UNIVERSITY OF CALIFORNIA, MERCED

Supramodal Theory: Unifying Visual Similarity and Categorization

A dissertation submitted in partial satisfaction of the
requirements for the degree Doctor of Philosophy

in

Cognitive and Information Sciences

by

Michael Robert Romano

Committee in charge:

Professor Michael J. Spivey, Chair

Professor Teenie Matlock

Professor David C. Noelle

2011

ACKNOWLEDGEMENTS

I am grateful to my parents, for always supporting my decisions and for encouraging me to pursue happiness in life. I thank my advisors, Michael Spivey and Teenie Matlock, for their guidance and mentoring in research and teaching. I thank the interdisciplinary community of professors and graduate students at the University of California, Merced, for always being willing and interested in sharing ideas across traditional walls of academia. I thank all the mentors and teachers in my life, both intentional and unintentional, that have supported my path to this stage and beyond.

TABLE OF CONTENTS

LIST OF FIGURES

LIST OF TABLES

ABSTRACT OF THE DISSERTATION

Supramodal Theory: Unifying Visual Similarity and Categorization

By

Michael Robert Romano

Doctor of Philosophy

University of California, Merced, 2011

Professor Michael J. Spivey
Professor Teenie Matlock
Professor David C. Noelle

Similarity and categorization are central phenomena in cognitive science. Despite its relevance, similarity is a poorly understood process. The current work proposes a new theory of similarity and how to unify it with categorization. Similarity is defined as the overlap of neuronal population codes that represent features of objects being compared. The experiments investigate how perceptual and conceptual information help to constrain similarity. A neural network model is designed that accurately predicts human performance on similarity judgments.

CHAPTER 1: INTRODUCTION

1.1 Introduction

Categories and concepts take a central role in the field of cognitive science, in that their significance is strong across even the diverging paradigms of symbolic and connectionist approaches (Lakoff, 1987; Lamberts & Shanks, 1997; Markman & Ross, 2003; Murphy 2002; Rogers & McClelland, 2004). Their relevance can be understood in nearly every thought we have, feeling we experience, and action we take. Categories and concepts have this central role because they allow us to organize and structure our mental spaces. It is because of this organization that we are able to recognize new objects as members of previously known categories, such as classifying a hanging fruit as an apple instead of an orange. And after we are able to classify that fruit as an apple, categories help in drawing inferences based on our knowledge of that category, such that it is likely to be tart if it is green or sweet if it is red. And in comparing that category to other categories, such as other types of fruit, we can make generalizations such as juices made from fruit are usually refreshing, and that ripe fruit tastes better than unripe fruit. Categorical structure provides us information for making generalizations about members within a category, discriminating between different categories, and organizing categories in relations such as a hierarchy (e.g., subordinate-level “pink lady”, basic level “apple”, superordinate-level “fruit”).

But the question of exactly how do categories structure our mental spaces, has given rise to several competing theories. Among the most influential theories of categorization, historical and contemporary, are the *classical theory*, *prototype theory*, *exemplar theory*, and the *theory theory*, and hybrid combinations. Each theory has its strengths and weaknesses, but none are able to account for all the effects discovered in experimental cognitive science and psychology. As we will see, at the heart of each of these theories (with the exception of the classical theory) is a strong dependence on the idea of similarity. If categories and concepts are central to our understanding of human cognition, and if similarity is central to theories of categories and concepts, then the natural conclusion is that similarity is central to human cognition. Yet despite the obvious relevance, similarity is a poorly understood phenomenon.

Similarity as a cognitive phenomenon also has many competing theories that attempt to both define and explain it, including *spatial models*, *feature models*, and *structural alignment models*. As with the theories of categorization, competition among theories of similarity arises because none are able to account for all of the many findings in the empirical data (see Chapter 2 for a detailed discussion). It has become a serious gap in our understanding of human cognition if two of the most central phenomena, similarity and categorization, remain unexplained to the extent that a coherent theory has yet to account for the empirical data collected across the many experimental studies and computational models of cognitive science and psychology.

The lack of unity between theories of similarity and theories of categorization has created another problem within cognitive science; how can categorization be grounded in similarity if the theories of categorization do not specify the process of similarity? The problem of explaining categorization is simply removed by one step until the process of similarity is properly specified. Similarity is simply a placeholder in this situation, a “black box” in a diagram of how the mind allegedly works.

Current theories of similarity and categorization fall victim to the “chicken and the egg” problem (Hahn & Chater, 1997). Typically categorization theories are explained by an underspecified process of similarity. But then the amount of similarity of one object/concept to another object/concept is typically explained by the learned categorical and conceptual knowledge. Hahn & Chater argue that this circularity, combined with a lack of specification in how the process of similarity works, is a central problem in cognitive science. Clearly a legitimate theory of categorization needs not only to properly explain similarity, but it also needs to *unify* similarity and categorization into a single framework.

Another central phenomenon in human cognition is metaphor. In their groundbreaking book *Metaphors We Live By*, Lakoff and Johnson (1980) explore the ubiquity of metaphor in everyday language. Independent of culture and particular language use, human language use is permeated with metaphor. There are two key claims to conceptual metaphor theory. The first is that metaphor is ubiquitous, in that a large proportion of everyday speech is metaphoric. The second is that we rarely recognize metaphoric language for being such. But the importance of metaphor goes beyond the surface of language use; the main argument of Lakoff and Johnson is that metaphors are conceptual and they shape the way we think about and act in the world.

And it is precisely because conceptual metaphors are at the metaphorical heart of human cognition that we rarely recognize when we are using them.

But if similarity, categorization, and conceptual metaphor are all central to human cognition, then a framework must be developed to unify them, or at the very least we must discover how information can flow across and between these seemingly separated phenomena. Therefore, the framework of *supramodal saliency maps* is proposed to not only resolve the gap between theories of similarity and categorization, but also to unify them with other central phenomena in cognitive science, such as metaphor. The supramodal saliency map theory, here called the *supramodal theory*, has its roots primarily in the bold framework of cognitive science called *the continuity of mind* (Spivey, 2007), and in *embodied cognition* (Gibbs, 2006; Lakoff, 1987; Lakoff & Johnson, 1980; Varela, Thompson, & Rosch, 1991). The supramodal theory relies on a continuous, dynamical flow of cognition that is situated in the interaction of an embodied entity and its environment.

1.2 Supramodal Theory

The goal of the supramodal theory is to unify the central processes of human cognition, including similarity, categorization, and metaphor. The proposal is both simple and complex: that the framework for understanding human cognition, both in its mathematical description and in its mechanistic explanation, needs to rely on a single method for representing the multitude of cognitive phenomena. Human cognition is a complex (Gazzaniga, Ivry, & Mangun, 1998), continuous and dynamical

(Spivey, 2007), parallel processing (Rumelhart & McClelland, 1986), distributed (O'Reilly & Munakata, 2000), embodied (Gibbs, 2006) system, and at some level each neuron is connected to every other neuron through *N*-links. Given these factors, our goal ought to be the search for a common method for representing the interactions of different cognitive processes of the same system, especially at the level of the central nervous system post-transduction, where neural population codes are the understood basis for perception, cognition, attention, emotion, and other mental processes.

The core framework of the supramodal theory is that human cognition emerges from the collision of saliency maps that continuously flow along the neuronal pathways of the brain. A saliency map (Koch & Ullman, 1985) is a topographical map that represents the saliency of features (how likely each feature will attract attention) for a particular field of attention. Saliency maps have been used to describe unimodal processing in visual attention (Itti & Koch, 2000; Itti & Koch, 2001; Li, 2002) and auditory attention (Kayser, Petkov, Lippert, & Logothetis, 2005), and they have been proposed to explain supramodal processing (for a review, see Spivey, 2007, Chapter 5).

For example, the collision of saliency maps can be demonstrated in the McGurk effect (McGurk & MacDonald, 1976). In the set up of this experiment, participants are first shown a video of a person saying "ga-ga" repeatedly. The information in the visual stream (the visible face and mouth movements) are compatible with the information in the auditory stream (the phonemes being heard), and so participants correctly have the perception of the person saying "ga-ga" repeatedly. However, in another trial of the experiment, the information from the visual scene is fixed with the original “ga-ga” mouth movements, and the information in the auditory stream is changed to the phonemes “ba-ba” (played in synchrony with the mouth movements of the video). The result is the participant’s experience of hearing the phonemes “da-da” repeatedly. That is, rather than perceiving the accurate information from either the visual information (“ga-ga”) or the auditory information (“ba-ba”), participants perceive an illusory phoneme (“da-da”). The visual and auditory perceptual systems interact and generate an experience not contained in either perceptual system, arguably at the supramodal level. The supramodal representation is a combination of the salient features that were attended in both modalities.

This idea of colliding saliency maps in the supramodal theory has similarities with the *conceptual blending theory* (Coulson, 1995; Fauconnier, 1997; Fauconnier & Turner, 1994; Turner & Fauconnier, 1995). In conceptual blending, there are mental spaces that are constructed whenever we think and use language. This theory proposes that a mental space “is a (relatively small) conceptual packet built up for purposes of local understanding and action” (Turner & Fauconnier, 1995, p. 184). For example, when understanding a phrase such as “my hike along the Yosemite Falls trail in Yosemite National Park in 2011,” my mental space will contain *hike, the hiker, the year, the location*, … , and Turner and Fauconnier argue that it will also has partial structure from the concept of *journey*. My “journey” as we might call it, is nothing concrete, such as the hiker or the location, but rather it is an abstraction that emerges from the combination of these concrete items in the sentence.

Mental spaces are typically constructed from multiple conceptual domains. According to the conceptual blending theory, the blended space has less information than the input spaces because it recruits only a partial structure from the input spaces, but what emerges is something new, which also provides more information than the input spaces. This is not unlike what happens during the McGurk effect, where an illusory phoneme is perceived from contradictory phonemic information entering from two distinct sensory organs and combining in the perceptual-conceptual systems of the brain. And again, with the "journey" in Yosemite, distinct concrete items are combined and something new, and not directly observed, takes shape.

In a similar fashion, colliding saliency maps in the supramodal map can be considered "blends" of unimodal maps, and what emerges is more than the sum of its parts. They are maps because of their sensory origin, as our visual and auditory sensory organs have spatial structure in their afferent signals (topographic in the eyes, tonotopic in the ears). However, there is a very important difference between these two theories. The spaces in the supramodal theory are describable at many levels, from the sensorimotor to conceptual, while the spaces in the conceptual blending theory can only describe the conceptual level. The conceptual structures employed in blending may very well have their origins in the sensorimotor system, but ultimately the conceptual blending theory is biased to describe higher-level cognitive processes such as language. Turner and Fauconnier are direct about this problem, saying that the blending theory sacrifices high variability and low parsimony for better sensitivity and generality. The supramodal theory, on the other hand, can describe both low- and high-level cognitive processes as we will see below, all while maintaining parsimony.

Saliency maps have primarily been used to model bottom-up perceptual processes such as visual attention. A visual scene is reduced to a two-dimensional image, and a third dimension allows for a topographical saliency map, where hills represent areas of more attraction for attention and valleys represent areas of less attentional attractiveness. For example, an object's luminance, motion, color, or texture might act as strong bottom-up attractors of attention. Or one might experience a "pop-out effect" in a visual search task (Desimone & Duncan, 1995), such as observing a single unique object of high distinctiveness among a sea of competitor objects of low distinctiveness.

It has been proposed that bottom-up saliency can be combined with and modified by the top-down influences of attention, an interaction that alters the saliency of locations, objects, or features based on behavioral relevance (Treue, 2003). In his review of recent research, Treue reports that attention has been shown to enhance the saliency of stimulus contrast, which is a natural and strong bottom-up salience factor (Reynolds, Pasternak, & Desimone, 2000; Reynolds & Desimone, 2003). The enhancement of top-down attention on bottom-up salience was strongest for stimuli of intermediate salience and weaker for stimuli of highest salience. Explaining this finding in the context of other research, Treue argues that stimuli of highest salience will already attract attention without needing help from top-down influence, which is compatible with the "pop-out effect", while stimuli of intermediate salience (such as an object among many similar distractors) might need top-down attention to make them salient. The question of how top-down attention interacts with the process of similarity during visual search will be explored later.

Saliency maps have been proposed to exist in several localized brain areas (Gottlieb 2007; Koch & Ullman, 1985; Kustov & Robinson, 1996; Li, 2002; Mazer & Gallant, 2003; Robinson & Petersen, 1992). The lateral intraparietal area has been shown to encode a topographical saliency map of the external world, and bind it with information relevant to tasks involving sensory, motor, and cognitive processes (Gottlieb, 2007). These findings have been demonstrated across a number of behavioral tasks, including eye-movements and motor responses, and goal-based responses. Gottlieb proposes that this brain region allows for interactions between low-level spatial-processing and high-level abstract cognitive processing, creating representations that guide spatial behavior based on behavioral needs. She extends this position further, by relating it to the embodied cognition paradigm, which claims that these abstract and high-level cognitive processes emerge from sensorimotor processes, not centralized computational modules in the brain. The intersection of high-level processing and sensorimotor processing provides a bridge for the processes of perception, action, and cognition to interact with each other, argues Gottlieb.

The current work builds upon these findings and arguments, and presents the idea that saliency maps are useful representations for unifying perception, attention, and action. More so, they are useful representations of abstract and high-level cognitive processes such as similarity, categorization, and metaphor. If high-level cognition is grounded in low-level sensorimotor processing, and if saliency maps can explain sensorimotor processing, then it would seem a natural fit to extend saliency maps to conceptualization. Our perceptual systems, goal-driven motor systems, and other online processes operate together in the transient activation of online tasks and strategies. And as we have seen, saliency maps can be used to integrate bottom-up and top-down processing. Therefore, it would seem that saliency map representations can easily be extended to explain these transient activations.

Transient activation can be a supramodal saliency map. It is supramodal when the online processing of multiple (if not all) unimodal perceptual systems are actively engaged simultaneously at any given moment, and they interact with the goal-, contextual-, and knowledge-driven processes of cognition to produce relevant and adaptive motor responses in behavior. An abstract supramodal saliency map is a useful description of the abstract state space that can represent an online task because it can incorporate the external world, the body's perceptual systems, and the internal processing of the central nervous system into a single meaningful descriptive representation and explanatory mechanism.

While all of these constructs (perception, action, cognition) are useful for theoretical purposes, in the end they are all part of the same unified, whole dynamical system that includes mind, body, and environment, and so it is important to have a single representational form that can be used to explain both the parts and the whole when we are investigating high-level phenomena such as similarity and categorization that span these parts. In short, the supramodal theory of cognition is a mathematical and mechanistic approach to embodied cognition. The theory is mathematical in the state space description of how information across seemingly distinct neural systems can interact and combine, and mechanistic in the neural explanation of how information is processed and represented.

Top-down components of online tasks and their transient activities have been demonstrated to interact with bottom-up processes explainable with saliency maps. Salience information has been demonstrated to extend from bottom-up visual attention to higher-level cognitive processes such as spatial working memory (Fine & Minnery, 2009). The main finding, that the ability to recall the location of targets increased as saliency increased, suggests that memories are encoded by prioritizing objects and features by their task relevance, especially as task difficulty increases or cognitive capacity is otherwise reduced. Another study shows that saliency information from semantic knowledge influences the speed and accuracy of detecting changes of objects in natural scenes by interacting with the bottom-up visual saliency map (Stirk & Underwood, 2007).

Taken together, these two studies demonstrate a bidirectional influence of saliency maps between bottom-up and top-down processes of visual perception and transient activation of online tasks. Thus, treating saliency maps as a common language between lower- and higher-level cognition is a useful framework for describing and explaining perception, action, and cognition. There is a reciprocal relationship in that visual saliency helps prioritize encoding memories, and memories help influence the saliency of visual attention.

The classic theory of working memory in cognitive psychology proposes that there are multiple modules that function together, including a central executive that controls attention, a visuospatial sketchpad that processes and stores visual information, and a phonological loop that processes and stores auditory information (Baddeley & Hitch, 1974). This theory is riddled with problems at both the theoretical and implementation level. To begin, it suffers from the homunculus problem, as with any theory that proposes a central executive model of cognition. There is also no reference to how the motor system interacts with attention or the perceptual systems (visuospatial sketch and phonological loop). Finally, it is inherent in the design of the theory that information in different modalities is represented in different formats, begging the question of how information is integrated at the supramodal level, such as in the McGurk effect. At the very least, a translation module would be necessary as an intermediary between the other modules of this compartmentalized approach, so that, for example, the visuospatial sketch and phonological loop could communicate to enable the McGurk effect.

Transient activation is a dynamic, continuously morphing and shifting topographical landscape at both the unimodal and supramodal levels. This transient neural activity of a cognitive entity will require continuous updating because the patterns of input it receives from the environment are naturally in constant change, and so the feedback processes of that entity will require a smooth exchange between its biological internal states and its external environment. This helps to explain the extraordinary level of flexibility and adaptability in online processing, but also introduces cognitive capacity limits such as limited storage in working memory, both temporally across time and in quantity for a single moment. A cognitive system ought to have enough flexibility to maintain the relevant information it already has stabilized while acquiring information it requires to perceive and act quickly in a dynamic reality, including biological beings such as ourselves that evolved and continue to exist in a world that demands action in real time.

It is also proposed that long term memory can store supramodal saliency representations. Simply, long term memories are stabilized echoes of transient activation; frames of the continuous flow of the unimodal and supramodal saliency maps of online tasks. A "frame" in the supramodal theory should not be confused with semantic frames in linguistics (Fillmore, 1982) or schemas in cognitive psychology. Simply, a frame of the supramodal saliency map is a momentary time slice from the continuous dynamical flow of that map. For example, if a participant in a visual attention study is presented a highly salient stimulus such as a flashing light at time t = 100ms, then a frame would be the saliency map at that moment t.

O'Reilly and Munakata (2000) support the claim that long term memory is the consequence of changes in connection strengths between neurons, while working memory is the pattern of activation that flow across connections between neurons. Long term memory, then, can be considered saliency maps that were strong enough to change the connections between neurons, and the salient features become encoded in those connections while the less salient features fade away. In the Leabra architecture, a tripartite modeling system (of the posterior cortex, hippocampus, and prefrontal cortex) that uses different representations for memory (Petrov, Jilk, & O'Reilly, 2010). This is consistent with the supramodal theory because the tripartite architecture avoids the problems of modularity while accounting for a broad range of diverse sensorimotor and cognitive phenomena in human behavioral research, and it does this by having a common, biologically-plausible implementation that ultimately utilizes the same neural code, albeit in different connectivity patterns.

For example, categorical knowledge is based on grouping objects by their perceptual, behavioral, and/or functional features, which can be represented as saliency maps, unimodal or supramodal. Long term memories of events typically contain the features that were most attended. I can remember where I first ate sushi, but I don't remember what I was wearing, what the music was playing in the background, how many people were in the restaurant, or what color were the walls. Perhaps I might if I stained my shirt with soy sauce, or some other experience made my clothes more salient. It is a natural consequence that we only remember the salient features of a category or an event, because the saliency maps of our perceptual, cognitive, and sensorimotor systems guide the attention of our online processing to focus on features through the interface of the body and its environment.

So if this event of eating sushi were topographically represented as a dynamically changing supramodal saliency map, a single "time slice" would cut horizontally across this topography. Anything above the threshold of this slice, such as the hills of the more salient features, would be preserved in long term memory, while anything below the threshold would fade as transient activation is updated.

These slices accumulate into long term memory, allowing for stored representations based on concrete experiences.

The classic theory of long term memory in cognitive psychology proposes two main types, declarative (explicit) memory and procedural (implicit) memory (Graf & Schacter, 1985). Declarative memories are like the ones just discussed, such as categorical knowledge and episodic events. Procedural memories are how we perform actions, such as riding a bicycle or tying our shoes. The supramodal theory would account for this distinction on the basis of the proportion of unimodal saliency composition in the saved frames of the supramodal map into long term memory. That is, long term memory in the supramodal theory are saved frames from the continuous flow of saliency maps in transient activations, and if maps from a particular time series (such as learning to tie shoes) have features highly salient in one modality vs. the others (such as the visual and motor actions in tying shoes, but not auditory or olfactory), then the saved frames will favor the features from a subset of modalities that have the highest salience in the supramodal map (the birds singing in the background are not relevant to tying shoes, and so that auditory experience is likely to not be salient in learning to tie shoes). This eliminates the need to categorize memory in different formats (explicit vs. implicit, declarative vs. procedural, etc.) and avoids the problem of explaining how different formats of memory communicate and interact, because instead we can base memory on the strongest and most influential components of the supramodal saliency maps, a singular format.

Earlier it was stated that the core framework of the supramodal theory is that human cognition emerges from the collision of saliency maps that continuously flow along the neuronal pathways of the brain. It is theorized that there may be several to many saliency maps simultaneously propagating their activations through the nervous system at any particular moment. And when they meet, they interact, collide, become integrated, and what emerges is the supramodal map (or perhaps a unimodal map, such as component feature maps becoming integrated in early stages of vision processing that complete the visual saliency map). This idea is inspired by another suggesting that working memory is the result of coherent, semi-synchronous oscillations of neurons, in the range of 40-70 Hz, whose function is to bind the relevant features of a visual object or scene (Crick & Koch, 1990a; Crick & Koch, 1990b). In representing modalities, attention, and memory as saliency maps, and integrating them based on their "collisions" – the timing, coherence, and synchrony of their activation patterns in the nervous system – the binding problem of how information can be both separated into features and also structured and organized into meaningful wholes becomes less problematic. Features are both contained in and structured by the continuous flow of saliency maps in the nervous system, because the topographic structure binds features together.

Consider split-brain patients, whose corpus callosum (the bundle of neurons that connect the left and right hemispheres of the brain) has been severed, typically as a remedy to repeated massive epileptic seizures. Although the sub-cortical pathways remain intact and connect the two hemispheres, in some cases of split-brain patients the two hemispheres function completely independently, such that each hemisphere maintains an independent focus of attention (Gazzaniga, 2005). Gazzaniga goes on to describe examples of split-brain patients that manipulate an object using the left hand (the right hemisphere) while the left hemisphere expressed confusion. Another case involved the experience of belligerence in one hemisphere and calm in the other.

Gazzaniga concludes that these cases raise interesting questions on the nature of consciousness, the self, and whether each hemisphere has an independent sense of self. While it is pure speculation, these effects might be explainable with the supramodal theory. If perception, action, and cognition are the collisions of saliency maps propagating across various neural pathways, and those collisions are no longer possible at the cortical level across hemispheres, then it may be the case that different perceptions, actions, and cognitive processes will emerge independently between hemispheres.

For example, component features (such as contour, color, brightness, etc.) are saliency maps that combine into a coherent visual perception. This saliency map combination can occur at a higher level, such as the visual and auditory perceptions combining in the McGurk effect. Thus, we can have lower level maps and higher level maps, where higher level maps are more likely to enter conscious awareness (that is, we are generally more aware of the higher-level visual experience than the lower-level components, such as a visual contour map). So in the case of the split-brain patients, there may be multiple higher level supramodal maps that are generated, and each has its own perception-action loop occurring. This type of speculation begs others: Is unconscious processing the result of saliency maps that never collided with maps that become overt behavioral actions? Is priming a saliency map that never collided with maps that generate awareness?

In the current work, the supramodal theory has the lofty goal of explaining similarity and categorization in terms of the saliency information gathered by and filtered on the interface between the human body and its environment. In other words, the goal is to unite the sensorimotor systems of perception and action with the conceptualization systems of cognition in a bidirectional relationship. While it is expected that this theory will not be able to explain all the effects discovered in the literature of empirical data, it upholds the principle that any theory of cognitive science ought to have both descriptive and explanatory power in its account. In this work, the supramodal theory will be used to explore the relationship between the perceptual systems (vision in particular) and the conceptual systems (similarity, and categorization), and will provide a meaningful account of the findings, as well as a computational demonstration as a proof of concept to demonstrate this theory's feasibility.

1.3 Unifying Similarity and Categorization

One of the principle questions for cognitive science to answer is what is the human conceptual system, and how is it organized (Lakoff, 1987)? Currently there are two main paradigms within cognitive science. The first is called *functionalism*, which makes the claim that cognitive processes are defined by their function independent of the architecture that implements those functions. In the functionalist account, human cognition is implemented by the human body, which avoids the problem of Cartesian dualism. But functionalism also holds that human cognition can be just as well implemented by a computer, by transplanting a human brain into a non-human biological or robotic body, or by many other physical systems. Ultimately in functionalism, the implementation architecture is irrelevant to the functions of cognition, human cognition included.

The second paradigm of cognitive science is called *embodiment*. Embodied cognition claims that human cognition emerges from the interaction of the human body and its environment – that cognition emerges from and makes sense of the world in terms of bodily experience (Gibbs, 2006). This does not preclude the idea that other types of cognition exist, such as computer cognition or animal cognition, but embodiment makes the claim that how the body interfaces and interacts with the environment is critical for understanding cognition.

Both functionalism and embodiment consider categorization to be an organizing principle of experience (Lakoff, 1987). As previously discussed in the introduction, there are several competing theories that attempt to explain categorization and similarity. In the next two chapters we will explore these theories and the relationship between similarity and categorization. As we will see, the theories currently competing to explain categorization all depend on the process of similarity, yet none adequately mechanize that dependency. This is a serious problem if categorization is to be considered the organizing principle of experience, because if the core of that principle has not been properly explained, then we have not yet understood the structure of experience. That is why the current work explores the interdependent relationship between similarity and categorization, and attempts to describe and mechanize them both within the supramodal theory.

Hahn and Chater (1997) argue that the relationship between similarity and categorization in the cognitive psychology literature presents us with a classic "chicken and egg" problem. First, categorization theories rely on the process of similarity as the basis for grouping objects together into a category. Hahn and Chater invented a novel category called "drib" that is composed of a lightbulb, Polly the pet parrot, the English channel, and the ozone layer. Comparing the category "drib" to the category "bird" we notice that the latter is a more coherent category because the members of the category "bird" have a higher degree of similarity than do members of "drib", which allows for better generalizations and inferences within the "bird" category. However, if we then ask why are members of "bird" similar to each other, we might say because they lay eggs. But Hahn and Chater argue that this is simply categorizing them as "egg-layer" instead of "bird", which suggests that similarity is rooted in categorization. This is the "chicken and egg" problem of similarity and categorization.

The nature of the supramodal saliency map places it firmly in the embodied approach to cognitive science. The parsimonious benefit of this theory is that, with a single unified system, it can represent both sensorimotor and cognitive processes, as well as the environment in how it is represented mentally. It uses the same data format to describe both transient activation and long term memory, as well as seamlessly spanning and integrating low- and high-level processes, and bottom-up and top-down processes.

Similarity and categorization will also be explained in terms of the supramodal theory. Both will be explained using saliency maps, with features being contributed from the sensorimotor systems. The goal is to resolve the “chicken and the egg” problem described by Hahn and Chater. When similarity and categorization are unified and explained in a single system, the problems that arise from mere description of the phenomena will disappear. Just as there is no “chicken and the egg” problem with vision and action (that is, we now understand the perception-action loop of attention-guided behavior as two sides of the same coin), we should not worry about an interdependent nature of similarity and categorization phenomena. The current work, then, will abandon the worries of the chicken and egg metaphor, and align with the metaphor of similarity and categorization being two sides of the same coin.

CHAPTER 2: SIMILARITY

2.1 Introduction

The phenomenon of similarity received a harsh criticism from philosopher Nelson Goodman (1972) when he said, “Similarity, I submit, is insidious…. ever ready to solve philosophical problems and overcome obstacles, [it] is a pretender, an impostor, a quack” (pp. 437). He argues that similarity is too relative and variable to be useful, and its predictability is not dependable. This is because the similarity of two objects, Goodman claims, is inextricably linked with the context and circumstance of those objects and the person making the judgment. Goodman uses the analogy of relativity in motion in physics, saying that similarity in psychology also needs a frame of reference to add specification as to *what* is similar between two objects. The result, he argues, is the reduction of ambiguity, but at the expense of reducing the definition similarity to specifying features common between objects.

At this point, similarity becomes unnecessary, because the frame of reference is doing all the work. Indeed, empirical work has shown that similarity judgments are very sensitive to context, suggesting that Goodman’s concerns might be justified (Goldstone, Medin, & Halberstadt, 1997). But this is simply the other side of the coin. Our context – the frame of reference for similarity, according to Goodman – is just as flexible, relative, and variable as the idea of similarity that he argues against (Medin, Goldstone, & Gentner, 1993). So even if the frame of reference replaces similarity, it still falls victim to Goodman’s own criticisms and undermines the basis for that replacement. Furthermore, nearly all cognitive processes are inextricably linked with the context and circumstance of one's situation. So this argument would apply to nearly all perceptual, cognitive, and motor phenomena, thereby pushing cognitive science into a radical and limited approach. It would mean not only abandoning reference to similarity, but abandoning reference to all phenomena that are context-sensitive. All phenomena would be described solely in terms of frames of reference.

Murphy and Medin (1985) also criticize the phenomenon of similarity, arguing that similarity is too unconstrained to be the basis for conceptual coherence. Two objects can be philosophically similar in an infinite number of ways, so in order for similarity at the psychological level to function there must be psychological constraints on similarity. And it is these constraints that ground categorization, not similarity, they argue. Hahn and Chater (1997) reply that because the human cognitive system is finite in scale, it is meaningless to argue that infinite similarities at the philosophical level is a criticism against psychological similarity. Hahn and Chater suggest that we should not be considering the similarity of objects as they exist out in the world, but rather the similarity relations between the *mental representations* of those objects in the world. Then the problem is not about finding the frames of reference of similarity in the world, but a problem of understanding what is mentally represented about objects and how features are selected in similarity relations.

The supramodal approach will take precisely this distinction between objects as they exist objectively in the world and a person's subjective mental representations of those objects. If similarity is a cognitive phenomenon, then it must be described and mechanized as a cognitive process, not as a disembodied philosophical construct. But before going into the details of the supramodal theory of similarity, it is important to discuss other theories, and their strengths and weaknesses in accounting for the empirical data. In this section we will explore three main theories of similarity in cognitive science as they relate to the mental representation of objects.

2.2 Feature Theory

In feature-based theories of similarity, objects are represented by sets of features or attributes, often weighted. For example, features can represent properties such as size, shape, and color; or features can be components of a face like eyes, nose, and mouth; or features can be abstract attributes such as quality; or features can be perceptual units such as phonemes or just noticeable differences in color (Tversky, 1977). There is great flexibility in what can be characterized as a feature, which makes them extremely capable of accounting for empirical data, but also fairly unconstrained as we will see.

The most influential feature-based model is Tversky's (1977) contrast model. The basic assumption is that the similarity between two objects is a linear combination that increases as the number of common features increases, and decreases as the number of distinctive features increases. Each object being compared is broken into its constituent features, and those features can be weighted based on the context. Color, for example, is relevant in some contexts and irrelevant in others, and so that feature would be weighted accordingly.

The contrast model incorporates task dependencies such as feature salience and type of comparison being asked. For example, a comparison can be directional, such as “How similar is x to y?” Tversky explains that people consider North Korea to be more similar to China than China is to North Korea, and people consider a child to be more similar to his or her parents than vice versa. There are many asymmetry and directionality effects in similarity judgments, so it is useful to have weights in the contrast model to account for that. But comparisons can also be non-directional, such as “How similar are x and y?”

Hahn and Chater (1997) point out that a strange outcome of the contrast model is that there is no upper bound to similarity. Common features can be continuously added making two objects infinitely similar, or distinctive features can be continuously added making two objects infinitely different. This goes back to the earlier criticism against the idea that similarity is too unconstrained (Goodman, 1972), and that it is insufficient for explaining conceptual coherence (Murphy & Medin, 1985). However, as was discussed earlier, physical systems, whether biological or robotic, are constrained and finite. Consequently, they cannot process infinite features, and so this criticism is not very relevant.

The contrast model also has difficulty representing continuous features, as it was designed for handling binary features. Tversky and Gati (1982) provide some workarounds for this, but it remains a serious problem for the contrast model.

Another problem for the feature theory of similarity is that there are no explanations for how features are related to one another (Goldstone, Medin, & Gentner, 1991).

Categories cannot be just an amorphous bag of features, even if they are weighted for saliency like on the contrast model or for typicality like in the prototype model of categorization (discussed in Chapter 3). There needs to be a structure to the category, and features need to have relations. Otherwise, how are we to know that, for features of a dog, the wagging tail is usually not found on the head, or that it is not the dog's ears that do the barking?

A possible solution might be to consider relations between features to be features themselves. But this is a cumbersome solution because it would need to propose two types of feature representations and a mechanism for how relational-features and feature-features combine, creating a new binding problem. This brings us back to the "chicken and egg" problem for similarity and categorization that Hahn and Chater argue exists among theories in the literature. How can features of an object be bound together without relying on a circular definition of basing similarity on categorization?

2.3 Spatial Theory

Spatial theories of similarity represent objects as points in a psychological space, where similarity increases as distance decreases, and difference increases as distance increases (Shepard, 1987). Multidimensional scaling (MDS) can be used to generate this spatial representation of objects in human performance data (Shepard, 1962). Tversky (1977) argues against the spatial theory because its core assumptions are psychologically invalid based on the empirical evidence. For example, spatial theory assumes symmetry in similarity relations between objects. But human judgments show asymmetric relations, such as a child being more similar to his parents than vice versa. One by one, Tversky shows how the metric axioms of spatial theory are psychologically implausible to account for similarity in human cognition.

The spatial theory also suffers from the same limitations as Tversky's contrast model. There are no structural relations to link objects together in psychological space, and so it cannot be the sole basis for explaining categorization. Again, Hahn and Chater point out the "chicken and egg" problem with the spatial theory. It is difficult for this theory to explain how new objects find their position in psychological space, especially if an object has not yet had its properties analyzed, which requires categorization. And so a spatial theory of similarity would then be grounded in categorization, when typically the data suggest it is the other way.

Spatial theories are very powerful descriptive models because of their statistical technique of using MDS. However, where it has power in description, it lacks in explanation, particularly in how its proposed mechanism is implemented in biological systems such as us. And even with all of this descriptive power and a broad range of free parameters, spatial theories fall short in accounting for many of the empirical results discussed thus above. The goal of the current work is to strike a balance of descriptive *and* explanatory power.

2.4 Structural Alignment Theory

The structural alignment theory of similarity proposes that mental representations have structure, and that the similarity relations between those representations are based in their alignability (Markman & Gentner, 1993a, 1993b, 1993c). This theory has its roots in models of analogical reasoning (Gentner, 1983, 1989). It builds upon previous theories that consider similarity relations to only involve individual saliency relations between objects or features, but it adds the participation of representational structure as part of that matching system. One of the key predictions in this theory is rather counterintuitive. The structural alignment theory predicts that it is easier to find differences between similar pairs of objects than between dissimilar pairs of objects, and this prediction plays out accurately in the empirical data (Gentner & Markman, 1994). The effect of structure has also been demonstrated with the process of difference. Previous research has shown many asymmetries in similarity and difference judgments (Tversky, 1977), and one explanation is simply that similarity relations are based on common features while difference relations are based on distinctive features. But the structural alignment theory has been shown to handle the asymmetry in similarity and difference in a single model (Markman, 1996). Additional evidence suggests that structural alignment facilitates the detection of differences even in high-similarity pairs (Gentner & Gunn, 2001).

Structural alignment theory has been proposed to be a unifying principle for similarity, categorization, analogy, and metaphor. It improves upon the spatial and feature theories of similarity by adding organizing structure to objects. But in doing so it introduces yet another "chicken and egg" problem. Are objects similar because of their structure, or do they have structure because they are similar? How does one even learn these structures? The problem of learning is also linked to this theory's "chicken and egg" problem with categorization. As Keane and Costello (2001) argue, the structural alignment theory is unable to explain conceptual combination. So if the structural alignment theory cannot explain whether similarity is grounded in structure or structure is grounded in similarity, and the theory cannot explain how structure is learned, and the theory cannot explain conceptual combination, then it is creating more problems it is solving.

Medin et al. (1993) defend the structural alignment approach against attacks that claim the theory relies on *a priori* structures when it accounts for the empirical data. The features are not aligned by *a priori*, they say, but on a particular comparison being made. The comparison provides the constraint satisfaction for the alignment process, and the items being compared select and constrain the features that are activated or inferred. A criticism of the feature theory was that it does not explain how features are related. Here, Medin et al. say that the comparison process constrains both the features and their structure. At best, this claim simply removes the problem by one step. Having the "comparison process" constrain the features is no better than having "similarity" constrain the features. What is providing the structure used in the alignment process? The structural alignment approach has powerful descriptive accounts of the data, but it is underdeveloped in its explanatory power.

2.5 Discussion

Similarity is in a position of simultaneously being taken for granted and implicitly incorporated into nearly every phenomenon of cognition while also being disregarded and outright banished from a set of credible theories of cognition. It is a deceptively simple concept that appears to generate more questions than it provides answers. Certainly in this chapter we have seen the spectrum of views of the value of similarity as a construct in cognitive science.

There are many approaches to similarity, and all have their strengths and weaknesses in accounting for the wide range of empirical findings. In Chapter 4, we will explore what is the supramodal approach to similarity, and how similarity might be unified with categorization. The goal is to create a parsimonious explanation for how similarity appears to be operating as a central process in human cognition.

CHAPTER 3: CATEGORIZATION

3.1 Theories of Categorization

For much of the last 30 years, classification learning has been considered the central strategy to forming concepts (Barsalou, 1990; Chin-Parker & Ross, 2004; Estes, 1994; Kruschke, 1992). Yet outside the laboratory, people do not always learn categories as a main goal. People learn and use categories in support of some other goal (Brooks, 1999). Markman and Ross (2003) argued that traditional classification learning studies overemphasize explicit classifications, even though people often make implicit classifications outside the laboratory. For example, when using a subway token at a turnstile entrance, people do not stop and explicitly ask, “Is this a subway token or a button?” Rather, they implicitly classify the item as part of the main goal of commuting to another location.

In this light, classification learning can be split into (at least) two types: explicit classification and implicit classification. Explicit classification is when the act of categorizing an item is the main goal. For example, bird watching involves explicit classification in that a person identifies particular species for observation. Implicit classification is when the act of categorizing an item is in support of a separate main goal, and often we are attending more to this goal than to the act of categorization. Implicit classification is more similar to everyday categorization, like when people cook, drive, and play games, among other activities. For example, prior to typing these words, I was not aware of the moment I categorized my computer as such a device to be used for word processing. Rather, the goal of writing these words was explicit, and the supporting categorization processes were implicit.

Previous research has demonstrated that different learning strategies can lead to differences in acquired mental representation of novel categories (Chin-Parker & Ross, 2004). Several studies have also shown that differences in categorization and reasoning occur between groups of different culture (Lopez, Atran, Coley, Medin, & Smith, 1997; Medin, Ross, Atra, Cox, Coley, Proffitt, & Blok, 2006) and level of expertise (Medin, Lynch, Coley, & Atran, 1997).

In studies involving explicit and implicit categorization, it has been shown that implicit learners are more likely than explicit learners to claim that they discovered a single defining feature that perfectly predicted category membership even when none existed (Brooks, Squire-Graydon, & Wood, 2007). This "simpler than it is" phenomenon demonstrates that a person's belief about the nature of categories is not necessarily consistent with actual categorization behavior (Murphy, 2002). Differences in memory recollection also show that implicit classification learners acquire both diagnostic and non-diagnostic category information, while explicit classification learners acquire only diagnostic category information (Romano, 2006).

The current work, outlined in Chapter 5, will investigate these phenomena in categorization and similarity. For now, we will explore other theories of categorization and cover their strengths and weaknesses in accounting for the empirical data.

3.2 Classical Theory

The classical theory (Smith & Medin, 1981) has historically been the longest-lasting theory of categorization, having been traced back to Aristotle (384 – 322 BC) with its influence lasting until the 1970s (Murphy, 2002). While its influence in cognitive science is now minimal, the intuitiveness of the theory continues to pervade naïve introspection of category structures (Brooks, Squire-Graydon, & Wood, 2007), a telling indication of why it persisted for thousands of years. It is exactly why the classical theory continued unchallenged for so long, because it is so intuitive, scientists and philosophers incorporated it in their arguments as a base assumption without question (Lakoff, 1987).

In the classical theory, categories and concepts are represented by linguistic definitions, which list the necessary and sufficient conditions for the members of any category or concept. The claim to power of such a representational system is that every object can easily be labeled by the categories that it has membership, and in no cases can an object only partially belong to a category. Moreover, all members of each category must be equally good examples of that category, if each member necessarily has the required features of membership (Murphy, 2002). Carving up the world into discrete categories that are simply lists of necessary and sufficient conditions is a simplistically powerful method for organizing the world. However, when one begins to consider the effect of typicality in categorization, the classical theory collapses.

For example, the category “dog” in the classical theory might be defined by a list of necessary and sufficient conditions, such as “has four legs, barks, has fur, drools when hungry, wags its tail when happy, …” and so on. But what happens if an animal that has been classified as a dog loses a leg in a car accident, and becomes three-legged? Is it no longer a dog? What about a hairless dog, or a dog that had its tail removed, or one that doesn’t bark? And what happens when a breeder crosses a wolf and a dog, creating a wolf-dog hybrid? Lions and tigers do not breed in the wild, but they have been mated in captivity to produce a liger, a genetic lion-tiger hybrid. I can use a book by reading it, or I can place it under a hot tea cup to protect my desk, or use it as a hammer to hang a picture. When one considers atypical natural kinds such as unusual examples like hairless dogs or genetic hybrids, or atypical uses of artifacts such as transforming the function of a book into the function of a hammer, the classical theory of categorization falls apart.

But the typicality problems of the classical theory are not limited to the extreme cases of genetic hybridization or altered functions of artifacts. Consider whales and dolphins as members of the category “mammals”. Nearly all mammals are land animals, while these atypical examples live in the oceans and seas. Typicality effects are what destroyed the classical theory of categorization, as we will see below. Lists of defining necessary and sufficient conditions are unable to explain the graded nature of category membership, because definitions mandate that all category members are equally valid and representative (Murphy, 2002).

While there are many problems with the classical theory of categorization, it does not have a “chicken and egg” problem (Hahn & Chater, 1997) with similarity because the category representations in this view make no reference to similarity. The act of categorization is reduced to a process of identity, not similarity, because category membership is an all-or-nothing state; there can be no partial-membership, and hence similarity is useless in this regard. The classical theory is the most parsimonious representation of categories because it unifies categorization and similarity, but it is also the theory least able to account for the empirical data. Consequently, its theoretical relevance rarely extends beyond its role in setting up the historical context of the development of categorization research in cognitive science.

3.3 Prototype Theory

Eleanor Rosch is credited with pioneering the empirical study of categorization and for promoting it to become a central issue in cognitive science (Lakoff, 1987). Inspired by the limitations of the classical theory of categorization, Rosch and colleagues (Rosch & Mervis, 1975; Rosch, Simpson, & Miller, 1976) developed the prototype theory of categorization. Rather than represent each category with a linguistic definition that contains a list of the necessary and sufficient features of membership, the prototype theory represents categories as probabilistic prototypes. Prototypes are representations stored in one’s memory, and those representations list the features that are most typical of the members of a category. Each feature is weighted based on its probability of it occurring among the members of that category.

One of the key strengths of prototype theory is that, like in the classical view, categories are represented as lists of features. But unlike the classical view that considers features in absolute terms of necessity and sufficiency, prototypes are lists of probabilistic features. This allows for the high level of specification for category members found in the classical theory, while preserving the graded category membership observed in typicality effects (e.g., in the category "bird" pigeons are more typical than penguins, at least for people that live in New York City). In this balance, prototype theory borrows the intuitiveness of the classical theory, and improves upon it by fitting the theory to empirical data. It also fits many real world effects in categorization, such as biases toward high frequency members of categories and the fuzzy nature of category boundaries.

The process of categorization in the prototype theory involves the following steps. A new object is perceived, and it is broken down into its relevant features. The features of that object are then compared against the features of the category prototypes that person already knows, and a similarity rating is generated for each comparison. The category prototype that has the highest similarity rating wins, and the new object is classified as a member of that category. If all similarity ratings across all prototypes are below a certain threshold, then the new object might be considered an object of a new but unknown category.

For example, the prototype for the category "dog" might have the following probabilistic features:

Sheds its fur	0.8

Wags its tail	0.9
Barks	0.5
Hunts for truffles	0.1
Responds to a name	0.7
...	...
...	...
Born in live birth	1.0

Imagine you visit a friend’s home for the first time and you see she has a pet animal. In the prototype view, you would compare the features of that animal against the features of animal prototypes that you know. Perhaps the animal is barking at you, then the owner calls its name and offers it a treat, and it wags its tail. With these features, you would be more likely to classify the animal as a dog rather than a cat or a turtle. Like the classical theory, the prototype theory has a certain level of intuitiveness. If you are asked to imagine a bird, you will likely visualize a typical bird such as a robin or pigeon, not a penguin or an ostrich. This is because, as the argument goes, the most typical members of a category are best represented by the category prototype, while members at the fringe are less represented by the prototype. In fact, when prompted with a category label, participants find it easier to produce typical category members than atypical ones, presumably because typical members are better represented by the category prototype (Mervis, Catlin, & Rosch, 1976). Murphy (2002) calls this a summary representation because a category is defined by a list of probabilistic features, not a singular “best example”, and those features might even be contradictory (for example, the category “weapon” might have a strongly weighted “harmful” feature, but weaker weighted features such as “made of metal”, “made of stone”, “made of wood”, and so on).

Unlike the classical theory, which is grounded in the process of identity for classifying category members, the prototype theory is grounded in the process of similarity. The problem in this approach is that exactly how the process of similarity operates remains unspecified. In particular, how is a similarity rating generated when the features of a newly perceived object are compared against the features of category prototypes? Solving the problem of categorization by reducing it to similarity simply removes the problem by one step until the process of similarity is properly explained. The prototype theory greatly advanced cognitive science and accounts for a large amount of empirical data. But the fact that the theory is based on an unspecified process (similarity) begs the question of whether the theory is succeeding in its role of explaining categorization as a central phenomenon in human cognition.

3.4 Exemplar Theory

The exemplar theory dramatically departs from both the classical theory and prototype theory in that it does not attempt to have a single representation that best covers all members of a category (Murphy, 2002). Rather, the view proposed by Medin and Schaffer (1978) is that there are no definitions or probabilistic feature-list representations to categories; a category is merely all of the previously experienced members of that category. For example, your category “apple” is the collection of all of the individual apples you have experienced in your lifetime and labeled as “apple”. Each experience is an exemplar of an apple. The act of categorization, then, is matching a newly perceived object to groups of previously seen and categorized objects. If the new object is most similar to all the previously seen apples, then it is labeled as an apple.

In this way, there is no longer a need to find the “best” representation for a category, because category membership is defined by the individual members. The work of the definition is spread among many exemplars. Categorization in exemplar theory works in the opposite direction of the prototype and classical theories, because category membership is defined by the category members themselves, not by an abstract “best-fit” representation. The typicality information is implicit in the exemplars, and similarity utilizes that information.

As in the prototype theory, the exemplar theory depends on the notion of similarity in its calculations of category membership. Murphy (2002) explains that the first step in determining similarity between exemplars in the exemplar theory is to determine the features that are relevant in the comparison, and the second step is to determine the amount of similarity for the relevant features. For example, when deciding if an object is an apple or an orange, size is not as important as when deciding if an object is an apple or a watermelon, but color, shape, taste, and other features might be important across both comparisons. After the feature importance is ranked, similarity scores can be calculated.

Murphy points out that this can be problematic from an explanatory perspective if one wanted to know the influence of similarity of features apart from the influence of attention, because the exemplar model has no method for separating the similarity score from the amount of attention paid to each feature. He suggests a solution might be to choose stimuli that are known to be equally different in their feature dimensions. For example, similarity ratings can be collected across stimuli features to generate new stimuli with features that have psychologically equivalent differences among their features (shape, size, color, etc.). These psychologically equivalent distances stimuli can then be used in categorization experiments with new participants. Unfortunately, as Murphy explains, this trick still tells us nothing about a participant's attention to features during category learning, it only tells us the perceptual similarity of the stimuli.

Murphy's problem of separating similarity from attention is not so problematic under the supramodal theory. First, it is nonsensical to consider the similarity ratings between features that do not receive any attention during category learning, because we can only process the similarity of features that we attend to. So there is little need to separate similarity from attention, because any processing of similarity will be a direct result of attending to the features being compared. Second, when Murphy says, "but sometimes one does wish to know what the real similarities are, separately from knowing which dimensions subjects paid more attention to" (2002, pp. 57), it suggests that there are objective similarities out there in the world. The psychological distances of similarity that Murphy was so interested in identifying do not exist in a vacuum, and by definition cannot exist outside of human (or animal, computer, etc.) cognition.

As we have seen in the introduction to the supramodal theory, higher-level cognitive processes involving context and goal-driven behavior interact with lower-level perceptual processes to produce saliency maps. The salience of a feature is its attentional attractiveness, and a feature will only be processed in the cognitive system if it enters that system by means of attention when the salience of that feature passes a given threshold (either bottom-up, top-down, or a combination of the two).

The exemplar theory seems correct to think that similarity calculations ought to involve attention, but just like with the prototype theory, how similarity really works is less specified. Both theories can take a feature-based approach to similarity (Tversky, 1977), but they do not attempt to explain how different effects in similarity judgments affect those theories of categorization. Unlike the prototype theory, the exemplar theory can use multidimensional scaling to find similarities, but again this statistical method lacks explanatory power. So we are still at the “chicken and egg” problem discussed earlier. But in addition, now we have the fragmentation of similarity and attention. The exemplar theory has values for both attention (the relevance of a feature in a similarity comparison) and similarity (the degree of overlap between relevant features). This presents yet another “chicken and egg” problem. Are features relevant because they are similar? Or are features similar because they are relevant?

3.5 Theory Theory

The theory theory (here called the theory view, for simplicity) claims that categories are coherent because the background knowledge one has about the world provides the internal structure for concepts (Murphy & Medin, 1985). In this view, feature-based similarity is too flexible to explain conceptual coherence because stimulus context and experimental task introduce more free parameters than degrees of freedom, as Murphy and Medin claim, rendering similarity too flexible to be the basis for categorization. After casting aside similarity, Murphy and Medin propose two components of conceptual coherence. The first claims that categories are structured by causal relationships that are either empirical or driven by expectations or hypotheses. The second claims that a category's coherence is structured by its position in the complete knowledge base of an individual, not by the category's internal structure. After background knowledge has provided structure for categories, within- and between-category similarity is then considered to be simply a by-product of that structure.

Murphy and Medin admit that the theory view appears circular, because theories are composed of concepts, but then theories are argued to be the basis for concepts. But Murphy and Medin say that they do not propose to reduce categories to theory-based representations, but rather suggest a bidirectional relationship between categories and theories. In the same spirit that the supramodal theory attempts to dissolve the distinction between similarity and categorization, the theory view attempts to harmonize categories and theories in the same representation.

Later, Murphy (2002) acknowledges that the theory view cannot rely solely on background knowledge to explain categorization, and admits that there must be a learning mechanism involved. However, he confesses that there is not yet any empirical learning mechanism proposed for the theory view. This poses a major problem for this theory of categorization, because if a new category is explained only in terms of one's prior knowledge and beliefs about the world, then it begs the question about how conceptual development works in one's first years of life. In other words, how are the first categories learned in the absence of prior knowledge? A proposal of some innate knowledge is required.

Murphy and Medin argue that similarity has little role to play in structuring our categories. Hahn and Chater (1997) counter that this claim is misleading because even if similarity is constrained by our knowledge and theories about the world, then the phenomenon of similarity still exists. Therefore, it is not that the theory view demands no account of similarity in explaining concepts, but rather the theory view demands a better understanding of similarity in conceptual coherence.

3.6 Discussion

All contemporary theories of categorization discussed here fall victim to various problems in the relationship between similarity and categorization. In the prototype theory, there is insufficient explanation of the exact mechanism of how the level of similarity between an item's features and the features of a category prototype. While this view resolves the impossibilities of the classical view, it fails to ground categorization in similarity when similarity is left unexplained. But this lack is also found in the exemplar theory, with a similarity mechanism that is not fully explained. Moreover, the exemplar theory fragments similarity and attention in ways that are not very useful, especially when the goal is to understand similarity as a cognitive process, not an objective process that is “out there” in the world. Finally, the theory view has problems between similarity and categorization, similarity and theories, and categories and theories. Plus, the theory view does not even propose a learning method, leaving open the question of how one acquires new categories, and how one acquires the initial background knowledge / theories necessary for categorization.

The exemplar theory in particular has great descriptive power. Its statistical approach is able to account for a broad range of the empirical data from human behavioral studies. However, it lacks biological realism in its explanation. A more biologically-plausible model is McClelland and Roger’s (2003) parallel distributed processing (PDP) approach. This neural network approach allows for a learning algorithm to acquire categorical knowledge, and the resulting connections among those neurons are the category representation. This connectivity pattern allows for a state space representation, and similar concepts tend to be near each other in this space. The PDP approach has both descriptive and explanatory power, and the current work follows in this endeavor.

As theories of categorization have aimed to explain more of the empirical findings and phenomena involved in category research, it seems that theories of categorization have become increasingly problematic, creating more problems than they are solving. A more parsimonious solution would be to approach similarity, categorization, attention, and perception in a more integrated approach, such as the supramodal theory. In the next chapter, exactly how the supramodal theory explains similarity and categorization will be explored.

CHAPTER 4: SUPRAMODAL THEORY

4.1 Introduction

As we have seen, there are serious problems in the major theories of similarity and categorization. Both the prototype and exemplar theories of categorization rely on feature-based similarity calculations. But as we have seen in Tversky's (1977) contrast model, feature-based similarity relations do not have a structure that can bind those features together in a meaningful way. This means that a barber pole and a zebra might be more similar than a horse and a zebra, based on their "stripes" feature (Murphy & Medin, 1985).

The theory view of categorization attempts to compensate for the lack of structure among an object's features in competing theories by claiming that conceptual coherence is grounded in one's prior knowledge and theories about the world. But the proponents of this theory readily admit that there is no learning method proposed. And it suffers from the infinite regression problem, because if new categories are grounded by prior knowledge, then it must explain when during conceptual development are the first categories learned and how are they learned in the absence of prior knowledge? Some innate knowledge must first be specified to provide the grounding of all the experience-based concepts.

In a similar manner, the structural alignment theory of similarity attempts to resolve the lack of coherence in the spatial and feature-based similarity theories by adding an organizing structure for objects. This structure is proposed to facilitate the similarity and difference relations between objects. But just like the theory view in categorization, the structural alignment theory in similarity creates more problems than it solves. How are structures learned? What is the difference between the structure and its features, and how are these distinct properties represented? For example, why can't structure also be a feature? And most importantly, are objects similar because of their structure, or do they have structure because they are similar?

The grand claim of the structural alignment theory is to unify similarity, categorization, analogy, and metaphor in a coherent framework. Yet it ignores the most basic question that inspired the research from which it developed: What is similarity and how is it related to human cognition? If learning is not part of the explanation, then it is an insufficient theory. And if circularity cannot be avoided, as in the structure vs. similarity causal loop, then the structural alignment theory offers no real solution.

The theories discussed so far all fall within the functionalist paradigm of cognitive science. And it is argued here that this is precisely why they all have problems with circularity. Rather than grounding these theories in the mind-body-environment relationship, they are based on a computer metaphor of mind. Cognitive processes are proposed as centralized computational modules, such as an attentional system that is somehow separate from processing similarity, a structural mapping of objects that is separate from categorization, and knowledge that is separate from learning. Any coherent theory of cognition should be able to explain how attention and cognition are linked through perception and action. That is why, in the current work, cognition will be explored under the embodied paradigm of cognitive science.

4.2 Unifying Similarity and Categorization

It is at this point that the supramodal theory is proposed to unify similarity and categorization by reimagining some of the most basic elements of both theories. Any theory of similarity and/or categorization will need to incorporate attention, perception, learning, and action. For the purposes of the current work, attention is not an action, it is a potential. Or perhaps a better explanation for attention is that it is a virtual action. Attention occurs when the hills of the topographic saliency maps extend above a certain threshold, thereby creating attractors for the real motor actions (e.g., eye fixations). This threshold is task- and context-dependent.

Saliency maps are generated by the mixture of bottom-up perceptual systems and top-down conceptual systems, both of which are grounded in sensorimotor systems. The perceptual and conceptual systems are generated by the body and its relationship with its environment. In this way we can use a single representational form – saliency maps – to talk about attention, perception, conceptualization, action, the cognitive entity, and its environment.

But already here we have a distinction between perceptual and conceptual. These constructs can be problematic in any theory. In the supramodal theory, ultimately all saliency maps have their origin in the sensorimotor systems, which include the sensory modalities and bodily action. This means that both perceptual and conceptual saliency maps are generated from the sensorimotor forces in the interface of an embodied entity with an environment. This runs the risk of sounding like saliency maps (and thus cognition) can only be bottom-up. But as we have seen in Chapter 1, saliency maps can be bottom-up, top-down, and a mix of the two.

In the supramodal theory, it is useful to think of the perceptual systems as carrying information in its saliency maps that originates from the spatially and temporally local environment. In other words, it is useful to think of perception as information coming from the immediate environment into the cognitive system. And it is useful to think of the conceptual systems as carrying information in its saliency maps that originates from any combination of the transient activation and the long term memory of the cognitive entity. And the information of conceptual saliency maps in working and long term memory has its origin in the spatially and temporally local environments of the past. In other words, it is useful to think of conceptualization as information originating from sensorimotor experiences of the past; experiences that once were but are no longer local to the cognitive entity. While perceptions are the local experiences of the present.

Saliency maps from the perceptual systems contain information originating in the environment, and the saliency of that information can still be influenced by the top-down processes of the conceptual systems. Perception is not a purely bottom-up process. But the top-down processes of the conceptual systems should not be considered something abstract and computational like in functionalism. Rather, the top-down processes of the conceptual systems are saliency maps generated from the activation of long term memory (the patterns of connections among neurons), and those patterns originate from transient neural activity. Finally, transient activation is the interaction of the bottom-up and top-down processes of perception and conceptualization, both of which are grounded in sensorimotor experience. The consequence of this interpretation of perception and conceptualization is that all cognitive processes are embodied, and cognitive information is purely sensorimotor in origin. Long term memories are sensorimotor experiences of the past that can influence and alter the sensorimotor experiences of the current transient activity.

Therefore, the supramodal theory unifies attention, perception, action, and cognition within a single framework that can be represented both descriptively and mechanistically. The descriptive representation has already been given. The mechanistic explanation is based on neural network theory. A full mechanistic account can be found in Chapter 5, in the neural network model of similarity.

Emergent properties may be created when saliency maps collide. The supramodal maps may have features not present in the maps that helped to create them. Phenomena in cognitive psychology that are usually called "conceptual" will often have emergent properties. Consider the category "furniture". Try to imagine a piece of furniture without visualizing a subordinate category like chair, desk, table, sofa, and so on. It is impossible to visualize the category "furniture" because it is a supramodal concept. The category "furniture" is not simply visual, and it is probably more tactile in nature than any of the other senses. But in our minds, how do we imagine the somatosensory experiences of lying on a bed, relaxing on a sofa, arms resting on a desk, or working in a chair?

The category “furniture” contains emergent properties that go above and beyond the visual and somatosensory experiences of the individual subcategories and exemplars within those subcategories. The visual and somatosensory saliency maps collide to generate the superordinate category “furniture”. This can be likened to the McGurk effect, where visual and auditory perceptions collide to create the perception of a phoneme that did not exist in either of the two signals. Our abstract categories can also be thought of as sensorimotor signals that collide and mix with memory (which, again, are simply stored sensorimotor signals and/or results of their combinations).

What occurs during these collisions is that the maps (say, a visual map and an auditory map) are integrated and what emerges is a supra-map. A supra-map can be supramodal if it combines more than one modality, or it can be unimodal, in that vision is a supra-map of multiple visual component maps (e.g., contours, brightness, orientation, motion, etc.). Collisions of maps occur when their transient activations combine and integrate to form new maps with new transient neural activations.

There is still the question of how the topography of a saliency map is formed. Ultimately the perceptual/conceptual interactions played out over a time series need to be constrained. Several factors can help with this. Our embodied situation naturally makes certain things salient to us, such as a flashing light and motion. Our knowledge can also make things salient, like patterns of flashing light translating to Morse Code.

Our goals promote the salience of objects and features, such as noticing a Post Office when there is a need to mail a letter, and not noticing it on other days when there are no mailing tasks on the agenda. The constraints on salience are a result of both the environment and the body of the cognitive entity (that is, its sensory organs and central nervous system), and how the two relate with each other. This means that salience is necessarily subjective process, albeit based on objective properties, and as we will see, this is critically relevant to any theory of similarity and categorization.

Goodman (1972) was correct in his attack on objective similarity, that it is useless until we state the frame of reference for a similarity relation. And Hahn and Chater (1997) were correct in saying that Goodman's critique on similarity is meaningless because we should care about similarity between cognitive representations of objects, not similarity between objects as they exist in the world. And while representations are grounded in sensorimotor experiences of a body existing in and interacting with an objective world, those representations are subjective to that cognitive entity. Simply not existing in the same exact location in space and time will necessarily generate two distinct and therefore subjective saliency maps for two people looking at the same object because there will be a multitude of differences in their perceptions, such as lighting and shadowing differences that might alter the perception of brightness or hue for that object.

Often we need to remind ourselves that cognitive science should not be interested in asking how objects and features are similar or salient as they exist "out there" in the world. We are interested in how they are similar or salient in our representations of those objects. And if those representations, both current (transient activity) and past (long term memory), are based on sensorimotor processes, it allows for both the connection to and grounding in the objective world, but also an emergent feedback loop onto itself. Saliency maps from the past can influence ones in the present. This feedback loop provides the coherence for cognition and action because of the mechanics of the body and the emergence of behavior.

Objects and features in the world can become salient because they are naturally attractive to our sensory systems (e.g., motion), and they can also become salient because our feedback processes are naturally attracting the objects and features to the focus of sensory systems (e.g., searching for the “You are here” marker of a map). Consequently, objects and features are always considered representations in the supramodal theory, and are not referred to by their objective physical nature. Thus, objects and features can attract the focus of our sensory systems (via bottom-up processes), but also our sensory systems can attract the focus of the objects and features (via top-down processes). But in the end, it is always the sensorimotor system doing the attracting. For example, the visual system is naturally attracted to features of brightness or motion, and our goal-directed behavior is guided by cognitive processes that are themselves grounded in sensorimotor experiences of the present and past.

Let us consider the visual search task described in the introduction. In a single-feature search, the target item is immediately detected among distractors in a pop-out effect. The target object and its features are so different from those around it that it naturally attracts our vision and “pops out” to us. The target’s salience (attention) is so much higher than everything around it that our eyes cannot help but fixate on it. In a conjunctive-feature search, it becomes more difficult to find the target. We require our online cognitive processes to increase the saliency of the target item’s features and to decrease the saliency of the distractor items’ features. Our goal-directed behavior is guiding vision to make certain objects and features more salient, and in doing so, attracting those objects and features to the focus of vision. Bottom-up salience occurs when a mental representation of a physical object in the world attracts the focus of one or more sensory modalities. Top-down salience occurs when one or more sensory modalities attract the features of the mental representation of an object in the world to our focus.

Goodman (1972) was correct; we do need to think of similarity as needing a frame of reference in the same way relative motion in physics needs a frame of reference. But why should we think that objects in the physical world are doing all the attracting? Why not also have subjective processes also attracting physical objects and features? In a mental space landscape of continuously flowing saliency maps, the collisions of maps should affect each other and sometimes even generate a new emergent map with new properties and features. So to be more specific, let us again consider the visual search task with distractor items.

Imagine you are looking for a green horizontal bar among a field of red and green bars that are both vertical and horizontal. To find the green horizontal bar, you must do a conjunctive search, and eliminate all the red horizontal bars and all the green vertical bars until you just have the green horizontal bar. If a top-down map flowed from the prefrontal cortex to the vision and motor action brain regions, telling it to look for “green” and “horizontal”, then that map would start attracting those features to the focus of vision. The topography of visual saliency maps would begin to change such that objects with one or both of those features would begin to heighten and form salience hills. Eventually in this dynamic process the object with both features would receive the highest amount of saliency and it would be attracted to the goal-based saliency map from prefrontal cortex. The top-down, goal-based saliency map attracts the bottom-up visual saliency map, and when they collide the goal concludes. Presumably that goal-based map would dissolve if enough time passed without finding the visual target.

So the problem of attention is resolved in thinking of cognition as a mental space of continuously flowing saliency maps that interact with each other to mutually excite and inhibit the features among those maps. Perceptual and cognitive processes are maps in mental space, and objects in the environment are also maps in mental space. This allows for one common space to describe all bottom-up and top-down interactions involved in attention that are really a combination of both the objective environment and the subjective mind.

Now that we have accounted for Goodman’s critiques of similarity, we can explain how the supramodal theory explains similarity. Imagine that you have the task of deciding whether two objects are similar. A saliency map for the task goal is created, and maps of the objects are created. The goal map collides with the two object maps, and the bottom-up features that are naturally salient (such as brightness) will attract attention, while the top-down goal map will attract attention to features relevant to the goal (such as shape, size, and color) and increase the salience of those features. In this sense, features are drawn up the topographic hill by the combination of the goal map and object maps, and are modulated by saliency.

What qualifies as a feature depends on both bottom-up processes (such as the biology of our sensory organs) and top-down processes (such as the task and context). For example, we might consider a face to be a single feature when looking for a friend in a crowd. Or we might break up that holistic nature of a face into its components, such as in trying to identify identical twins based on a minor difference in nose tilt. It may be more appropriate to capture the essence of the definition of a feature with a philosophical description. For the current work, features emerge from the combination of the environment, our biology, and our goals and actions.

The identity of two objects can be defined as sharing equivalent feature sets. Similarity, then, can be defined as partial identity, or partially sharing feature sets. We measure this partial-sharing by integrating the common features of the two objects being compared, just as in the contrast model (Tversky, 1977). For example, a cat and a dog share many common features (such as having four legs, tails, fur, etc.), and they also have many distinctive features (such as eating habits, the sounds they make, social behavior, etc.). In determining how similar a cat is to a dog, the supramodal theory proposes that each animal has its own saliency map, and those maps collide with each other (and with the goal map of the similarity judgment) to create an integrated map. If the integrated map can sufficiently activate the neural population codes that represented the cat and dog, then the two animals are similar. That is, the combination of the two animals overlaps sufficiently the population of neurons that encoded for the original stimuli. This account of similarity will be fully established and demonstrated in the next chapter.

This leads us into the process of categorization. Categories are collections of maps in long term memory; collections that are based on similarity. Because similarity can be applied to all kinds of maps, including supramodal maps, this allows for a broad range of categories from the concrete to the abstract. For example, when a child sees a wolf for the first time he might categorize it as a dog because its features will likely activate population codes of category members of “dog”. When the child sees enough wolves, the collection of wolf saliency maps in long term memory will accumulate to the point that the diagnostic features start spreading into different population codes between the “dog” category and the emerging “wolf” category. That is, over time the child will acquire more features of wolves that are different from dogs. As the number of distinctive features grows, the amount of perceived similarity will decrease (as long as the similarity threshold remains constant). When the overlap of population codes is below threshold, they break off into two distinct categories.

A key prediction, based on this idea of being above or below a similarity threshold, is that similarity should not be affected very much by category boundaries. That is, if there is sufficient overlap in neuronal population codes, then two objects are similar, and otherwise they are not. This may seem counterintuitive, especially to many of the competing theories discussed in earlier chapters. But in the following chapter we will see empirical results that support this prediction. This prediction also allows room for spontaneous, goal-related categories that are created dynamically, and rule-based categories.

New categories are created when members of a specific category have too much incoherence in their population codes. In other words, similarity is based on population codes, and when there is not enough overlap in similarity within a category, the category will fracture into two or more categories and organize in a way that maximizes similarity by economizing the shared population codes within and between each category. Superordinate categories are created in the opposite manner. When there are between category similarities (when different categories start sharing some population codes but not enough to merge together) then a superordinate category might emerge to allow generalization to flow more efficiently between those categories.

4.3 Discussion

The supramodal theory of cognition is proposed to unify some of the core phenomena of cognitive science, including similarity and categorization. The idea is simple: cognitive processes are saliency maps that can have bidirectional influence on one another. This theory is primarily concerned with how the world becomes represented subjectively in mental state space, not necessarily how the world is “out there” objectively. In this way, salience (attention) should be explained in terms of its frame of reference. An object in the world might have strong bottom-up salience, and thereby attract the focus of vision. But goal directed behavior can increase the salience of certain features, and thereby attracting features to the focus of vision. The frame of reference, therefore, is truly relative, meaning that it can original in either a bottom-up or top-down process or a combination of the two.

Saliency maps exist physically in the brain, in the form of transient activation patterns across a neural network (biological or artificial), or in the form of long term memory in the weighted connections between neurons. Activated population codes represent these salient features, and when the same feature activates the same population code, they can be considered identical. This provides us with a relatively concrete definition of identity. And if we define similarity as partial identity, we are beginning to ground similarity as a mechanistic account. If two objects have shared population codes, then they share features. If they share enough features to pass a task-dependent threshold, then they are considered similar.

Categories organize members by maximizing the amount of shared population codes (similarity) while minimizing the differences. Categories can be supramodal in that different modalities can contribute to the composition of that category. Many of our higher level categories can be considered supramodal, like furniture, animal, vehicle, flower, etc. But even more basic level categories such as bird and dog are supramodal. We know what they look like, sound like, feel like, smell like, and sometimes, taste like. And because the supramodal map loses much information when it is projected down into a single modality, we can never really visualize any of our more abstract categories, such as furniture, without drawing upon a concrete instance of that abstract category. The consequence of this theory is that embodied experience is central to the phenomena of similarity and categorization.

The need to study and understand the similarities and differences between human cognition and the one that will emerge from artificial sources (or other animal species) becomes critical in our understanding of what is cognition. This will fundamentally shape our understanding of embodied cognition more than anything else, because only then will we truly be faced with the problem of not being able to share concepts with a cognitive entity that has a body so radically different from our own. What is the threshold for differences that result in incompatible concepts in one cognitive embodied agent vs. another? How will we communicate with these entities? Certainly human cognition is helping to shape one that might emerge from robotic and virtual agents. But ultimately, whether through our similarities or our differences, those emergent agents will be a mirror for us to understand what is unique to human cognition, and what may be universal to all cognition.

CHAPTER 5: SUPRAMODAL MODEL OF VISUAL SIMILARITY

5.1 Experiment 1: Judgments of Similarity and Difference

Introduction

Similarity is a process that is central to human cognition. It facilitates the formation of concepts and categories by associating commonalities among objects; to transfer knowledge from one domain to another; to create metaphors and analogies; to make inferences and predictions; and to prime and cue memory and behavior. Similarity as a cognitive phenomenon is strongly connected to both high- and low-level processes. But despite its relevance by connectedness to so many theories, similarity remains a poorly understood aspect of human cognition.

Several models provide accounts of the role that similarity has in human cognition (e.g., Edelman, 1998; Shepard, 1962). One of the most influential is the contrast model (Tversky, 1977). In this model, objects are represented as collections of features, and the process of similarity is reduced to one of matching features between objects. Between any pair of objects, either perceptual or conceptual, features are weighted for salience and importance, and the common features are contrasted with the distinctive features, and a similarity rating is calculated. A key claim in the contrast model is that similarity increases linearly as the number of common features increases and the number of distinctive features decreases. The contrast model is the inspiration for the connectionist model in the next section.

The contrast model has been criticized as being unrealistic through over simplification. For example, a barber pole and a zebra would be more similar than a horse and a zebra if certain features, such as black and white stripes, were weighted as more salient or important (Murphy & Medin, 1985). And because salience and importance are highly context-dependent, there are no unique answers to how similar one object is to another. As a result, Murphy and Medin argue that similarity becomes unrealistically flexible to explain conceptual coherence because “there are more free parameters than degrees of freedom” (p. 292). In addition to being too flexible, the contrast model fails to capture how features are related to one another (Medin, Goldstone, & Gentner, 1993).

In the earlier chapters, we explored in depth many theories of similarity and their strengths and weaknesses. No theory has really become a leader within the field of cognitive science. Due to this situation, the purpose of this study is to investigate the processes of similarity and difference in visual cognition to collect empirical data that is useful to compare with other theories. A deliberate effort was made to study similarity in visual objects that normally have no associated background knowledge, so that it could be better understood how similarity operates for perceptual objects without conceptual influence.

An effort was also made to remove the influence of environmental context as much as possible. The goal of the experiment was to understand whether and how perception helps to constrain similarity and difference judgments. Of course, this does not preclude the possibility that context can interact with perception, and in fact it is expected that it does in many cases. But the current work seeks to understand similarity as a phenomenon by starting the investigation at a low-level and adding additional influences, such as context and category boundaries, in additional studies. This strategy will support an incremental development of an explanatory model of similarity that grows over time to account for more complex empirical data as the phenomena involved become higher-level.

To facilitate this goal of studying the constraints of low-level perception on similarity judgments, no definitions of similarity or difference were given to participants, nor were participants informed of any expectations of right or wrong answers. In fact, they were told that there were no right or wrong answers, and that they should respond based on their own judgments. In understanding how perceptual similarity and difference operate within these conditions, it may be possible to understand how these two phenomena are constrained at the perceptual level. When it becomes possible to isolate the constraints of these processes at the perceptual level, we can better understand the extra effects and interactions of conceptual and contextual influences on the perception of similarity and difference. The current study aims to understand similarity as an embodied phenomenon that is naturally constrained by our perceptual systems.

In particular, key predictions of the contrast model (Tversky, 1977) will be tested, and a neural network model will be constructed to test those claims. The criticism of how features are related to one another and bound together (Murphy & Medin, 1985) will also be explored in the model. Other explanations of similarity, such as the Structural Alignment Model (Markman & Gentner, 1993a, 1993b, 1993c), propose solutions to these problems. However, there is also evidence that the structural alignment theory, just like the contrast model, is unable to explain conceptual combination (Keane & Costello, 2001). The necessity of structural relations in resolving the binding problem will be also explored in the proposed model.

The complement to the cognitive process of similarity is the ability to perceive differences. Indeed, it should be expected that when comparing any two objects, decreasing the amount of perceived similarity should proportionally increase the amount of perceived difference. However, it has been found that similarity and difference are not always inversely related (Medin, Goldstone, & Gentner, 1990; Estes & Hasson, 2004). This non-inversion effect creates explanatory problems for many models of similarity. The models discussed in Chapter 2 do not fully explain the non-inversion effect of similarity and difference. Nor do the models explain or predict how the perceptual system constrains similarity. These models, and the experiments that test their predictions, have generated an excellent framework for understanding similarity. However, they do not do an adequate job of controlling for effects of prior knowledge, environmental context, or what, if any, differences exist in judgments of similarity between perceptual information and conceptual information.

Experiment

In this experiment, participants were presented with pairs of visual stimuli and were prompted to make judgments of the perceived similarity or difference of each pair. Two types of feature-dimensions were used in the visual stimuli in this experiment. The set of object pairs always had the same abstract structure in feature-space, but the feature dimensions were manipulated between participants. The goal of this experiment was to investigate whether the perception of similarity and difference varies between these different types of feature-dimensions.

The first type was separable dimension stimuli (Shepard, Hovland, & Jenkins, 1961). An object with separable dimensions has features that are able to be analyzed independently of one another. For example, an object's shape, size, and color can all be independently manipulated without affecting the other dimensions. One can imagine a small blue coffee mug, and then imagine the same mug only larger, or maybe red. Changing the color of the mug does not affect the perception of its size and shape.

Contrasting with separable dimension stimuli, the second type of feature dimension was integral dimension stimuli. An object with integral dimensions has features that cannot be analyzed independently of one another. For example, a color is defined by its hue, saturation, and brightness. But changing one of those features influences how the other two features are perceived in the integral whole (the color). Thus, for integral dimension stimuli, features cannot be changed independently of one another.

Method

Participants. A total of 80 undergraduate students from the University of California at Merced participated in the experiment for course credit.

Materials. The stimuli in the experiment were pairs of objects presented on a computer monitor. The stimuli were designed to measure the effects of feature dimension types as well as the degree of feature overlap between objects in a pair. Each object in the pair had 3 dimensions consisting of binary features. Pairs of objects had either all separable or all integral feature dimensions. The binary features between the objects in a pair could be manipulated such that the pair was either completely identical, more similar / less different, less similar / more different, or completely different. Numerically, this is translated as 100% identical, 67% similar, 33% similar, or 0% similar. The 3 binary feature dimensions for object pairs with separable dimensions (Figure 5.1.1) were shape (square or triangle), size (large or small), and color (blue or orange), from Shepard, Hovland, & Jenkins (1961). The binary features for object pairs with integral dimensions (Figure 5.1.2) were hue, saturation, and brightness, from Nosofsky & Palmeri (1996).

Design. Each participant was assigned to 1 of 2 groups, either seeing only separable stimuli or only integral stimuli. This created a between-participants factor of stimuli type. There were 40 participants in each group.

Procedure. The experiment consisted of multiple trials of the same task. In each trial, a participant was presented with a pair of objects and a question about that pair, which prompted a judgment response of the similarity or difference of that pair.

There were 2 blocks of trials. Between the 2 blocks, the set of object pairs was held constant, but the question varied. In one block the question was “Are the two objects similar?”, and in the other block the question was “Are the two objects different?” The question order between blocks was randomized, and the presentation order of object pairs within each block was also randomized. Each trial within a block repeated the same question but with different pairs of objects. In this way, participants were asked to judge the similarity and difference for each object pair.

For each trial, upon presentation of an object pair and a question about that pair’s similarity/difference, participants were prompted to respond either “Yes” or “No” by pressing one of two buttons on the keyboard. Participants were asked to respond as fast as possible, but to also be as accurate as possible. Instructions were elaborated that participants should give themselves enough time to read each question and look at each pair of objects before responding, but to still respond as fast as possible.

Before starting the experiment, it was explained to each participant what types of objects they would be viewing. Participants in the separable dimensions group were shown the array of objects of different shapes, sizes, and colors. Participants in the integral dimensions group were shown the array of objects of different hue, saturation, and brightness. This was done so that participants understood the range of the stimuli in similarity-space from the moment they were first asked to make similarity and difference judgments.

Participants were never instructed on any definition of similarity or difference. If participants asked for a definition of similarity or difference, or for more elaboration on any expectations of right or wrong answers, the participants were told that there were no right or wrong answers in the experiment, and they were asked to simply respond based on what they understood those definitions to be from their own experience. It was a deliberate choice to not instruct participants on definitions of similarity and difference. This experiment was investigating how constrained is the perception of similarity and difference between participants, and whether there would be a high or low degree of reliability in their judgments. Each object in every pair had 3 dimensions with binary features.

Results and Discussion

The data of interest were the responses to each stimulus pair for each question. There were 4 types of feature overlap (100% similar, 67% similar, 33% similar, and 0% similar), and 2 types of questions (similarity and difference). The two groups being compared are those participants that viewed separable-dimension stimuli vs. those participants that viewed integral-dimension stimuli.

Several factors are created from this data set. There is a between-subjects factor of stimuli type (separable vs. integral). Within-subjects, there is the type of question being asked (is the pair similar vs. is the pair different), and there is the degree of feature overlap in a pair of objects for each trial (100%, 67%, 33% or 0%).

The following analyses use the proportion of yes-responses for each conjunction of feature overlap and question type. However, the yes-responses for questions of difference were inverted. When an object pair is 100% similar, it should be expected that participants would respond 100% "Yes" for questions of similarity, and 0% "Yes" for questions of difference, simply because those pairs are completely identical. This expected inverse relationship needed to be represented in the data for comparative analyses between questions. Therefore, in the following analyses, the yes-responses for questions of difference were transformed by (1 – response). In this way, when the patterns of responses are graphed for questions of similarity and difference, the values are commensurable and should overlap in the graph instead of displaying their inverse relationship. It also permits a visual observation of the non-inversion effect, described in the current experiment's introduction.

The distinctiveness in the patterns of perceived similarity and difference between separable (Figure 5.1.3) and integral (Figure 5.1.4) dimensions is quite apparent. For both types of feature dimensions, as the number of overlapping features decreases, the perception of similarity decreases while the perception of difference increases. However, the rate and pattern of this decrease / increase for similarity / difference varies between types of feature dimensions.

Object pairs with separable dimensions tended to have a more gradual change in the perception of similarity and difference. Object pairs with integral dimensions tended to have a rather dichotomous perception of similarity and difference (e.g., pairs were either similar or not similar, with little variation). This observation may be indicative of the more holistic nature of the visual processing of integral dimensions. Specifically, changing just one feature can disrupt the whole object and alter the perception of the other two features even when they remain unchanged physically.

Another potential explanation is that participants have more experience with color categorizations than with the objects used in the separable-features condition. This is a valid criticism, and future work will need to tease apart this confound of perceptual constraints and expertise effects. Additionally, the dichotomous ratings in the integral group may be due to participants not being able to select out individual features, thereby reducing the efficacy of comparing overlapping features. But this explanation of the result is still compatible with the goal of the study, which was to investigate the constraints on judgments of similarity.

A repeated-measures analysis of variance (ANOVA) was performed, with stimulus dimension type (separable vs. integral) as the between-subjects factor. There were 2 within-subjects factors: the degree of feature overlap of stimulus pairs (100%, 67%, 33%, 0%), and the question that was asked for each stimulus pair (similar or different). The results showed a significant main effect of feature overlap, $F(3, 234) = 446.441$, $p < .001$; a significant main effect of question-type, $F(1, 234) = 5.590$, $p < .05$; and a significant interaction of feature-overlap and question-type, $F(3, 234) = 8.914$, $p < .001$.

Between-participants there was a significant interaction of dimension type and feature overlap, $F(3, 234) = 11.996$, $p < .001$; and a significant interaction of dimension type and feature overlap and question type, $F(3, 234) = 4.638$, $p < .01$. The interaction of dimension type and question type did not reach statistical significance, *p* = .063. The main effect of feature overlap shows that decreasing the number of common features between objects in a pair decreases the perception of similarity while increasing the perception of difference, consistent with Tversky's (1977) contrast model.

The significant interaction of feature overlap and question type is observable in Figure 5.1.3, replicating the non-inversion effect (Medin, Goldstone, & Gentner, 1990). When stimulus pairs are completely matching (100% feature overlap) or completely mismatching (0% feature overlap), there appears to be agreement in response proportions between questions of similarity and difference such that the non-inversion effect disappears. However, when stimulus pairs are only partially similar (67% or 33% feature overlap), response proportions appear to be disproportionate between questions of similarity and difference, such that these stimulus pairs underwent a greater increase in perceived difference than their decrease in perceived similarity. Future work will need to explore the effect of the task itself on the non-inversion effect. For example, would we get the same effect if we used the word "dissimilar" rather than "different" in the question participants received?

Follow-up independent samples t-tests, with an adjusted α = .625E-2, showed a significant difference between dimension types for questions of similarity at 67% feature overlap, $t(78) = 3.562$, $p = .001$; and for questions of similarity at 33% feature overlap, $t(78) = 5.561$, $p < .001$. All other t-tests comparing the conjunction of feature overlap and question type between groups were not significant, $p >$ .625E-2.

The results of these t-tests suggest that the main difference between the separable / integral groups was in the perception of similarity for object pairs with partially overlapping features (67% and 33% overlap). Participants that viewed separable-dimension stimuli were much more likely to judge partially-overlapping features as being similar than were participants that viewed integral-dimension stimuli, and participants viewing separable stimuli preserved the non-inversion effect for partially-overlapping features while the integral-dimension group did not.

One-sample t-tests were performed with an adjusted α = .625E-2, and a test value of .5. The purpose was to test whether participants' response proportions differed significantly from chance. Only participants that viewed separable dimension, partially overlapping pairs (67% or 33% feature overlap) during questions of similarity responded at chance p = .166 (67% overlap), p = .173 (33% overlap). All other combinations of dimension-type, feature-overlap, and question-type were significantly different from chance, p < .001.

The results of these t-tests suggest something distinct about separable-features from integral-features in the perception of similarity, specifically when the pairs of objects have both common and distinctive features. Tversky's contrast model does not account for this finding. To investigate this issue in more depth, 2 separate repeated-measures ANOVAs were performed on the two types of stimuli, both with 2 within-subjects factors: the degree of feature overlap, and the question type.

In the first repeated-measures ANOVA, response proportions from only the separable-dimension participants were analyzed. The results showed a significant main effect of feature overlap, *F*(3, 117) = 169.420, *p* < .001; a significant main effect of question type, *F*(1, 117) = 11.915, *p* < .01; and a significant interaction of feature overlap and question type, *F*(3, 117) = 16.288, *p* < .001. In the second repeated-measures ANOVA, response proportions from only the integral-dimension participants were analyzed. The results showed a significant main effect of feature overlap, *F*(3, 117) = 321.967, *p* < .001; no main effect of question type, *p* = .764; and no interaction, *p* = .351.

The results of these two ANOVAs suggest important distinctions in the perception of similarity and difference between separable and integral dimension stimuli. Participants that viewed separable-dimension stimuli perceived the similarity and differences between pairs of objects in such a way that resulted in the non-inversion effect. The non-inversion effect appears to have taken place while viewing partially similar stimulus pairs (pairs with 67% or 33% feature overlap). However, participants that viewed integral-dimension stimuli did not display the non-inversion of similarity and difference.

General Discussion

The main finding of this study suggests that the visual perception system helps to constrain judgments of similarity and difference. Different patterns of judgments resulted between groups that viewed separable- or integral-features. As has been mentioned, separable features can be analyzed independently of one another, while integral features cannot (changing one feature alters the perception of the other two features). This finding suggests that Goodman’s (1972) and Murphy and Medin’s (1985) critiques of similarity are not as threatening as once believed. They argued that similarity is too unconstrained and cannot serve as a basis for other cognitive processes such as concepts and categories. Here, however, we offer support for a perceptual constraint on similarity.

The main distinctions between the separable and integral groups occurred with partially-overlapping stimuli. Specifically, pairs that overlapped 67% or 33% yielded significant differences between groups, while pairs with 100% or 0% overlap did not. This is a very reasonable result and should be expected. At the extremes of 100% (completely identical objects in the pair) and 0% (completely different objects in the pair), the appropriate response should be fairly obvious regardless of the type of stimuli that was viewed. But the partially-overlapping pairs allowed for the independent and non-independent nature of the features between separable and integral features, respectively, to become evident in the data.

Consider a partially-overlapping integral pair of objects. At 67% overlap, the pair has 2 features in common and 1 feature that is different. But this 1 feature that is different will affect the perception of the other 2 features, even if they are in common. A single distinctive feature will affect the integral whole of the objects in the pair. And in fact, the judgment of similarity for partially-overlapping pairs is significantly lower for integral pairs than separable pairs. This is because changing 1 feature in a separable pair will do little to affect the perception of the other 2 features. Each feature can be analyzed independently, so the mismatch does little to affect the whole object.

It may be possible to generate a separable stimuli set that captures the results seen in the integral stimuli group of the current experiment. The separable stimuli in the current example were balanced for saliency (Shepard, Hovland, & Jenkins, 1961), but one can imagine a situation where saliency was manipulated such that the features that received a lot of attention may be more independent than those that received little attention and blend together. Also, faces can be simultaneously considered separable stimuli and integral stimuli, in that we often perceive them holistically but we are also able to attend to specific features. Future work is needed to explore the effect of salience.

Tversky's (1977) contrast model predicts the basic findings of this experiment, that similarity increases as the number of common features increases, and decreases as the number of distinctive features increases. The assumptions of that model predict that, "similarity between objects is a linear function of the measure of their common features" (Tversky, 1977, pp. 345). But this linear trend was not found in the results of either group, suggesting that the contrast model might not be a full account of feature-based similarity. This prediction will be tested with the neural network model in the next section.

Finally, participants in the separable stimuli group replicated the non-inversion effect found in other similarity studies. The judgments of similarity and difference were not proportional with each other, such that when added together they did not sum to 1.0. For example, for any given pair of objects, if a person perceives the pair to be 70% similar, the pair ought to be 30% different. However, people over-valued their difference judgments relative to their similarity judgments for separable stimuli, demonstrating the non-inversion effect. But the integral group did not replicate the non-inversion effect. This further suggests that there is a qualitative distinction in processing these two types of stimuli in the perception of visual similarity and difference.

5.2 Modeling Visual Similarity

A connectionist model was developed to explore possible mechanisms for the experimental results and to make predictions for future studies. The current model attempts to replicate human performance data for similarity judgments collected in Experiment 1. The results of both the human data and the model are compared with predictions of the contrast model (Tversky, 1977) from which the current model was based.

The current version of this model uses the supervised learning algorithm of backpropagation (Rumelhart & McClelland, 1986). There were 3 layers in the network: input, hidden, and output. The input layer had 6 units which represented all possible features; there were 3 dimensions and 2 features per dimension. The hidden layer had 3 units, representing the 3 dimensions that all 6 features filtered through. The output layer had 8 units; 1 unit for each possible object. The units in the hidden layer and output layer had associated bias weights. The network's layers were fully connected with randomized initial weights and biases. It used a learning rate of .1 and an alpha of .9.

The network was trained on a set of 8 input and output patterns. Each input pattern was a vector of features that represented a single object. If the object had a feature, the associated feature-unit in the input layer received an input of ‘1’. If the feature was absent, the input for that unit was ‘0’. Each output pattern was a vector of objects, with each unit representing a single stimulus object. The network was trained to receive a set of features and to associate those features with a unique output unit. The learning criterion was reached when the mean squared error reached the threshold of 0.01 or lower. A learning rate of 0.1 was used, with a momentum of 0.9.

The network easily learned to associate distinct sets of features with unique output units, essentially becoming a features-to-object classifier. After reaching the learning criterion, the network was tested on its ability to make judgments of similarity. Due to its training, the network was already capable of making judgments of identity. That is, the network received a set of features and correctly activated the output unit that represented those features. But a method of judging similarity needed to be developed.

The process for judging the similarity between two objects in a pair was computed in the following steps. First, two input vectors, each containing the set of features representing an object in a pair, were normalized. An integration vector was then calculated based on the type of feature-dimensions in the object pair (see Eq. 1 & 2). The integration vector was normalized, and then weighted based on the number of dimensions of the object pair. Finally, the integration vector was used as an input vector for the trained network previously described. The integration vector was fed through the hidden layer, and the output layer's activations were recorded. This essentially allowed us to present two objects simultaneously (instead of just one object as done in training) to the fully trained network.

The output unit with the maximum activation was selected in a winner-take-all fashion. The network declares that 2 input vectors are similar if two qualifications are met. The first qualification is that the most active output unit matches one of the 2 original input vectors that were used to create the integration vector. That is, based on its prior training, the features of at least 1 of the 2 input vectors correspond exactly to the most active output unit. The second qualification is that the most active output unit has an activation ≥ 0.75.

For example, in the learning phase, the network learned to associate feature vector F_1 with output unit U_1, and feature vector F_2 with output unit U_2. An integration vector is calculated from F_1 and F_2, and fed into the network's input layer. If either U_1 or U_2 has an activation ≥ 0.75, then the input vectors are treated as similar. In other words, if the combined integration vector is able to substantially activate at least one of the two objects being compared, then the two objects are considered similar. Just like with the human responses, the network's similarity judgments for stimulus pairs are simple "Yes/No" responses.

There were two integration equations used in this network, with one equation for each dimension-type: separable and integral. Tversky's (1977) contrast model prescribes that similarity calculations ought to contrast common features with distinctive features in a comparison. Following this claim, we test the theory that separable features can be analyzed independently, unlike integral features. Therefore, Equation 1 factors out the distinctive features in the similarity calculation for separable stimuli, while Equation 2 integrates all features, regardless if they are common or distinctive between objects in the comparison.

The integration equation for separable stimuli was:

EQ. 1 $I_{1,2} = F_1 + F_2 – \mathrm{abs}(F_1 – F_2)$

where $I_{1,2}$ is the integration vector of objects 1 and 2, F_1 is the feature vector of object 1, F_2 is the feature vector of object 2, and abs(X) is the absolute value of X.

The integration equation for integral stimuli was:

EQ. 2 $I_{1,2} = F_1 + F_2$

where $I_{1,2}$ is the integration vector of objects 1 and 2, F_1 is the feature vector of object 1, F_2 is the feature vector of object 2.

Because of the assumption that separable features are able to be processed independently, Equation 1 integrates the features in common but factors out the distinctive features. In this way, the integration equation puts attention weights solely on the features common to both objects in the stimulus pair. And because of the assumption that integral features are unable to be processed independently, Equation 2 integrates all the features, both common and distinctive, between the two input vectors. In this way, the integration equation puts attention weights on all features of both objects in the stimulus pair.

The results of the model can be seen in Figures 5.2.1 (separable features) and 5.2.2 (integral features). The network is able to produce similarity judgments that match the human data for both types of feature-dimensions. The data set produced by the model was collected and averaged over 40 repeats to match the number of human participants. The network produced the step function response pattern for separable features seen in the human data, as well as the asymptotic response pattern for integral features seen in the human data. In addition to producing the qualitative pattern of human responses, the model is able to closely replicate the approximate quantitative values for averaged response proportions.

A key part of the network is that it does not learn anything specifically relating to separable or integral features. In fact, all it learns is an abstract feature set of binary features (0, 1). In other words, the network simply learns to integrate collections of abstract binary feature sets into meaningful whole objects in the output. It is in the integration vector, derived after learning is complete, that does all the work of distinguishing separable from integral feature integration.

Experiment 1 and the model simulation suggest that different types of feature dimensions in visual perception can result in different patterns of responses for judgments of similarity, even when holding constant the abstract structure of the stimuli. When viewing object pairs with separable dimensions, people compare only the common features by factoring out the distinctive features. This is consistent with the belief that separable features can be analyzed and manipulated independently of one another. However, when viewing object pairs with integral dimensions, people are believed to be unable to separate common and distinctive features. This is consistent with the view that integral features are not independent but processed holistically. The current model and its integration equations for different dimension types provide a mechanistic account of these findings.

Tversky's (1977) contrast model predicts the most basic findings of the current model. As the number of common features increases, so does the perceived similarity. And as the number of distinctive features increases, so does the perceived difference. A key claim in the contrast model is that similarity is a linear function of the number of common features. However, the rate at which similarity decreases as distinctive features increase (or vice versa, the rate at which difference increases as the common features decrease) is not consistent between groups. That is, similarity has a much steeper decline for integral pairs than for separable pairs. And neither group shows linear changes in their similarity judgments.

While the contrast model does not accurately predict the results of Experiment 1 and the ability of the current model to fit the data, it did provide a crucial insight in the design of the current model. In the contrast model, objects are represented as collections of features, and similarity is reduced to the process of matching features between objects. Between a pair of objects, the common features are contrasted with the distinctive features, and a similarity rating is calculated. The integration equation used in the current model for separable stimuli did precisely what the contrast model prescribes. Despite that, both the human data and the model's data produced results that go against what the contrast model predicts: that similarity increases linearly as the number of common features increases and the number of distinctive features decreases. Moreover, the contrast model does very poorly in explaining both human and model judgments of similarity for integral stimuli pairs. First, people are unable to factor out distinctive features, as the contrast model prescribes. Second, the pattern of changes is not linear as the contrast model predicts.

The current model accounts for the empirical data without the need for structural relations among the objects' features. The structural alignment model argues for relations among features to resolve the binding problem. However, the network learns to integrate the features at the input layer into a coherent object representation at the output layer, without needing to learn any explicit structural relations among those features. While the structural alignment model may have its usefulness in explaining similarity relations among more conceptual, abstract, or rule-based objects, the current study demonstrates it is possible to account for humans' similarity judgments without them. This lends to a more parsimonious solution to the problem of representing relations of features, because was discussed in Chapter 2, it is not fully explained how structures are actually learned in the structural alignment model.

The non-linear structure of both the human brain and the current connectionist model produces non-linear similarity responses, even as the number of common and distinctive features changes linearly. As a result, both systems can integrate objects in similarity judgments precisely as the contrast model instructs, but they do not produce the output that the contrast model predicts. Feature-based models, such as the contrast model, need to be updated to reflect the effects that qualitatively different types of feature dimensions have on the perception of similarity and difference. The non-independent nature of integral dimensions might be corrected for by considering them to be linked by structural relations. A change in one integral feature has an impact on the other features within an object. Such models also need to be adapted to account for the non-linear dynamics of human cognition (Spivey, 2007).

The current model is able to match extremely well to the human data. It accomplishes this by integrating the features of objects into a blended object and passing that feature vector through the network. If it is able to sufficiently activate one of the two output units whose features were used to create the integration vector, then those two objects are considered to be similar. In other words, similarity can be considered to be partial identity. Because this network is a features-to-object classifier, it essentially is processing object identity based on the observed features.

Identity is binary, either two objects are identical or they are not. Similarity, however, can be graded. The graded nature of similarity is implemented by the population codes of units that translate the collections of features into coherent objects. If enough of the population codes are excited by an integration vector, then the activation of appropriate output units increases past a threshold for saying “yes” to the question of similarity. This threshold can be task dependent, meaning that goals and context-driven behavior can raise or lower the threshold for judging similarity between two objects. But in the end, it will be a process of evaluating the proportion of overlapping among population codes that represent the objects being evaluated.

Reducing similarity to partial identity resolves many of the theoretical complexities that cause other theories to be problematic. In the current model, the integration vector is constrained by the visual perceptual system. Pairs of objects with separable features can integrate only their common features and disregard distinctive features, while pairs of objects with integral features are forced to integrate all features regardless of their commonality and distinctiveness. Allowing the integration equation to be constrained by our embodied perceptual systems resolves Goodman's (1972) critique of similarity.

Yet the current model also allows room for context and goals to help choose with what respects two objects are similar (Medin, Goldstone, & Gentner, 1993). The threshold value allows for a continuum in determining exactly what proportion of partial identity is necessary to judge two objects as being similar. The balance between the integration equation and similarity threshold provides parsimony for the current model in that it formalizes the natural constraints of the biology of our sensory organs with the cognitive flexibility needed in accounting for the varied effects of different contexts and goal-driven behaviors.

Grounding similarity in partial identity is an easy to understand description for a highly flexible and often abstract phenomenon. While other models and theories propose abstract geometric comparisons, linear combinations of features, multiple primitives such as features and their organizing structures, the current model proposes that similarity is simply the degree of overlap between the population codes of biological neurons or artificial units that encode for objects. Mechanistically, similarity as partial identity is understood in terms of the network implementation. Descriptively, similarity as partial identity is understood in terms of overlapping mental state spaces. Any theory in cognitive science ought to provide both descriptive and mechanistic explanations that are parsimonious and easily communicated. The current work continues that tradition in exploring similarity as partial identity for simple visual objects.

The results of Experiment 1 and the current model are both well accounted for under the supramodal theory. The integration vector can be considered as the result of two saliency maps that collide with each other. Which input vector is selected (separable or integral) would depend on some prior, currently unspecified layer. A biologically-plausible layer that parses visual features, such as area V1 in the visual cortex parsing out contours, is a likely candidate. Each input vector is a map of the salient features (the features that enter the network's input layer), and the integration vector is their combination. Treating similarity as partial activation of overlapping population codes allows for a host of different types of similarity judgments, from the unimodal to the supramodal. Also the saliency of the features is a combination of the biology of the visual system and the context of the current goal-driven behavior. The supramodal theory can simultaneously serve as both a useful and simple description of the observed data, and as a basis for explaining the mechanism and implementation of similarity as partial identity in the overlap of population codes of neurons.

CHAPTER 6: SIMILARITY AND CATEGORIZATION

6.1 Experiment 1: Separable Features

Introduction

In Chapter 5, we explored whether the perceptual system can constrain the cognitive phenomena of similarity and difference. Two types of feature dimensions were explored. Separable features are ones that can be analyzed independently of one another, such as an object’s shape, size, and color. Integral features are ones that cannot be analyzed independently of one another, such as a color’s hue, saturation, and brightness.

The human visual system processes these distinct feature types differently, and consequently the cognitive processes that utilize those features will process them differently. This includes our judgments of similarity and difference. It was concluded that the perceptual system helps to constrain our judgments of similarity, and that the cognitive process of similarity is not a wild and free, unbounded and therefore irrelevant process, as some researchers have feared (Goodman, 1972; Murphy & Medin, 1985). Rather, similarity is bounded, at least in part, by perception, and it is easily modeled by a relatively simple neural network model. People did not exhibit an infinite number of degrees of freedom, and their judgments were very accurately predicted by the model.

The conclusion in Chapter 5 was that similarity can be defined as the partial overlap of populations of neurons that encode for the two objects being compared. If the integration of the two saliency maps is able to sufficiently activate the populations of neurons that generated that integration map, then the two objects are similar. If the integration falls below the threshold, then the objects are not similar. The threshold can change based on context and online task demands. This allows for both a structured organization (in terms of encoding of information in a neural network) and a certain degree of flexibility based on the current situation. And a particular strength of this model is that this flexibility is extremely parsimonious, in that all the flexibility arises from just a single variable, reducing the degrees of freedom of the model. The result is that we can have a predictable yet still flexible process of similarity.

If Chapter 5 sought to investigate whether the perceptual system constrains similarity, the current chapter seeks to investigate how similarity is related to categorization. Of particular interest is whether category boundaries affect the perception of similarity. For example, if we have a set of 10 items, and 5 belong to Category A and the other 5 belong to Category B, does our perception of similarity (both within-category and between-categories) change as a result of learning that category boundary that divides the 10 items? Or is visual similarity largely unaffected by category boundaries and conceptual knowledge?

Experiment

The current experiment had 3 phases: a pre-test, a learning phase, and a post-test. The pre- and post-tests were replications of the experiment described in Chapter 5. In these testing phases, participants were presented with pairs of visual stimuli and were prompted to make judgments of the perceived similarity or difference of each pair. Participants in this experiment only viewed stimuli with separable dimensions. The goal of this experiment was to investigate whether the perception of similarity and difference changes between pre- and post-tests as a result of learning category boundaries.

The stimuli in this experiment had separable features along the dimensions of shape, size, and color. The features along each dimension varied continuously across 4 equidistant points, which allowed for dimensions of 4 possible features, rather than binary features in Chapter 5. However, each participant only saw 2 of these possible features for each dimension, so in effect they learned binary features.

The purpose of having 4 possible and equidistant features for each dimension was to create a between-subjects factor of category distance. That is, participants can either learn categories with prototypes that are close to each other, or categories with prototypes that are far from each other. Imagine we have 4 category prototypes, consisting of linearly separable feature sets (prototypes A, B, C, and D). These 4 prototypes are equidistant from each other, such that the distance between A & B is the same as the distance between B & C, and C & D. In this way, we can have two groups of participants. One group can learn near categories (prototypes B & C), and another group can learn far categories (prototypes A & D).

This arrangement allows us to explore not only whether category boundaries affect the perception of similarity and difference, but also whether the distance of categories (the inherent similarity between categories) also affects those perceptions.

Of particular interest is whether learning close category boundaries will distort perceptions of similarity more than far category boundaries. If two categories are very close, there should inherently be a higher degree of similarity among members of both categories than for two far categories. And the other question of interest is whether there will be divergent results for within-category pairs of objects vs. between-categories pairs of objects. Specifically, will members within a single category increase their perceived similarity, while members between two categories decrease their perceived similarity?

Method

Participants. A total of 30 undergraduate students from the University of California at Merced participated in the experiment for course credit.

Materials. The stimuli in the experiment were pairs of objects presented on a computer monitor. The stimuli were designed to measure the effects of the degree of feature overlap between objects in a pair, as well as distance between category boundaries. Each object in the pair had 3 dimensions consisting of binary features. The binary features could either be close to (see Figure 6.1.1) or far from (see Figure 6.1.2) each other in terms of between-category members. All objects had separable dimensions of shape, size, and color (Shepard, Hovland, & Jenkins 1961). The binary features between the objects in a pair could be manipulated such that the pair was either completely identical, more similar / less different, less similar / more different, or completely different. Numerically, this is translated as 100% identical, 67% similar, 33% similar, or 0% similar.

Design. Each participant was assigned to 1 of 2 groups, either seeing objects that were far from each other, or seeing objects that were close to each other, based on the features along each dimension. This created a between-participants factor of category distance. There were 15 participants in each group.

Procedure. The experiment had 3 phases: a pre-test, a learning phase, and a post-test. The pre-test and the post-test were exactly the same, and they both consisted of multiple trials of the same task. In each trial, a participant was presented with a pair of objects and a question about that pair, which prompted a judgment response of the similarity or difference of that pair.

There were 2 blocks of trials in these testing phases. Between the 2 blocks, the set of object pairs was held constant, but the question varied. In one block the question was “Are the two objects similar?”, and in the other block the question was “Are the two objects different?” The question order between blocks was randomized, and the presentation order of object pairs within each block was also randomized. Each trial within a block repeated the same question but with different pairs of objects. In this way, participants were asked to judge the similarity and difference for each object pair.

For each trial, upon presentation of an object pair and a question about that pair’s similarity/difference, participants were prompted to respond either “Yes” or “No” by clicking on either box on the screen with the mouse. Participants were asked to give themselves enough time to read each question and look at each pair of objects before responding, but to still respond as fast as possible. Both the participant's explicit response ("yes" or "no") and implicit response (their mouse-tracked trajectory) were recorded for each trial.

Mousetracking involves recording the X-Y pixel coordinates of the monitor as a participant moves the mouse to make a selection (Dale, Kehoe, & Spivey, 2007; Dale, Roche, Snyder, & McCall, 2008; Spivey, Grosjean, & Knoblich, 2005). When there are two alternatives that a participant must choose between, the amount of curvature of the mouse trajectory is an inferred measure of the amount of competition occurring during the time course of that online perception-action event. If the mouse trajectory is relatively direct, such that the trajectory shows little curving toward the competing (and non-selected) option, then it is inferred that there was little competition between those options during that time course.

Before starting the experiment, it was explained to each participant what types of objects they would be viewing. This was done so that participants understood the range of the stimuli in similarity-space from the moment they were first asked to make similarity and difference judgments. Participants were never instructed on any definition of similarity or difference. If participants asked for a definition of similarity or difference, or for more elaboration on any expectations of right or wrong answers, the participants were told that there were no right or wrong answers in the experiment, and they were asked to simply respond based on what they understood those definitions to be from their own experience.

It was a deliberate choice to not instruct participants on definitions of similarity and difference. This experiment was investigating how constrained is the perception of similarity and difference between participants, and whether there would be a high or low degree of reliability in their judgments. Each object in every pair had 3 dimensions with binary features.

In the learning phase, participants learned to categorize the objects they saw in the testing phases into two categories. This was done with a standard classification learning design (Markman & Ross, 2003). In each trial during the learning phase, participants saw only one object, and they were asked to decide if the object belonged to Category A ("Lokad") or Category B ("Koozle"). After making a choice, the participant received feedback whether they were correct or incorrect. Upon meeting the learning criterion, participants then proceeded to the post-test. The learning criterion was met with a minimum of 4 blocks and a maximum of 30 blocks, with 2 consecutive blocks of 90% correct responses per block.

In this design, we were able to collect participants' perceptions of similarity and difference of the pairs of objects in the pre-test without the influence of category knowledge. And then in the post-test, we can compare whether perceptions changed as a result of learning category boundaries. Because both explicit and implicit measures were recorded, we can explore both aspects of the perception. Both the explicit judgment and the implicit mouse movement trajectory allow for two views of the same process, where the explicit judgment focuses on the end result and the mouse trajectory focuses on the process of that judgment over time. Mousetracking has been shown to be an effective measure for understanding the time course of a cognitive process, in that cognitive processes that involve competition leak into motor action and influence the curvature of the mouse trajectory (Dale, Kehoe, & Spivey, 2007; Dale, Roche, Snyder, & McCall, 2008).

Results and Discussion – Explicit Responses

The data of interest were the responses to each stimulus pair for each question in the pre- and post-tests. There were 4 types of feature overlap (100% similar, 67% similar, 33% similar, and 0% similar), and 2 types of questions (similarity and difference). The two groups being compared are those participants that learned close categories vs. those participants that learned far categories. In addition, responses between the two testing phases were compared. The results can be viewed in Figures 6.1.3 – 6.1.10.

Several factors were created from this data set. There is a between-subjects factor of category distance (close vs. far). Within-subjects, there is the test phase (pre- vs. post-test), there is the type of category comparison (within category or between categories), there is the type of question being asked (is the pair similar vs. is the pair different), and there is the degree of feature overlap in a pair of objects for each trial (100%, 67%, 33% or 0%).

The following analyses use the proportion of yes-responses for each conjunction of category distance, test phase, category membership, question type, and feature overlap. However, the yes-responses for questions of difference were inverted. When an object pair is 100% similar, it should be expected that participants would respond 100% “Yes” for questions of similarity, and 0% “Yes” for questions of difference, simply because those pairs are completely identical. This expected inverse relationship needed to be represented in the data for comparative analyses between questions. Therefore, in the following analyses, the yes-responses for questions of difference were transformed by (1 – response). In this way, when the patterns of responses are graphed for questions of similarity and difference, the values are commensurable and should overlap in the graph instead of displaying their inverse relationship. It also permits a visual observation of the non-inversion effect. The non-inversion effect occurs when there is a disproportionate response between similarity judgments and difference judgments.

A repeated measures analysis of variance (ANOVA) was performed, with Category Distance as the between-subjects factor (near or far categories). There were 4 within-subjects factors: Learning (pre-test vs. post-test), Category Membership (within or between categories), Question (similar or different), and Feature Overlap (100%, 67%, 33%, or 0%). Results showed a significant main effect of Category Membership, $F(1, 28) = 598.392$, $p < .001$; a significant main effect of Question, $F(1, 28) = 9.417$, $p < .01$; a significant main effect of Feature Overlap, $F(3, 84) = 55.690$, $p < .001$; a significant interaction of Feature Overlap and Category Distance, $F(3, 84) = 5.171$, $p < .01$; a significant interaction of Learning and Question, $F(1, 28) = 5.852$, $p < .05$; a significant interaction of Category Membership and Feature Overlap, $F(3, 84) = 1.151\text{E}3$, $p < .001$; a significant interaction of Question and Feature Overlap, $F(3, 84) = 6.568$, $p < .001$; a significant interaction of Learning and Question and Feature

Overlap, *F*(3, 84) = 4.340, *p* < .01; and an approaching significant interaction of Learning and Question and Feature Overlap and Category Distance, *F*(3, 84) = 2.600, *p* = .057. All other effects and interactions were not significant.

To take a deeper look into these main effects and interactions, follow-up ANOVAs were performed with fewer factors in each analysis. In the following ANOVAs, we considered looking only at within-category comparisons, and split up the data between groups (near and far categories).

In the first ANOVA, we looked at the near category group. There were 3 within-subjects factors: Learning, Question Type, and Feature Overlap. There was an approaching significant main effect of Question, *F*(1, 14) = 3.791, *p* = .072; a significant main effect of Feature Overlap, *F*(1, 42) = 114.413, *p* < .001; a significant interaction of Learning and Question, *F*(1, 14) = 5.199, *p* < .05; and a significant interaction of Learning and Question and Feature Overlap, *F*(3, 42) = 4.198, *p* < .05. All other effects and interactions were not significant.

In the second ANOVA, we looked at the far category group. There were 3 within-subjects factors: Learning, Question Type, and Feature Overlap. There was a significant main effect of Question, *F*(1, 14) = 7.858, *p* < .05; a significant main effect of Feature Overlap, *F*(3, 42) = 100.083, *p* < .001; and a significant interaction of Question and Feature Overlap, *F*(3, 42) = 5.647, *p* < .01. All other effects and interactions were not significant.

The results of these ANOVAs suggest that learning category boundaries affected the perception of similarity and difference for participants that learned near categories. There was a significant interaction involving Learning for both Question and Feature Overlap, which implies learning category boundaries is influencing how participants are responding. However, for participants that learned far categories, there was no significant effects or interactions related to the Learning factor, suggesting that learning categories did not influence the perceptions of similarity and difference for this group.

To understand the source of the variance, more ANOVAs were computed with even fewer factors. This time, only 2 within-subjects factors were used: Learning and Feature Overlap.

The first ANOVA looked at near categories, within category comparisons, for questions of similarity. There was a significant main effect of Feature Overlap, $F(3, 42) = 32.487$, $p < .001$; and no other significant effects or interactions.

The second ANOVA looked at far categories, within category comparisons, for questions of similarity. Again there was a significant main effect of Feature Overlap, $F(3, 42) = 66.293$, $p < .001$; and no other significant effects or interactions.

The third ANOVA looked at near categories, between category comparisons, for questions of similarity. Again there was a significant main effect of Feature Overlap, $F(3, 42) = 7.391$, $p < .001$; and no other significant effects or interactions.

The fourth ANOVA looked at far categories, between category comparisons, for questions of similarity. Again there was a significant main effect of Feature Overlap, $F(3, 42) = 34.728$, $p < .001$; and no other significant effects or interactions.

The result of these four ANOVAs is that perception of similarity does not seem to be affected by learning category boundaries. That is, there was no main effect or interaction of Learning on any of the analyses. But in the earlier, larger analyses we found an effect of Test Phase for participants that learned near categories. This suggests that we ought to look for the source of that variance for questions of difference, not for questions of similarity. So we repeat the previous four ANOVAs, this time for questions of difference.

The first ANOVA looked at near categories, within category comparisons, for questions of difference. There was a significant main effect of Learning, *F*(1, 14) = 8.682, *p* < .05; a significant main effect of Feature Overlap, *F*(3, 42) = 157.979, *p* < .001; and a significant interaction of Learning and Feature Overlap, *F*(3, 42) = 5.021, *p* < .01.

The second ANOVA looked at far categories, within category comparisons, for questions of difference. There was an approaching significant main effect of Learning, *F*(1, 14) = 4.317, *p* = .057; a significant main effect of Feature Overlap, *F*(3, 42) = 75.436, *p* < .001; but no interaction.

The third ANOVA looked at near categories, between category comparisons, for questions of difference. There was an approaching significant main effect of Learning, *F*(1, 14) = 3.827, *p* = .071; a significant main effect of Feature Overlap, *F*(3, 42) = 8.591, *p* < .001; and a significant interaction between Learning and Feature Overlap, *F*(3, 42) = 4.850, *p* < .01.

The fourth ANOVA looked at far categories, between category comparisons, for questions of difference. There a significant main effect of Feature Overlap, *F*(3, 42) = 8.591, *p* < .001; and no other significant effects or interactions.

The results of these four ANOVAs suggest that the main source of variance due to learning categories occurs for people learning near categories and in questions of difference. That is, when participants learn two categories that are near each other, it influences their perception of the differences in pairs of objects. Specifically, learning categories that are near each other caused an increase in the perception of differences of object pairs between the pre- and post-tests. There was a significant increase in difference perceptions after learning when both objects in a pair were members of the same category.

The results of participants' explicit responses suggest that learning category boundaries does not appear to affect the perception of similarity for either near or far categories. However, learning category boundaries does appear to affect perceptions of difference. See Table 6.1.1 for a summary of the effects. For both near and far category boundaries, there was a common trend to increase perceptions of difference between objects in the pair after learning categories. The trend was significant for near categories in within- and between-category comparisons, and approaching significance for far categories in within-category comparisons.

Perhaps this result is due to the nature of the category learning task in this experiment. It has been found that classification learning tasks tend to focus attention on diagnostic features (Chin-Parker & Ross, 2004; Markman & Ross, 2003), and biases the encoded category representation in memory to emphasize those features (Romano, 2006). In the case of this experiment, when categories are close together, the learning task appears to increase the perceived differences between objects in a pair. So if the nature of the classification learning task is to emphasize the diagnostic features, then it makes sense that learning a category boundary between two near categories affects the perception of difference, in that it increases the perceived differences. Close categories are naturally more similar than far categories. And if a person wants to learn to separate those objects into two categories, attention must be paid to the diagnostic features, which in turn would bias perception to focus on differences between objects.

Results and Discussion – Implicit Responses

In the previous section we explored participants' explicit responses when questioned about their perceptions of similarity and difference for various pairs of objects. In this section we consider their implicit responses, measured with mousetracking. Mouse movement trajectories have been demonstrated to provide informative data which indicates the level of competition during the execution of an online cognitive process (Dale, Kehoe, & Spivey, 2007; Dale, Roche, Snyder, & McCall, 2008).

For example, imagine there are two images at the top left and top right of a computer monitor. One is an image of a candle, and one is an image of a candy. The participant's task is to use the mouse to click on the picture of the word s/he hears. If the word stimulus has initial phonemes that match both competitor options (such as candle and candy), then those options will be highly competitive. But if the competitor options were not very competitive initially, such as having ending phonemes that match (such as candle and pickle), then there would not be much competition at the onset of the trial. This is an example of attractor dynamics of two competing choices involved in language processing that leaks into motor output (Spivey, Grosjean, & Knoblich, 2005).

In fact, when a participant is presented with choices that have high competition, the cognitive process of settling into an appropriate choice leaks into the motor output of clicking on an option with the mouse, and the trajectory of that mouse movement becomes more curved toward the competitor. However, when competition is low, there is little curvature in the trajectory, as the participant makes a more direct motion to the choice. In this way, we have an implicit measure of how much competition is occurring during the process of making a decision. Whereas with the explicit responses, we only get the final output of that decision, not the process over the time course of the decision.

In the case of this experiment, we are interested to know if learning category boundaries affects mouse movement trajectories. Specifically, we can test the hypothesis that learning category boundaries will reduce the amount of competition between objects such that categorical knowledge helps to structure the mental organization of these objects. Also, we can test whether we get divergent results between explicit and implicit responses. In the case of explicit responses, learning influenced only the perception of difference, not similarity, and then only for learning near categories. But again, this is only considering the final, end product of the decision making process. Perhaps during that process, there was a change in the level of perceived competitiveness as a result of learning categories, which would in turn affect mouse movements.

For each trial, we can extract two key pieces of information from the mouse movement trajectory. First, we can calculate the total area under the curve, which gives us an indication of how much competition, non-competition, or even repulsion there was between objects in each trial. Recall that mouse trajectories can vary in how direct vs. how curved was the motion from the starting position to the selection during the time course of the trial. If there was a lot of competition, the participant will ultimately choose a response, but the mouse trajectory will have curved toward the other response during that process. When there is no competition the trajectory ought to be a direct line from the initial position to the response. And when there is repulsion the mouse trajectory will curve away from the competing object. The other dependent variable that we get from mouse tracking is the maximum deviation of the curve. This informs us of the highest peak of the curved mouse trajectory.

Area under the curve and maximum deviation are calculated in a number of steps. First, trajectories are translated into a coordinate system such that the origin of the mouse trajectory is at the coordinate (0, 0). When a trajectory landed on the response option on the left of the screen, that trajectory was reversed horizontally so that they could be compared equivalently with trajectories that curved to the right. All trajectories were normalized and interpolated to be standardized as 50 point vectors

(Dale, Kehoe, & Spivey, 2007; Dale, Roche, Snyder, & McCall, 2008; Spivey, Grosjean, & Knoblich, 2005). This allowed for equivalent comparisons of all trajectories from all trials, and for the computation of area under the curve and maximum deviation.

Looking at Figures 6.1.11 – 6.1.18, which all show area under the curve for the mouse movement trajectories, we can compare the pre- and post-tests and evaluate the effect of learning category boundaries in regard to a variety of factors. All conditions involve separable stimuli, and we can examine the data based on category boundary distance, the type of question, and whether we are comparing within or between categories. A cursory view of the figures suggests that there is no clear-cut story, and that the data is very noisy. A reasonable hypothesis is that learning category boundaries should help to reduce competition between alternatives, thereby reducing curvature in the mouse trajectories. But the results show that sometimes learning results in less curvature in the post-test, and sometimes it results in more. The standard errors around the means between pre- and post-test values overlap in nearly all conditions, suggesting a very noisy sampling of the data.

Recall that for the explicit responses, the significant effects of learning occurred for questions of difference involving near categories, where the perception of differences increased as a result of learning. When examining Figures 6.1.12 and 6.1.14, which plot area under the curve for those same conditions that gave rise to significant explicit distinctions, we notice that there is no hint of implicit effects in the mouse trajectories as a result of learning. Both the means are rather close, and the standard errors bleed together. If there is an effect to be found, the current data contain too much noise to discern it.

Although there is a lack of visible differences among the figures that graph the implicit results (Figures 6.1.11 – 6.1.18), a quantitative view may yield more subtle effects. A repeated measures ANOVA was performed, with Category Distance as the between-subjects factor (near vs. far categories). There were 4 within-subjects factors: Learning (pre-test vs. post-test), Category Membership (within vs. between categories), Question (similar vs. different), and Feature Overlap (100%, 67%, 33%, 0%). Results showed no significant main effects or interactions for any combination of factors, all $p > .05$. This result shows that there is no overall trend to be generalized for implicit measures of mousetracking.

To take a deeper look, follow-up ANOVAs were performed with few factors in each analysis. In the following ANOVAs, we considered looking only at within-category comparisons, and split up the data between groups (near and far categories).

In the first ANOVA, we looked at the near category group for within-category comparisons. There were 3 within-subjects factors: Learning, Question, and Feature Overlap. Results showed no significant main effects or interactions for any combination of factors, all $p > .05$.

In the second ANOVA, we looked at the near category group for between-category comparisons. There were 3 within-subjects factors: Learning, Question, and Feature Overlap. Results showed no significant main effects or interactions for any combination of factors, all $p > .05$.

In the third ANOVA, we looked at the far category group for within-category comparisons. There were 3 within-subjects factors: Learning, Question, and Feature Overlap. Results showed no significant main effects or interactions for any combination of factors, all *p* > .05.

In the fourth ANOVA, we looked at the far category group for between-category comparisons. There were 3 within-subjects factors: Learning, Question, and Feature Overlap. Results showed no significant main effects or interactions for any combination of factors, all *p* > .05.

Results of these four analyses suggest that there really is no effect of learning on implicit measures of mousetracking trajectories when considering area under the curve. We performed even more ANOVAs, this time with only 2 within-subjects factors: Learning and Feature Overlap. For all possible combinations of analyses, no significant effects or interactions were discovered, *p* > .05.

What we can conclude is that learning category boundaries does not appear to have an effect on the motor output of making similarity and difference judgments. For questions of similarity, matches the explicit response data, where we found that category learning did not affect explicit perceptions of similarity. That is, similarity largely remained constrained only by vision perception, not conceptual knowledge. Also in the explicit responses, we discovered that perceptions of difference changed as a result of learning category boundaries. But this influence of categorical knowledge appears to be limited to the explicit choice, and not in competition leaking into motor output.

Let us now consider also the maximum deviation of the curved mouse tracking trajectories. Running all of the same ANOVAs with the maximum deviation as our dependent variable, we arrive at the same non-significant results, with all $p > .05$. This makes sense considering the values of area under the curve and maximum deviation should be highly correlated.

General Discussion

In this experiment we investigated the effect of learning category boundaries on the perception of similarity and difference. First we examined participants' explicit responses in their judgments of the similarity and difference among pairs of objects. We found that learning categories had no effect on perception of similarity, while learning did affect the perception of difference. In particular, when learning categories that are close to each other, participants tended to perceive object pairs as more different after learning category boundaries. This trend was strongest for within-category comparisons.

This finding may be influenced by the fact that the learning task itself emphasizes finding differences between category members. That is, in a standard classification task, attending to the diagnostic features is the most efficient method of learning. And it has been demonstrated that different category learning strategies result in different category representations (Chin-Parker & Ross, 2004; Markman & Ross, 2003; Romano, 2006). Acquiring category representations that are biased toward diagnosticity may be the basis for the effect of category learning (in this case, classification learning) on the perceptions of difference pre- and post-learning.

The most curious finding is the exclusive influence of learning on difference judgments, while similarity judgments remained unchanged. This raises several important questions.

First, how related are the cognitive processes of similarity and difference? Specifically, if conceptual knowledge influences only one of the two processes, then it suggests that there is some level of independence between similarity and difference. This would not be the first finding that suggests some level of independence. As mentioned in Chapter 5, it has been found that similarity and difference are not always inversely related (Medin, Goldstone, & Gentner, 1990; Estes & Hasson, 2004).

Second, what is the relationship between similarity and categorization? Nearly all theories of categorization rely on some process of similarity, either explicitly or implicitly. But in this experiment, similarity judgments remained the same both before and after learning categories. In Chapter 5, it was suggested that similarity is constrained by visual perception, in that qualitatively different visual features resulted in different patterns of responses. Perhaps similarity remained unaffected by learning categories because both processes are grounded and constrained by perception. But this does not explain why difference was affected by learning and not similarity. Clearly another explanation is needed.

It may also be that the post-test did not make use of the learned category labels. Category labels have been shown to strongly affect both the perceived similarity within a category and the perceived differences between categories (Goldstone, Lippa, & Shiffrin, 2001). And when categorization experience is equated, such as in locating a specific type of object, people that have a label for those category members have an advantage in correctly classifying them (Lupyan, Rakison, & McClelland, 2007). Future work will need to investigate the role of category labels, and how those effects interact with separable and integral features, as well as near and far category boundaries.

In Chapter 3 we explored many theories of categorization. All theories of categorization employ some process of similarity in explaining how those theories function and result in successful category learning and use. If this is truly the case, that categorization is grounded in similarity, then it should be expected that learning categories does not influence the perception of similarity. Similarity ought to be stable enough to provide the support necessary for conceptual coherence. If similarity was as unconstrained and unstable as some researchers fear, then it would validate the argument that similarity is useless and unworthy of study (Goodman, 1972; Murphy & Medin, 1985).

Recall that the supramodal theory defines similarity as simply the overlap of neuronal population codes that represent two or more objects. When a certain proportion of those populations are activated above a context-dependent threshold, then the objects are similar, otherwise they are not. This means that similarity operates on the encoded representations in the brain. Categories are themselves representations of an organizing principle governing a collection of objects, and that representation also exists in the brain. Similarity should remain unaffected by category learning, so long as the category learning does not significantly affect how the features of category members are encoded. In other words, similarity is partially-overlapping population codes, and as long as category representations do not significantly alter those codes, similarity will stay constant.

Perhaps if category learning significantly changed the neuronal representation of visual features, then judgments of similarity would change. For example, if acquiring some kind of conceptual information influences the perception of a particular feature in a top-down manner, then similarity judgments might be affected. Consider context effects, where the same object placed in different scenes results in different classifications of that object (Palmer, 1975). The supramodal theory proposes that similarity operates on population codes that act as feature detectors. If the perceived features of an object change based on its context, then similarity judgments will also change. In the case of this experiment, there were no contextual influences, and so the perceived features ought to have been very stable for participants. Consequently the similarity judgments that utilize those stable features will also be stable even when learning category boundaries.

We also examined implicit measures. In looking at participants' mouse movement trajectories, it was possible to obtain implicit measures for the amount of competition occurring during the time course of a cognitive event such as making judgments of similarity and difference. The idea is that as a mental process unfolds over time, and a participant settles on one of two possible responses ("yes" or "no"), the competition between those two alternatives will play out during that process and leak into the motor action of moving the mouse. How competitive the two alternatives were with each other is inferred by the amount of curvature in the mouse trajectory. That is, when the mouse curves more to the alternative before making the final choice, then it is said that the two objects were highly competitive. When there is little curvature and the trajectory is more direct, then there was little competition involved during the cognitive process of judging similarity or difference.

Unlike with the explicit responses, there was no influence of learning on the implicit measures. It appears that the curvature of the mouse trajectories remained the same both before and after learning categories. This makes sense in the case of questions about similarity, given that learning also did not affect the explicit judgments of similarity. However, it is somewhat strange there was no effect for questions about difference. Examining Figures 6.1.12 and 6.1.14, we see the data are overwhelmed with noise, as the standard error stretches well beyond the means and overlap between conditions. If there is an effect to be found, the current sample is not precise enough to cut through the noise.

Recall that we found that for near categories, learning resulted in increased perceptions of difference, especially for within-category pairs. Currently the supramodal theory does not have a strong account of the phenomenon of difference. Based on the neural network model in Chapter 5, it appears that difference relies on something more than just perceptual features. This is especially true for separable features, which replicate the non-inversion effect between similarity and difference.

Because of these effects, it is proposed that difference judgments are utilizing some other information. Whatever that may be, it has been found to be involved with learning categories, which suggests that it has something to do with memory.

In particular, the effect of learning on difference perceptions was to increase the perceived difference between objects in a pair. The learning task itself emphasized attention on diagnostic features, and repeated learning trials until participants met the learning criterion. Memory of focusing on differences during learning may be playing a role in increasing difference perceptions immediately after learning. Future work is needed to explore in more depth the process of difference, and whether there are factors involved that make it a phenomenon independent of similarity.

6.2 Experiment 2: Integral Features

In Experiment 1 of this chapter, we explored how similarity is related to categorization. Of particular interest was whether category boundaries affect the perception of similarity and difference. For example, if we have a set of 10 items, and 5 belong to Category A and the other 5 belong to Category B, does our perception of similarity (both within-category and between-categories) change as a result of learning that category boundary that divides the 10 items? Or is visual similarity largely unaffected by category boundaries and conceptual knowledge?

It was discovered that for classification learning, perceptions of similarity remained unchanged. The supramodal theory argues that similarity is the result of partially overlapping feature encoders in a neural network, and the theory predicts that similarity should remain stable if the perceived features remain stable (that is, if the neural representations are stable). The lack of influence of category learning on similarity judgments fits well, then, under the supramodal theory.

However, it was also discovered that perceptions of difference changed as a result of category learning. Specifically, participants that learned near categories resulted in increased perceptions of difference after learning category boundaries. This effect was strongest for within-category comparisons. It remains unclear exactly why difference perceptions were affected by category learning. One hypothesis is that the process of difference uses additional information, unlike the process of similarity that is proposed to simply be using activations of feature encoders. For example, if difference perceptions are also partially based on memory, then category learning could potentially influence changes in difference perceptions.

The questions that we asked in Experiment 1 are again asked in the current experiment, but this time with integral feature stimuli. Particular attention will be paid to whether we replicate the results of Experiment 1, or if integral features lead to divergent results as they did in the results in Chapter 5.

Experiment

The current experiment had 3 phases: a pre-test, a learning phase, and a post-test, replicating Experiment 1. In these testing phases, participants were presented with pairs of visual stimuli and were prompted to make judgments of the perceived similarity or difference of each pair. Participants in this experiment only viewed stimuli with the integral dimensions of hue, saturation, and brightness. The goal of this experiment was to investigate whether the perception of similarity and difference changes between pre- and post-tests as a result of learning category boundaries.

The stimuli in this experiment had integral features along the dimensions of hue, saturation, and brightness. The features along each dimension varied continuously across 4 equidistant points, which allowed for dimensions of 4 possible features, rather than binary features in Chapter 5. However, each participant only saw 2 of these possible features for each dimension, so in effect they learned binary features. To generate the continuous points across each dimension, we started with the integral stimuli from Experiment 1 (Nosofsky & Palmeri, 1996) and varied them by a constant proportion across 4 points.

This arrangement allows us to explore not only whether category boundaries affect the perception of similarity and difference, but also whether the distance of categories (the inherent similarity between categories) also affects those perceptions. Of particular interest is whether learning close category boundaries will distort perceptions of similarity more than far category boundaries. If two categories are very close, there should inherently be a higher degree of similarity among members of both categories than for two far categories. And the other question of interest is whether there will be divergent results for within-category pairs of objects vs. between-categories pairs of objects. Specifically, will members within a single category increase their perceived similarity, while members between two categories decrease their perceived similarity?

Method

Participants. A total of 29 undergraduate students from the University of California at Merced participated in the experiment for course credit.

Materials. The stimuli in the experiment were pairs of objects presented on a computer monitor. The stimuli were designed to measure the effects of the degree of feature overlap between objects in a pair, as well as distance between category boundaries. Each object in the pair had 3 dimensions consisting of binary features. The binary features could either be close to or far from each other in terms of between-category members. All objects had integral dimensions of hue, saturation, and brightness (Nosofsky & Palmeri, 1996). The binary features between the objects in a pair could be manipulated such that the pair was either completely identical, more similar / less different, less similar / more different, or completely different. Numerically, this is translated as 100% identical, 67% similar, 33% similar, or 0% similar.

Design. Each participant was assigned to 1 of 2 groups, either seeing objects that were close to each other (see Figure 6.2.1), or seeing objects that were far from each other (see Figure 6.2.2), based on the features along each dimension. This created a between-participants factor of category distance. There were 14 participants in the near category group, and 15 participants in the far category group.

Procedure. The current experiment replicated the procedure of Experiment 1. The experiment had 3 phases: a pre-test, a learning phase, and a post-test. The pre-test and the post-test were exactly the same, and they both consisted of multiple trials of the same task. In each trial, a participant was presented with a pair of objects and a question about that pair, which prompted a judgment response of the similarity or difference of that pair.

In this design, we were able to collect participants' perceptions of similarity and difference of the pairs of objects in the pre-test without the influence of category knowledge. And then in the post-test, we can compare whether perceptions changed as a result of learning category boundaries. Because both explicit and implicit measures were recorded, we can explore both aspects of the perception. Both the explicit judgment and the implicit mouse movement trajectory allow for two views of the same process, where the explicit judgment focuses on the end result and the mouse trajectory focuses on the process of that judgment over time. Mousetracking has been shown to be an effective measure for understanding the time course of a cognitive process, in that cognitive processes that involve competition leak into motor action and influence the curvature of the mouse trajectory (Dale, Kehoe, & Spivey, 2007; Dale, Roche, Snyder, & McCall, 2008).

Results and Discussion – Explicit Responses

The data of interest were the responses to each stimulus pair for each question in the pre- and post-tests. There were 4 types of feature overlap (100% similar, 67% similar, 33% similar, and 0% similar), and 2 types of questions (similarity and difference). The two groups being compared are those participants that learned close categories vs.

those participants that learned far categories. In addition, responses between the two testing phases were compared. The results can be seen in Figures 6.2.3 – 6.2.10.

Several factors were created from this data set. There is a between-subjects factor of category distance (close vs. far). Within-subjects, there is the test phase (pre- vs. post-test), there is the type of category comparison (within category or between categories), there is the type of question being asked (is the pair similar vs. is the pair different), and there is the degree of feature overlap in a pair of objects for each trial (100%, 67%, 33% or 0%).

The following analyses use the proportion of yes-responses for each conjunction of category distance, test phase, category membership, question type, and feature overlap. However, the yes-responses for questions of difference were inverted. When an object pair is 100% similar, it should be expected that participants would respond 100% “Yes” for questions of similarity, and 0% “Yes” for questions of difference, simply because those pairs are completely identical. This expected inverse relationship needed to be represented in the data for comparative analyses between questions. Therefore, in the following analyses, the yes-responses for questions of difference were transformed by (1 – response). In this way, when the patterns of responses are graphed for questions of similarity and difference, the values are commensurable and should overlap in the graph instead of displaying their inverse relationship. It also permits a visual observation of the non-inversion effect, described in the current experiment’s introduction.

A repeated measures analysis of variance (ANOVA) was performed, with Category Distance as the between-subjects factor (near or far categories). There were 4 within-subjects factors: Learning (pre-test vs. post-test), Category Membership (within or between categories), Question (similar or different), and Feature Overlap (100%, 67%, 33%, or 0%). Results showed a significant main effect of Category Membership, *F*(1, 27) = 671.582, *p* < .001; a significant main effect of Question, *F*(1, 27) = 14.052, *p* < .01; a significant main effect of Feature Overlap, *F*(3, 84) = 79.780, *p* < .001; a significant interaction of Learning and Category Membership, *F*(1, 27) = 4.704, *p* < .05; a significant interaction of Learning and Question, *F*(1, 27) = 4.287, *p* < .05; a significant interaction of Category Membership and Question, *F*(1, 27) = 5.215, *p* < .05; a significant interaction of Category Membership and Feature Overlap, *F*(3, 81) = 755.548, *p* < .001; a significant interaction of Category Membership and Feature Overlap and Category Distance, *F*(3, 81) = 2.798, *p* < .05; a significant interaction of Learning and Category Membership and Feature Overlap and Distance, *F*(3, 81) = 2.929, *p* < .05; a significant interaction of Question and Feature Overlap, *F*(3, 81) = 10.079, *p* < .001; a significant interaction of Category Membership and Question and Feature Overlap, *F*(3, 81) = 7.102, *p* < .001; and a significant interaction of Category Membership and Question and Feature Overlap and Category Distance, *F*(3, 81) = 3.136, *p* < .05. All other effects and interactions were not significant.

To take a deeper look into these main effects and interactions, follow-up ANOVAs were performed with fewer factors in each analysis. In the following

ANOVAs, we split up the data between groups (near and far categories) and analyzed them separately with the factor of Category Membership.

In the first ANOVA, we looked at the near category group for within-category comparisons. There were 3 within-subjects factors: Learning, Question, and Feature Overlap. There was a significant main effect of Question, $F(1, 13) = 7.982$, $p < .05$; a significant main effect of Feature Overlap, $F(3, 39) = 225.242$, $p < .001$; and a significant interaction of Question and Feature Overlap, $F(3, 39) = 7.131$, $p < .01$. All other main effects and interactions were not significant.

In the second ANOVA, we looked at the far category group for within-category comparisons. There were 3 within-subjects factors: Learning, Question, and Feature Overlap. There was a significant main effect of Feature Overlap, $F(3, 42) = 140.592$, $p < .001$; a significant interaction of Learning and Question, $F(1, 14) = 10.012$, $p < .01$; and a significant interaction of Learning and Question and Feature Overlap, $F(3, 42) = 3.907$, $p < .05$. All other main effects and interactions were not significant.

In the third ANOVA, we looked at the near category group for between-category comparisons. There were 3 within-subjects factors: Learning, Question, and Feature Overlap. There was a significant main effect of Question, $F(1, 13) = 9.795$, $p < .01$; a significant main effect of Feature Overlap, $F(3, 39) = 12.614$, $p < .001$; and a significant interaction of Question and Feature Overlap, $F(3, 39) = 8.728$, $p < .001$. All other main effects and interactions were not significant.

In the fourth ANOVA, we looked at the far category group for between-category comparisons. There were 3 within-subjects factors: Learning, Question, and Feature Overlap. There was a significant main effect of Question, *F*(1, 14) = 5.228, *p* < .05; a significant main effect of Feature Overlap, *F*(3, 42) = 10.810, *p* < .001; and an approaching significant interaction of Question and Feature Overlap, *F*(3, 42) = 2.751, *p* = .054. All other main effects and interactions were not significant.

The results of these four analyses suggest that the Learning factor only played a significant role when it involved far categories for within-category comparisons. To explore more into this effect of learning, follow-up ANOVAs were performed with fewer factors in each analysis. In the following ANOVAs, we split up the data between groups (near and far categories) and analyzed them separately with the factor of Category Membership and Question.

In the first ANOVA, we looked at the near category group for within-category comparisons for similarity judgments. There were 2 within-subjects factors: Learning and Feature Overlap. There was a significant main effect of Feature Overlap, *F*(3, 39) = 95.227, *p* < .001; and no other significant main effects or interactions.

In the second ANOVA, we looked at the far category group for within-category comparisons for similarity judgments. There were 2 within-subjects factors: Learning and Feature Overlap. There was a significant main effect of Feature Overlap, *F*(3, 42) = 132.647, *p* < .001; and no other significant main effects or interactions.

In the third ANOVA, we looked at the near category group for between-category comparisons for similarity judgments. There were 2 within-subjects factors: Learning and Feature Overlap. There was a significant main effect of Feature Overlap, *F*(3, 39) = 14.027, *p* < .001; and no other significant main effects or interactions.

In the fourth ANOVA, we looked at the far category group for between-category comparisons for similarity judgments. There were 2 within-subjects factors: Learning and Feature Overlap. There was a significant main effect of Feature Overlap, *F*(3, 39) = 12.425, *p* < .001; and no other significant main effects or interactions.

The results of these four analyses suggest that there is no effect of learning on questions of similarity for either near or far categories, within- or between-category comparisons, or any combination of those factors. The next four analyses will investigate whether learning affects perceptions of difference.

In the first ANOVA, we looked at the near category group for within-category comparisons for difference judgments. There were 2 within-subjects factors: Learning and Feature Overlap. There was a significant main effect of Feature Overlap, *F*(3, 39) = 303.186, *p* < .001; and no other significant main effects or interactions.

In the second ANOVA, we looked at the far category group for within-category comparisons for difference judgments. There were 2 within-subjects factors: Learning and Feature Overlap. There was a significant main effect of Learning, *F*(1, 14) = 5.688, *p* < .05; a significant main effect of Feature Overlap, *F*(3, 42) = 114.947, *p* < .001; a significant interaction of Learning and Feature Overlap, *F*(3, 42) = 3.327, *p* < .05; and no other significant main effects or interactions.

In the third ANOVA, we looked at the near category group for between-category comparisons for difference judgments. There were 2 within-subjects factors: Learning and Feature Overlap. There was a significant main effect of Feature Overlap, *F*(3, 39) = 3.574, *p* < .05; and no other significant main effects or interactions.

In the fourth ANOVA, we looked at the far category group for between-category comparisons for difference judgments. There were 2 within-subjects factors: Learning and Feature Overlap. There was a significant main effect of Feature Overlap, *F*(3, 42) = 6.515, *p* < .01; and no other significant main effects or interactions.

The results of these four analyses show that Learning had a significant influence on the perception of difference only for far categories involved within-category comparisons. In fact, looking at Figure 6.2.8, we can see clearly that for questions of difference involving far categories and within-category comparisons, the perception of difference is decreased after learning. In other words, participants were less likely, after learning category boundaries, to say that two objects within the same category were different.

Across all of the analyses, the trend was that learning category boundaries did not affect the perception of similarity, but it did affect the perception of difference for a very particular case: for participants that learned far categories and were making difference judgments involving pairs of objects from the same category. Again, perhaps this result is due to the nature of the category learning task in this experiment. It has been found that classification learning tasks tend to focus attention on diagnostic features (Chin-Parker & Ross, 2004; Markman & Ross, 2003), and biases the encoded category representation in memory to emphasize those features (Romano, 2006).

In the case of this experiment involving integral features, the effect of learning was to decrease the perception of difference for far/within-category pairs. Perhaps this is due to the nature of integral features. As we saw in Chapter 5, when just one feature in an integral stimulus is changed, it affects the perception of the other features. It follows that decreasing the perceived difference among integral stimuli would facilitate learning.

Given the nature of integral stimuli, the within-category differences ought to be perceived as greater than the within-category differences for separable stimuli, for the same abstract feature set of linearly separable features. That is, each member within a linearly separable category will be probabilistically similar to its category prototype. The category prototype 111 will have 3 members, 011, 101, and 110. If it is true that changing 1 feature of an integral stimulus will affect the perception of the other 2, and if each category member has 1 mismatch with its category prototype, then there is inherently a greater amount of within-category difference for integral stimuli. And given the nature of the classification learning task, which focuses attention of the diagnostic features of objects, it makes sense that decreasing the perceptions of difference within integral categories would facilitate learning.

Results and Discussion – Implicit Responses

In the previous section we explored participants' explicit responses when questioned about their perceptions of similarity and difference for various pairs of objects. In this section we consider their implicit responses, measured with mousetracking. Mouse movement trajectories have been demonstrated to provide informative data which indicates the level of competition during the execution of an online cognitive process (Dale, Kehoe, & Spivey, 2007; Dale, Roche, Snyder, & McCall, 2008; Spivey, Grosjean, & Knoblich, 2005). In this way, we have an implicit measure of how much competition is occurring during the process of making a decision. Whereas with the explicit responses, we only get the final output of that decision, not the process over the time course of the decision.

In the case of this experiment, we are interested to know if learning category boundaries affects mouse movement trajectories. Specifically, we can test the hypothesis that learning category boundaries will reduce the amount of competition between objects such that categorical knowledge helps to structure the mental organization of these objects. Also, we can test whether we get divergent results between explicit and implicit responses. In the case of explicit responses, learning influenced only the perception of difference, not similarity, and then only for learning far categories. But again, this is only considering the final, end product of the decision making process. Perhaps during that process, there was a change in the level of perceived competitiveness as a result of learning categories, which would in turn affect mouse movements.

Just as we did in Experiment 1, for each trial, we can extract two key pieces of information from the mouse movement trajectory. First, we can calculate the total area under the curve, which gives us an indication of how much competition, non-competition, or even repulsion there was between objects in each trial. The other dependent variable that we get from mouse tracking is the maximum deviation of the curve. This informs us of the highest peak of the curved mouse trajectory.

Area under the curve and maximum deviation are calculated in a number of steps. First, trajectories are translated into a coordinate system such that the origin of the mouse trajectory is at the coordinate (0, 0). When a trajectory landed on the response option on the left of the screen, that trajectory was reversed horizontally so that they could be compared equivalently with trajectories that curved to the right. All trajectories were normalized and interpolated to be standardized as 50 point vectors. This allowed for equivalent comparisons of all trajectories from all trials, and for the computation of area under the curve and maximum deviation.

Looking at Figures 6.2.11 – 6.2.18, which all show area under the curve for the mouse movement trajectories, we can compare the pre- and post-tests and evaluate the effect of learning category boundaries in regard to a variety of factors. All conditions involve integral stimuli, and we can examine the data based on category boundary distance, the type of question, and whether we are comparing within or between categories. A cursory view of the figures suggests that there is no clear-cut story, and that the data is very noisy. A reasonable hypothesis is that learning category boundaries should help to reduce competition between alternatives, thereby reducing curvature in the mouse trajectories. But the results show that sometimes learning results in less curvature in the post-test, and sometimes it results in more. The standard errors around the means between pre- and post-test values overlap in nearly all conditions, suggesting a very noisy sampling of the data.

Recall that for the explicit responses, the significant effects of learning occurred for questions of difference involving far categories, where the perception of differences decreased as a result of learning. When examining Figure 6.2.16, which plots area under the curve for those same conditions that gave rise to significant explicit distinctions, we notice that there is no hint of implicit effects in the mouse trajectories as a result of learning. Both the means are rather close, and the standard errors bleed together. If there is an effect to be found, the current data contain too much noise to discern it.

Perhaps with a larger sample size the standard error would be reduced, with more accurate sample means. Or perhaps the nature of the task elicits too much variability across participants' mouse movements. The relative simplicity and open-endedness of the experiment instructions were heralded as a strength earlier, but perhaps this lack of constraints also opens the door to too many uncontrolled factors becoming involved in the perception-action loop. For example, a participant may evolve a definition of similarity and difference over the course of the experiment, and this dynamic process would likely impact the mouse trajectories.

Although there is a lack of visible differences among the figures that graph the implicit results (Figures 6.2.11 – 6.2.18), a quantitative view may yield more subtle effects. A repeated measures ANOVA was performed, with Category Distance as the between-subjects factor (near vs. far categories). There were 4 within-subjects factors: Learning (pre-test vs. post-test), Category Membership (within vs. between categories), Question (similar vs. different), and Feature Overlap (100%, 67%, 33%, 0%). Results showed no significant main effects or interactions for any combination of factors, all $p > .05$. This result shows that there is no overall trend to be generalized for implicit measures of mousetracking.

To take a deeper look, follow-up ANOVAs were performed with few factors in each analysis. In the following ANOVAs, we considered looking only at within-category comparisons, and split up the data between groups (near and far categories).

In the first ANOVA, we looked at the near category group for within-category comparisons. There were 3 within-subjects factors: Learning, Question, and Feature Overlap. Results showed no significant main effects or interactions for any combination of factors, all $p > .05$.

In the second ANOVA, we looked at the near category group for between-category comparisons. There were 3 within-subjects factors: Learning, Question, and Feature Overlap. Results showed no significant main effects or interactions for any combination of factors, all $p > .05$.

In the third ANOVA, we looked at the far category group for within-category comparisons. There were 3 within-subjects factors: Learning, Question, and Feature

Overlap. Results showed no significant main effects or interactions for any combination of factors, all *p* > .05.

In the fourth ANOVA, we looked at the far category group for between-category comparisons. There were 3 within-subjects factors: Learning, Question, and Feature Overlap. Results showed no significant main effects or interactions for any combination of factors, all *p* > .05.

Results of these four analyses match our assumptions based on a visual overview of the figures, and suggest that there is no clear effect of learning on implicit measures of mousetracking trajectories when considering area under the curve. We performed even more ANOVAs, this time with only 2 within-subjects factors: Learning and Feature Overlap. For all possible combinations of analyses, no significant effects or interactions were discovered, *p* > .05.

What we can conclude is that learning category boundaries does not appear to have a clear effect on the motor output of making similarity and difference judgments. For questions of similarity, this corroborates the story gathered from the explicit response data, where we found that category learning did not affect perceptions of similarity. That is, similarity largely remained constrained only by vision perception, not conceptual knowledge. Also in the explicit responses, we discovered that perceptions of difference changed as a result of learning category boundaries. But this influence of categorical knowledge appears to be limited to the explicit choice, and not in competition leaking into motor output.

Let us now consider also the maximum deviation of the curved mouse tracking trajectories. Running all of the same ANOVAs with the maximum deviation as our dependent variable, we arrive at the same non-significant results, with all $p > .05$. This makes sense considering the values of area under the curve and maximum deviation should be highly correlated.

General Discussion

In this experiment we investigated the effect of learning category boundaries on the perception of similarity and difference. First we examined participants' explicit responses in their judgments of the similarity and difference among pairs of objects. We found that learning categories had no effect on perception of similarity, while learning did affect the perception of difference. In particular, when learning categories that are far from each other, participants tended to perceive object pairs as less different after learning category boundaries for within-category comparisons.

Again, the selective influence of learning on perception of difference and not perception of similarity raises the question of what is the nature of the relationship between similarity and difference. If conceptual knowledge influences only one of the two processes, then it suggests that there is some level of independence between similarity and difference. According to the supramodal theory, similarity should remain unaffected by category learning, so long as the category learning does not significantly affect how the features of category members are encoded. In other words, similarity is partially-overlapping population codes, and as long as category representations do not significantly alter those codes, similarity will stay constant.

Perceptions of difference of integral pairs decreased for within-category comparisons, after learning far category boundaries. The human visual system is inherently more sensitive to differences involving integral stimuli. In order to assist conceptual coherence, the differences among members within a category must be reduced to facilitate learning. Again, it appears as though perception of difference is related to memory, but also the manner in which difference perceptions are affected is based on the type of stimuli that is being learned.

We also examined implicit measures. In looking at participants' mouse movement trajectories, it was possible to obtain implicit measures for the amount of competition occurring during the time course of a cognitive event such as making judgments of similarity and difference. How competitive the two alternatives were with each other is inferred by the amount of curvature in the mouse trajectory. That is, when the mouse curves more to the alternative before making the final choice, then it is said that the two objects were highly competitive. When there is little curvature and the trajectory is more direct, then there was little competition involved during the cognitive process of judging similarity or difference.

Unlike with the explicit responses, there was no influence of learning on the implicit measures. It appears that the curvature of the mouse trajectories remained the same both before and after learning categories. This makes sense in the case of questions about similarity, given that learning also did not affect the explicit judgments of similarity. But the explicit judgments of difference in certain cases were significantly influenced by learning. Given that the implicit effects were not significant, it suggests that whatever is causing the influence on explicit difference responses is not influencing in a way that is causing any more or less competition between alternatives. The implicit response is just measuring how much competition is involved during the cognitive process of judging difference, and if learning does not significantly change it, then it simply means that category learning did not alter the competitiveness of objects in the comparisons.

6.3 General Discussion

The experiment and computational model in Chapter 5 investigated whether similarity is constrained, at least in part, by visual perception. Results indicate that perceptions of similarity are influenced by the type of features that are observed, suggesting that the perceptual system helps to ground judgments of similarity. In Experiments1 and 2 of the current work, we sought to understand what are the influences of category learning on the perceptions of similarity and difference, both explicitly and implicitly. For explicit measures, we used participants' explicit responses in their judgments of similarity and difference for various pairs of objects. For implicit measures, we used the mouse movement trajectories that were generated as a result of participants responding to stimuli and making online judgments.

The supramodal theory proposes that similarity is partial identity. In other words, similarity is the partial overlap of the population codes of neurons that encode for particular objects. When there is sufficient overlap among the neurons that represent objects, then the two objects are perceived as similar. Because the population codes are acting as feature detectors and actively represent observed features, similarity is easily represented as sharing these features and mutually activating represented features above a certain threshold to achieve similarity.

Based on these assumptions, the supramodal theory of similarity predicts that similarity ought not be influenced significantly by learning category boundaries, so long as that category learning does not influence the encoding of the observable features. In the case of the features involved in Experiments 1 and 2, learning category boundaries likely did not influence how those low-level perceptual features (such as shape, size and color; and hue saturation, and brightness) were encoded in participants' brains. But if participants were presented with features that were more ambigious, such as features that can be perceived in more than one way based on context, then categorical knowledge might affect perceptions of similarity by altering which of the ambiguous representations a particular feature settles. But in the end, the process of similarity remains the same; it simply acts on the set of features currently activated in the neural network.

This account offers an explanation for the lack of influence of categorical knowledge on the perception of similarity. But why was there an influence on the perception of difference after learning categories? In the case of Experiment 1, perceptions of difference *increased* as a result of learning category boundaries for near categories, especially for within-category comparisons. But in Experiment 2, perceptions of difference *decreased* as a result of learning category boundaries for far categories involving within-category comparisons. Both experiments involve changes in difference judgments and both involve within-category comparisons, but for separable stimuli this occurred with near categories and increased difference perceptions while for integral stimuli this occurred with far categories and decreased difference perceptions.

The non-inversion effect suggests that there is some level of independence between similarity and difference perceptions, as they are not always inversely related (Medin, Goldstone, & Gentner, 1990; Estes & Hasson, 2004). Based on the neural network model in Chapter 5, it appears that difference relies on something more than just perceptual features. It is proposed that difference judgments are utilizing some other information. In particular, if difference judgments are influenced by acquiring categorical knowledge, then it is proposed that difference perceptions are influenced in part by memory.

The nature of classification learning is to focus on the diagnostic features of the objects being categorized. It has been demonstrated that different category learning strategies result in different category representations (Chin-Parker & Ross, 2004; Markman & Ross, 2003; Romano, 2006). Acquiring category representations that are biased toward diagnosticity may be the basis for the effect of category learning (in this case, classification learning) on the perceptions of difference pre- and post-learning. Specifically, the perception of difference among members within a category was influenced by this classification learning strategy.

For separable stimuli, the effect was to increase the perceived difference within categories for near categories. For integral stimuli, the effect was to decrease the perceived difference within categories for far categories. Considering these two effects together, we can infer that the cognitive process of difference is utilizing both perceptual and conceptual factors, as the directionality of the effect is mediated by stimulus type (perceptual influence) and the type of comparison affected is mediated by categorical boundaries (conceptual influence).

What is the benefit of increasing difference perceptions both within and between categories for near categories involving separable features? Members of near categories will have a high degree of similarity to other objects both within and between categories. The result is that it becomes more challenging to correctly classify an object as belonging to Category A if that object is also highly similar to members of Category B. But if perceptions of difference are increased, then it becomes easier to differentiate an object and correctly classify it.

A real world example would be correctly classifying cars based on make and model. Cars are highly similar objects, yet experts can easily classify even minor annual changes in the same make and model car. It is likely that this is due to an emphasis on the features that are distinct. In fact, the structural alignment theory predicts that it is easier to find differences between similar pairs of objects than between dissimilar pairs of objects, and this prediction plays out accurately in the empirical data (Gentner & Markman, 1994). Experiment 1 replicated this effect.

And what is the benefit of decreasing difference perceptions within categories for far categories involving integral features? The human visual system is inherently more sensitive to differences involving integral stimuli. Changing just one feature will dramatically affect the perception of the other two, such that a single disturbance can affect the perception of the whole object. This consequence is even more exaggerated for far categories. In order to assist conceptual coherence, the differences among members within a category must be reduced to facilitate learning. Again, it appears as though perception of difference is related to memory, but also the manner in which difference perceptions are affected is based on the type of stimuli that is being learned.

Finally, category learning in these experiments had no significant effect on the implicit measures involved in perceptions of similarity and difference. Both area under the curve and maximum deviation of the curve of mouse movement trajectories were calculated, and neither showed significant results. These implicit variables measure the competitiveness of the options given in each trial. In this case, the competitiveness of judging whether a pair of objects was similar or different.

The data may just be too noisy to really draw any conclusion regarding the implicit effects of category learning on judging similarity and difference. Perhaps the task of this experimental design was too open-ended, which allowed for too much influence of individual differences to leak into the trajectory data. Or we may simply require a larger sample size to reduce the noise. As it stands, the only real conclusion we can draw from the implicit data is that it has been influenced by a strong source of noise that has not yet been identified. If learning did not significantly change the curvature of mouse trajectories, then it simply means that category learning did not alter the competitiveness of objects in the comparisons. In other words, learning can result in the acquisition of categorical knowledge, but it is not necessary for that knowledge to affect how competitive two options are when processing the perceived similarity or difference of pairs of objects. Future research is needed to investigate methods for reducing noise in the motor task of this experiment so that we may better understand the relationship between category learning and similarity and difference perceptions.

CHAPTER 7: GENERAL DISCUSSION

Similarity is a process that is central to human cognition. It facilitates the formation of concepts and categories by associating commonalities among objects; to transfer knowledge from one domain to another; to create metaphors and analogies; to make inferences and predictions; and to prime and cue memory and behavior. Similarity as a cognitive phenomenon is strongly connected to both high- and low-level processes. But despite its relevance by connectedness to so many theories, similarity remains a poorly understood aspect of human cognition.

Several models provide accounts of the role that similarity has in human cognition (e.g., Edelman, 1998; Shepard, 1962). In Chapter 2, we discussed the more influential models of similarity, and covered their respective strengths and weaknesses. A key argument against similarity is that it becomes unrealistically flexible to explain conceptual coherence because “there are more free parameters than degrees of freedom” (Murphy & Medin, 1985, p. 292), and that many models fail to explain how features are related to one another and bound together (Keane & Costello, 2001; Medin, Goldstone, & Gentner, 1993). While several of the theories discussed have strong descriptive power, especially the statistical models, no theory has really struck a balance of descriptive and explanatory power such as providing a biologically plausible implementation.

A new theory of similarity has been proposed, one that unites it with categorization in a parsimonious manner. The supramodal theory of similarity and categorization proposes that collections of features are best represented with saliency maps that can interact and collide together to create supramaps that can contain emergent properties that were not in the original maps. Mechanistically this theory is easily implementable and testable in artificial neural networks.

In the supramodal theory, similarity is defined as partially overlapping population codes of neurons that are acting as feature detectors. When the similarity between two objects is being evaluated, the criterion is the proportion of mutual activation, or the amount of overlap between activated neurons that represent those two objects. The threshold of that criterion can be task-dependent. This allows for a great deal of stability in the process of similarity, which satisfies the criticisms of Goodman (1972) and Murphy and Medin (1985), who argue that similarity is too unconstrained. And the threshold allows for the flexibility in similarity perceptions by having context and goals influence what proportion of overlap is necessary for a useful judgment of similarity. With just this one threshold variable, this is a parsimonious solution to a flexible similarity mechanism that avoids the free parameters problem in many models of similarity.

Of course, this removes the problem by just one step, by pushing the problem back to one of selecting features. However, the model can be updated such that the integration vector is selected by the network automatically detecting the type of features being viewed. Standard backpropagation networks, just like the one used in the current work, have been paired with image-processing Gabor filters that can process integral stimuli differently from separable stimuli, and the output can be used by the neural network (Tijsseling & Gluck, 2002). This pre-processing layer can be thought of as a sensory layer, or perhaps an early vision processing layer such as area V1 parsing contours. Expanding the current model to use this method of automatically detecting and distinctly processing separable and integral stimuli will resolve this issue of pushing the problem onto features, and doing so in a manner that has both descriptive power and is biologically plausible.

Another problem among models of similarity is that there are few explanations for how features are related to one another (Goldstone, Medin, & Gentner, 1991). Categories cannot be just an amorphous bag of features, even if they are weighted for saliency like on the contrast model (Tversky, 1977) or for typicality like in the prototype model (Rosch & Mervis, 1975). The supramodal theory accounts for the binding problem simply by employing population codes of neurons that act as feature detectors. In Chapter 5 we explored a connectionist model that accurately predicts human behavior in similarity judgments for both separable and integral features. It accomplishes this simply by learning to associate an abstract feature set with a whole object. In other words, the network learns to bind the features together without needing to specify a post-hoc structural alignment framework, in the same way that the human neural network system can bind features together.

The theory view discussed in Chapter 3 claims that categories are coherent because the background knowledge one has about the world provides the internal structure for concepts (Murphy & Medin, 1985). In this view, feature-based similarity is too flexible to explain conceptual coherence because stimulus context and experimental task introduce too many free parameters, rendering similarity too flexible to be the basis for categorization. But as we just discussed, the supramodal theory accounts for similarity with just one free parameter. Here, features emerge from the patterns of connectivity among units. Considering features in this way allows for both innate features, such as center-surround receptive fields detecting contrast (Schiller, 1992), and learned features, such as door knobs, to be utilized in similarity perceptions, categorizations, and any other feature-based task.

The experiments in Chapter 6 demonstrate that similarity can be unaffected by learning categories, which suggests that similarity perceptions can remain stable even with background knowledge. The supramodal theory argues that similarity is very stable because the features of the world are very stable. When dealing with abstract features or features that are ambiguous, those features are likely to change based on context, and similarity perceptions will consequently change. But this is a matter of featural flexibility, not an unbounded process of similarity.

A key prediction of the supramodal theory, based on this idea of being above or below a similarity threshold, is that similarity should not be affected very much by category boundaries. That is, if there is sufficient overlap in neuronal population codes, then two objects are similar, and otherwise they are not. This may seem counterintuitive, especially to many of the competing theories discussed in earlier chapters. But similarity in the supramodal theory is more concerned with how features are represented than how categorical knowledge is organized.

However, category labels have been shown to strongly affect both the perceived similarity within a category and the perceived differences between categories (Goldstone, Lippa, & Shiffrin, 2001), suggesting that categories affect the perception of similarity. And people that have a label for those category members have an advantage in correctly classifying them (Lupyan, Rakison, & McClelland, 2007). Yet people who speak different languages can have strikingly different patterns of naming for a set of objects while at the same time seeing the same similarities among those objects (Malt, Sloman, Gennari, Shi, & Wang, 1999). These findings suggest two things. First, category labels play an important role in the recognition of objects and the perception of similarity. Second, there is some level of independence between categorization and similarity perception in the cross-linguistic study, because linguistic categories varied while similarity perceptions were consistent across speakers. Perhaps the lack of effect on similarity perception by learned category boundaries was a result of not referencing category labels in the post-test.

The main finding in Chapter 5 suggests that the visual perception system helps to constrain judgments of similarity and difference. Qualitatively different patterns of judgments resulted from qualitatively different types of features, separable vs. integral. As has been mentioned, separable features can be analyzed independently of one another, while integral features cannot (changing one feature alters the perception of the other two features). The results suggest something distinct about separable-features from integral-features in the perception of similarity, specifically when the pairs of objects have both common and distinctive features. Tversky's (1977) contrast model does not account for this finding. While it predicts that similarity will increase with common features and decrease with distinctive features (as the result show), it does not predict the qualitatively different patterns of increases and decreases that is seen in the data. The supramodal connectionist model, however, is able to account for both the claims of the contrast model, and predict the nonlinear response patterns found in the empirical results.

It appears as though similarity is fairly stable, while difference is influenced by both perception and conceptual knowledge. There is some evidence to suggest that the processes of similarity and difference are somewhat independent. The non-inversion effect demonstrates that, for some types of features, perceptions of similarity and difference are not proportional. For example, for any given pair of objects, if a person perceives the pair to be 70% similar, the pair ought to be 30% different. However, people over-valued their difference judgments relative to their similarity judgments for separable stimuli, demonstrating the non-inversion effect.

In addition to the non-inversion effect, difference judgments in some cases are influenced by conceptual knowledge. The experiments in Chapter 6 sought to understand the influences of category learning on the perceptions of similarity and difference, both explicitly and implicitly. In the case of Experiment 1, perceptions of difference *increased* as a result of learning category boundaries for near categories, especially for within-category comparisons. But in Experiment 2, perceptions of difference *decreased* as a result of learning category boundaries for far categories involving within-category comparisons. Both experiments involve changes in difference judgments and both involve within-category comparisons, but for separable stimuli this occurred with near categories and increased difference perceptions while for integral stimuli this occurred with far categories and decreased difference perceptions.

The classification learning task focuses attention on the diagnostic features of the objects being categorized. Acquiring category representations that are biased toward diagnosticity may be the basis for the effect of category learning on the perceptions of difference pre- and post-learning. Specifically, the perception of difference among members within a category was influenced by this classification learning strategy. For separable stimuli, the effect was to increase the perceived difference within categories for near categories. For integral stimuli, the effect was to decrease the perceived difference within categories for far categories. Considering these two effects together, we can infer that the cognitive process of difference is utilizing both perceptual and conceptual factors, as the directionality of the effect is mediated by stimulus type (perceptual influence) and the type of comparison affected is mediated by categorical boundaries (conceptual influence).

The effect of increasing perceived difference for near categories with separable features is compatible with other findings. The structural alignment theory predicts that it is easier to find differences between similar pairs of objects than between dissimilar pairs of objects, and this prediction plays out accurately in the empirical data (Gentner & Markman, 1994). So if participants are learning categories through classification, which focuses attention on differences, and it is easier to find differences between similar pairs of objects than dissimilar pairs, then it follows that perceptions of difference should increase when the encoding of categorical knowledge is biased to focus on that diagnostic information.

And what about the apparent decrease in difference perceptions involving far categories and members within a category that have integral features? As we saw in the results of experiments both Chapters 5 and 6, the human visual system is inherently more sensitive to differences involving integral stimuli. Changing just one feature will dramatically affect the perception of the other two, such that a single disturbance can affect the perception of the whole object. This consequence is even more exaggerated for far categories. In order to assist conceptual coherence, the differences among members within a category must be reduced to facilitate learning. Again, it appears as though perception of difference is related to memory, but also the manner in which difference perceptions are affected is based on the type of stimuli that is being learned.

The implicit measures were designed to test whether acquiring category knowledge affects the time course of processing similarity and difference. However, learning did not significantly change the curvature of mouse trajectories, which suggests that category learning did not alter the competitiveness of objects in the comparisons. In other words, learning can result in the acquisition of categorical knowledge, but it does not necessarily follow that category knowledge will affect how competitive two options are when processing the perceived similarity or difference of pairs of objects. The noisiness of the sample precludes a more accurate picture and a definitive evaluation of the effect of category learning on the time course of making similarity and difference judgments.

Future research is needed to explore in more depth the process of difference, and to tease apart what are the factors involved that make it a somewhat independent process from similarity. Because category learning involves both similarity and difference, the supramodal theory is incomplete without an accurate understanding of how difference works. Of particular importance is to develop a model that explains the mechanism of difference within the supramodal theory, just as we have an explained mechanism for similarity.

The experiments in Chapters 5 and 6 used stimuli that were carefully crafted to balance salience among features. Treue (2003) argues that stimuli of highest salience will already attract attention without needing help from top-down influence, which is compatible with the "pop-out effect," while stimuli of intermediate salience (such as an object among many similar distractors) might need top-down attention to make them salient. Future work should explore categories that require more or less attention to learn based on the saliency of the features of the category members. This will help test in more depth the claims of saliency maps in the supramodal theory. In particular, to be able to tease apart the effects of saliency on the perceptions of similarity and difference both before and after learning categories would be especially useful.

The goal of the current work was to take a different approach to studying similarity than the approach used in many other studies. A deliberate effort was made to study only very simple perceptual features, and to directly ask participants about their perceptions of similarity and difference. In Chapter 5, the goal was to understand how similarity might be constrained by visual perception without the influence of background knowledge. In Chapter 6, the goal was to understand what role background knowledge had on perceptions of similarity, such as category boundaries.

The goal of the supramodal theory was to design an approach to similarity that was parsimonious, unifying with other cognitive processes, makes clear predictions, is testable, and can be successfully modeled and compared with empirical results. In Chapter 5 we saw that indeed, the supramodal model did an excellent job of accounting for the human data. While this theory certainly cannot account for all of the phenomena involved in similarity and categorization, the goal was to create a simplified approach that can easily evolve to accommodate new findings and expand its predictive and explanatory power.

REFERENCES

Baddeley, A. D., & Hitch, G. (1974). Working memory. In G. A. Bower (Ed.), The Psychology of Learning and Motivation: Advances in Research and Theory, Vol. 8. New York, NY: Academic Press.

Brooks, L. R., Squire-Graydon, R., & Wood, T. J. (2007). Diversion of attention in everyday concept learning: Identification in the service of use. Memory & Cognition, 35, 1-14.

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

Coulson, S. (1995). Analogic and metaphoric mapping in blended spaces: Mendez brothers virus. The Newsletter of the Center for Research in Language, 9, 1.

Crick, F., & Koch, C. (1990a). Towards a neurobiological theory of consciousness. Seminars in the Neurosciences, 2, 263-275.

Crick, F., & Koch, C. (1990b). Some reflections on visual awareness. Cold Spring Harbor Symposia on Quantitative Biology, LV, 953-962.

Dale, R., Kehoe, C., & Spivey, M. J. (2007). Graded motor responses in the time course of categorizing atypical exemplars. *Memory & Cognition, 35,* 15-28.

Dale, R., Roche, J., Snyder, K., & McCall, R. (2008). Exploring action dynamics as an index of paired-associate learning. *PLoS ONE, 3*, 1-10.

Desimone, R., & Duncan, J. (1995). Neural mechanisms of selective visual attention. Annual Review of Neuroscience, 18, 193-222.

Edelman, S. (1998). Representation is representation of similarities. *Behavioral and Brain Sciences, 21*, 449-467.

Fauconnier, G. (1997). Mappings in Thought and Language. Cambridge, UK: Cambridge University Press.

Fauconnier, G., & Turner, M. (1994). Conceptual projection and middle spaces. UCSD Cognitive Science Technical Report, 9401.

Fauconnier, G., & Turner, M. (1998). Conceptual integration networks. Cognitive Science, 22, 133-187.

Fillmore, C. J. (1982). Frame semantics. In Linguistic Society of Korea (Ed.), Linguistics in the Morning Calm. Seoul, South Korea: Hanshin Publishing Company.

Fine, M. S., & Minnery, B. S. (2009). Visual salience affects performance in a working memory task. The Journal of Neuroscience, 29, 8016-8021.

Gazzaniga, M. S. (2005). Forty-five years of split-brain research and still going strong. Nature Reviews Neuroscience, 6, 653-659.

Gazzaniga, M. S., Ivry, R. B., & Mangun, G. R. (1998). Cognitive Neuroscience: The Biology of the Mind. New York, NY: W. W. Norton & Company, Inc.

Gentner, D. (1983). Structure-mapping: A theoretical framework for analogy. Cognitive Science, 7, 155-170.

Gentner, D. (1989). The mechanisms of analogical learning. In S. Vosniadou & Ortony (Eds.), Similarity and Analogical Reasoning. New York, NY: Cambridge University Press.

Gentner, D., & Gunn, V. (2001). Structural alignment facilitates the noticing of differences. Memory & Cognition, 29, 565-577.

Gentner, D., & Markman, A. B. (1994). Structural alignment in comparison: No difference without similarity. Psychological Science, 5, 152-158.

Gentner, D., & Markman, A. B. (1997). Structure mapping in analogy and similarity. American Psychologist, 52, 45-56.

Gibbs, R. W. Jr. (2006). Embodiment and Cognitive Science. New York, NY: Cambridge University Press.

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

Goldstone, R. L., Medin, D. L., & Gentner, D. (1991). Relational similarity and the nonindependence of features in similarity judgments. Cognitive Psychology, 23, 222-262.

Goldstone, R. L., Medin, D. L., & Halberstadt, J. (1997). Similarity in context. Memory & Cognition, 25, 237-255.

Goodman, N. (1972). Problems and Projects. New York, NY: The Bobbs-Merrill Company, Inc.

Gottlieb, J. (2007). From thought to action: The parietal cortex as a bridge between perception, action, and cognition. Neuron, 53, 9-16.

Graf, P., & Schacter, D. L. (1985). Implicit and explicit memory for new associations in normal and amnesic subjects. Journal of Experimental Psychology: Learning, Memory, and Cognition, 11, 501-518.

Hahn, U., & Chater, N. (1997). Concepts and Similarity. In K. Lamberts & D. Shanks (Eds.), Knowledge, Concepts, and Categories (pp. 43-92). Cambridge, MA: The MIT Press.

Itti, L., & Koch, C. (2000). A saliency-based search mechanism for overt and covert shifts of visual attention. Vision Research, 40, 1489-1506.

Itti, L., & Koch, C. (2001). Computational modeling of visual attention. Nature Reviews Neuroscience, 2, 1-11.

Kayser, C., Petkov, C. I., Lippert, M., & Logothetis, N. K. (2005). Mechanisms for allocating auditory attention: An Auditory Saliency Map. Current Biology, 15, 1943-1947.

Keane, M. T., & Costello, F. (2001). Setting limits on analogy: Why conceptual combination is not structural alignment. In D. Gentner, K. J. Holyoak, & B. N. Kokinov (Eds.), The Analogical Mind: Perspectives from Cognitive Science. Cambridge, MA: The MIT Press.

Koch, C., & Ullman, S. (1985). Shifts in selective visual attention: towards the underlying neural circuitry. Human Neurobiology, 4, 219-227.

Kustov, A. A., & Robinson, D. L. (1996). Shared neural control of attentional shifts and eye movements. Nature, 384, 74-77.

Lakoff, G. (1987). Women, Fire, and Dangerous Things. Chicago, IL: The University of Chicago Press.

Lakoff, G., & Johnson, M. (1980). Metaphors We Live By. Chicago, IL: The University of Chicago Press.

Lamberts, K., & Shanks, D. (Eds.). (1997). Knowledge, Concepts, and Categories. Cambridge, MA: The MIT Press.

Li, Z. (2002). A saliency map in primary visual cortex. TRENDS in Cognitive Sciences, 6, 9-16.

Lupyan, G., Rakison, D. H., & McClelland, J. L. (2007). Language is not just for talking. *Psychological Science, 18,* 1077-1083.

Malt, B. C., Sloman, S. A., Gennari, S., Shi, M., & Wang, Y. (1999). Knowing versus naming: Similarity and the linguistic categorization of artifacts. *Journal of Memory and Language, 40,* 230-262.

Markman, A. B. (1996). Structural alignment in similarity and difference judgments. Psychonomic Bulletin & Review, 3, 227-230.

Markman, A. B., & Ross, B. H. (2003). Category use and category learning. Psychological Bulletin, 129, 592-613.

Markman, A. B., & Gentner, D. (1993a). Structural alignment during similarity comparisons. *Cognitive Psychology, 25*, 431-467.

Markman, A. B., & Gentner, D. (1993b). Splitting the differences: A structural alignment view of similarity. *Journal of Memory and Language*, *32*, 517-535.

Markman, A. B., & Gentner, D. (1993c). All differences are not created equal: A structural alignment view of similarity. *Proceedings of the Fifteenth Annual Conference of the Cognitive Science Society,* 682-686.

Mazer, J. A., & Gallant, J. L. (2003). Goal-related activity in V4 during free viewing visual search: Evidence for a ventral stream visual salience map. Neuron, 40, 1241-1250.

McClelland, J. L., & Rogers, T. T. (2003). The parallel distributed processing approach to semantic cognition. *Nature Reviews Neuroscience, 4,* 1-14.

McGurk, H., & MacDonald, J. (1976). Hearing lips and seeing voices. Nature, 264, 746-748.

Medin, D. L., Goldstone, R. L., & Gentner, D. (1990). Similarity involving attributes and relations: Judgments of similarity and difference are not inverses. *Psychological Science, 1,* 64-69.

Medin, D. L., Goldstone, R. L., & Gentner, D. (1993). Respects for similarity. Psychological Review, 100, 254-278.

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

Mervis, C. B., Catlin, J., & Rosch, E. (1976). Relationships among goodness-of-example, category norms, and word frequency. Bulletin of the Psychonomic Society, 7, 283-284.

Murphy, G. L. (2002). The Big Book of Concepts. Cambridge, MA: The MIT Press.

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

Ortony, A. (1979). Beyond literal similarity. Psychological Review, 86, 161-180.

Palmer, S. E. (1975). The effects of contextual scenes on the identification of objects. *Memory & Cognition, 3,* 519-526.

Petrov, A. A., Jilk, D. J., & O'Reilly, R. C. (2010). The Leabra architecture: Specialization without modularity. *Behavioral and Brain Sciences, 33,* 286-287.

Reynolds, J. H., Pasternak, T., & Desimone, R. (2000). Attention increases sensitivity of V4 neurons. Neuron, 26, 703-714.

Reynolds, J. H., & Desimone, R. (2003). Interacting roles of attention and visual salience in V4. Neuron, 37, 853-863.

Robinson, D. L., & Petersen, S. E. (1992). The pulvinar and visual salience. Trends in Neuroscience, 15, 127-132.

Rogers, T., & McClelland, J. (2004). Semantic Cognition: A Parallel Distributed Processing Approach. Cambridge, MA: The MIT Press.

Romano, M. (2006). Intentional and incidental classification learning in category use. *Proceedings of the 28th Annual Conference of the Cognitive Science Society* (pp. 2047-2052). Austin, TX: Cognitive Science Society.

Rosch, E., & Mervis, C. (1975). Family resemblances: Studies in the internal structure of categories. Cognitive Psychology, 7, 573-605.

Rosch, E., Simpson, C., & Miller, R. S. (1976). Structural bases of typicality effects. Journal of Experimental Psychology: Human Perception and Performance, 2, 491-502.

Rumelhart, D. E., & McClelland, J. L. (1986). Parallel Distributed Processing: Explorations in the Microstructure of Cognition. Cambridge, MA: The MIT Press.

Schiller, P. H. (1992). The ON and OFF channels of the visual system. *Trends in Neuroscience, 15,* 86-92.

Shepard, R. N. (1962). The analysis of proximities: Multidimensional scaling with an unknown distance function: Part I. Psychometrika, 27, 125-140.

Shepard, R. N. (1987). Toward a universal law of generalization for psychological science. Science, 237, 1317-1323.

Shepard, R. N., Hovland, C. I., & Jenkins, H. M. (1961). Learning and memorization of classifications. *Psychological Monographs, 75,* (13, Whole No. 517).

Smith, E. E., & Medin, D. L. (1981). Categories and Concepts. Cambridge, MA: Harvard University Press.

Spivey, M. (2007). The Continuity of Mind. New York, NY: Oxford University Press.

Spivey, M. J., Grosjean, M., & Knoblich, G. (2005). Continuous attraction toward phonological competitors. *Proceedings of the National Academy of Sciences, 102,* 10393-10398.

Stirk, J. A., & Underwood, G. (2007). Low-level visual saliency does not predict change detection in natural scenes. Journal of Vision, 7, 1-10.

Tijsseling, A. G., & Gluck, M. A. (2002). A connectionist approach to processing dimensional interaction. *Connection Science, 14,* 1-48.

Treue, S. (2003). Visual attention: the where, what, how, and why of saliency. Current Opinion in Neurobiology, 13, 428-432.

Turner, M., & Fauconnier, G. (1995). Conceptual integration and formal expression. Metaphor and Symbolic Activity, 10, 183-204.

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

Tversky, A., & Gati, I. (1982). Similarity, separability, and the triangle inequality. Psychological Review, 89, 123-154.

Varela, F. J., Thompson, E., & Rosch, E. (1991). The Embodied Mind: Cognitive Science and Human Experience. Cambridge, MA: The MIT Press.

Figure 5.1.1: Separable Feature Stimuli Set

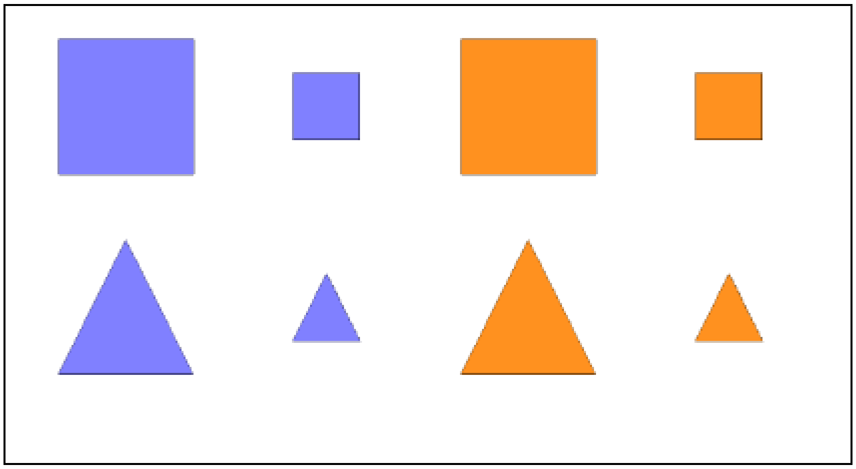

Figure 5.1.2: Integral Feature Stimuli Set

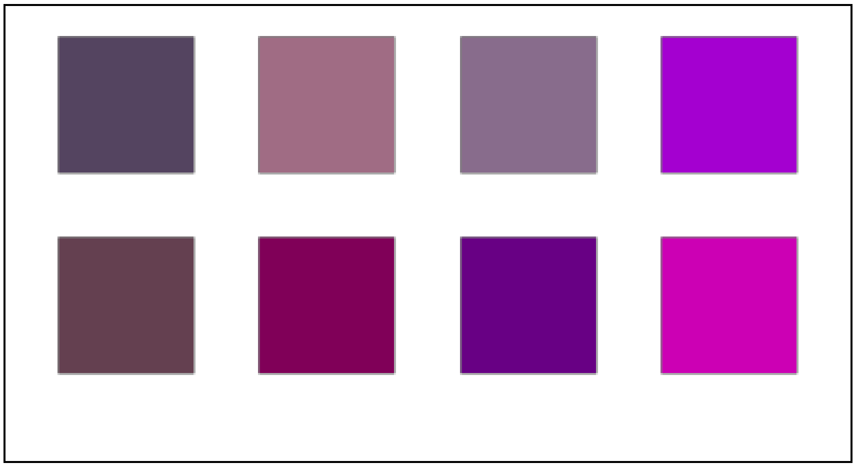

Figure 5.1.3: Results: Separable Features

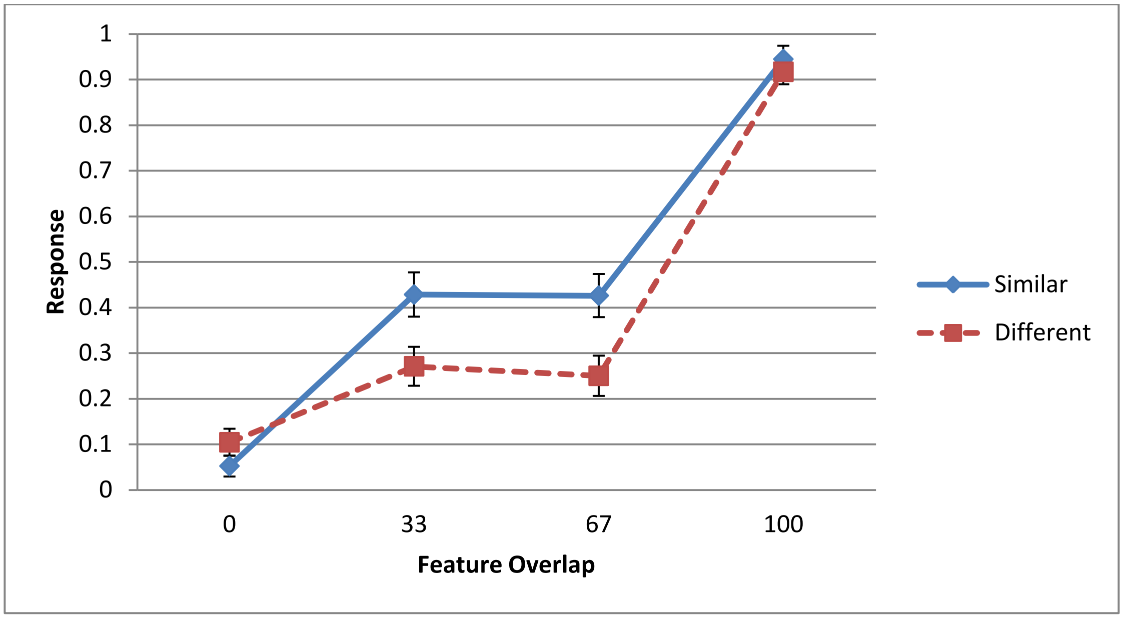

Figure 5.1.4: Results: Integral Features

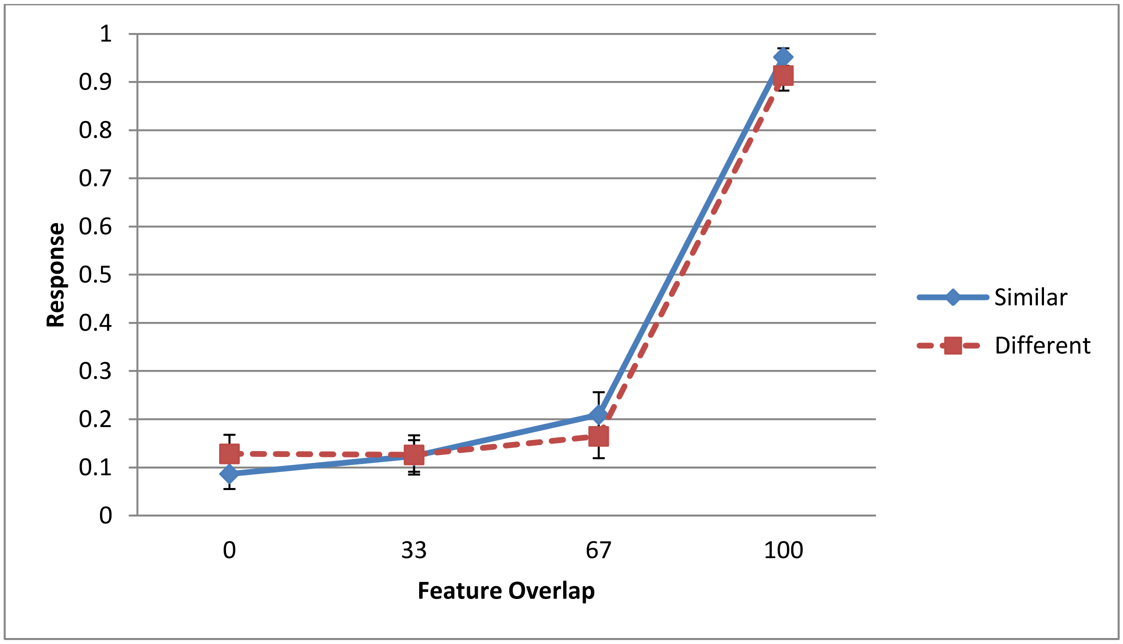

Figure 5.2.1: Model: Separable Features

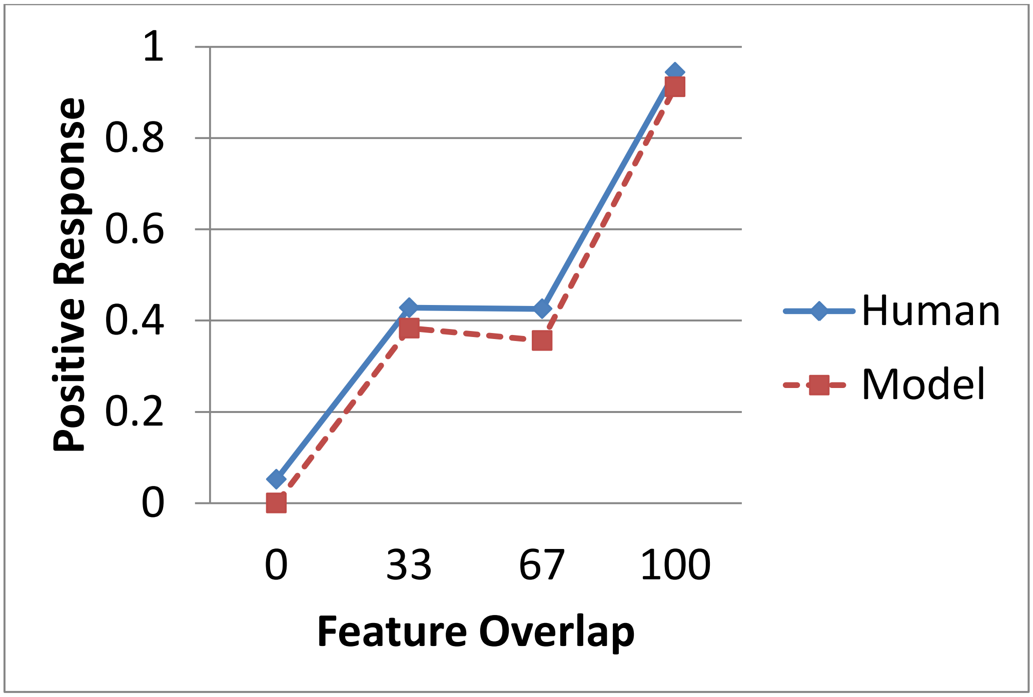

Figure 5.2.2: Model: Integral Features

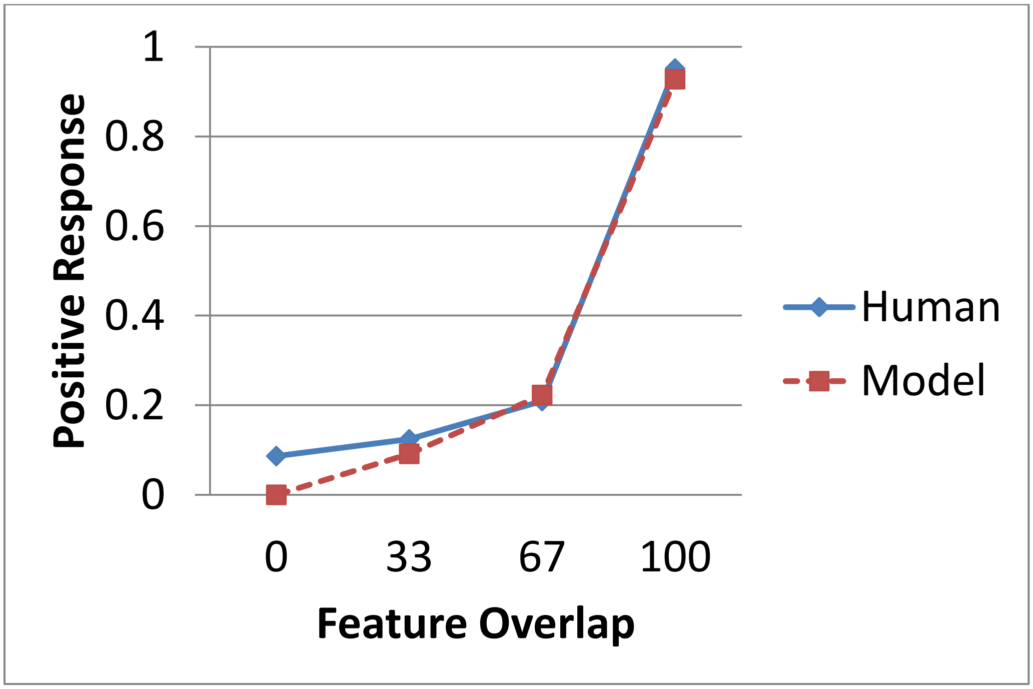

Figure 6.1.1: Separable Feature Stimuli, Near Categories

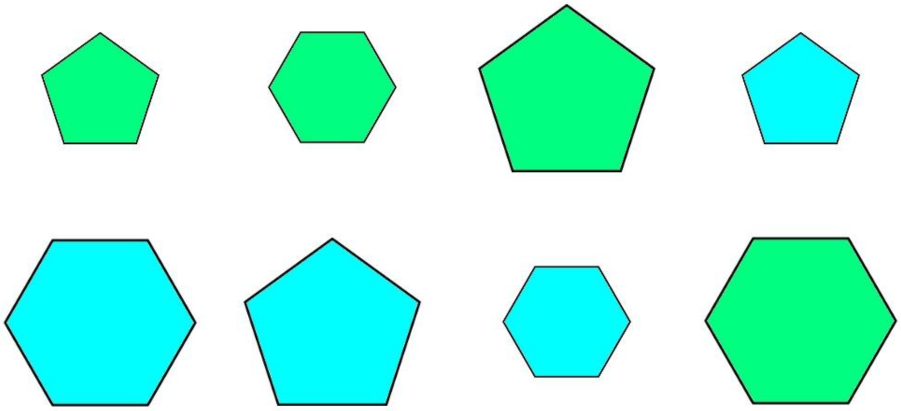

Figure 6.1.2: Separable Feature Stimuli, Far Categories

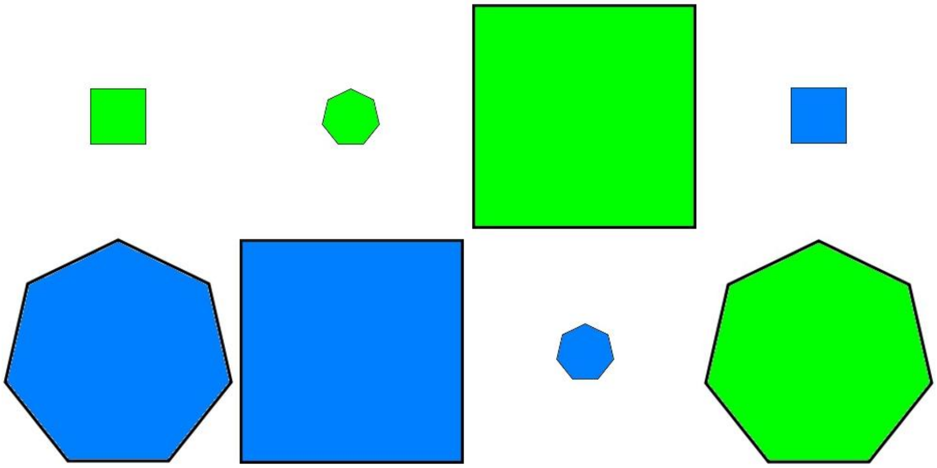

Figure 6.1.3: Similarity Ratings for Separable Features, Near Categories, Within-Category Comparisons

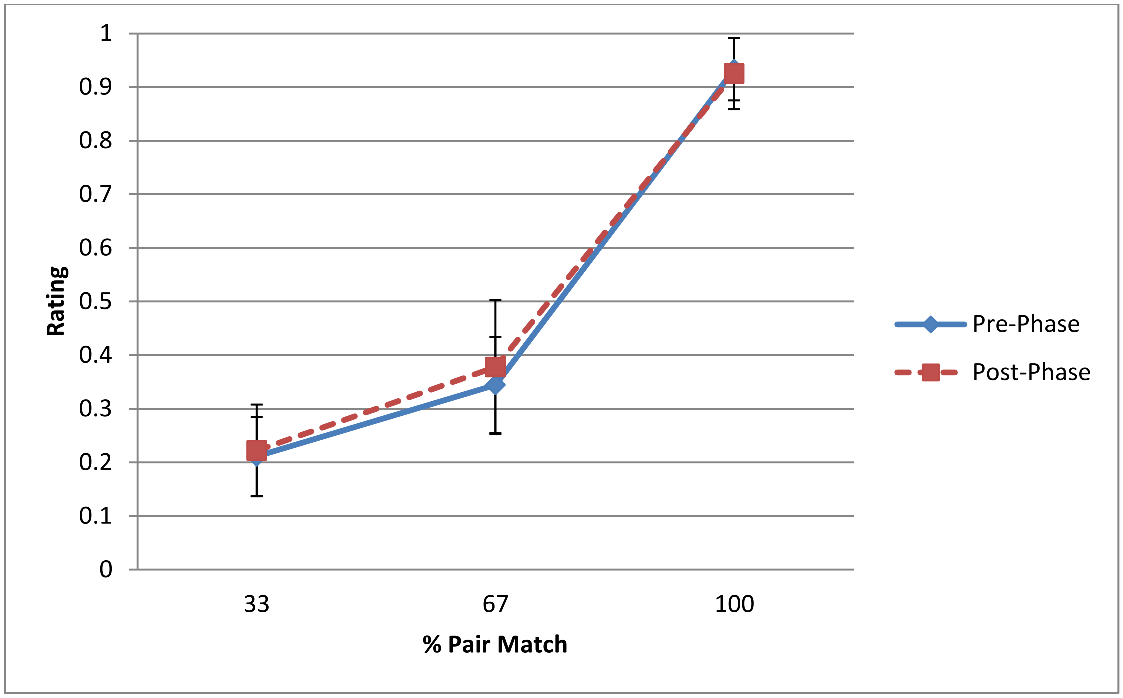

Figure 6.1.4: Difference Ratings for Separable Features, Near Categories, Within-Category Comparisons

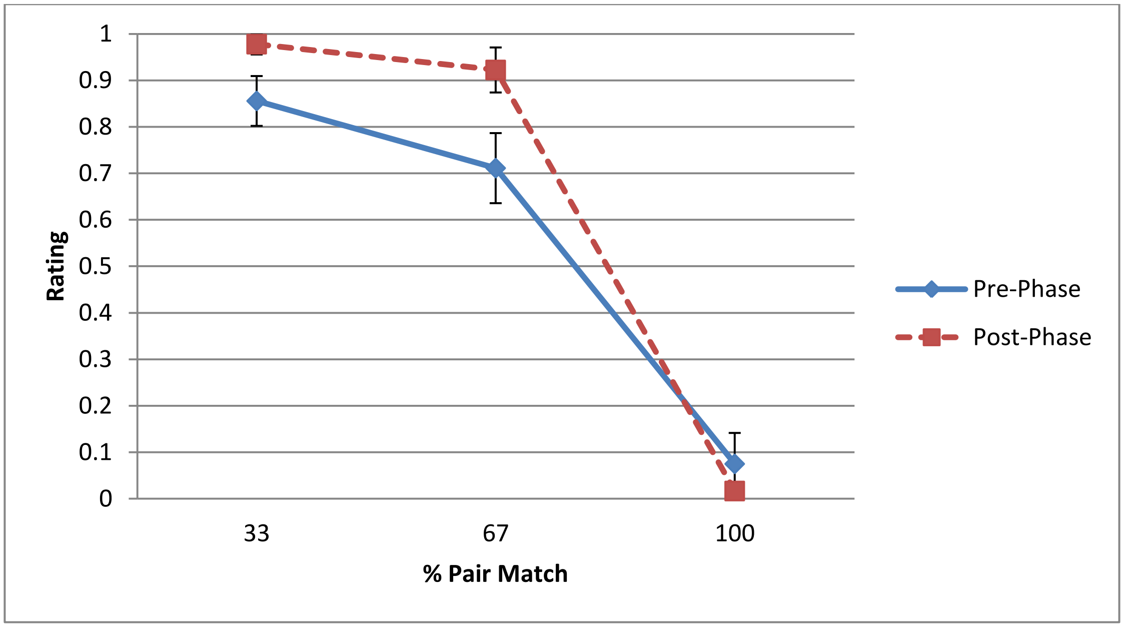

Figure 6.1.5: Similarity Ratings for Separable Features, Near Categories, Between-Categories Comparisons

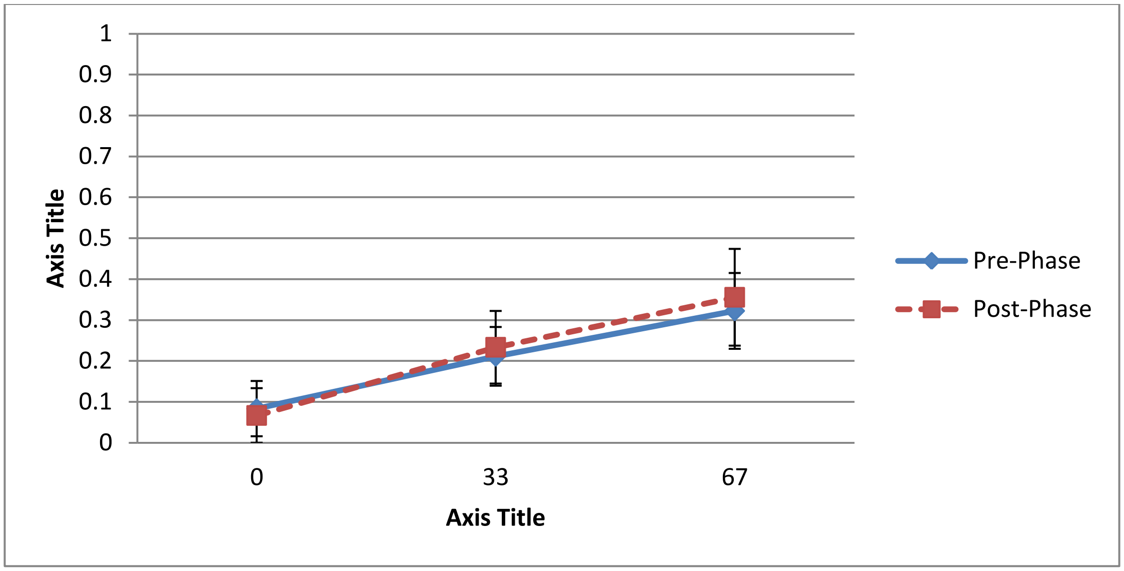

Figure 6.1.6: Difference Ratings for Separable Features, Near Categories, Between-Categories Comparisons

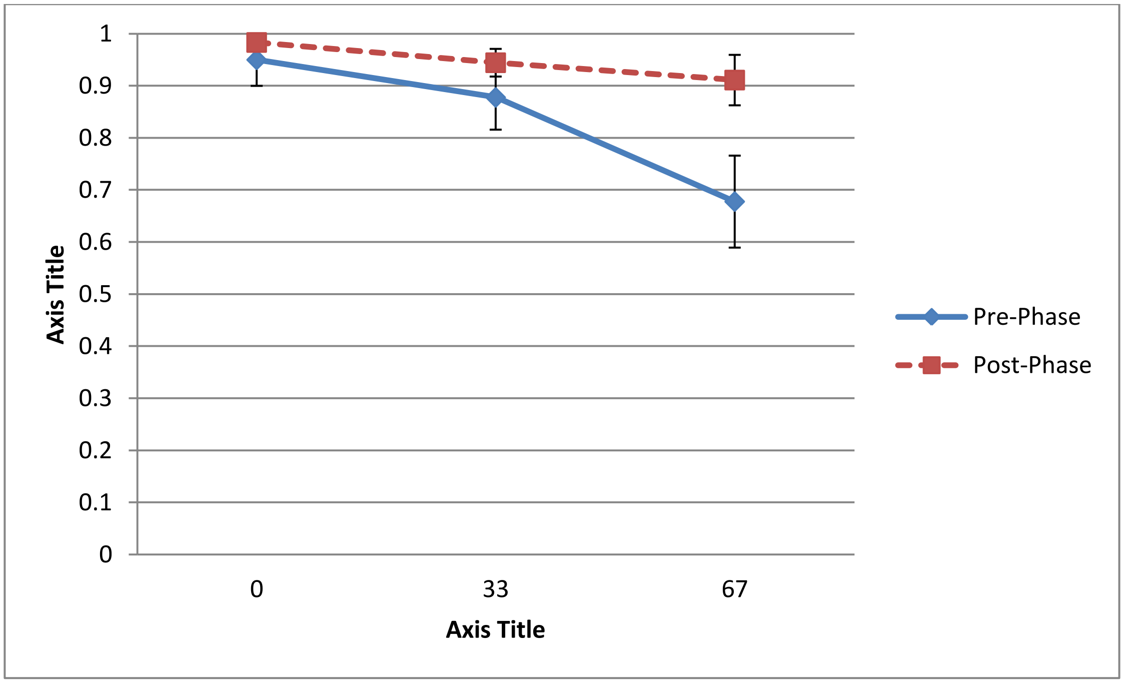

Figure 6.1.7: Similarity Ratings for Separable Features, Far Categories, Within-Category Comparisons

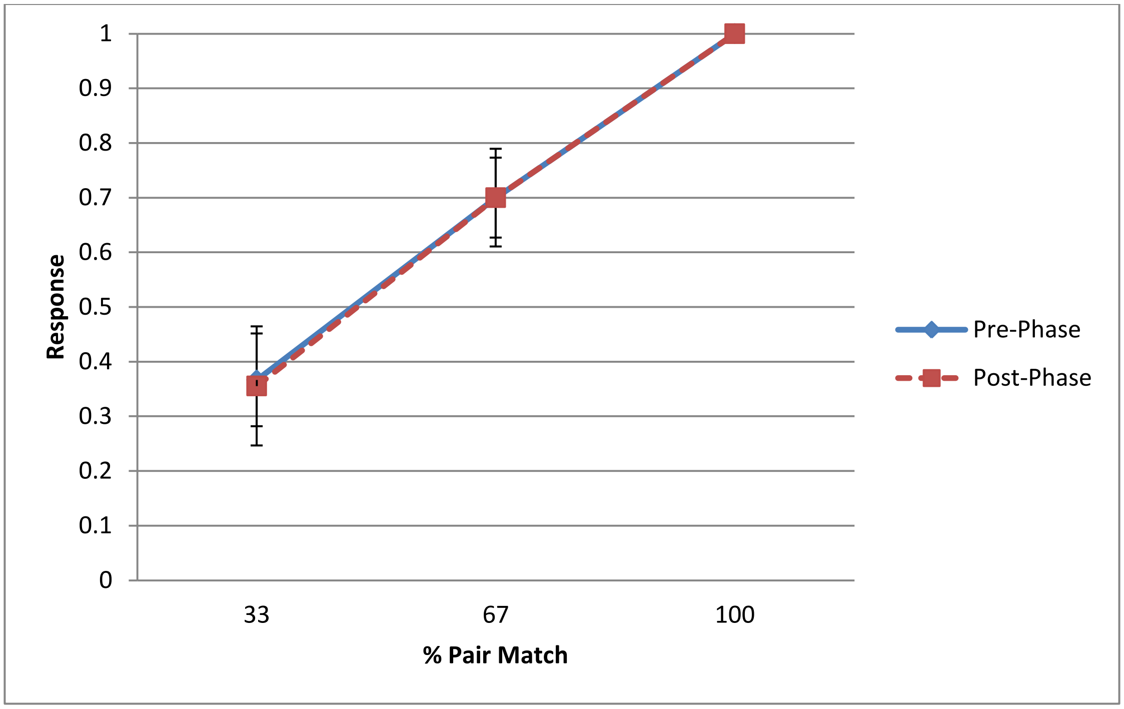

Figure 6.1.8: Difference Ratings for Separable Features, Far Categories, Within-Category Comparisons

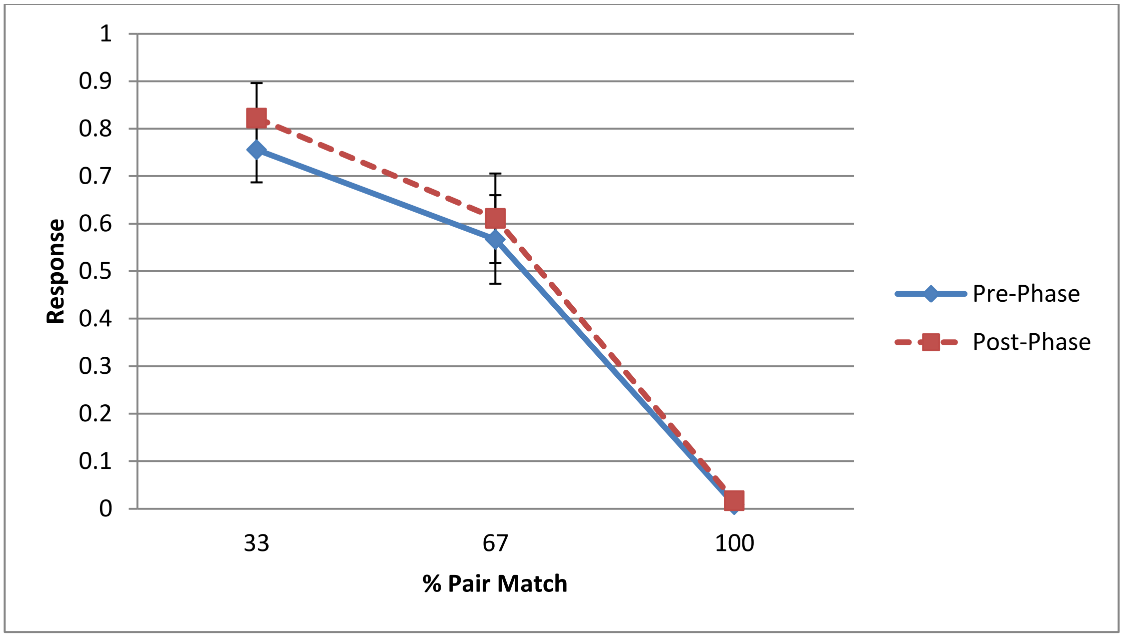

Figure 6.1.9: Similarity Ratings for Separable Features, Far Categories, Between-Categories Comparisons

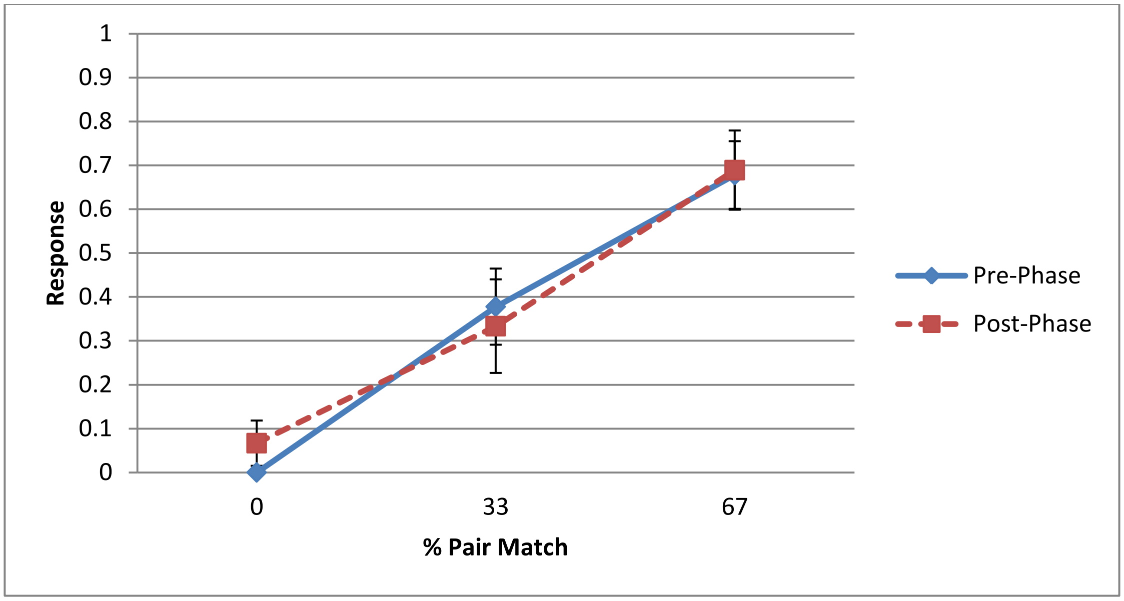

Figure 6.1.10: Difference Ratings for Separable Features, Far Categories, Between-Categories Comparisons

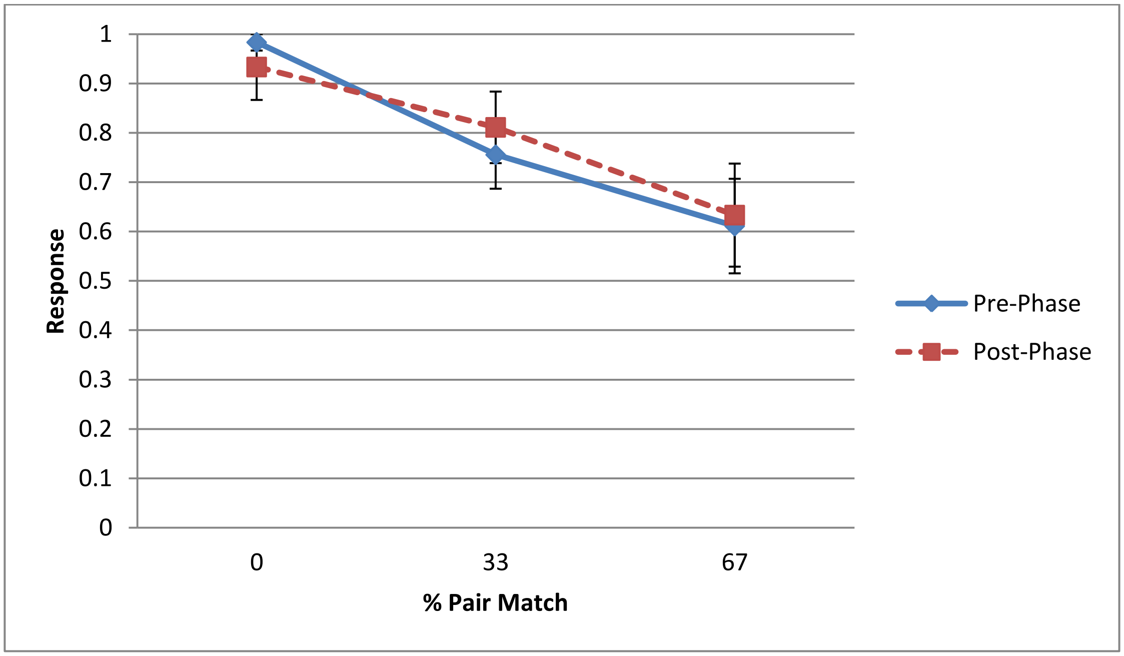

Figure 6.1.11: Similarity Area Under the Curve for Separable Features, Near Categories, Within-Category Comparisons

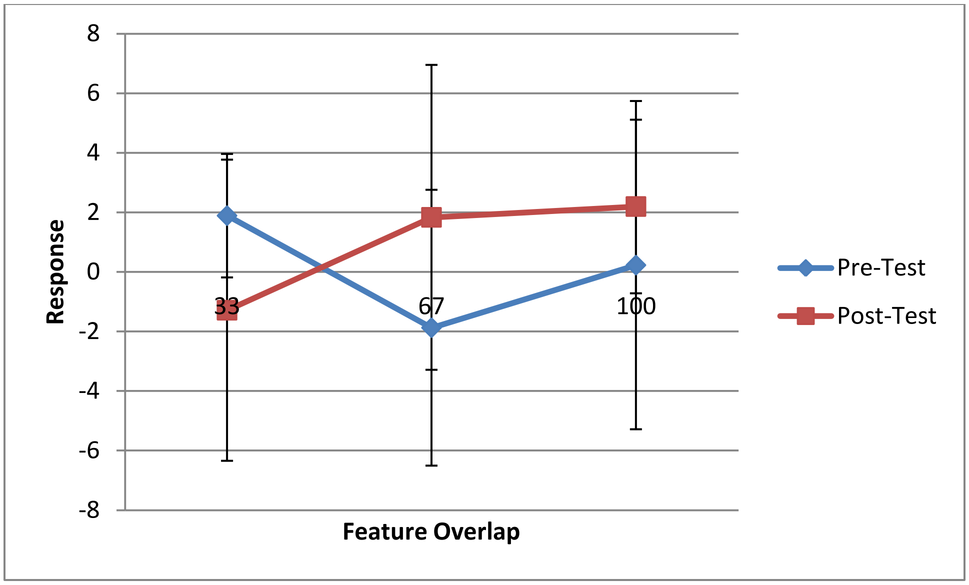

Figure 6.1.12: Difference Area Under the Curve for Separable Features, Near Categories, Within-Category Comparisons

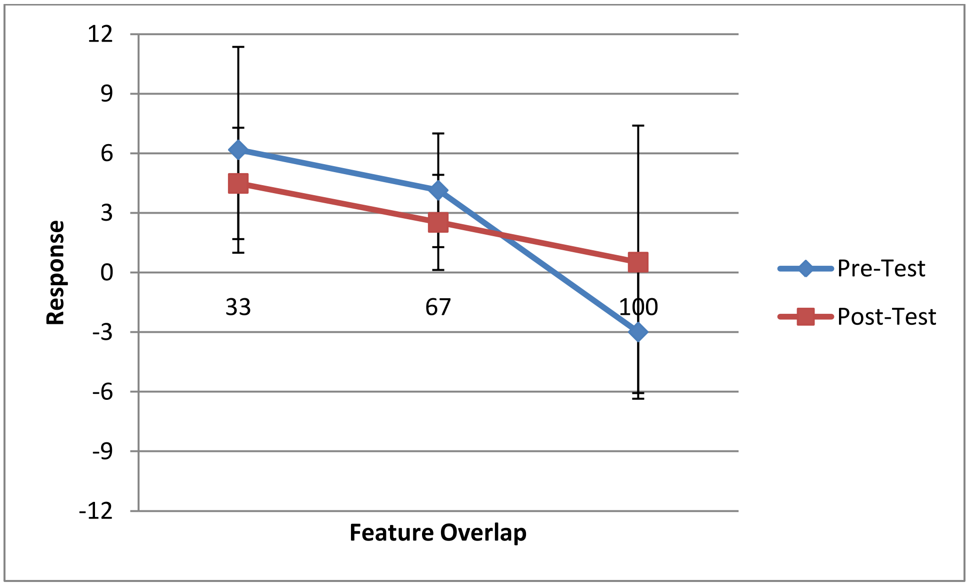

Figure 6.1.13: Similarity Area Under the Curve for Separable Features, Near Categories, Between-Categories Comparisons

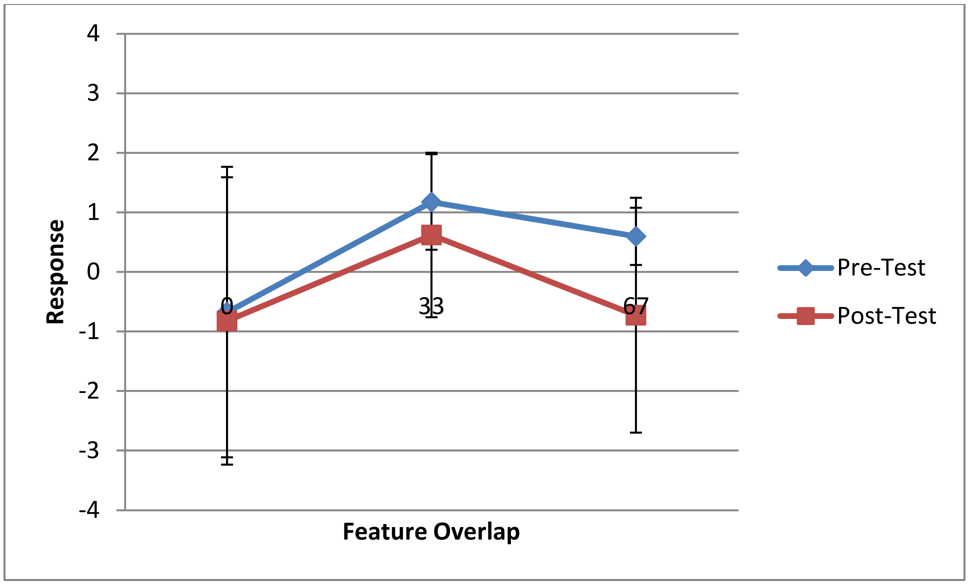

Figure 6.1.14: Difference Area Under the Curve for Separable Features, Near Categories, Between-Categories Comparisons

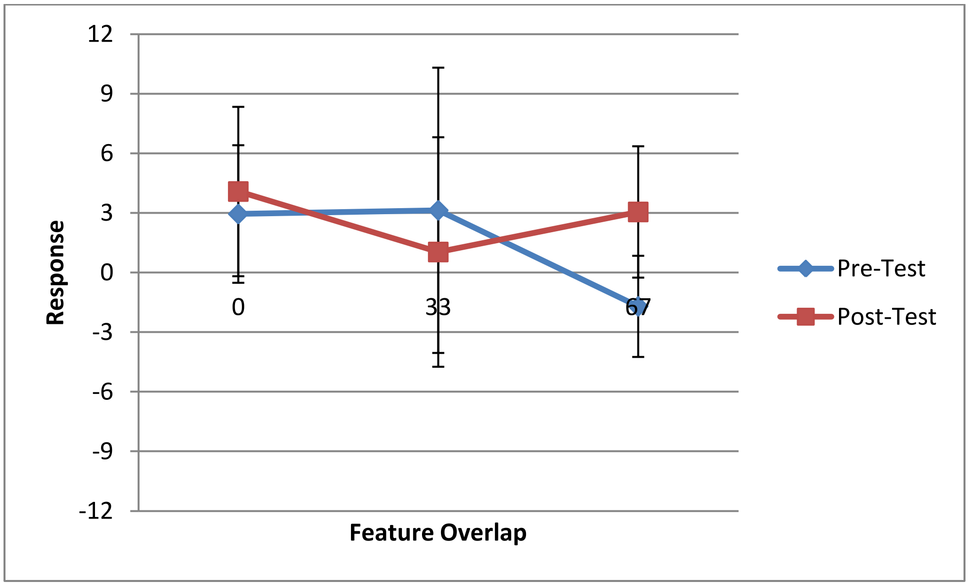

Figure 6.1.15: Similarity Area Under the Curve for Separable Features, Far Categories, Within-Category Comparisons

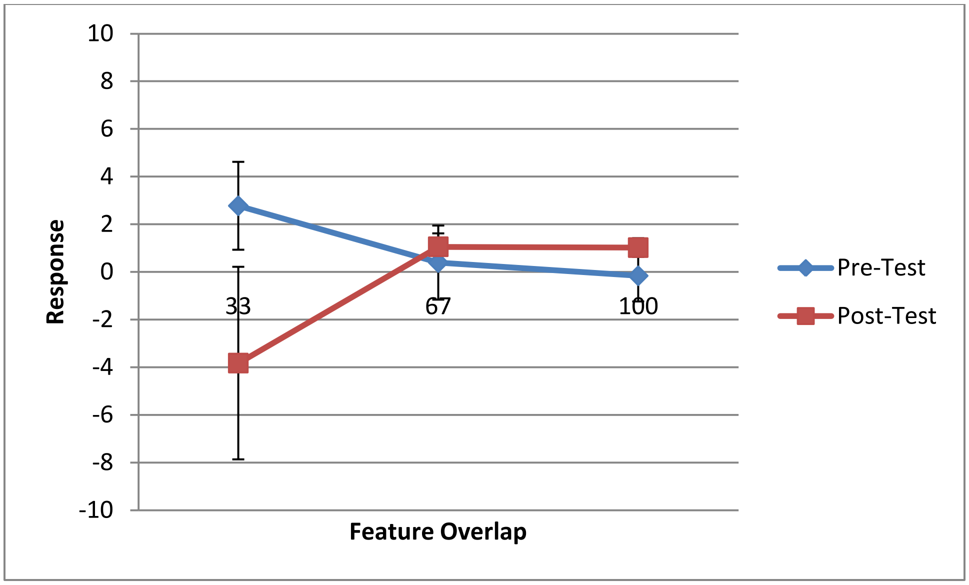

Figure 6.1.16: Difference Area Under the Curve for Separable Features, Far Categories, Within-Category Comparisons

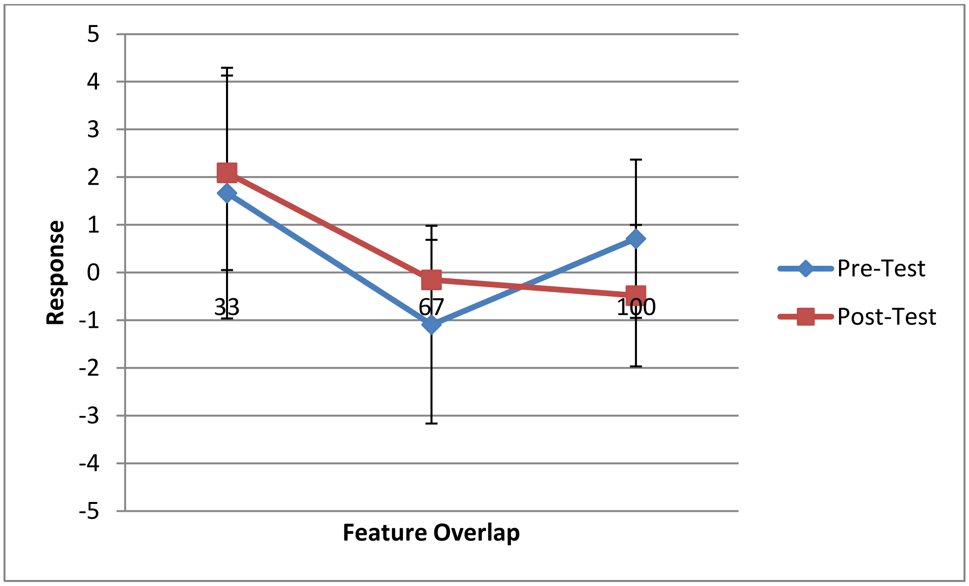

Figure 6.1.17: Similarity Area Under the Curve for Separable Features, Far Categories, Between-Categories Comparisons

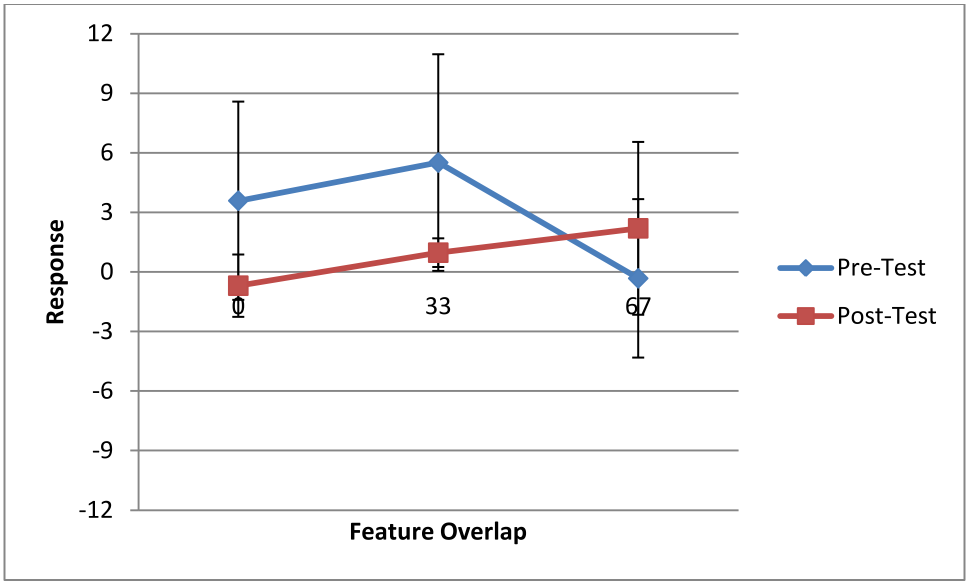

Figure 6.1.18: Difference Area Under the Curve for Separable Features, Far Categories, Between-Categories Comparisons

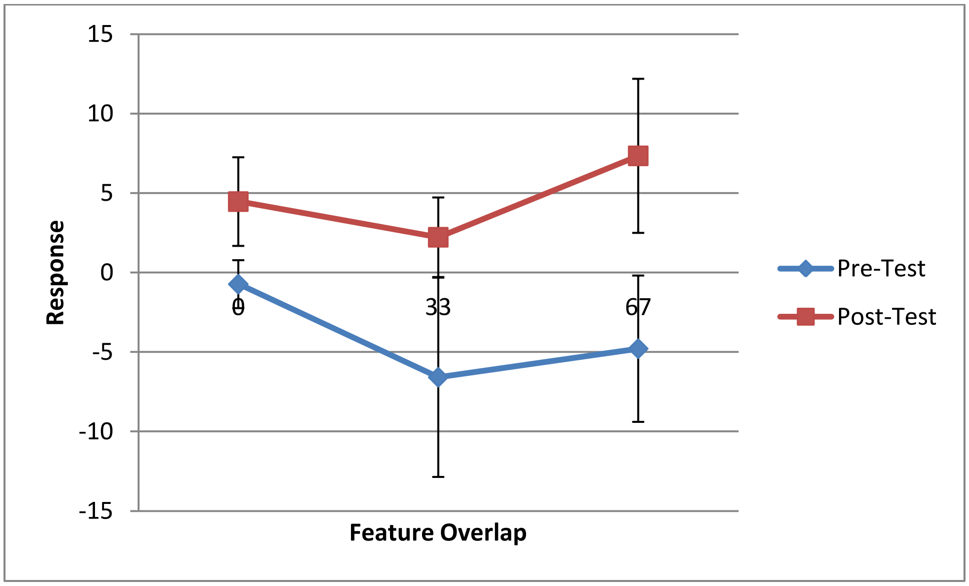

Figure 6.2.1: Integral Feature Stimuli, Near Categories

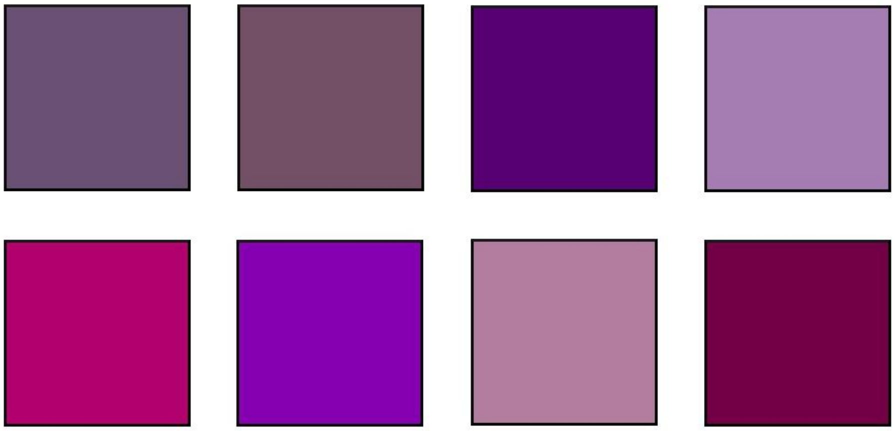

Figure 6.2.2: Integral Feature Stimuli, Far Categories

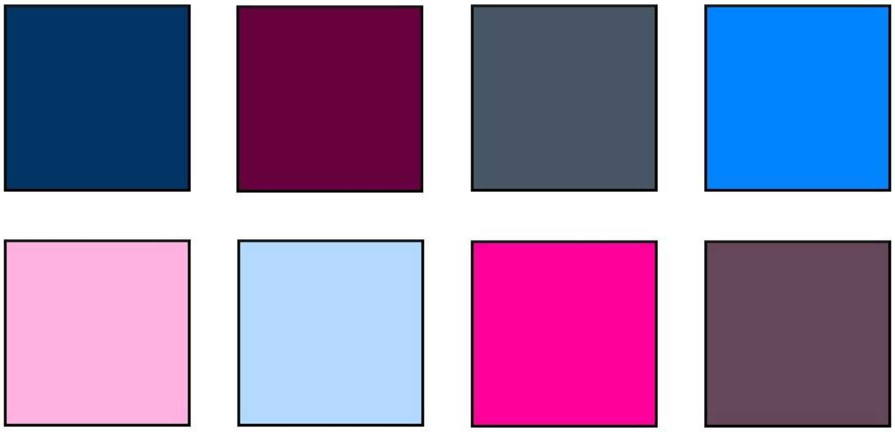

Figure 6.2.3: Similarity Ratings for Integral Features, Near Categories, Within-Category Comparisons

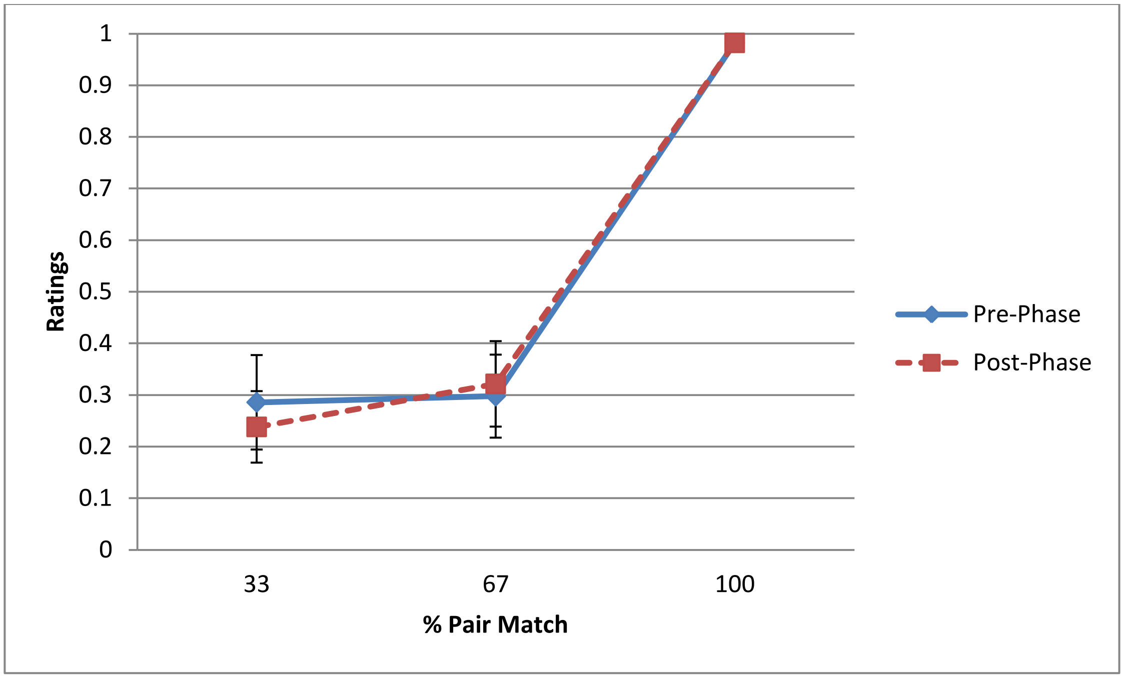

Figure 6.2.4: Difference Ratings for Integral Features, Near Categories, Within-Category Comparisons

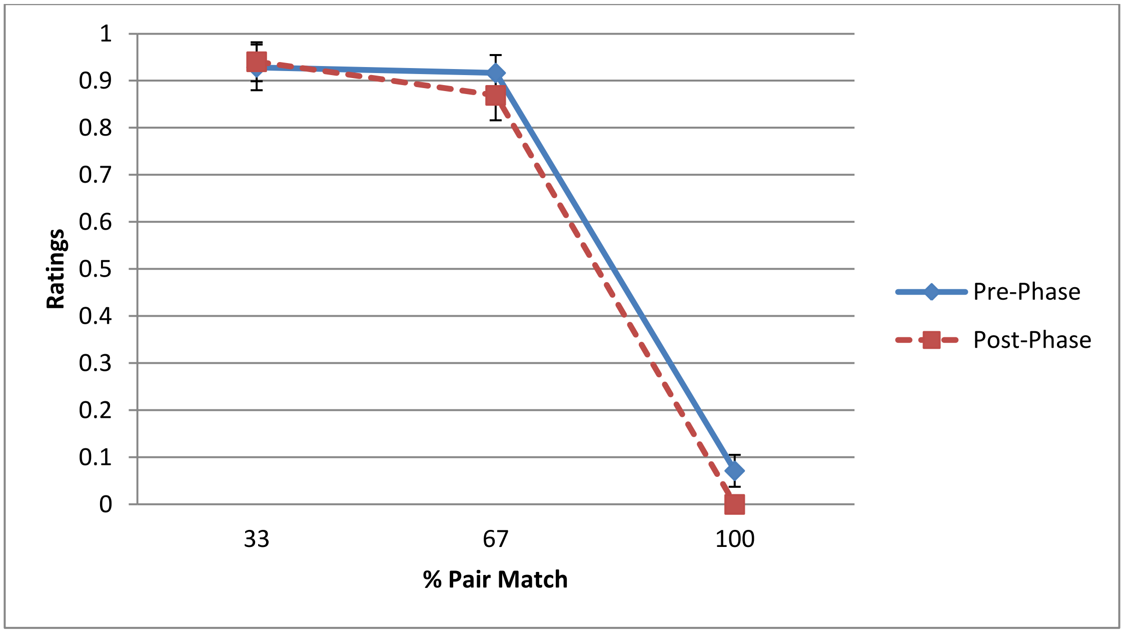

Figure 6.2.5: Similarity Ratings for Integral Features, Near Categories, Between-Categories Comparisons

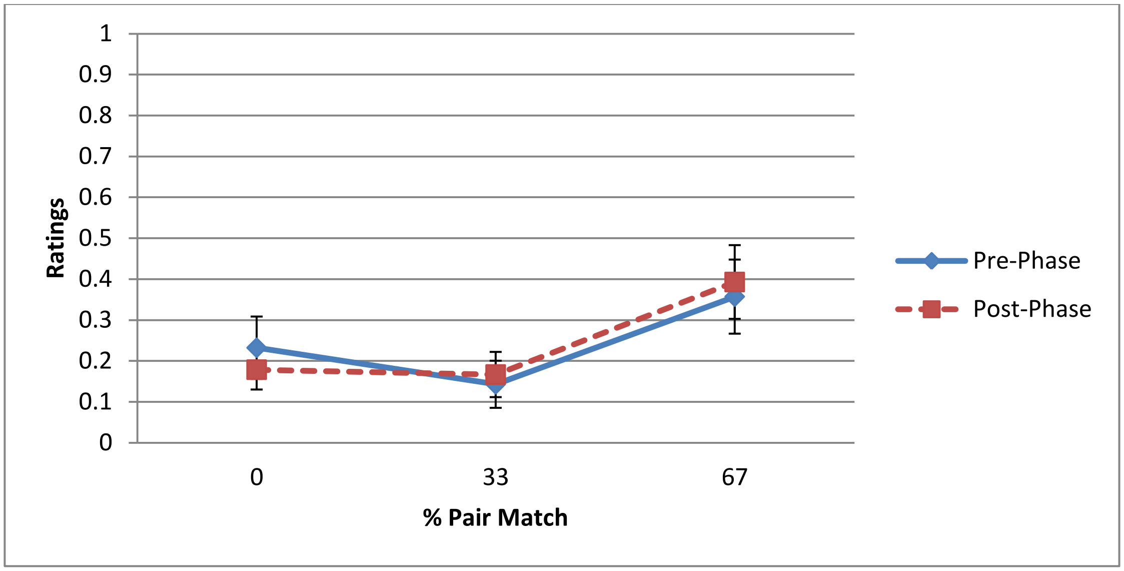

Figure 6.2.6: Difference Ratings for Integral Features, Near Categories, Between-Categories Comparisons

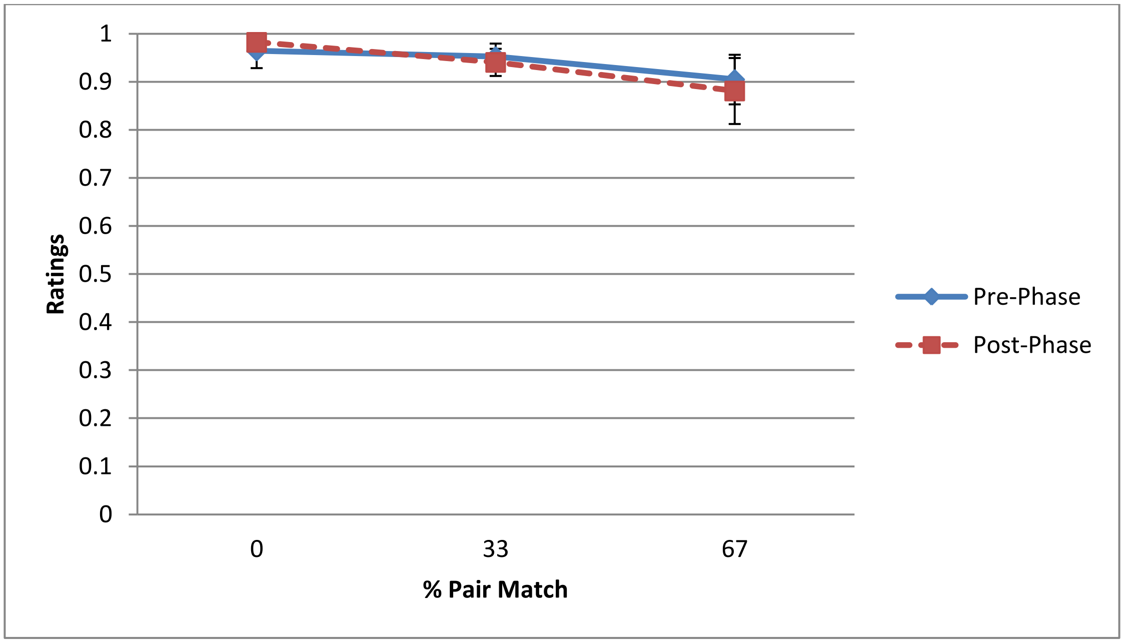

Figure 6.2.7: Similarity Ratings for Integral Features, Far Categories, Within-Category Comparisons

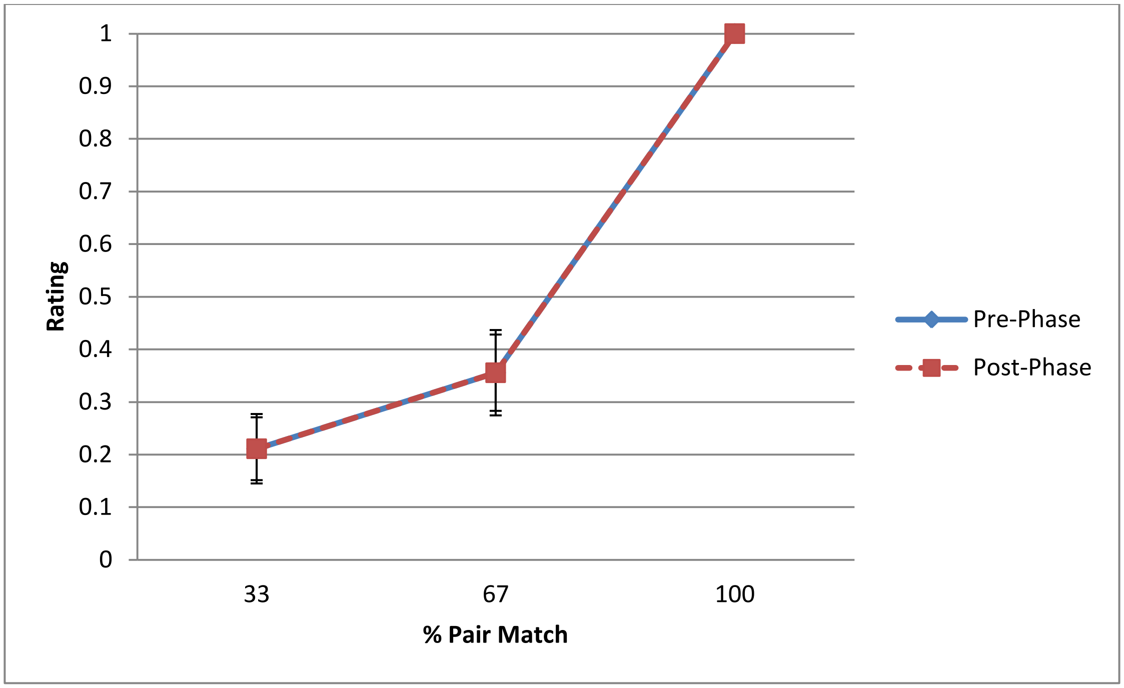

Figure 6.2.8: Difference Ratings for Integral Features, Far Categories, Within-Category Comparisons

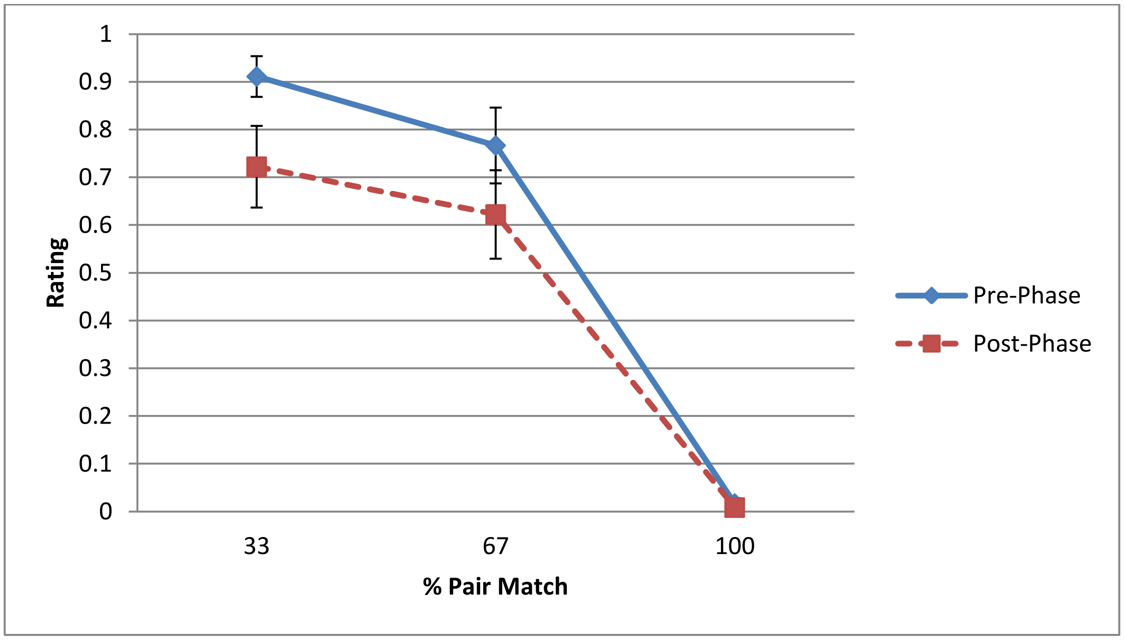

Figure 6.2.9: Similarity Ratings for Integral Features, Far Categories, Between-Categories Comparisons

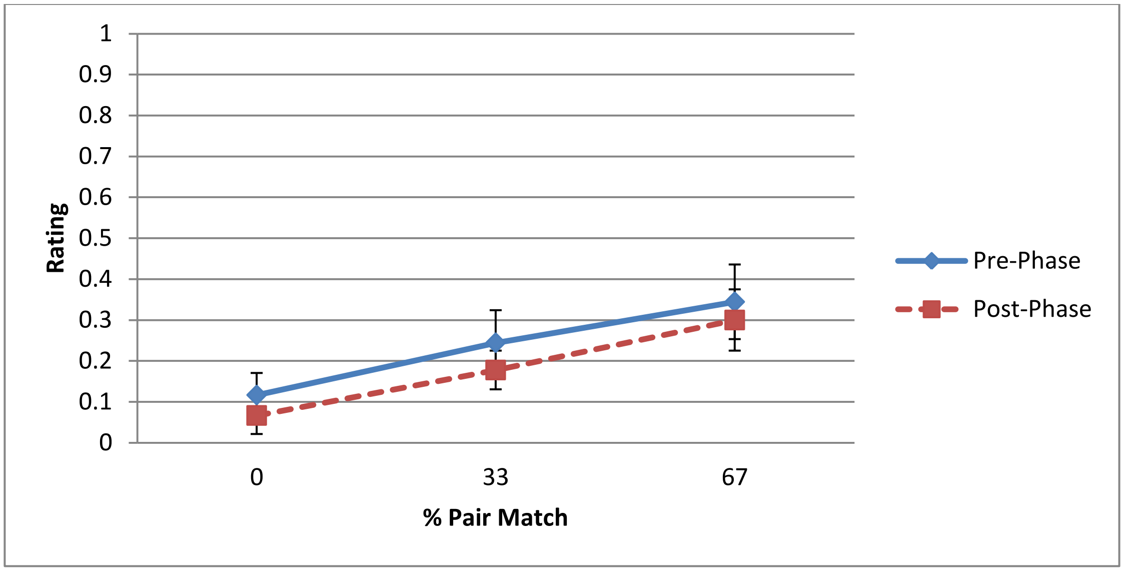

Figure 6.2.10: Difference Ratings for Integral Features, Far Categories, Between-Categories Comparisons

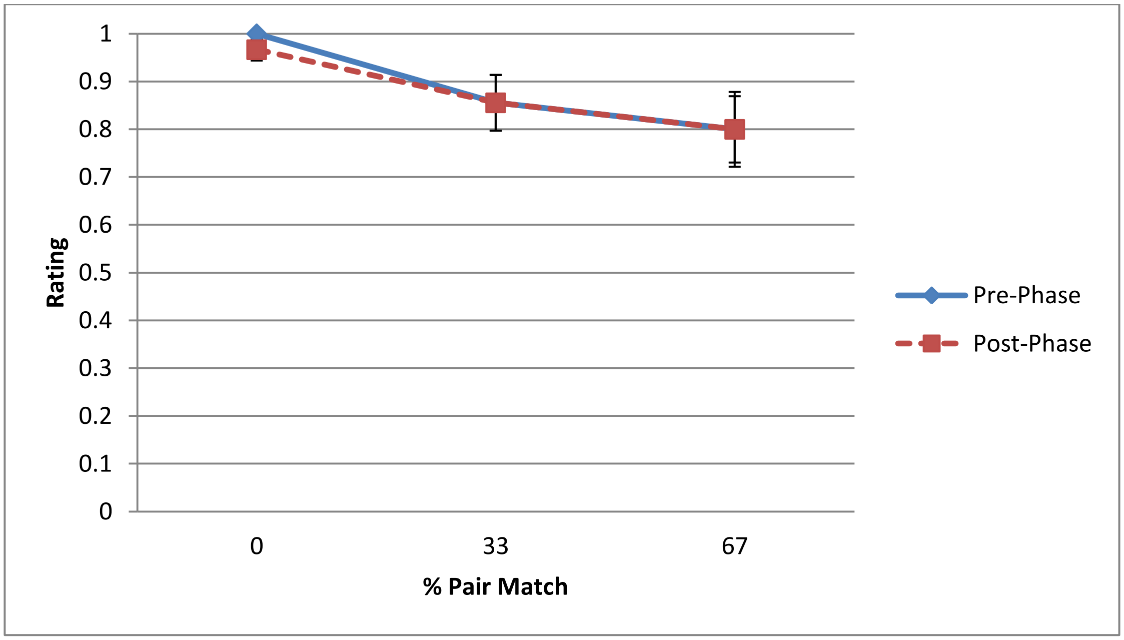

Figure 6.2.11: Similarity Area Under the Curve for Integral Features, Near Categories, Within-Category Comparisons

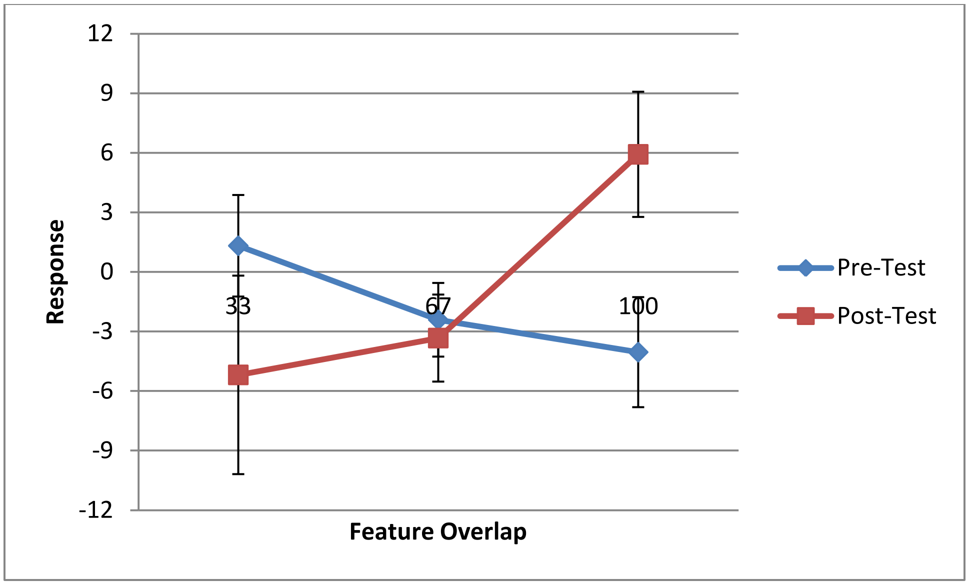

Figure 6.2.12: Difference Area Under the Curve for Integral Features, Near Categories, Within-Category Comparisons

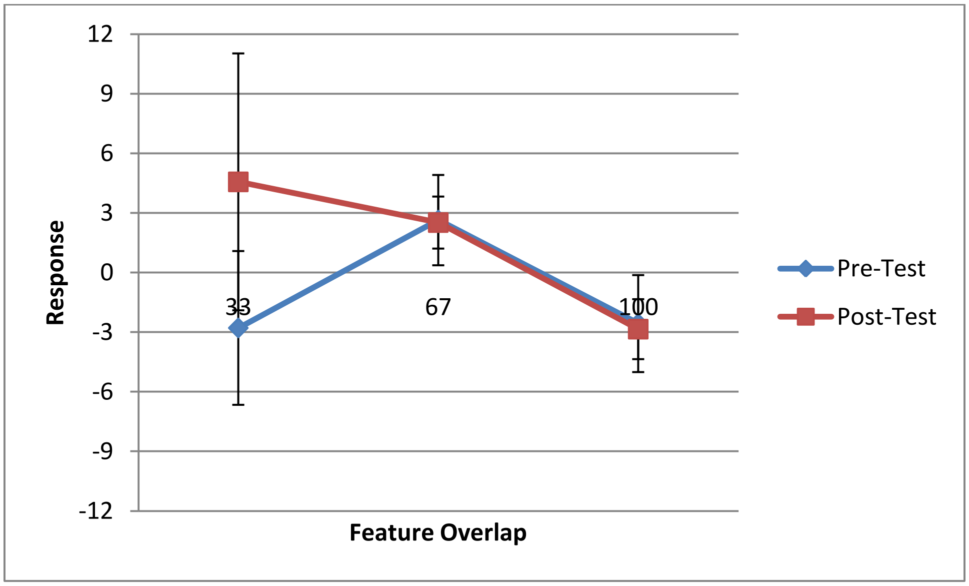

Figure 6.2.13: Similarity Area Under the Curve for Integral Features, Near Categories, Between-Categories Comparisons

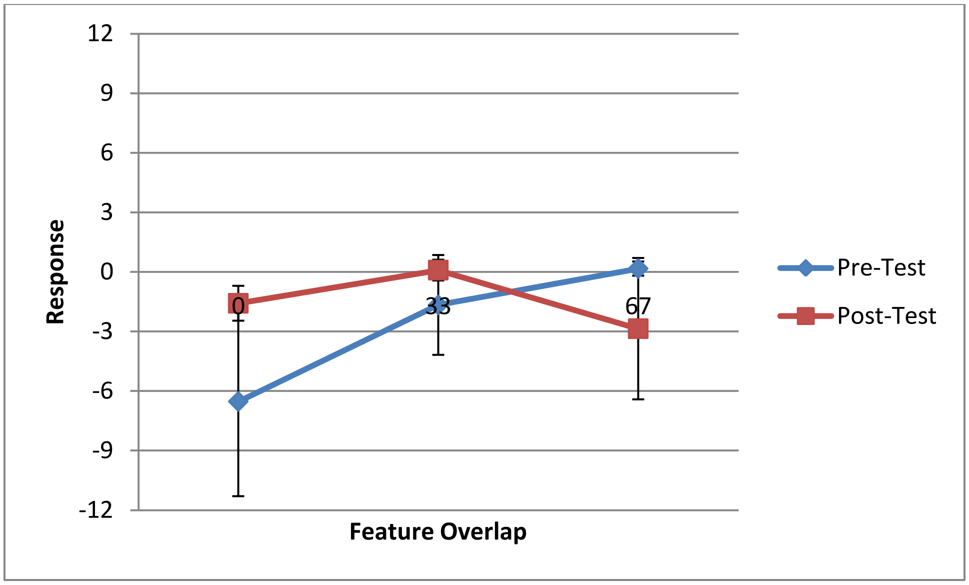

Figure 6.2.14: Difference Area Under the Curve for Integral Features, Near Categories, Between-Categories Comparisons

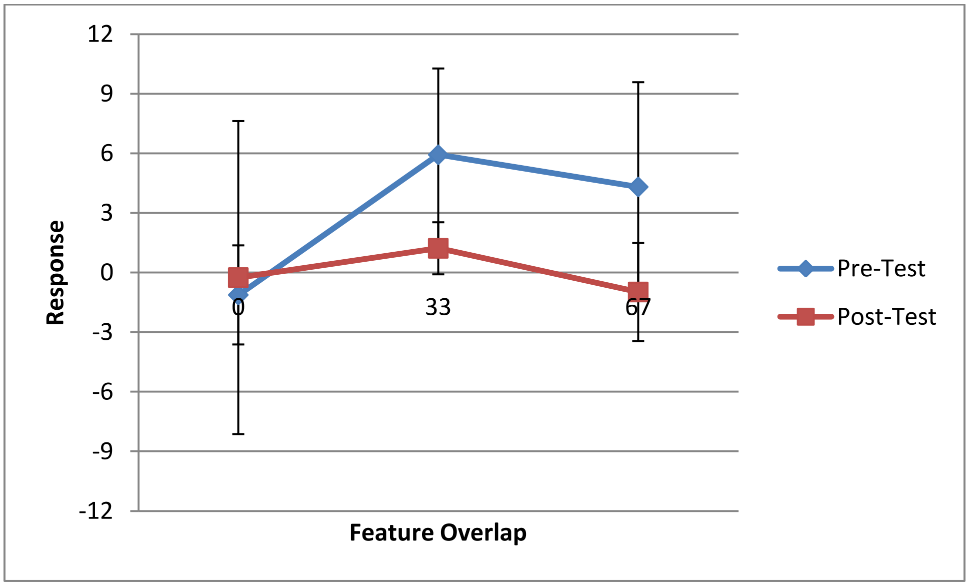

Figure 6.2.15: Similarity Area Under the Curve for Integral Features, Far Categories, Within-Category Comparisons

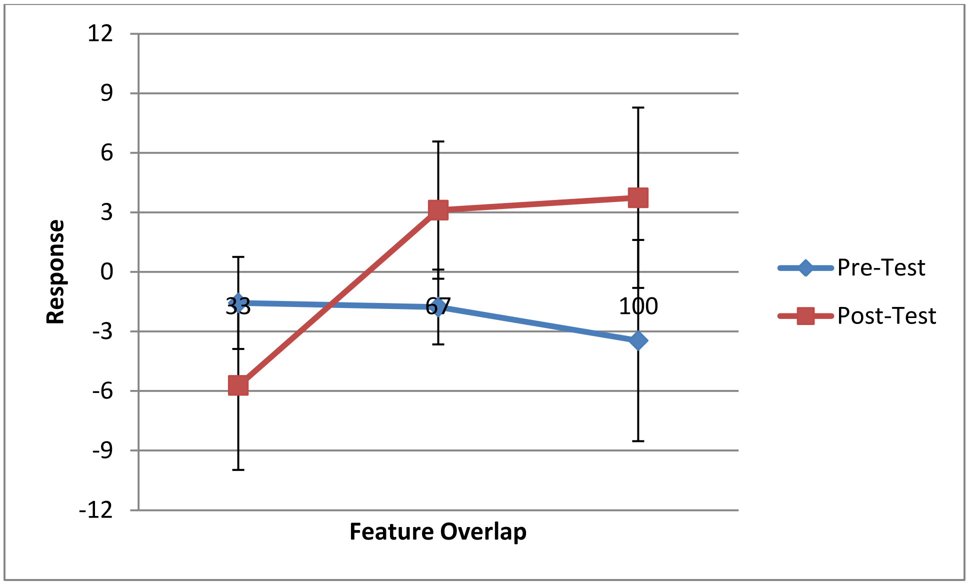

Figure 6.2.16: Difference Area Under the Curve for Integral Features, Far Categories, Within-Category Comparisons

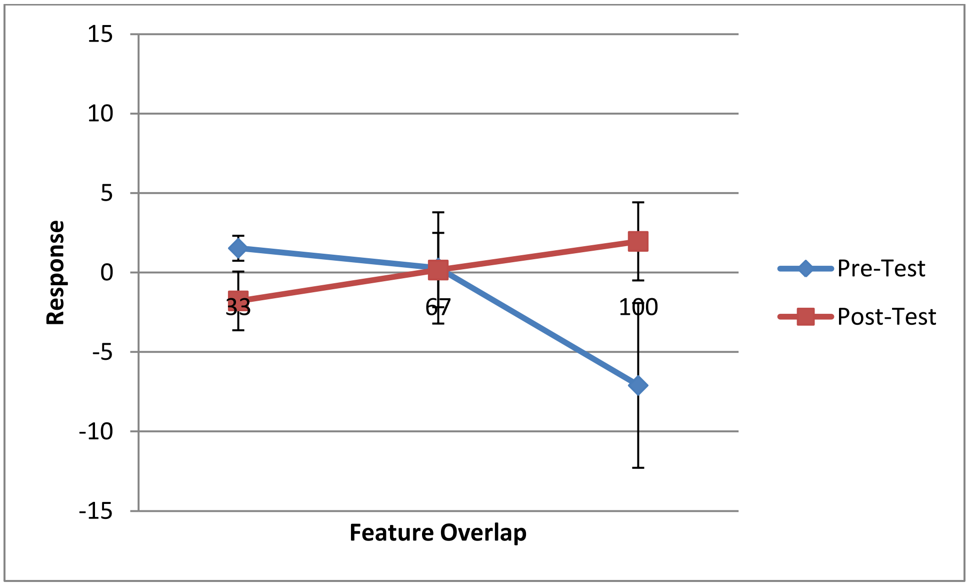

Figure 6.2.17: Similarity Area Under the Curve for Integral Features, Far Categories, Between-Categories Comparisons

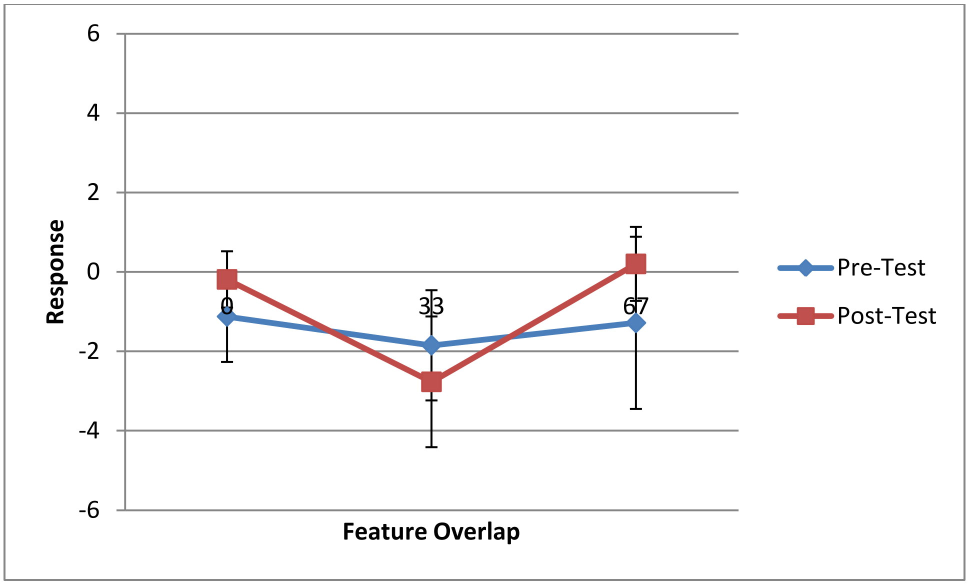

Figure 6.2.18: Difference Area Under the Curve for Integral Features, Far Categories, Between-Categories Comparisons

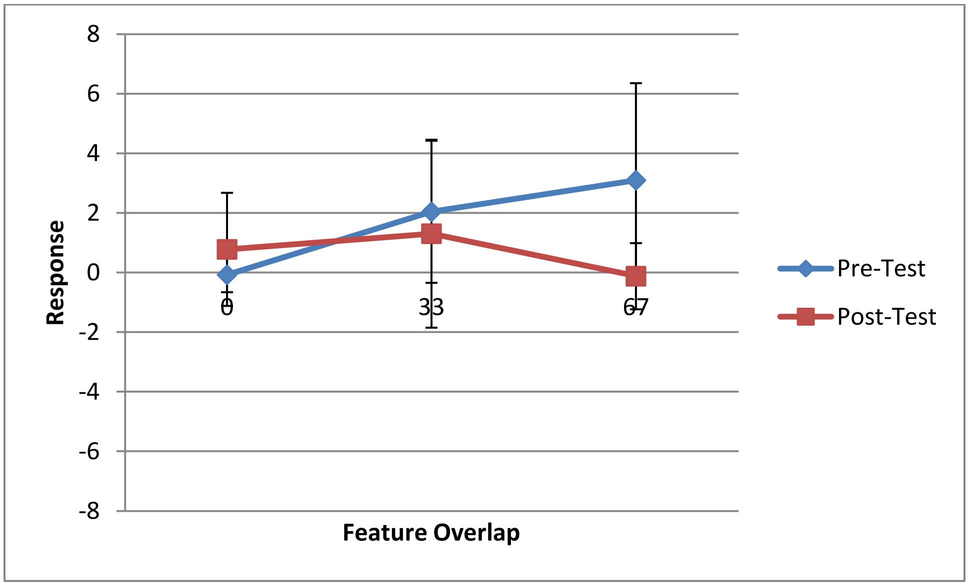

Table 6.1.1: Significant results for questions of Difference

	Within-Category Comparisons	**Between-Categories Comparisons**
Near Category Boundaries	Main effect of learning, $p < .05$ Interaction of learning and feature overlap, $p < .01$	Approaching main effect of learning, $p = .071$ Interaction of learning and feature overlap, $p < .01$
Far Category Boundaries	Approaching main effect of learning, $p = .057$ No interaction	No effect of learning No interaction