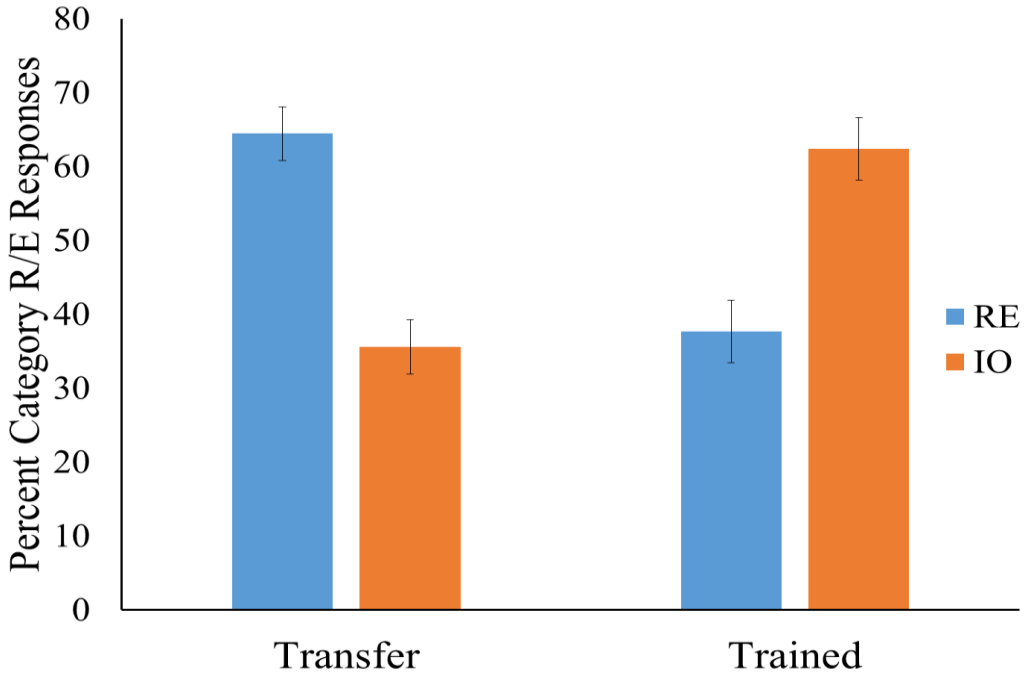

Figure 5: Mean percent category R/E responses for trained and transfer stimuli when the RI category labels are reversed

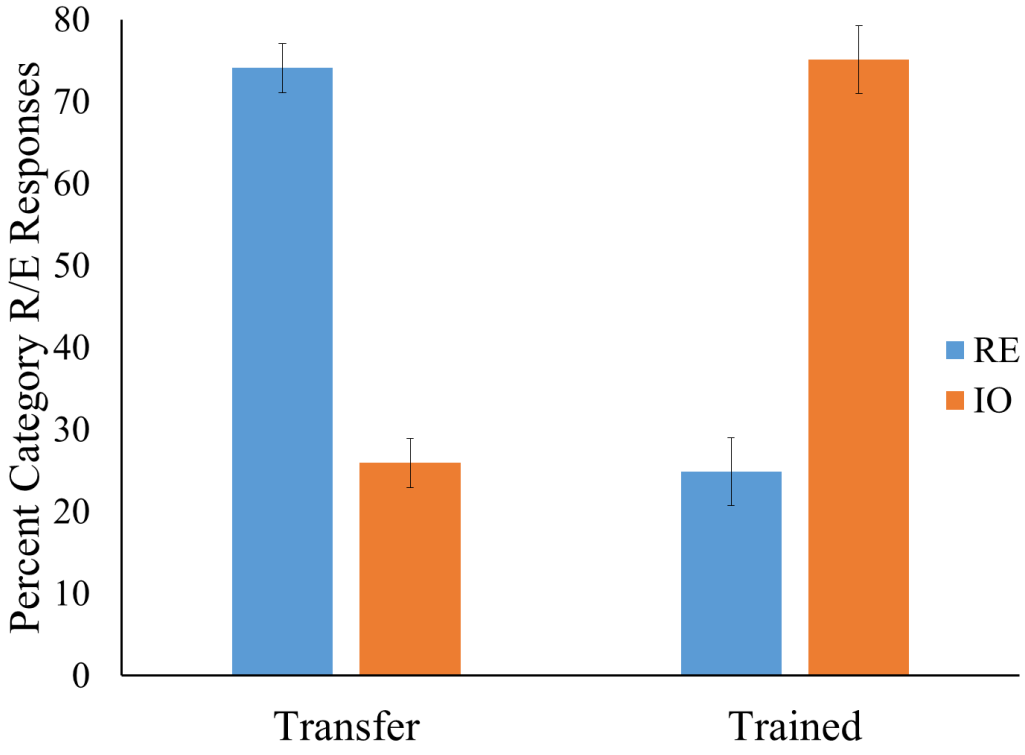

Figure 6: Mean percent category R/E responses for trained and transfer stimuli when the EO category labels are reversed

stimuli continued to receive reversed responses, while transfer stimuli received responses in-line with how they were originally trained. These data showed a non-significant amount of reversal transfer, consistent across both label reversal conditions.

Individual participant performance was examined, to ensure there was no masking of reversal behavior through group analysis. Only 2 participants were identified who appeared to extend the reversal to the transfer stimuli (one from each label reversal condition).

Discussion

My hypothesis that reversal training for one set of labels would transfer to the non-reversed label sets was not supported. Instead, participants continued to respond to each set of labels consistent with how they had most recently been trained; the label set trained with reversed responses continued to receive reversed responses, and the label set that was not subjected to reversal training maintained the responses consistent with initial training. These results indicated that participants may not have formed a single category representation. Additionally, accuracy for the transfer phase dropped below the criterion they had been trained to, although why this occurred is unknown. It is possible that the removal of feedback led to an overall decrease in accuracy due to a lack of any sort of response monitoring system for performance. Furthermore, if they treated each set of category labels as separate categorization problems, some sort of extra-logical steps may have also been implemented. These steps could be in response to having to transition between label sets, in combination with having to account for trail-by-trial alternations of response locations. There could be several possible reasons for the lack of reversal of the second label set. One possibility is that, even though the stimuli for both sets of labels were identical, participants could have learned two separate category representations, and therefore each label set could have a separate category representation – category label association. In that case, when one set of labels were reversed, it did not change the association for the other set because they had separate associative links. Another explanation could be that participants did learn a single, shared representation, but they may have learned multiple separate category label associations for the category representation. Therefore, reversing one set only altered that particular category representation – category label association. A third

explanation is that there could have been a shared representation learned during initial training, but that reversal may have led the to the acquisition of a new category representation for the labels presented during the reversal training due to participants partitioning each portion of the task separately. Finally, there may not be a category representation formed in information-integration tasks. Without an intermediate category representation, reversing one set of labels had no effect on the other because the stimulus – category label associations for each set of labels are entirely separate.

Another possibility is that subjects may not have completed enough training to solidify a category representation. Kruschke (1996), in their paper describing the AMBRY model, suggested that training strengthens a category representation. It is possible that the 5 blocks necessary to progress through each segment of the task were insufficient. In the Wills et al. (2006; Experiment 2) study, participants took an average of 27 blocks to reach criterion in the initial training phase, and an average of 3 blocks in the partial reversal phase. However, the average number of blocks for training and reversal in Study 1 was 36 and 10 blocks respectively.

Alternatively, participants may not have had enough reversal practice to significantly alter the category representation, if there was one shared between both sets of labels, and may have instead responded in the transfer phase based on however they had most recently been trained for a given label set. It could therefore have been a combination of difficulty and length of practice that precluded the formation of a strong enough category representation during the reversal phase. Finally, it is possible that the difference in results is due to differences in the category learning task itself, as the Wills et al. (2006) study used a family resemblance task.

The difference in results between Study 1 and the study by Wills and colleagues (2006) could also be due, in part, to the difference in methods. Their study used two unique sets of stimuli within the category, and only trained participants on a partial reversal for one set of stimuli, whereas this study used two sets of labels, applied to a common set of stimuli, and participants were trained using a full reversal for all stimuli with one set of category labels. With that in mind, Study 2 mirrored much more directly the Wills et al. (2006) study by examining a partial reversal of a subset of the stimuli within the category.

CHAPTER 3: STUDY 2

In Experiment 1, I tested whether a category representation stage might be formed when learning information integration tasks and be shared between two sets of category labels. In this study, I used a method which was simpler, in that it used a single set of category labels, which is more similar to the methods used by the Wills et al. (2006) study. This study examined if reversing the category labels for a subset of stimuli would transfer to the remaining category members. If participants transferred the reversal to untrained stimuli, it would suggest that a single category representation had been learned, and that the category label association with the category representation had been altered. Participants were trained on one information integration categorization problem with a single set of category labels. After they reached a predetermined accuracy criterion, one half of the stimuli within each category (clustered together in perceptual space, see Figure 8) received reversal training. In the final phase, all stimuli were presented again without feedback, resulting in two types of stimuli: those that had received reversal training, and those that had not.

In addition to manipulating stimulus-label reversal, I also manipulated the consistency of the label-response mappings. In the Consistent condition, the category labels on the bottom of the screen indicating the appropriate button press response remained on the same side on every trial, creating a consistent category label- motor response relationship. In the Inconsistent condition, the labels alternated sides in a pseudorandom sequence as in Study 1. I hypothesized that this manipulation might affect the degree to which subjects learned category labels and formed an accompanying abstract category representation. In the Inconsistent mapping condition, subjects cannot perform the task without learning the category labels; this condition at the least forces learning of a category label representation, and may provide the best condition within which to identify category representation formation. However, under Consistent mapping conditions, participants may ignore the labels and instead learn direct stimulus-response relationship, effectively bypassing learning category labels and not forming abstract category representation.

The primary hypothesis was that abstract category representations are developed during information integration tasks. The primary prediction was that reversal of a subset of category label – category mappings for a subset of category members would be extended to the remaining members.

The secondary hypothesis was that the consistency of category label – response mapping might modulate learning of the category representation and category label – category mappings. The COVIS theory (Ashby et al., 1998), based on data from experiments using consistent response mappings, found that learning of information integration categories is based on direct stimulus-response relationships which would preclude developing both a mediating category representation and a category label association. If participants in the consistent group only learn direct stimulus-response relationships, there should be no reversal during the testing phase for transfer stimuli. If reversal is found in the consistent mapping condition, that would imply that subjects did learn category labels and an abstract category representation, in opposition to the COVIS model of category learning. Unlike consistent response mappings, training with inconsistent response mappings forces subjects to learn the category labels. If learning category labels increased abstract category representation learning it should result in greater transfer of reversal in inconsistent mapping conditions. In contrast, if there is a similar amount of transfer of reversal for untrained stimuli in both mapping conditions, it would suggest that each group learned all associative stages equally.

An exploratory hypothesis was that participants might utilize different category learning models, which would predict that the reversal manipulation might h*a*ve different effects (see Figure 7 for the three models of interest). For instance, if participants only learn direct stimulus – response relationships, the reversal phase would require relearning of every stimulus – response association. During the transfer phase, reversal should only occur for stimuli trained in reverse, since the stimulus – response association for the transfer stimuli has not been altered. Furthermore, both training phases should take longer for participants in the Inconsistent mapping group, due to the inability of participants to associate a stimulus

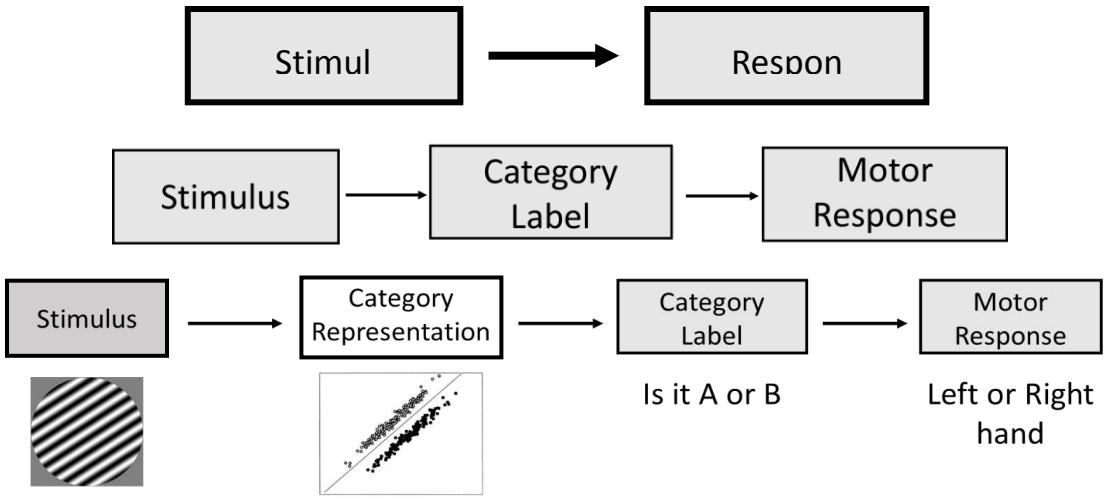

Figure 7: Diagrams of each separate category learning model. From top to bottom: 1-stage, with a consistent response (see Ashby et al., 2003 and Maddox et al. 2004 for examples of this effect). For the two-stage model, during reversal training, only the stimulus – category label association would be changed, leaving the category label – motor response association intact. Since the stimulus – label association is separate from the label – response association, only the stimuli which have had their stimulus – label association reversed should continue that response pattern in the transfer phase. There should be a similar decrement in performance for the Inconsistent group, again, due to the inability to associate a category label with a motor response. Finally, the predictions for the three-stage model are as mentioned above: that the reversal training should extend to the transfer stimuli, and that this effect should be greatest for participants in the Inconsistent mapping group.

Methods

Participants

Participants were recruited from the PSY100 and PSY250 research pool. Each student in these courses is required to participate in research studies for class credit. An a priori power analysis using G*Power 3.1.9.2 (Faul, Erdfelder, Lang, & Buchner, 2007), allowing for a moderate effect size (f=.25; based on data from Cantwell et al., 2015) with 0.95 power, suggested 36 participants per mapping condition.

Stimuli

The stimuli were identical to those from study 1, except for a few changes to the stimulus space parameters: the distance from the bound was reduced (Category A: mean X = 270, Category B: mean X = 430), the range of perceptual space sampled was increased from 425 to 624, and the number of stimuli generated was doubled (from 1000 to 2000; see Figure 8). During the partial reversal, one half of each cluster (top or bottom, counterbalanced) was selected for retraining (see boxes on Figure 8 for example)

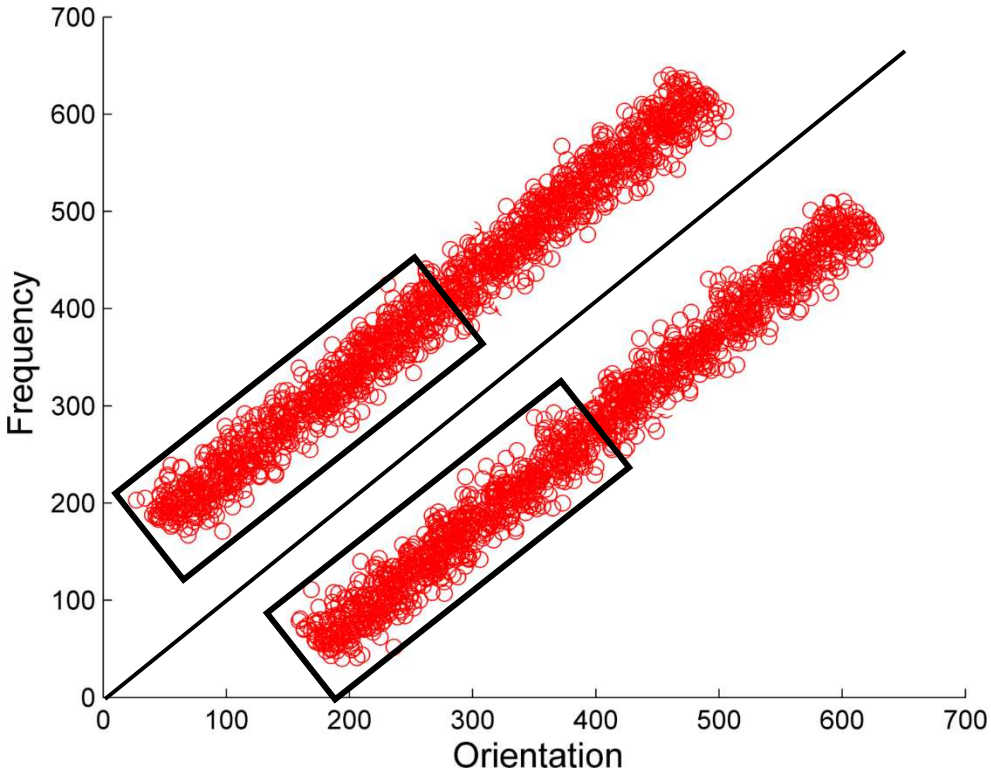

Figure 8: Proposed sampling space of stimuli. Black boxes represent areas of perceptual space to be trained in reverse during reversal phase

Procedure

Participants were told that they were participating in a visual categorization task, and that their goal was to perform as accurately as possible. They were also told that they will have to learn through trial and error, but that it is possible to perform the task with a high degree of accuracy. Details were given on the response keys they should use, and for the inconsistent group, the alternating-response nature

of the task was elaborated. Participants were assigned to either use consistent or inconsistent response mappings throughout the task; these mappings did not change.

After an initial instruction screen, which reiterated the verbal instructions, participants began the training portion of the task. Each trial proceeded like this: a stimulus was presented in the center of the screen, with the two category labels presented beneath it and nearer to the edge of the screen. Participants had two seconds to respond, and after a half-second delay, feedback was presented for a half-second indicating whether they were: "correct," "incorrect," or "too slow." This section was divided into thirty-trial blocks, and at the end of each block, the participants were told whether or not they met the 80% accuracy criterion. During this time, they were also able to take a break, as the task only began again once they hit a response key. Once they reached this criterion ten times, the training phase ended.

In the reversal phase, a sub-set of stimuli from each category (refer to black boxes in Figure 8) was presented again, however, the category labels for the stimuli were reversed. Whether the top or bottom half of the stimuli space was sampled was counter-balanced across participants, resulting in two separate groups for each mapping condition. As in the training phase, stimuli were presented in 30 trial blocks, and at the end of each block, accuracy was assessed. Once subjects reached the 80% accuracy criterion on ten blocks, they moved on to the final phase. In this final, transfer phase, stimuli were drawn from the full distribution, including both regions that underwent reversal, and regions that were not reversed. As in Study 1, feedback was longer given, and stimuli were also not presented in blocks. Participants completed 300 trials in this manner.

Results

Overall, 239 subjects were recruited, of which 87 had complete data which could be analyzed. 128 subjects failed to complete the task within the two hours allotted, and data from an additional 24 subjects were lost due to technical problems. 40 subjects were in the consistent mapping group (22 with top quadrant reversed, ConTop, 18 with bottom quadrant reversed, ConBot), and 47 were in the inconsistent mapping group (24 with top quadrant reversed, IncTop, and 23 with bottom quadrant

reversed, IncBot). The average number of blocks to complete each training phase for each condition is listed in Table 2. As can be seen, the range of blocks necessary to meet criterion in each training phase

Table 2
Number of blocks to reach accuracy criterion in each training phase

Training			
Group	Average blocks	SD	Range
ConTop	47	18.7	17-84
ConBot	52.2	15.8	21-81
IncTop	45.6	13.2	22-76
IncBot	53.6	16.3	28-83
Reversal Training			
Group	Average blocks	SD	Range
ConTop	17	8	11-46
ConBot	17.5	9.3	11-47
IncTop	15.3	8.2	10-51
IncBot	15	3.6	10-24

varied greatly between participants. There were, however, no significant group differences between blocks to criterion for either training phase.

For the analysis of the transfer phase (phase 3), the dependent variable was the percentage of stimuli categorized as belonging to category A within each of the originally trained categories A and B. This measure was chosen to avoid the ambiguity inherent in judging which categorization choice is correct (in accordance with original training, or the reversal training). For each subject, proportion of A responses was calculated separately for stimuli from originally trained categories A and B (factor 1), and for stimuli from regions in phase 2 that subjects trained to reverse [trained stimuli], and regions that were not trained to be reversed [transfer stimuli]; (factor 2). Data were collapsed across reversal region (e.g. top quadrant vs bottom quadrant) since initial examination of the data indicated that the results for each reversal condition did not differ. Separate 2x2 ANOVAs were conducted on factors 1 and 2 (within-subjects) for the Consistent and Inconsistent response mapping condition (between-subjects; see Figures 9 and 10 respectively). For both mapping conditions, there was a significant interaction effect

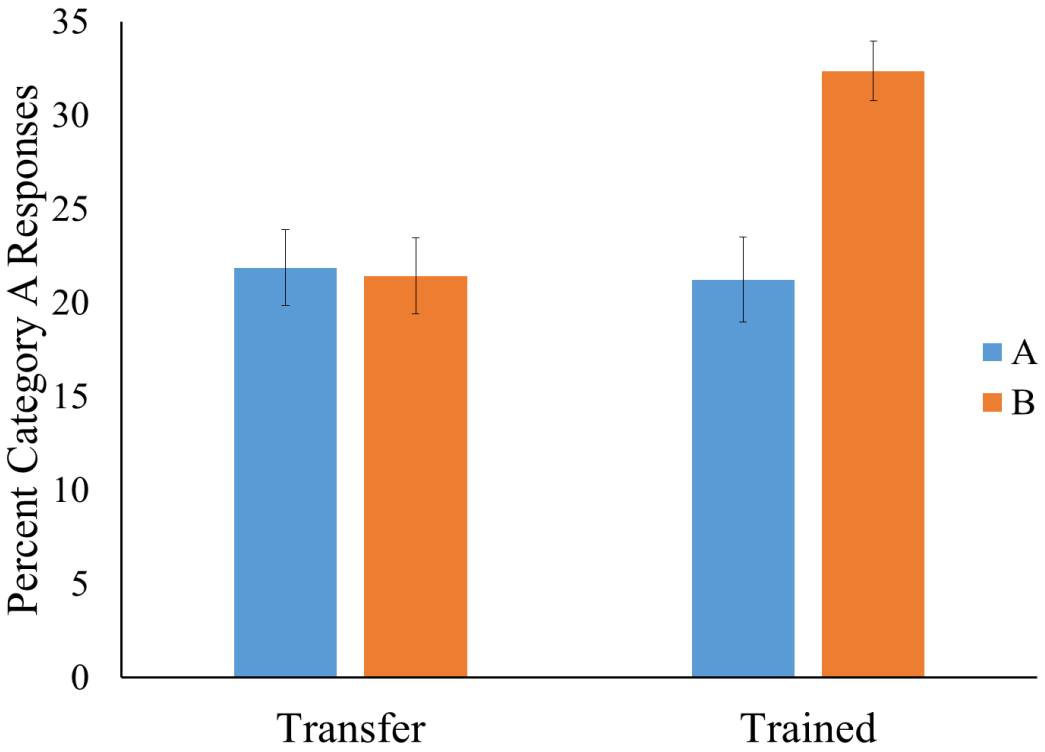

Figure 9: mean percent category A responses in the Consistent mapping group for Transfer and Trained stimuli

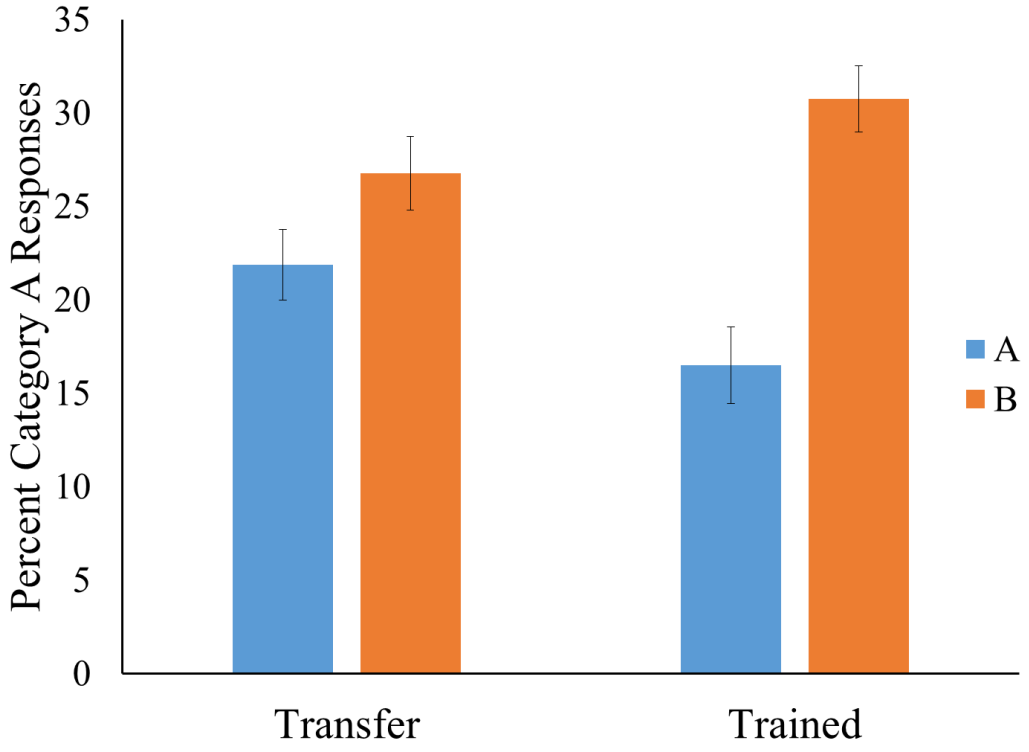

Figure 10: mean percent category A responses in the Inconsistent mapping group for Transfer and Trained stimuli

(Consistent, $F(1,39) = 21.78$, $p<.001$; Inconsistent, $F(1,46) = 26.38$, $p<.001$), with a higher proportion of stimuli from category B being categorized as members of category A for trained stimuli than for transfer stimuli.

There was a main effect for training only in the Consistent mapping condition, $F(1,39) = 4.68$, $p<.05$. The main effects for category were significant in both response mapping conditions (Consistent, $F(1,39) = 4.56$, $p<.05$; Inconsistent, $F(1,46) = 10.92$, $p<.01$), with both groups responding to category B with an A response significantly more often than responding to category A with an A response. Post-hoc tests (see Table 3), indicated that the only pairwise significant differences were within the trained

Table 3

Pair-wise comparison by category, for stimuli trained in reverse

Mapping Group	Mean Difference	SE	p-value	95% CI
Consistent	-11.14	2.9	0.001*	-17 | -5.28
Inconsistent	-14.24	3.26	0.0001*	-20.79 | -7.69

Pair-wise comparison by category, for transfer stimuli

Mapping Group	Mean Difference	SE	p-value	95% CI
Consistent	0.44	2.56	0.865	-4.75 | 5.63
Inconsistent	-4.88	2.79	0.087	-10.51 | .743

Note: mean difference is calculated by subtracting the percent endorsement for category A when presented with category A stimuli from the percent endorsement of category A for category B stimuli. Negative values indicate reversal, positive values indicate maintenance of original category

* Significant at 0.05 level

conditions. This suggests that, on average, both Consistent and Inconsistent groups maintained the reversal training they experienced prior to the final transfer phase, but that extension of the reversal to the transfer region was not reliable within each group. Across all three phases of the task, there was a significant difference in response time between the Consistent and Inconsistent mapping groups (phase 1: $F(1,83) = 50.32$, $p < .001$; phase 2: $F(1,83) = 40.68$, $p < .001$; phase 3: $F(1,83) = 49.2$, $p < .001$), with those in the Consistent group responding about 180 ms faster on average (this was relatively constant across all 3 phases). This is not surprising, given that those with Inconsistent mappings had to take a brief amount of time each trial to ascertain which side of the screen each category label was positioned in.

The large amount of variability in the group analyses suggested that there may be significant individual differences in strategy. I examined each individual's performance and identified four qualitatively different patterns of reversal responses (see Table 4). One group, Reversal Transfer,

Table 4
Reversal strategies

	Consistent Mapping	**Inconsistent Mapping**
Strategy	*n*	*n*
Reversal transfer	14	27
Reversal training only	12	7
Response reversion	5	8
Indeterminate strategies	9	5

Note: reversal transfer represents a strong transfer of reversed responses to the untrained category, partial reversal represents a moderate transfer, reversal training is for participants who continued to reverse the stimuli they had just been trained to reverse, but did not effectively transfer to the untrained stimuli, response reversion is participants who seemingly ignored the reversal training they had just completed, and those with indeterminate strategies were either guessing or using some unknown strategy

continued to consistently respond with the reversed labels on the trained stimuli and completely transferred the reversal training to the transfer stimuli; subjects in this group met a criterion of at least a 60% reversal responses in both trained and transfer regions. These reversal patterns are in line with the hypothesis that if subjects learn a category representation, any change in the relationship between the representation and category labels should alter the relationship for other stimuli that belong to the same category representation.

The second group of participants, Reversal Training Only, continued to reverse the trained stimuli, but did not reverse transfer stimuli, as evidenced by a skewed reversal percent in favor of the trained stimuli (at least a 20% difference in reversal rate between trained and transfer). This group's responses are not consistent with the hypothesis, which stated that if a single category representation was learned, reversal of part of the stimuli the should be extended to all. A third group of participants, Response Reversion, reverted to their original training, seemingly ignoring the reversal training they had

just completed, as judged by a less than 35% reversal response rate on both the trained and transfer stimuli. Finally, the fourth group of participants appeared to either be using no strategy, or one which was not obvious. They had a reversal rate between 35% and 60%, suggesting that they may have been guessing, or not attempting to respond accurately.

While the number of participants who reverted to responding based on their original training, as well as those with indeterminate strategies, was similar across mapping conditions, the other strategies had a greater variety of endorsement. There were almost twice as many participants who fell into the "Reversal transfer" group in the Inconsistent relative to the Consistent mapping group, suggesting that something about how they learned the task was more conducive to extending the learned category representation – category label reversal. Furthermore, fewer participants in the Inconsistent group maintained only the reversal training ("Reversal training only") relative to the Consistent group. Taken together, these data seem to suggest that something about how the Inconsistent mapping group had to learn the task allowed for easier transfer of the reversed associations relative to the Consistent group.

Discussion

Overall, the results indicate that for trained stimuli, both Consistent and Inconsistent mapping groups maintained the reversal training they had just experienced, as evidenced by a greater proportion of responses endorsing category A when presented with category B stimuli, relative to category A, for trained stimuli. However, the group results for the transfer phase showed no overall significant reversal or maintenance of original category membership. This pattern may be due to subsets of subjects using different strategies during the transfer phase. The most relevant strategy relates to the initial hypothesis: that the reversal of a sub-set of each category would alter the category representation – category label association for the untrained, transfer stimuli (i.e. "reversal transfer" strategy), with this transference manifesting behaviorally as a reversal of category membership. Within the consistent/inconsistent mapping groups, this was the most common strategy (40% and 66% of all participants for each mapping group respectively). These participants responded to the transfer stimuli at a similar rate as the trained stimuli (and at a rate mostly approaching the 80% training criterion; approximately 70% reversals for

both categories on average, for each mapping condition), suggesting that the reversal training did alter the category label – category associations for all stimuli, trained and untrained both. That this strategy was the most common (54% of all participants, across conditions) directly supports the hypothesis that reversing a sub-set responses for each category would alter the category representation – category label association for the remaining stimuli.

Furthermore, the much higher rate of the Reversal Transfer strategy within the Inconsistent mapping condition provides evidence in favor of the secondary, exploratory hypothesis that consistency of response mapping might modulate degree of transfer. Specifically, the lower rate of transfer in the Consistent mapping condition may be due to subjects learning direct stimulus – response associations rather than learning the category labels or creating a separate category representation. Alternatively, the consistent mapping condition could also be accounted for by the creation of a separate "category representation" for the reversed sub-set during reversal training, which did not extend to the entire set of stimuli. Subjects in the Inconsistent group, consistent with the secondary hypothesis, were required to learn the category labels as they could not rely on a stimulus – response relationship only. Focus on the category labels may have facilitated learning an abstract category representation linked to each category label.

Further supporting the hypothesis that Inconsistent mapping would facilitate learning of a abstract category representation, the inconsistent mapping group also had a lower proportion of participants who reversed only the trained stimuli, relative to the consistent mapping group (15% versus 30%). This may indicate that those in the inconsistent mapping condition formed a stronger abstract category representation than in the consistent mapping condition.

The "response reversion" strategy is especially interesting. Each participant had to complete at least 10 blocks of 30 trials at 80% accuracy to proceed through each successive phase of the task so this strategy cannot be attributed to subjects merely failing to learn the reversal. Why these participants reverted to responding based on their original training can only be speculated. In a strict stimulus – response model, the reversal training should have completely reset the learned response associations,

making reversion to phase 1 category endorsement impossible, so the existence of this strategy is also inconsistent with simple stimulus-response category theories. It is possible that there was some "participant bias" from these subjects, who may have thought that the reversal was some sort of "trick" manipulation to affect their responding on the last block. To these participants, the reversal must have been quite obvious (and to anybody who spent much time in the first training phase, it should have been), so they may have purposefully reverted to their initial training. If so, this indicates that categorization is subject to executive control by subjects. Alternatively, subjects may have treated the reversal training phase as a novel categorization learning task, and partitioned their learning in this phase in a way which resulted in formation of a new category representation that did not interfere with the category representation formed during in the first phase, allowing this representation to reemerge to control responding in the final phase. Regardless, it is unclear why this occurred to the degree that it did.

Those with 'indeterminate strategies' may have been merely guessing. The maximum time allowed for completion of the task was 2 hours, and many participants took almost that entire time. Many of them became audibly frustrated (e.g. sighing, asking how long the task was), and undergraduate students at this university in general are not necessarily always the most motivated, as they are only incentivized with class credit. Furthermore, as evidenced by the high rate of participants who did not complete the task, this was a difficult task. Making the task difficult was necessary to prevent participants from being able to use a rule-based strategy, and force an information integration strategy.

Overall, this study provide evidence in favor of the primary hypothesis: that at least some subjects learn a mediating "category representation" association in implicit, information integration tasks, and that altering the relationship between this category representation and the associated category label may alter the same relationship for all other stimuli which belong to the altered category representation.

CHAPTER 4: GENERAL DISCUSSION

For Study 1, I hypothesized that reversing one of two sets of category labels, which were shared by one set of stimuli, would cause participants to reverse the other. For Study 2, I hypothesized that reversing a subset of a category would cause participants to extend the reversal to the remaining stimuli. For Study 1, the hypothesis that a shared, mediating category representation would form for both sets of labels, and that reversing one set of labels would reverse the other, was not supported. While participants continued to reverse the trained set, this reversal did not transfer to the untrained set, inconsistent with the results found by Wills and colleagues (2006). One interpretation is that subjects failed to acquire a shared category representation. However, other interpretations are also possible. Participants in Experiment 1 may have partitioned each set of labels and the associated stimuli into separate category representations, even though the stimuli from both sets of labels were identical, rather than acquiring a shared category representation. It is also possible that no abstract category representation was acquired, consistent with early S-R theories of information integration learning such as COVIS. These results led to Experiment 2, in which more closely followed the method used by Wills et al. (2006; Experiment 2). The results from this experiment supported the hypothesis that subjects can acquire a mediating category representation, and that altering its association with a category label could alter the association for all stimuli belonging to that category. Previous studies of information integration task have not examined whether mediating category representations might be learned. In fact, one study which was examining behavioral dissociations for the two-stage model (i.e. stimulus – category label – motor response) mentioned that they explicitly attempted to design their study to control for the possibility of a category representation association (Maddox et al., 2010).

Although there was no significant reversal transfer effect in Study 1, a small number of participants did appear to either transfer the reversal (n=2), or ignore the reversal entirely and revert to their original training (n=5). One participant even reversed the reversal (reverted their responses for the reversed label set, and reversed their responses for the transfer stimuli). In Study 2, although more than

half of the participants transferred the reversal, there was also a wide variety of other strategies (see Table 3). The reason for the variety of strategies is difficult to explain. Failure to transfer the reversal, as shown by the majority of participants in Study 1 and the subset of participants in Study 2 who fell into the "reversal training only" category, can be accounted for by the original COVIS model (Ashby et al., 1998), the other strategies do not fit any established theories so neatly.

Previous theories on category representations suggest that, during a full reversal, a new association is formed between the implicit cue and appropriate motor response (Kendler & Kendler, 1962), and that furthermore, it is the presence of this cue which triggers the reversed behavioral response. In information-integration tasks, the stimulus *and* the labels may be a part of the implicit cue. Therefore, in Experiment 1, the presence of the labels may have been as important as the presence of the stimulus itself, and it is this entire cue "package" which had its association changed. Thus, when one set of labels was reversed, the other set may not have been affected, which would not prompt the reversed response. It could also be the case that only the label – response association was altered, and only then for the one set trained in reverse. This might suggest a shared category representation, but separate category label associations.

In the Wills et al. (2006; Experiment 2) study, participants were presented with two separate categorization problems with an identical set of labels (i.e. A and B). They reasoned that if a category representation did not exist, reversing the label – response association for a subset of one categorization problem might alter it for the other categorization problem, due to the shared category labels. However, they found that it only altered responses for the categorization problem that had the subset trained, and that furthermore, the reversal extended to the untrained stimuli. They reasoned that each separate categorization problem developed its own category representation, and that even though the labels were identical, the reversal did not extend to the other categorization problem due to the fundamental difference in the associative properties of each set of categories (i.e. separate problems, separate representations; Wills et al., 2006). In Experiment 1, I attempted to see if a category representation could

be shared somehow with separate labels; a design opposite of that done by Wills and colleagues. Perhaps it is simply not possible to share a category representation between separate sets of labels.

The secondary hypothesis in Study 2 that there would be a difference in likelihood that the reversal would extend to transfer stimuli for each mapping condition was also supported. Although simple S-R based category learning theories such as COVIS imply that information-integration learning should be impossible under conditions of inconsistent response mappings, there is some empirical support for learning under these conditions. Spiering and Ashby (2008) suggested that much of the previous literature which suggested information-integration learning requires consistent mapping only attempted to introduce inconsistent mapping after a period of training (often several hundred trials) *exclusively* with consistent mapping. In their study, they had participants *begin* the task with inconsistent mappings (or a consistent control), and although they found block 1 differences in accuracy, with the control group scoring higher, asymptotic accuracy did not significantly vary across conditions. The authors suggested that a consistent *feature* identity (category label; they used letters or colored circles) was sufficient for learning to take place, although a consistent *spatial* identity (response location) and a consistent feature identity promoted slightly quicker learning. The authors further suggested that previous studies demonstrating performance issues when there was a switch from consistent to inconsistent mapping may have had to do with forcing participants to suddenly switch their reliance from primarily spatial cues to feature cues only (Spiering & Ashby, 2008). Additionally, this theory could possibly explain the results of Study 1 as well; participants had no consistent spatial *or* feature identity to rely on (different labels, alternating locations), therefore, a shared representation may have been impossible to develop for that reason alone.

One additional explanation from the SRT literature concerns awareness of the procedural element of the task (i.e. the pattern of stimulus presentation). It has been suggested that sequence learning can rely on different internal representations depending on whether or not there is conscious awareness of the pattern (Willingham, Wells, Farrell, & Stemwedel, 2000). Given that SRT tasks and information-integration tasks rely on the procedural system and the basal ganglia to primarily guide learning, it is

possible that strategy differences in Study 2 could have had to do with varying levels of conscious awareness participants may have had in regards to the "rule" that guided proper categorization.

Implications for theories of category learning

The original COVIS model (Ashby et al., 1998) predicted information integration learning would be based on direct stimulus – response relationships; this theory predicts that complete reversion, or a full reversal, should not be possible. While there is still debate over single- versus multiple-system explanations for category learning (Ashby et al., 2011; Zaki & Kleinschmidt, 2014), several different updates have been proposed to the COVIS model, most notably, the addition of a mediating category label association, and the inclusion of a multiple-systems view for explicit and implicit systems (Ashby et al., 2003; Maddox et al., 2004; Maddox et al., 2010). The data from Study 2 compliment this, and add further evidence to the multiple-systems theory of category learning.

However, it is unclear how to incorporate learning with inconsistent mappings into these recent extensions of the COVIS model. Although Spiering and Ashby (2008) published results demonstrating that information-integration category learning is as equally possible with inconsistent as with consistent response mappings, these data have received practically no acknowledgement. Relatively recent reviews of COVIS have not discussed the role of inconsistent mapping in implicit category learning, choosing instead to focus on the role of consistent mapping in the two-stage associational model (Ashby et al., 2011; Ashby & Maddox, 2011). Outside of Spiering and Ashby's (2008) study which demonstrated information-integration learning is possible under inconsistent response mapping conditions, the topic has been largely avoided. Even within their paper, however, they only sought to demonstrate that previous dissociation studies that suggested information-integration learning was only possible with consistent mapping may have overlooked some methodological issues in their tasks which incidentally biased their results. This makes interpreting the results of Study 1, but especially Study 2, difficult. Clearly, across Study 1 and Study 2, participants were able to learn the task to a high-degree of accuracy despite inconsistent response mappings. The current results highlight the need to extend COVIS to be able to account for learning under inconsistent response mapping conditions.

Limitations and future directions

One of the biggest limitations of these Experiments, especially Experiment 2, is participant motivation. This limitation is present in almost any research conducted with unpaid, undergraduate volunteers, but it seems especially relevant in this case. To ensure that participants could not rely on a rule-based strategy, the task was made exceptionally difficult, with the sampling space for the stimuli being extended parallel to and close enough to the decision bound that a unidimensional rule was not possible to achieve the accuracy criterion. It was necessary to prevent rule-based strategies because in rule-based tasks response location or category label manipulations have little effect on performance, due to the explicit nature of the rule-based system (Ashby et al., 2003; Ashby & Maddox, 2005; Maddox et al., 2004). However, as a result, the perceptual differences between category A and category B were very small, and within category variability was relatively large, making the task very difficult. More participants failed to complete the task than did finish. Even those who did finish often became frustrated or despondent near completion of the task (which often took participants the full 2 hours for Study 2).

Furthermore, it is difficult to assess whether individual differences played a role in the probability that a participant would complete the task. While working memory differences affect performance on rule-based tasks (working memory is important for developing and maintaining complex rules), there is no such effect for information-integration tasks (Ashby & O'Brien, 2005). Since the procedural system is heavily implicated in information-integration tasks, any possible differences may be related to the basal ganglia; specifically, individual differences in the strength of dopamine mediated learning. However, little to no research has studied differences in the basal ganglia system with regards to information-integration tasks.

In future studies, it would be informative to collect a variety of individual difference measures (e.g., working memory capacity, cognitive flexibility, or depressive symptoms, which at high levels, have been shown to enhance reflexive-optimal category learning tasks [Maddox, Gorlick, Worthy, & Beevers, 2012]) during a task similar to Study 2. Since there has not been much research on individual differences in information-integration tasks, and due to the wide variety of strategies present in Study 2, identifying

the individual difference factors correlating with strategy could potentially help elucidate the reported results. It is unclear, currently, why some participants followed one strategy over another. Debriefing questionnaires could be used in future studies as well which could ask participants about their strategy for the task, and their thoughts on the various phases (e.g. "did you purposefully choose to revert to your original training"). It might also be informative to see if a particular task manipulation could induce certain strategies in participants.

In general, there is little to no research on the existence of abstract category representations in information-integration tasks. While research has been conducted using other tasks which suggest that abstract representations underlie performance (Kendler & Kendler, 1962; Sanders, 1971; Wills et al., 2006), the only information-integration task that acknowledged that abstract representations might play a role at all treated them merely as a possibly confound that they needed to control for (Maddox et al., 2010). The data from Study 2 provide the only evidence specifically addressing abstract category representations and behavioral response mapping dissociations. I believe this perspective is valuable, and that follow-up research studies may further elucidate the exact nature of how these results fit into currently established theories of category learning.

CHAPTER 5: REFERENCES

Ashby, F.G., & Townsend, J.T. (1986). Varieties of perceptual independence. *Psychological Review, 93(2),* 154-179.

Ashby, F.G., & Maddox, W.T. (1990). Integrating information from separable psychological dimensions. *Journal of Experimental Psychology: Human Perception and Performance, 16(3),* 598-612.

Ashby, F.G., & Maddox, W.T. (1992). Complex decision rules in categorization: Contrasting novice and experienced performance. *Journal of Experimental Psychology: Human Perception and Performance, 18(1),* 50-71.

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

Ashby, F.G., Alfonso-Reese, L.A., Turken, A.U., & Waldron, E.M. (1998). A neuropsychological theory of multiple systems in category learning. *Psychological Review, 105(3),* 442-481.

Ashby, F.G., Queller, S., & Berretty, P.M. (1999). On the dominance of unidimensional rules in unsupervised categorization. *Perception & Psychophysics, 61(6),* 1178-1199.

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

Ashby, F.G., Ell, S.W., Waldron, E.M. (2003). Procedural learning in perceptual categorization. *Memory & Cognition, 31(7),* 1114-1125.

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

Ashby, F.G., & Maddox, W.T. (2005). Human category learning. *Annual Reviews in Psychology, 56,* 149-178.

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

Ashby, F.G., Paul, E.J., & Maddox, W.T. (2011). COVIS. In E.M. Pothos & A.J. Wills (Eds.). *Formal approaches in categorization.* New York: Cambridge University Press.

Bourne, L.E. (1970). Knowing and using concepts. *Psychological Review, 77(5),* 545-556.

Brainard, D. H. (1997). The Psychophysics Toolbox, *Spatial Vision, 10,* 433-436.

Cantwell, G., Crossley, M.J., & Ashby, F.G. (2015). Multiple stages of learning in perceptual categorization: Evidence and neurocomputational theory. *Psychonomic Bulletin Review.* doi:10.3758/s13423-015-0827-2

Chandrasekaran, B., Koslov, S.R., & Maddox, W.T. (2014). Toward a dual-learning systems model of speech category learning. *Frontiers in Psychology, 5(825),* 1-17.

Cohen, A., Ivry, R.I., & Keele, S.W. (1990). Attention and structure in sequence learning. *Journal of Experimental Psychology: Learning, Memory, and Cognition, 16(1),* 17-30.

Crossley, M.J., Ashby, F.G., & Maddox, W.T. (2014). Context-dependent savings in procedural category learning. *Brain and Cognition, 92,* 1-10.

Curran, T. (1995). On the neural mechanisms of sequence learning. Psyche [on line], 2(12).

Faul, F., Erdfelder, E., Lang, A.G., & Buchner, A. (2007). G*Power 3: A flexible statistical power analysis program for the social, behavioral, and biomedical sciences. *Behavior Research Methods, 39(2),* 175–191.

Goldman, D. & Homa, D. (1977). Integrative and metric properties of abstracted information as a function of category discriminability, instance variability, and experience. *Journal of Experimental Psychology: Human Learning and Memory, 3(4),* 375-385.

Grill-Spector, K., Henson, R., & Martin, A. (2006). Repetition and the brain: Neural models of stimulus-specific effects. *TRENDS in Cognitive Sciences, 10(1),* 14-23.

Gureckis, T.M., James, T.W., & Nosofsky, R.M. (2011). Re-evaluating dissociations between implicit and explicit category learning: An event-related fMRI study. *Journal of Cognitive Neuroscience, 23(7),* 1697-1709.

Kendler, H.H., & Kendler, T.S. (1962). Vertical and horizontal processes in problem solving. *Psychological Review, 69(1),* 1-16.

Kleiner, M., Brainard, D., Pelli, D. (2007). What's new in Psychtoolbox-3? *Perception, 36,* ECVP Abstract Supplement.

Kruschke, J.K. (1992). ALCOVE: An exemplar-based connectionist model of category learning. *Psychological Review, 99(1),* 2-44.

Kruschke, J.K. (1996). Dimensional relevance shifts in category learning. *Connection Science, 8(2),* 225-247.

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.

Maddox, W.T., Bohil, C.J., & Ing, A.D. (2004). Evidence for a procedural-learning-based system in perceptual category learning. *Psychonomic Bulletin & Review, 11(5),* 945-952.

Maddox, W.T., Ashby, F.G., Ing, A.D., & Pickering, A.D. (2004). Disrupting feedback processing interferes with rule-based but not information-integration category learning. *Memory & Cognition, 32(4),* 582-591.

Maddox, W.T., Gorlick, M.A., Worthy, D.A., & Beevers, C.G. (2012). Depressive symptoms enhance loss-minimization but attenuate gain-maximization in history-dependent decision making. *Cognition, 125,* 118-124.

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

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

Olson, C.R., & Gettner, S.N. (1999). Macaque SEF neurons encode object-centered directions of eye movements regardless of the visual attributes of instructional cues. *The Journal of Neurophysiology, 81,* 2340–2346.

Pelli, D. G. (1997). The VideoToolbox software for visual psychophysics: Transforming numbers into movies. *Spatial Vision, 10,* 437-442.

Poldrack, R.A., & Packard, M.G. (2003). Competition among multiple memory systems: Converging evidence from animal and human brain studies. *Neuropsychologia, 41,* 245-251.

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

Posner, M.I., & Keele, S.W. (1968). On the genesis of abstract ideas. *Journal of Experimental Psychology, 77(3),* 353-363.

Rand, M. K., Hirosaka, O., Miyachi, S., Lu, X., Nakamura, K., Kitaguchi, K., & Shimo, Y. (2000). Characteristics of sequential movements during early learning in monkeys. Experimental Brain Research, 131, 293–304.

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

Reed, S.K. (1972). Pattern Recognition and categorization. *Cognitive Psychology, 3,* 382-407.

Richler, J.J., & Palmeri, T.J. (2014). Visual category learning. *WIREs Cognitive Science, 5,* 75-94.

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

Seger, C.A., & Miller, E.K. (2010). Category learning in the brain. *Annual Review of Neuroscience, 33,* 203-219.

Seger, C.A., & Peterson, E.J. (2013). Categorization = decision making + generalization. *Neuroscience and Biobehavioral Reviews, 37,* 1187-1200.

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

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

Sutherland, N. S., & Mackintosh, N. J. (1971). Mechanisms of animal discrimination learning. New York: Academic Press.

Verwey, W.B., & Wright, D.L. (2004). Effector-independent and effector-dependent learning in the discrete sequence production task. *Psychological Research, 68,* 64-70.

Verwey, W.B., & Clegg, B.A. (2005). Effector dependent sequence learning in the serial RT task. *Psychological Research, 69,* 242-251.

Wächter, T., Lungu, O.V., Liu, T., Willingham, D.T., & Ashe, J. (2009). Differential effect of reward and punishment on procedural learning. *The Journal of Neuroscience, 29(2),* 436-443.

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

Willingham, D.B., Wells, L.A., Farrell, J.M., & Stemwedel, M.E. (2000). Implicit motor sequence learning is represented in response locations. *Memory & Cognition, 28,* 366-375.

Wills, A.J., Noury, M., Moberly, N.J., & Newport, M. (2006). Formation of category representations. *Memory & Cognition, 34(1),* 17-27.

Worthy, D.A., Markman, A.B., & Maddox, W.T. (2013). Feedback and stimulus-offset timing effects in perceptual category learning. *Brain and Cognition, 81,* 283-293.

Zaki, S.R., & Kleinschmidt, D.F. (2014). Procedural memory effects in categorization: Evidence for multiple systems or task complexity. *Memory & Cognition, 42,* 508-524.

Zeithamova, D., & Maddox, W.T. (2006). Dual-task interference in perceptual category learning. *Memory & Cognition, 34(2),* 387-398.