Neurobehavioral Mechanisms Supporting the Generalization of Learned Fear in Humans

by

Joseph E. Dunsmoor

Department of Psychology & Neuroscience
Duke University

Date: ______________________

Approved:

___________________________

Kevin S. LaBar, Supervisor

___________________________

Ahmad R. Hariri

___________________________

Stephen R. Mitroff

___________________________

Timothy J. Strauman

Dissertation submitted in partial fulfillment of
the requirements for the degree of Doctor of Philosophy in the Department of
Psychology & Neuroscience in the Graduate School
of Duke University

2012
ABSTRACT

Neurobehavioral Mechanisms Supporting the Generalization of Learned Fear in Humans

by

Joseph E. Dunsmoor

Department of Psychology & Neuroscience
Duke University

Date:_______________________
Approved:

___________________________
Kevin S. LaBar, Supervisor

___________________________
Ahmad R. Hariri

___________________________
Stephen R. Mitroff

___________________________
Timothy J. Strauman

An abstract of a dissertation submitted in partial fulfillment of the requirements for the degree of Doctor of Philosophy in the Department of Psychology & Neuroscience in the Graduate School of Duke University

2012
Abstract

An inescapable component to survival in a dynamic environment is detecting and reacting to signals of danger. One of the most elegant processes animals possess to handle this complex task is classical conditioning, wherein stimuli associated with an aversive event acquire the capacity to elicit defensive behaviors. This process helps ensure quick reactions prior to the occurrence of an imminent threat. A problem of living in a dynamic environment, however, is that reliable signals of danger are rarely re-encountered in the exact same form from one situation to the next. Thus, to be truly adaptive it is imperative for defensive responses to extend beyond a specific instance towards other exemplars that might portend the same negative outcome. While the phenomenon of stimulus generalization was recognized in the earliest studies of conditioning from Pavlov’s laboratory, a century of conditioning research has not resolved how humans and other animals actually meet this challenge. The research presented herein employs a combination of psychophysiological and functional imaging methods to examine how humans recruit neurocognitive systems to determine what stimuli do (and do not) pose a threat. Results show that human fear generalization is a complex phenomenon affected by the perceptual and conceptual nature of the stimulus. Brain regions and functional networks involved in fear generalization comprise cortical areas involved in coding the representation of conditioned stimuli and subcortical regions involved conditioned learning and the production of behavioral responses, most notably the amygdala. These results reveal the importance of stimulus-specific factors in fear learning and generalization, provide support for anatomically constrained models of fear generalization, and contribute to the development of model systems of fear generalization processes in human anxiety disorders.
Dedication

para mi madre
Contents

Abstract................................................................................................................................................iv

List of Tables ..........................................................................................................................................xii

List of Figures ........................................................................................................................................xiii

1. Introduction........................................................................................................................................1

  1.1 Aims ..................................................................................................................................................2

  1.2 Organization of thesis .....................................................................................................................4

  1.3 Background on Pavlovian fear conditioning .................................................................................5

      1.3.1 The fear conditioning preparation ..........................................................................................5

      1.3.2 Fear pathways in the brain ......................................................................................................8

          1.3.2.1 The amygdala ..................................................................................................................8

          1.3.2.2 Other brain structures involved in fear conditioning .........................................................10

      1.3.3 Stimulus generalization and Pavlovian fear conditioning ......................................................11

          1.3.3.1 Empirical studies in humans .............................................................................................11

          1.3.3.2 Neurobiological basis of fear generalization ..................................................................13

      1.3.4 Extending animal models to humans: fMRI of human fear conditioning ............................17

      1.3.5 Fear generalization in clinical anxiety disorders ..................................................................18

  1.4 Summary .......................................................................................................................................18

2. The concept of generalization in the conditioning literature ............................................................20

  2.1 Foundational research and early theory .......................................................................................20

      2.1.1 Ivan Pavlov and Clark Hull ....................................................................................................20

      2.1.2 A critical view of stimulus generalization ..............................................................................21

      2.1.3 The Guttman and Kalish experiment .....................................................................................23
2.2 The elements of stimulus generalization .......................................................... 24

2.2.1 The generalization gradient ........................................................................... 24

2.2.2 Peak shift effects and transposition ............................................................... 26

2.2.3 Intensity generalization .................................................................................. 29

2.3 Higher-order forms of conditioned learning and generalization ..................... 30

2.3.1 Mediated generalization .............................................................................. 30

2.3.2 Sensory preconditioning .............................................................................. 31

2.4 Associative learning models of stimulus generalization .................................... 31

2.4.1 Bush and Mosteller ....................................................................................... 31

2.4.2 Rescorla and Wagner (1972) ....................................................................... 32

2.4.3 Pearce (1987, 1994) ..................................................................................... 34

2.4.4 McLaren and Mackintosh (2002) ................................................................. 36

2.5 Is generalization a result of forgetting? ............................................................ 38

2.6 Summary ............................................................................................................ 39

3. Perceptual fear generalization along a dimension of emotional intensity .......... 41

3.1 Introduction ........................................................................................................ 42

3.2 Methods ............................................................................................................. 47

3.2.1 Participants .................................................................................................... 47

3.2.2 Stimulus set ................................................................................................... 47

3.2.3 Task design and procedures ....................................................................... 48

3.2.3.1 Preconditioning ..................................................................................... 49

3.2.3.2 Fear conditioning ................................................................................... 49

3.2.3.3 Generalization testing .......................................................................... 50

3.2.4 Psychophysiological methods and analysis .................................................. 50
3.3 Results .......................................................................................................................... 51

3.3.1 Experiment 1: Preconditioning ............................................................................. 51

3.3.2 Experiment 1: Fear conditioning ............................................................................. 54

3.3.3 Experiment 1: Generalization test ......................................................................... 54

3.3.3.1 SCRs ..................................................................................................................... 54

3.3.3.2 Pre-to-post SCR analysis .................................................................................. 55

3.3.3.3 RT ......................................................................................................................... 55

3.3.3.4 Subjective face ratings ....................................................................................... 55

3.3.3.5 Retrospective CS+ identification ....................................................................... 56

3.3.4 Experiment 2 .......................................................................................................... 56

3.3.5 Experiment 2: Preconditioning ............................................................................. 57

3.3.6 Experiment 2: Fear conditioning ............................................................................. 59

3.3.7 Experiment 2: Generalization test ......................................................................... 59

3.3.7.1 SCR analysis ....................................................................................................... 59

3.3.7.2 Pre-to-post SCR analysis .................................................................................. 59

3.3.7.3 RT ......................................................................................................................... 60

3.3.7.4 Subjective face ratings ....................................................................................... 60

3.3.7.5 Retrospective CS+ identification ....................................................................... 60

3.4 Discussion ................................................................................................................... 61

3.4.1 Relationship to sensory forms of intradimensional intensity generalization ....... 61

3.4.2 Contribution of discriminative fear learning on stimulus generalization .......... 62

3.4.3 Clinical implications ............................................................................................... 64

3.4.4 Conclusion ................................................................................................................ 65

4.1 Introduction ................................................................................................................. 67
4.2 Methods .................................................................................................................... 70
  4.2.1 Participants ........................................................................................................... 70
  4.2.2 Stimulus set .......................................................................................................... 70
  4.2.3 Task design and procedures ................................................................................. 70
  4.2.4 Psychophysiological methods and analysis ......................................................... 71
  4.2.5 Functional image acquisition, preprocessing, and GLM analysis .................... 72
  4.2.6 fMRI Regression analysis .................................................................................. 74
  4.2.7 Amygdala connectivity analysis ......................................................................... 75
4.3 Results ..................................................................................................................... 76
  4.3.1 Behavioral results ............................................................................................... 76
  4.3.2 Brain imaging results ......................................................................................... 78
  4.3.3 Brain-behavior interactions .............................................................................. 81
  4.3.4 Amygdala connectivity results ......................................................................... 83
4.4 Discussion ................................................................................................................. 87

5. Conceptual similarity promotes generalization of higher order fear learning .......... 96
  5.1 Introduction .............................................................................................................. 97
  5.2 Methods .................................................................................................................. 100
    5.2.1 Participants ....................................................................................................... 100
    5.2.2 Stimulus set ...................................................................................................... 101
    5.2.3 Task design and procedures .......................................................................... 101
    5.2.4 Psychophysiological analysis ........................................................................ 103
  5.3 Results .................................................................................................................... 104

ix
5.4 Discussion ........................................................................................................................................ 108

6. Aversive learning modulates cortical representations of object concepts ........................................ 112

   6.1 Introduction .................................................................................................................................... 112

   6.2 Methods ........................................................................................................................................... 118

      6.2.1 Participants .................................................................................................................................. 118

      6.2.2 Stimulus set .................................................................................................................................. 118

      6.2.3 Task design and procedures ........................................................................................................ 119

      6.2.4 Psychophysiological methods and analysis ................................................................................... 121

      6.2.5 Functional image acquisition, preprocessing, and GLM analysis ............................................. 122

         6.2.5.1 Independent animal-tool localizer ......................................................................................... 123

         6.2.5.2 Aversive learning .................................................................................................................... 123

         6.2.5.3 Representational similarity analysis ....................................................................................... 124

         6.2.5.4 Psychophysiological interaction (PPI) analysis ....................................................................... 125

         6.2.5.5 Resting state connectivity ......................................................................................................... 126

         6.2.5.6 Parametric modulation analysis of object typicality ................................................................ 126

   6.3 Results ............................................................................................................................................. 127

      6.3.1 Behavioral results ....................................................................................................................... 127

      6.3.2 Brain imaging results .................................................................................................................. 128

         6.3.2.1 Animal-tool functional localizer ............................................................................................... 128

         6.3.2.2 Aversive learning modulates activity in posterior category-selective cortical regions ............ 130

         6.3.2.3 Effects of aversive learning on representational similarity in object-selective cortex .............. 133

         6.3.2.4 Role of the medial temporal lobe and other brain regions ....................................................... 136
6.3.2.5 Testing the animacy hypothesis: Learning to fear animals is characterized by enhanced functional connectivity between the amygdala and lateral fusiform gyrus..... 141

6.3.2.6 Hippocampal signaling of object typicality contributes to category-based fear learning early in training through interactions with the amygdala .......................... 143

6.4 Discussion ....................................................................................................................................................... 145

7. Synthesis and conclusions..................................................................................................................................... 150

7.1 Fear generalization in human anxiety disorders............................................................................................. 150

7.1.1 Short history of conditioning-based models of anxiety................................................................................. 151

7.1.2 Overgeneralization of fear in Specific Phobia and PTSD .............................................................................. 152

7.1.2.1 Specific Phobia ........................................................................................................................................... 152

7.1.2.2 PTSD....................................................................................................................................................... 153

7.2 What factors promote fear generalization?........................................................................................................ 154

7.2.1 Physical similarity and fear generalization................................................................................................. 155

7.2.2 Intensity dimensions and fear generalization............................................................................................. 156

7.2.3 Higher-order forms of fear conditioning and fear generalization.............................................................. 157

7.2.4 Category-based fear generalization............................................................................................................. 158

7.2.5 Future directions............................................................................................................................................ 159

References.............................................................................................................................................................. 162

Biography............................................................................................................................................................ 183
List of Tables

Table 1: Fear learning-related activity ................................................................. 80

Table 2: Animal-tool localizer ........................................................................... 129

Table 3: Statistical analysis of aversive learning-related activity extracted from category-selective ROIs ................................................................. 131

Table 4: Activitation during concept-based fear learning .................................. 138
List of Figures

Figure 1: Standard Fear Conditioning Preparation ................................................................. 6
Figure 2: Simplified fear conditioning circuit centered on the amygdala ................................. 9
Figure 3: Generalization gradients ......................................................................................... 25
Figure 4: Perceptual generalization experimental design ......................................................... 45
Figure 5: Experiment 1 Results for preconditioning and the generalization test ...................... 53
Figure 6: Experiment 2 Results for preconditioning and the generalization test ...................... 58
Figure 7: Behavioral results ..................................................................................................... 77
Figure 8: Brain regions involved in differential fear learning show generalized patterns of activity. ............................................................................................................................... 81
Figure 9: Brain–behavior interactions in fear generalization ..................................................... 83
Figure 10: Learning-related and generalization-related increases in amygdala–visual cortex activity and its relation to trait anxiety ......................................................................................... 86
Figure 11: Schematic of experimental design ............................................................................ 100
Figure 12: Behavioral results ................................................................................................... 106
Figure 13: Correlations between fear generalization and trait anxiety scores for subjects in the conceptually similar condition ..................................................................................... 107
Figure 14: Fear conditioning procedure and behavioral results .................................................. 117
Figure 15: Activity in localized category-selective cortex during aversive learning .................. 132
Figure 16: Representational similarity analysis (RSA) ............................................................... 134
Figure 17: Representational similarity analysis for the amygdala ............................................. 136
Figure 18: Amygdala and whole-brain activity during aversive learning ................................. 140
Figure 19: Amygdala-lateral FFG functional connectivity at learning and at rest ...................... 142
Figure 20: Activity in the hippocampus is modulated by object typicality during early aversive learning. ......................................................................................................................... 144
Acknowledgements

I entered graduate school with an interest in human fear conditioning, and never once considered another area of research. It is probably quite normal for a graduate student to at some point (or at many points) become discouraged and consider another research track or total career overhaul. That this never happened for me is due in large measure to my PhD advisor Dr. Kevin LaBar, who I would like to express special thanks and gratitude. Dr. LaBar fosters a supportive environment that I have enjoyed as formative, productive, and never unduly stressful.

Between college and graduate school, I spent two years at the National Institute of Mental Health, where I gained my first exposure to functional imaging and fear conditioning research. This experience was foundational, and I would like to acknowledge my mentors at NIMH, Dr. Peter Bandettini and Dr. David Knight. When I contacted Dr. Bandettini about my nascent interest in neuroimaging, he referred me his postdoc, David Knight, who was using fMRI to explore human fear conditioning. This was a fortuitous connection, as I was able to learn about fMRI in a premier neuroimaging lab while simultaneously learning about a topic (fear conditioning) that would engage me for the next seven years and counting. I learned a great deal from Dr. Bandettini and Dr. Knight, who are exceptional neuroscientists and who helped significantly to get me into graduate school.

This work also benefited from my committee members, Drs. Stephen Mitroff, Ahmad Hariri, and Timothy Strauman. I thank Dr. Alex Martin from NIMH who contributed to the research presented herein on conceptual generalization. And my cohort - the class of 2012 - Deepu Murty and David Paulsen, deserve a shout-out as well. I would like to acknowledge support provided by a National Research Service Award (NRSA) from the National Institutes of Health. Finally, I am happy to acknowledge my wife, the newly minted Doctor Beth Widen.
1. Introduction

Detecting and reacting to environmental threats clearly serves a critical evolutionary function. From the conditioned withdrawal reflex in simple organisms like *Aplysia* to defensive conditioned responses in mammalian species, animals are capable of modifying behavior on the basis of learning in order to avert threatening stimuli and situations. Yet, the ubiquitous nature of this process belies its complexity. For instance, while fearful reactions to imminent threats are more or less hardwired reactions, organisms must learn to adapt these behaviors to predict and avoid a diversity of potential threats. One of the most pervasive questions in the domain of fear learning is how animals determine what stimuli do (and do not) pose a threat in a dynamic environment. Adding to this challenge of detecting threat, stimuli that are well-known to signal danger are often encountered under a variety of environmental conditions or in unique forms. Thus, just because we have learned the predictive value of certain signals for danger does not automatically ensure that we respond appropriately to these signals in the future.

To meet this challenge, fear learning must be flexible so that we can later generalize information learned in the past to make predictions in the future. For example, it is advantageous for prey surviving a close encounter with a predator to generalize fear responses to similar dangerous animals in the future, since stimuli that are alike probably herald a similar outcome. Of course, while the ability to generalize information across different experiences is paramount to adaptive behavior, this ability can prove *maladaptive* if acquired knowledge is applied too broadly. The maladaptive nature of overgeneralization of defensive behaviors is perhaps best exemplified by clinical anxiety disorders like post-traumatic stress disorder (PTSD), specific phobias, and panic disorder.

Generalization of learning and behavior has a surprisingly rich and interesting history in classical conditioning research, dating back to the earliest experiments in Pavlov’s dogs. Yet, the
lack of research on generalization in the fear learning literature (particularly in humans) is also surprising. Recently, however, there has been renewed interest in the overlap between these fields. This uptick in empirical research on fear generalization in humans has been spurred in no small part by continuing interest in animal models of fear conditioning, as well as increased use of conditioning-based models to explain the etiology of pathological fear and anxiety.

The research presented in this thesis utilizes both behavioral laboratory-based approaches and functional neuroimaging methods to detail the mechanisms of fear generalization in healthy adults. Much of this research is built on animal models of conditioning, which have provided key insights into the cellular neuroscience and key brain regions involved in the acquisition and expression of learned fear. In the past decade, these models have been extended to humans using noninvasive neuroimaging techniques like functional magnetic resonance imaging (fMRI). However, significant challenges exist in drawing parallels between conditioned learning in humans and nonhuman animals. In particular, humans are adept at combining conditioning principles with other cognitive operations to determine how we learn and generalize information about potentially threatening events. Moreover, while other species are capable of generalizing beyond perceptual regularities extracted from a learning experience, humans frequently utilize conceptual knowledge and inferential reasoning to guide learning and generalization. How perceptual and conceptual systems interact with evolutionarily conserved systems involved in conditioned fear learning has remained unclear, and is a focus of this research.

1.1 Aims

This thesis aims to address how it is that humans recruit perceptual and conceptual systems to meet the challenge of fear learning across stimuli that vary in form from a known threat. Fear learning is traditionally studied using Pavlovian conditioning procedures. These investigations reveal that an innocuous stimulus (e.g. a tone) that predicts an emotionally aversive
stimulus (e.g. electrical shock) will itself begin to elicit an emotional response (e.g. a change in heart rate). Fear generalization is demonstrated when a cue that has not directly entered into association with an aversive stimulus nonetheless elicits a learned fear response based on similarity or a previous relationship to the known threat.

The research presented in this thesis is divided between studies of fear generalization that predominately rely on either perceptual or conceptual processes. This division is based on the qualitative nature of the stimulus dimension. Perceptually-based fear generalization refers, in this case, to generalization between stimuli that share a physical resemblance. In the stimulus generalization literature, perceptual stimuli have included simple unimodal stimuli that encompass points along a simple sensory dimension (e.g. wavelengths of light), or complex multidimensional stimuli composed of numerous features that can vary in a parametric way (e.g. faces of different identities or expressing different emotions). Stimuli that vary along perceptual dimensions are by far the most commonly utilized for investigations of stimulus generalization. Conceptually-based fear generalization, on the other hand, goes beyond physical similarity to encompass what Medin (1989) has described as “an idea that includes all that is characteristically associated with it” (p. 1469). Studies on conceptual generalization have remained an exceedingly rare topic in the conditioning literature. The conceptually-based fear generalization studies in this thesis are built upon Pavlovian conditioning principles as well as a rich literature on concepts and categories (e.g. Murphy, 2002).

The specific aims are as follows: 1. Examine generalization of conditioned fear across a dimension of perceptually graded stimuli using psychophysiological and functional imaging methods. 2. Investigate fear generalization across conceptually-related and conceptually-unrelated stimuli. 3. Investigate whether fear conditioning modulates the neural representation of conceptual knowledge.
1.2 Organization of thesis

The remainder of Chapter 1 provides a comprehensive background on Pavlovian fear conditioning and reviews relevant information on the neurobiological basis of fear conditioning. This chapter also discusses work on fear generalization in nonhuman animal neurobiological investigations in order to provide a framework for investigations on the neurobehavioral mechanisms governing fear generalization in humans. It is worth noting up front that neurobiological research on fear generalization is scant relative to the number of standard Pavlovian fear conditioning studies.

Chapter 2 provides a background on stimulus generalization research and theory. Both historical and contemporary views are presented. Notably, the majority of this research has traditionally employed instrumental conditioning procedures with appetitive reinforcement, thus differentiating these studies from the Pavlovian fear conditioning research discussed in Chapter 1. Chapter 2 also outlines important elements of stimulus generalization. This includes the nature of the generalization gradient, peak shift effects, and an underappreciated concept known as intensity generalization that has particular implications for understanding fear generalization, per se.

The remaining chapters present research on fear generalization in humans that have been conducted by the author towards completion of his dissertation. Chapters 3 and 4 are focused on perceptually-based forms of fear generalization. While Chapter 3 details a laboratory based psychophysiological experiment on fear generalization, Chapter 4 delves into the neurobehavioral mechanisms of perceptual fear generalization. Experiments reported in Chapter 3 and Chapter 4 utilize a perceptual dimension of graded emotional expressions and provide compelling evidence that fear generalization in humans is influenced by the amount of emotional intensity expressed in generalized stimuli. Chapter 4 complements the behavioral findings detailed in the Chapter 3 by
showing that brain regions involved in the acquisition of fear support generalization to stimuli that are similar to a learned threat, but vary in emotional intensity value.

Chapters 5 and Chapter 6 build off of the previous fear generalization experiments by examining generalization as a function of conceptual, rather than perceptual, similarity. Chapter 5 tests the hypothesis that conceptual similarity promotes the generalization of fear. This study uses a procedure known as *sensory preconditioning*, which is a higher-order form of conditioning that is predicated on an association between stimuli acquired prior to fear conditioning. In Chapter 5 we show that conceptual similarity promotes between-stimulus fear generalization. Chapter 6 builds off of research detailed in Chapters 3, 4, and 5 by examining generalization of conditioned fears across a wide range of basic level exemplars that belong to the same category domain, but vary widely in perceptual features. Chapter 6 provides novel evidence that fear learning can engage inductive reasoning processes. Brain imaging results presented in Chapter 6 also shows that concept-based fear conditioning modulates the cortical representations of object categories coded in posterior occipital-temporal cortex.

Finally, Chapter 7 concludes the dissertation by summarizing and synthesizing these findings. The concluding chapter discusses the applications to fear and anxiety research in clinical disorders, and proposes that studies of fear generalization can be utilized in areas outside the affective domain to address questions in a range of cognitive disciplines.

### 1.3 Background on Pavlovian fear conditioning

#### 1.3.1 The fear conditioning preparation

To investigate how organisms learn about and respond to threats, researchers often use classical (or Pavlovian) fear conditioning paradigms (Figure 1). In this procedure, an emotionally neutral stimulus (conditioned stimulus, *CS*), such as a tone or light, predicts a naturally aversive or threatening stimulus (unconditioned stimulus, *US*), such as an electric shock. Just as Pavlov’s
experimental dogs would salivate to a tone that had been paired with food, animals that undergo fear conditioning will begin to express a number of autonomic responses (e.g., a change in heart rate, perspiration, respiration rate) to the presentation of the previously neutral CS. These fear conditioned responses (CR) reflect the well-known fight-or-flight behaviors, and indicate that an association has been formed between the CS and US.

In a typical differential fear conditioning experiment, a stimulus (CS+) is paired with an aversive US while another stimulus (CS-) is unpaired. The CSs are separated by an intertrial interval allowing time for sympathetic arousal to return to baseline. CS = conditioned stimulus; US = unconditioned stimulus, denoted by the lightning bolt.

Studies of fear conditioning traditionally utilize one stimulus that is to be conditioned (referred to as CS+), and a control stimulus that is explicitly never paired with the US (referred to as CS-). This differential fear conditioning procedure is necessary to validate specificity of learning, and rule out non-associative effects like pseudo-conditioning or sensitization, i.e. responding to cues merely because they occur in the context of the US but not because they have been associated with the US directly.

An important development in theories of fear acquisition concerns the qualitative nature of the CS. Animal research shows that learning occurs best to particular classes of stimuli that are, in many cases, species-specific. For instance, a rat can readily learn to associate a taste CS with illness US, or a noise CS with shock US, but cannot easily form the crossed association
(taste with shock, or noise with illness) (Garcia & Koelling, 1966). Seligman (1971) proposed that humans are also predisposed to form associations between particular classes of CSs and USs. As evidence for the selectiveness of stimuli to which people readily condition, Seligman noted that phobias generally fall under a select number of “fear-relevant” objects or situations that have served as evolutionarily significant throughout mammalian development. These include certain animals (e.g. snakes and spiders), environmental phenomenon (e.g. thunder and lightning), physical locations (e.g. being in an enclosed space or high off the ground), and social interactions (e.g. emotionally charged exchanges or evaluations). Notably, the list does not include fear-relevant items of ontogenetic origin, such as manmade weapons. Fear conditioning between CSs from phylogenetic prepared classes and an aversive US may occur more readily than between stimuli that are not related through fear-relevance. In support of this theory, fear-relevant CSs oftentimes lead to superior conditioning and delay extinction learning relative to fear-irrelevant environmental images (e.g. images of flowers or mushrooms), suggesting a strong selective association between fear-relevant CSs and aversive USs (Öhman & Mineka, 2001).

As investigations of generalization have tended towards appetitive learning (see Chapter 2), the dependent measures of generalization have typically been those associated with appetitive behaviors, such as key-pecking and lever pressing to receive a reward. However, Pavlovian fear conditioning involves a different set of behaviors that can be gauged in many cases by an increase in sympathetic arousal. The studies presented in this thesis measure the conditioned skin conductance response (SCR) as a primary dependent measure of fear learning and generalization. The SCR measures the change in electrodermal conductance on the palmar surface of the hands through eccrine sweat glands, which increases due to sympathetic arousal induced by feared stimuli (Ohman, 1971).
1.3.2 Fear pathways in the brain

Much of what we know about the neurobiology of Pavlovian fear conditioning has been informed by neurophysiological investigations in nonhuman animals. Decades of animal research has delineated the neural circuitry important for processing and responding to both acquired and innate fears. While an in-depth review of the neurobiological basis of fear conditioning is beyond the scope of this thesis (for detailed reviews see Davis, 1992; Kim & Jung, 2006; LeDoux, 2000; Pape & Paré, 2010), an overview of the essential findings from this line of research is important to fully appreciate investigations into human fear learning and generalization.

1.3.2.1 The amygdala

The neural pathways involved in forming the CS-US association and producing the CR have been traced predominately using rodent models. These investigations have shown that sensory information concerning the CS and US converges in the amygdala (LeDoux, 2000; Pape & Paré, 2010) (Figure 2). The amygdala receives extensive afferent projections from a number of brain systems, including all sensory systems and higher-order association cortex. A widespread distribution of efferent projections from the amygdala is important for modulating information processing broadly across the brain. For instance, projections from the amygdala to the ventral visual stream may be important for modulating the representation of sensory stimuli following emotional experiences (Vuilleumier, 2005).
Figure 2: Simplified fear conditioning circuit centered on the amygdala

This diagram depicts inputs and outputs to the amygdala, and the role of the output regions in fear conditioning. The diagram simplifies the layout of the amygdala subnuclei, combining several distinct nuclei that constitute the basolateral complex. The intrinsic connections between the amygdala subnuclei, and the ITC, are not detailed here. BLA = basolateral complex; CE = central nucleus; ITC = intercalated cells.

Importantly, the amygdala is not a homogenous structure, but is instead a collection of interconnected subnuclei. The nuclei most often implicated in fear conditioning include the basal (B), accessory basal (AB), lateral (LA), and central (CE) nuclei. The L, AB, and B are often referred to collectively as the basolateral complex (BLA), and the subnuclei within this complex can be further reduced into anatomically distinct regions. Interposed between the BLA and CE are clusters of intercalated cell masses that may serve a role in gating activity between these subnuclei of the amygdala (Ehrlich et al., 2009). In the classic anatomical model of fear

<table>
<thead>
<tr>
<th>Output connections</th>
</tr>
</thead>
<tbody>
<tr>
<td>Hypothalamus: sympathetic nervous activation</td>
</tr>
<tr>
<td>Bed nucleus of stria terminalis: stress hormone release, anxiety behaviors</td>
</tr>
<tr>
<td>Periaque ductal gray: freezing</td>
</tr>
<tr>
<td>Locus coeruleus: mediates arousal through release of norepinephrine</td>
</tr>
<tr>
<td>Ventral tegmental area: dopamine release important for associative fear learning</td>
</tr>
</tbody>
</table>

<table>
<thead>
<tr>
<th>Output (and reciprocal) connections</th>
</tr>
</thead>
<tbody>
<tr>
<td>Striatum: instrumental behaviors, e.g. conditioned avoidance</td>
</tr>
<tr>
<td>Medial PFC: extinction, recall of extinction</td>
</tr>
<tr>
<td>Hippocampus: context conditioning, return of fear, memory consolidation</td>
</tr>
<tr>
<td>Insula: interoception associated with psychophysiological arousal</td>
</tr>
<tr>
<td>Sensory cortex: modulate sensory processing of conditioned stimuli</td>
</tr>
</tbody>
</table>

<table>
<thead>
<tr>
<th>Input</th>
</tr>
</thead>
<tbody>
<tr>
<td>e.g. thalamus, sensory and association cortex, hippocampus, PFC</td>
</tr>
</tbody>
</table>
conditioning, the BLA is the principal site for receiving sensory information, and is important for forming the association between the CS and US. This region may also be involved in the consolidation and storage of conditioned fear memories (Schafe, Nader, Blair, & LeDoux, 2001). The CE receives projections from the BLA and is principally involved in initiating the conditioned fear response through output projections to the hypothalamus and brainstem structures.

A key feature of amygdala neurocircuitry is that sensory information reaches the amygdala from two partially separated neural pathways: a relatively slow cortical pathway that flows from the thalamus to primary sensory cortex and then to higher-level association cortex before reaching the amygdala, and a more rapid subcortical pathway that sends projections directly from the thalamus. While the cortical route supplies detailed information concerning a sensory stimulus, the subcortical route quickly detects potentially threatening objects and generates immediate fear responses (but see Pessoa & Adolphs, 2010 for an alternative view). These dual processing routes to the amygdala may have implications for understanding complex fear behaviors (LeDoux, 1996). For instance, due to the fact that the thalamo-amygdala pathway contains fewer synapses, information about potential threats is sped to the amygdala before the cortex has had time to process the stimulus more thoroughly. Therefore, an organism might react rapidly to a potential threat (e.g., initiating a freezing response), only to realize a split second later that the threat has passed or was nonexistent. Once a CR is initiated, however, it is difficult to quickly shut it off since the physiological response systems engaged have a relatively slow time course compared to the neural response.

1.3.2.2 Other brain structures involved in fear conditioning

A number of other brain regions have been consistently implicated in fear conditioning (Maren, 2001). For instance, animal models of context conditioning have focused on the role of
the hippocampus in encoding and representing the feared environment. The hippocampus is thought to provide a substrate for binding the numerous features of the context into a unitary representation, and may be specialized for calling to mind a full representation of a feared context from partial information through the process of pattern completion (O'Reilly & Rudy, 2001). In addition, the medial prefrontal cortex (mPFC) is critically involved in extinction learning. Extinction involves the reduction in the CR once the US is removed. The ventromedial PFC (vmPFC) in particular appears to be important for exerting inhibitory control over the amygdala in order to reduce fear expression (Milad & Quirk, 2002). The lateral PFC has been implicated in the voluntary regulation of emotion during fear conditioning through reappraisal strategies (Delgado, Nearing, LeDoux, & Phelps, 2008). In addition, studies by Norman Weinberger and colleagues (Weinberger, 2004, 2007) have demonstrated that fear conditioning alters the receptive field of neurons in primary sensory cortex, suggesting that conditioning modulates plasticity in regions that receive and project basic level sensory information.

1.3.3 Stimulus generalization and Pavlovian fear conditioning

1.3.3.1 Empirical studies in humans

Fear generalization was demonstrated in one of the earliest studies of human fear conditioning. In the infamous “Little Albert” experiment (Watson & Rayner, 1920), the subject – 9 month old Albert – was presented with several small objects and animals in the absence of a fear inducing stimulus. Two months later, the experimenters struck a steel bar with a hammer creating a loud aversive noise whenever Albert would reach out to touch a rat. After a few rat (CS) and noise (US) pairings, the 11-month-old would cry and withdraw from the rat (CR) in the absence of the noise. Moreover, Albert reportedly generalized this fear reaction to several other objects, including a dog, a rabbit, and white cotton. Of course, it should be mentioned that the Little Albert experiment is roundly criticized today both for the dubious ethical practice of fear
conditioning an infant and for an overall lack of experimental rigor (Field & Nightingale, 2009). Furthermore, this experiment had components of instrumental conditioning, as delivery of the US was contingent on Albert interacting with the rat.

Early studies of spatial generalization in humans (Bass & Hull, 1934) used an electric shock US and measured electrodermal activity, and thus may be considered an early example of fear generalization in classical conditioning. However, it should be noted that this study used procedures that are today considered antiquated; for instance, the experiment involved volunteers lying nude on an army cot while vibrating tactual stimulation was applied to different points along the back of the body. Auditory generalization studies by Hovland (1937) and Humphreys (1939) used procedures that are more in keeping with today’s fear conditioning methodologies; although attempts to replicate the gradients obtained by Hovland using the same procedures were unsuccessful (Epstein & Burstein, 1966). Notably, early studies were not expressly concerned with the emotion of fear, as such. In fact, as might be expected given the behaviorist zeitgeist, there was nary a mention of the word “fear” or “emotion” in early studies of stimulus generalization experiments in humans. It has not been until recently that well controlled fear generalization experiments have regained interest as providing potentially relevant insight into fear learning and memory processes in humans.

A much more recent study by Lissek et al. (2008) was one of the first to directly and systematically examine gradients of conditioned fear responses in humans. This study measured gradients of the fear-potentiated startle reflex, which is a rapid muscle contraction frequently indexed from facial electromyography (EMG) in response to a brief noxious startle-probe (often a 50 millisecond white noise of 100 dB or louder) that is more pronounced in the presence of the CS+ than the CS- or the probe-alone (Grillon & Baas, 2003). In this fear generalization study, subjects were conditioned to the image of a ring at one end of a continuum (CS+) through
repeated pairings with an aversive shock US, while the ring at the other end of the continuum served as the CS-. Following acquisition, subjects were presented with the CS+, CS-, and 8 novel rings that varied parametrically in size between the CS+ and CS-. Startle responses during the generalization test showed a classic generalization decrement that decreased as a function of similarity to the CS+. Fear generalization has also been reported along other simple visual dimensions using fear-potentiated startle (Hajcak et al., 2009), SCRs (Vervliet, Kindt, Vansteenswegen, & Hermans, 2010), and subjective reports of US expectancy (Vervliet, Vansteenswegen, & Eelen, 2006) as the primary measures of generalization.

1.3.3.2 Neurobiological basis of fear generalization

Given the behaviorist school of thought that permeated early-to-mid twentieth century psychological research, it is perhaps unsurprising that few neurocognitive studies of fear generalization were conducted in the first half of the twentieth century. Simple neural-network models in the mid-twentieth century conceptualized stimulus generalization in terms of connections between nodes (Bush & Mosteller, 1951; Wolpe, 1952). In the 1960s, neurophysiological studies began to examine the extent to which cortical ablations impaired discrimination abilities in monkeys (Thompson, 1965). However, few neurophysiological investigations explicitly focused on stimulus generalization, perhaps indicating that the phenomenon did not represent an area of broad neuroscientific interest.

One obvious reason for the lack of knowledge on stimulus generalization is that most laboratory studies seek to avoid generalization altogether, as it complicates interpretations of research findings. That is, if an animal responds with near equal levels to the experimental (i.e., CS+) and control conditions (i.e., CS-), then associative learning cannot be properly evaluated. Consequently, lengths are taken to select against generalization in typical fear conditioning experiments. Another important point is that animal models of fear generalization often examine
generalization across contexts (e.g. Biedenkapp & Rudy, 2007) or to a restricted set of test stimuli that might only include the CS- (e.g. Shaban et al., 2006), rather than employing a form test of generalization using several graded stimuli. Therefore, it is not always clear which features drive generalized patterns of responses.

As the amygdala is consistently implicated in fear conditioning processes, it stands to reason that this brain area serves a role in fear generalization as well. Baldi et al. (2008) showed that reversible inactivation of the BLA following re-exposure to a CS reduces fear generalization to a novel context. Conditioning was conducted using procedures that were known to induce strong generalization (tone-shock pair using a high intensity shock of 1.2 mA) (Baldi, Lorenzini, & Bucherelli, 2004). Reminder tones were then given to the rat in the conditioning context 96 hours after conditioning, immediately followed by reversible inactivation of the BLA or nucleus basalis. Generalization was tested either 72 or 96 hours post-reactivation by measuring freezing in a novel context in the presence or absence of the CS. In the control group- as well as a group of rats receiving post-reactivation inactivation of the nucleus basalis- robust generalization was observed both to the context (before presentation of the tone) and to the tone at both retention test intervals. In contrast, rats receiving bilateral inactivation of the BLA immediately after re-exposure showed reduced fear generalization in the novel context but showed similarly high levels of responses to the CS in the novel context during the retention test. This finding suggests that amygdala involvement during memory reconsolidation processes may influence fear generalization.

In a combined optogenetic, pharmacological, and electrophysiological fear conditioning study in mice, Ciocchi et al. (2010) revealed a role for subregions of the CE in regulating fear generalization. Neuronal activations in the medial subdivision of the CE (the primary output pathway for initiating fear behaviors) are tonically inhibited by the lateral subdivision the CE.
Ciocchi et al. (2010) found that changes in tonic activity between the CE microcircuit resulting from fear conditioning were related not only to the expression of fear to the CS+, but also to generalization to the CS-. This study builds off previous findings showing a role for other subnuclei of the amygdala (Shaban, et al., 2006), as well as the closely associated bed nucleus of the stria terminalis (Duvarci, Bauer, & Pare, 2009), in fear generalization.

As most fear conditioning studies with rodents utilize acoustic CSs, the contribution of auditory cortex to fear conditioning processes has been extensively studied. Although auditory cortex lesions do not abolish fear conditioning to acoustic stimuli (Romanski & LeDoux, 1992), it has been argued that the cortical pathway to the amygdala is important when fear conditioning involves complex stimulus discriminations (Weinberger, 2007). Receptive fields of neurons in auditory cortex tend to show sharp tuning curves, whereas neurons in the auditory thalamus are broadly tuned to respond to a wide range of frequencies (Bordi & LeDoux, 1994).

Early neurophysiological studies of stimulus generalization in cats showed that ablation of the auditory cortex prior to conditioning lead to broad stimulus generalization to acoustic stimuli, but leaves frequency discrimination intact (Thompson, 1965). In a similar way, lesions to the occipital and inferotemporal cortex in monkeys also lead to broad stimulus generalization across dimensions of visual stimuli, while leaving most perceptual capacities intact (Butter, Mishkin, & Rosvold, 1965). These early findings were important for distinguishing between generalization and discrimination processes, as it revealed broad behavioral generalization resulting from cortical ablation but did not interfere with the ability to perceptually distinguish between the stimuli.

However, the notion that the auditory cortex is necessary to reduce overgeneralization was challenged by Armony et al. (1997). In this study, auditory cortex lesions did not lead to overgeneralization between a CS tone and unreinforced tones that differed by an octave or more.
The authors suggested that the subcortical thalamic pathway to the amygdala is therefore sufficient to allow effective fear acquisition and expression without significant overgeneralization - at least between simple sounds. Yet, in another study increased levels of activity in the auditory thalamus – induced through infusions of an exogenous transcription factor – did lead to overgeneralization of freezing to unreinforced stimuli, suggesting that the thalamus may be involved in broad generalization (Han et al., 2008). The precise contribution for cortical and subcortical pathways to the amygdala in fear generalization is still not clear.

The role of the auditory cortex in fear conditioning is still generating interest. Studies by Norman Weinberger and colleagues frequently employ classic generalization tests to assess representational plasticity within primary auditory cortex following fear acquisition to an acoustic CS. In a typical experiment, electrophysiological recordings from primary auditory cortex (A1) are taken for a range of frequencies to map the receptive field. The stimulus that does not have the best frequency is used as the CS in order to test whether the receptive field shifts towards the CS after tone-shock pairing. While these experiments are not expressly interested in testing stimulus generalization for its own sake, these CR gradients often track the shift in auditory tuning along the auditory dimension (Weinberger, 2004, 2007).

Studies of context generalization have focused on the role of the hippocampus. In the conditioning literature, “context” refers to relatively static features of the environment, including temporal and sensory aspects (e.g. time of day, sights and sounds), as well as internal states, (e.g. moods and emotions) (Bouton, Westbrook, Corcoran, & Maren, 2006). The hippocampus is implicated in forming contextual representation during fear conditioning (Fanselow, 2010; O'Reilly & Rudy, 2001). Connections between the hippocampus and BLA are important for associating a CS with a particular context and for learning to fear the context itself. While lesions to the amygdala impair or abolish fear conditioning altogether, lesions to the hippocampus impair
context conditioning (Phillips & LeDoux, 1992). Hippocampal lesions also affect the ability for the CS to elicit a fear response after extinction when it is encountered in a new context, known as fear renewal (Bouton, et al., 2006). Extinction refers to the reduction in fear to a CS when it is presented in the absence of the US (Pavlov, 1927). In a typical fear renewal paradigm, fear to a cue (e.g., tone paired with a foot-shock) is acquired in context A, and then the tone is extinguished in context B. In most cases, an extinguished CR returns following extinction learning if the CS is re-presented in the acquisition context (ABA design) or in a novel context (ABC design). However, hippocampal inactivation following extinction impairs fear renewal (Ji & Maren, 2005). The fact that CSs can elicit fear under multiple contexts while extinction is context specific reveals how extinction is far less generalizable than acquisition.

1.3.4 Extending animal models to humans: fMRI of human fear conditioning

While discoveries of the neurophysiological substrates of fear conditioning have been made predominately in rodents using invasive procedures, the role of particular brain regions in fear conditioning are now being explored in humans using noninvasive functional brain imaging techniques. Findings from human patients with brain damage and fMRI studies have revealed a crucial role for the amygdala and other regions in the acquisition, expression, and control of fear (reviewed in LaBar & Cabeza, 2006; Phelps & LeDoux, 2005; Sehlmeyer et al., 2009). Although positron emission tomography (PET) studies on fear processes have also been conducted, the blocked designs inherent to the PET technique constrain interpretations. Specifically, because PET signals are integrated over many seconds, brain activity uniquely associated with CS and US presentations cannot be readily distinguished and different trial types cannot be intermixed in the experimental design. Pre- versus post-conditioning comparisons can be made with PET (e.g. J. S. Morris, Öhman, & Dolan, 1998), but event-related fMRI designs have proved optimal for examining the neural systems involved in fear conditioning not only because of these design
issues but also because of improved spatial resolution (e.g., for distinguishing activation in the amygdala and adjacent hippocampus).

1.3.5 Fear generalization in clinical anxiety disorders

Human anxiety disorders are characterized by an excessive, disabling, response to stress (DSM-IV Axis I Clinical Syndrome). The major categories of anxiety disorders are acute stress disorder, agoraphobia, generalized anxiety disorder (GAD), obsessive-compulsive disorder (OCD), panic disorder, specific and social phobias, and posttraumatic stress disorder (PTSD). These disorders are all marked by excessive and often disabling fear-related behaviors (i.e., increases in sympathetic “fight-or-flight” arousal, such as changes in heart rate, respiration, blood pressure, etc.). There is empirical evidence that fear generalization may be involved in the development and maintenance of many anxiety disorders. For instance, Lissek et al. (2010) showed broader generalization gradients of startle EMG in patients with panic disorder relative to healthy controls, providing key evidence that generalization testing procedures can be used to index heightened fear responses symptomatic of certain anxiety disorders. A more thorough discussion on how fear generalization research advances conditioning-based models of anxiety is presented in Chapter 7.

1.4 Summary

An understanding of fear conditioning across species has been fostered by advances in neuroscience. The anatomical fear circuit is centered on the amygdala, a small structure in the anterior portion of the medial temporal lobe with extensive connections with numerous brain regions. Due to an overall lack of research on fear generalization over the past several decades, however, the role of the amygdala and other brain regions on fear generalization is not entirely clear. The empirical research presented in this thesis extends the role of the fear conditioning
circuit to the realm of fear generalization, wherein stimuli that bear some similarity or relationship to a known threat elicit a generalized fear response.
2. The concept of generalization in the conditioning literature

The term “generalization” has been used to describe multiple facets of learning and behavior (Mostofsky, 1965). In the conditioning literature, the term can be used as a description of empirical data (e.g. “Bidirectional generalization gradients were obtained from measures of response rates” (Guttman & Kalish, 1956, Abstract), as well as to denote a process on the part of the animal, e.g. “Perhaps the most salient aspect…is the orderliness and reproducibility of the generalization process within the individual” (Guttman & Kalish, 1956, p. 84). Contemporary usage of the term connotes both an active process and a description of a behavioral finding (Shepard, 1987).

2.1 Foundational research and early theory

2.1.1 Ivan Pavlov and Clark Hull

The earliest studies to demonstrate generalization of learned behavioral responses utilized classical conditioning methods with an appetitive US (i.e., meat powder to elicit a salivary CR). Early studies from Pavlov’s laboratory revealed that once an animal has been conditioned to the CS, the learned response can be evoked by a range of other stimuli that have never been paired with the US (Pavlov, 1927). Remarkably, the magnitude of the CR (the amount of salivation) was in some cases proportional to the similarity to the trained stimulus: “[I]f a tone of 1000 d.v. is established as a conditioned stimulus, many other tones spontaneously acquire similar properties diminishing proportionally to the intervals of these tones from the one of 1000 d.v.” (Pavlov, 1927, p. 113). As these other stimuli have never directly predicted the US, Pavlov interpreted these responses as arising from a transfer of conditioned learning from the CS. He described the CR decrement as resulting from a spread of cortical excitation from the point activated by the CS to adjacent points along the cortex. This concept of neural irradiation did not catch on,
particularly in light of emerging evidence on how the nervous system operates (e.g. the Neuron Doctrine).

Although dismissive of Pavlov’s neurophysiological view of generalization, Clark Hull (1943) borrowed from Pavlov’s notion of a spread of effect to describe stimulus generalization as arising from a spread of habit strength along a physical dimension. Hull emphasized the importance of the just-noticeable-difference, postulating that generalization decreases as a function of degrees of similarity defined as the discriminable distance between points along a physical continuum.

In Hull’s view, generalization was an orderly, quantitative, phenomenon: “[T]he reaction involved in the original conditioning becomes connected with a considerable zone of stimuli other than, but adjacent to, the stimulus conventionally involved in the original conditioning” (Hull, 1943 p. 183). The idea that learning spreads in an orderly way to stimuli that are not actually present at the time of initial learning had far-reaching implications for emerging models of conditioning, and stirred a number of researchers into providing counter-arguments to this view.

2.1.2 A critical view of stimulus generalization

A highly critical examination of Hull’s “neo-Pavlovian” view of stimulus generalization was offered by Karl Lashley and Marjorie Wade (1946). They argued that generalization is not, in fact, a consequence of conditioning. Rather, generalization merely reflects the inability (or unwillingness) on the part of the animal to differentiate between the CS and other stimuli. In other words, generalization is not a special process, but simply a failure to discriminate. They called into question a number of findings from Pavlov’s laboratory, arguing that Pavlov had ascribed generalization to animals that simply hadn’t yet learned to differentiate between stimuli. Moreover, Lashley and Wade argued that human studies are not able to address questions of CR
generalization, as adult human subjects are already familiar with the physical stimuli prior to conditioning. Also, humans can use relational and inferential reasoning abilities in the course of generalization testing – which is not, strictly speaking, primary stimulus generalization as conceptualized by Pavlov and Hull. Others eschewed the use of the term generalization as well, arguing that the construct of generalization offered nothing beyond what was already described by processes of attention, orienting, and discrimination (Prokasy & Hall, 1963).

Gregory Razran (1949) offered another critical review of theories and studies of stimulus generalization. After reviewing a number of studies from Pavlov’s laboratory, Razran determined that, while dogs did respond to unreinforced stimuli early in training, generalization gradients showed only crude decrements. He was also exceedingly skeptical that generalization gradients could be replicated in humans. Indeed, upon closer examination of the results from human generalization experiments (Bass & Hull, 1934; Hovland, 1937), Razran remarked upon the lack of orderly and reliable generalization gradients at the single subject level. Razran (1949) concluded that stimulus generalization is not a result of the original conditioning experience, as asserted by Pavlov and Hull, but instead manifests during the subsequent generalization test as a result of similarity judgments.

It is important to emphasize here that the construct of generalization was historically being used interchangeably both in discussions on learning and behavior theory, as well as descriptions of an empirical phenomenon (Honig & Urcuioli, 1981; Mednick & Freedman, 1960; Prokasy & Hall, 1963). While both uses of the term were under fire, the methods being used in classical conditioning experiments were providing poor evidence of orderly and reproducible generalization gradients (G. Razran, 1949).
2.1.3 The Guttman and Kalish experiment

A seminal study by Guttman and Kalish (1956) helped to resolve many of the issues regarding the validity of stimulus generalization as an empirical phenomenon. This study borrowed from a design proposed by Skinner (1950): “By changing to several values of a stimulus in random orderly repeatedly during the extinction process, the gradient for stimulus generalization may be read directly in the rates of responding under each value” (p. 204). Using an instrumental conditioning paradigm, pigeons were initially trained to key peck to a light of a specific color value using an interval reinforcement schedule, which was well-known to produce high levels of responding. Generalization was measured in a subsequent extinction test in which other color values were presented above and below the wavelength of light of the training CS. Since pigeons possess excellent color vision, it was hypothesized that CR generalization would be observed to color values that were perceptually similar to the CS. The question was whether generalization would abruptly cease near the borders that divided discriminable color boundaries (e.g. between green and yellow), or extend as a function of the underlying physical dimension unimpeded by the categorical boundary. As Guttman and Kalish later wrote in a popular science magazine article, “Here discrimination might be expected to take the upper hand over generalization: the difference would seem greater than the kinship” (Guttman & Kalish, 1958). However, this is not what they observed. Instead, Guttman and Kalish (1956) plotted orderly generalization gradients that spanned across color boundaries: “like a mechanical frequency analyzer, the bird appears to recognize wavelength entirely without reference to color” (Guttman & Kalish, 1958). In effect, the experimenters demonstrated that generalization is not simply a failure to discriminate, since the pigeons were generalizing in an orderly fashion to stimuli that could be discriminated from the CS. Moreover, orderly gradients were obtained on a single
subject level, in contrast to Razran’s (1949) view that gradients could not be obtained for individual subjects.

The elegant and (perhaps more importantly) reproducible results from the Guttman and Kalish (1956) experiment ushered in decades of research on factors contributing to the generalization gradient (Ghirlanda & Enquist, 2003; Honig & Urcuioli, 1981; Mednick & Freedman, 1960). These included studies on the effect of discrimination training on subsequent generalization (Hanson, 1959; Jenkins & Harrison, 1960), generalization of inhibitory learning (Honig, Burstein, Boneau, & Pennypacker, 1963), and attentional factors (Honig & Urcuioli, 1981; Jenkins & Harrison, 1962). Contemporary models of associative learning have also built upon findings from stimulus generalization studies (McLaren & Mackintosh, 2002; Pearce, 1987), and the distinguished cognitive scientist Roger Shepard went so far as to deem generalization a “universal law” for psychological science (Shepard, 1987). In all, stimulus generalization is today regarded as an acceptable empirical phenomenon and is often used as an explanatory construct in theories of learning and behavior.

2.2 The elements of stimulus generalization

2.2.1 The generalization gradient

An orderly decremented gradient of responses tracking a well-defined sensory dimension is often considered the gold standard in empirical stimulus generalization research (see Figure 3). But it was not until the Guttman and Kalish (1956) study that examining the shape of the gradient was approached from a truly empirical standpoint. The method for analyzing these gradients was similar across different studies and usually included the presentation of several values that approximated the CS (often along the same sensory dimension) during extinction, when the CS is no longer reinforced by the US. What differed between these studies were various manipulations during or prior to generalization testing. By quantifying the effects of these numerous
manipulations on the generalization gradient, it became possible to test hypotheses on what mechanisms drive generalized responding.

Figure 3: Generalization gradients

Empirical studies of stimulus generalization have uncovered numerous types of gradients. (A) A typical excitatory gradient centered on the CS+ with an orderly response decrement. (B) An inhibitory gradient in which responses dip around the non-reinforced CS- and increase as similarity to the CS- decreases. (C) A peak shift induced by differential reinforcement during training. (D) Intensity gradient in which responses increase to values of greater intensity than the CS+. CS = conditioned stimulus; dB = decibel; red and blue circle indicates maximum and minimum response, respectively.
Honig and Urcuioli (1981) described four major properties to the stimulus generalization gradient: area, height, slope, and form. Area and height refer to the amount of responding that occurs during generalization testing to the entire range of values, with area referring to the amount of responding to the CS and all other values, and height referring to the amount of responding to the value with the greatest response magnitude. Slope is an index for the difference in responding between two values along the dimension, often measured as a rate of change between the CS and a value adjacent to the CS. Slope is often a qualitative assessment (Honig & Urcuioli, 1981), such that a “flat” gradient with no slope indicates a lack of stimulus control (broad generalization) while a “narrow” gradient with a steep slope indicates a high level of stimulus control. Form is highly subject to the procedures employed in training and test. A precise measurement for form is therefore problematic to determine across different studies that use various procedures and/or stimuli (Honig & Urcuioli, 1981). For example, if the perceptual difference between values along a dimension is variable, then the gradient will be asymmetrical, with more generalization to values that are closer in similarity to the CS. This is apparent along the color spectrum in which pigeons have more difficulty discriminating between certain color values than others, despite equal distances in terms of physical units (i.e., nanometers) (Haber & Kalish, 1963).

### 2.2.2 Peak shift effects and transposition

One of the most extensively studied elements of stimulus generalization is that of the peak shift effect. The peak shift effect can be described as follows: In a typical stimulus generalization experiment, the organism is first trained to respond to particular stimulus before it is tested on several unreinforced values that are often along the same sensory dimension as the CS. With the exception of intensity dimensions (see below), organism trained with a single CS
will evoke the greatest amount of responding to the CS during the subsequent generalization test (Guttman & Kalish, 1956). If instead the organism is trained to respond to one CS (CS+) but withhold responses to another CS (CS-), a phenomenon is observed during the subsequent generalization test such that the CS+ no longer evokes the greatest amount of responding. Instead, a novel stimulus that has never been reinforced by the US will often elicit more responding than the CS+. This value is often a point along the stimulus dimension that is relatively farther from the CS- than is the CS+. This phenomenon is known as the peak shift effect (see Figure 3c).

The first empirical demonstration of the peak shift effect was provided by Hanson (1959). Employing procedures similar to those of Guttman and Kalish (1956), several groups of pigeons were trained to respond to a 550 nm light (CS+). A control group was conditioned only to the CS+, while the experimental group was also presented with a CS- during training. During the generalization test the control group showed a standard generalization gradient that peaked at the CS+. The experimental groups, on the other hand, responded more to a novel stimulus, further from the CS- along the color spectrum. Notably, this particular study made use of intradimensional discrimination learning, such that the CS+ and CS- were values along the same dimension. Alternatively, if the stimuli are along different dimensions then the animal will form an interdimensional discrimination (Switalski, Lyons, & Thomas, 1966).

Prior to the empirical work by Hanson (1959), the peak shift effect had been predicted by the Kenneth Spence’s gradient-interaction theory (Spence, 1936, 1937). According to this theory, gradients of excitation and inhibition form around the CS+ and CS-, respectively. These gradients summate, producing a new gradient that peaks at a value further from the CS-. Gradient-interaction theory has been used to explain the results from several studies showing peak shift effects (Honig & Urcuioli, 1981). It has met with some difficulty, however. For instance, Hanson rejected gradient-interaction theory as an explanation for his peak shift results, as the gradients
did not show a sharp decrease to values around the CS-. Further, gradient-interaction theory has been challenged on the basis that inhibition does not appear to generalize in a manner similar to excitation. For instance, Rescorla (2006) showed narrower generalization around a conditioned excitor than a conditioned inhibitor, thus demonstrating that generalization is a function of the type of learning (e.g., excitatory versus inhibitory) and not merely a function of the sensory features of the stimulus dimension. Finally, Ghirlanda (2002) remarked that inhibitory gradients are difficult to empirically verify and gradient interaction theory fails to predict the large peak shift effect observed along perceptual dimensions of increasing physical intensity (e.g. volume or brightness).

The peak shift may also be explained by virtue of the fact that animals form a relational representation between differentially reinforced CSs (Lazareva, 2012). This process may be a result of transposition (Kohler, 1939). Simply put, when discriminating between stimuli, animals may learn more about the relationship between the two stimuli. For example, in the study by Hanson (1959) animals may have learned not only that a 550 nm light was paired with the US, but that a “more green” light was reinforced and another “less green” light was unreinforced. When presented with a light that was “greener” than the CS+, animals may have key-pecked more often based off this relational rule. Whether animals utilize relational properties (Kohler, 1939) between stimuli or are responding according to interacting gradients of inhibition and excitation (Spence, 1937) remains unclear, and may depend on whether the animal is making a choice between different values, or is responding to a series of individually presented values following conditioning. Skinner (1953) favored the relational concept of stimulus generalization, arguing that reinforcements in a natural environment are typically based on the relative properties (e.g. size) of a conditioned object rather than its absolute properties. Importantly, an organism can
also learn to ignore relational information and instead focus on the absolute qualities of the CS if these properties are reinforced.

### 2.2.3 Intensity generalization

Stimuli within certain sensory dimensions can also vary in physical intensity. For instance, sound and light dimensions can vary in volume and brightness, respectively. Variations in the intensity between a CS and a non-conditioned value have unique influences on how generalization is expressed (Ghirlanda & Enquist, 2003, 2007). In a comprehensive review, Guttman and Kalish (2003) arrive at the following conclusions regarding intensity versus non-intensity dimensions (Ghirlanda & Enquist, 2003):

1. Following acquisition to a CS+ of a given intensity, there is a strong shift in response magnitude to stimuli of greater intensity than the CS+. As opposed to peak shift observed in non-intensity dimensions, shifts along intensity dimensions often extend to values far past the CS+. This pattern may be reversed when the unreinforced CS- is a value of greater intensity than the CS+.

2. Gradients obtained from non-intensity dimensions are typically symmetrical around the CS+, whereas intensity dimensions result in asymmetrical gradients with more overall responses to values of greater intensity.

3. Studies examining generalization along dimensions of varying object size show effects of both intensity and non-intensity intensities; they show strong response biases to objects of larger size, but the gradients are typically peaked.

Models of stimulus-intensity generalization may be particularly well suited to describe fear behaviors, since following an aversive experience (e.g. encounter with a vicious dog) an intrinsically salient stimulus (e.g a forbidding dog like a doberman pinscher) may preferentially
evoke a greater fear response than a similar but less intense stimulus (e.g. a harmless dog like a chihuahua).

2.3 Higher-order forms of conditioned learning and generalization

2.3.1 Mediated generalization

In many cases, stimuli are treated with equivalence by the simple fact that the stimuli resemble one another in a definable way, e.g. along a sensory dimension. But in other cases, perceptually dissimilar stimuli are treated with equivalence despite the fact the stimuli do not bear any meaningful resemblance. The question of how organisms learn to treat physically dissimilar objects as equivalent has been approached primarily in terms of associative learning processes, such as mediated generalization. Mediated generalization (also referred to as acquired equivalence) can be achieved through a common association between perceptually dissimilar stimuli (Goldstone, 1998). For example, if you know that two of your friends enjoy the opera, and you later learn that one of the friends also likes Mexican food, then you may infer that the other friend likes Mexican food as well. Acquired equivalence has mostly been studied in laboratory animals. In a seminal conditioning experiment, Honey and Hall (1989) demonstrated that rats could switch from appetitive to aversive related behaviors by indirectly conditioning an associated stimulus. In this experiment, rats were initially conditioned to expect a food reward (US) following each of two discriminable tones of different frequency (CS1+, CS2+). Subsequently, when one of the tones (CS1+) was paired with a new aversive US (an electrical shock), the new CR (freezing) generalized to the cue that was never paired with the US (CS2+) (Honey & Hall, 1989). In this type of experiment, generalization between distinct cues is facilitated by a prior association to the same US. In other words, the original US served as an intermediary that allowed a new behavior to generalize.
2.3.2 Sensory preconditioning

In some cases, generalizing an acquired fear following an emotional learning episode is predicated on an association between stimuli formed before the emotional experience even occurred. This form of stimulus generalization is known as sensory preconditioning (Brogden, 1939; Gewirtz & Davis, 2000), and is an important process whereby the representation of a known stimulus is modified following a salient experience with a related stimulus. A typical sensory preconditioning procedure occurs in three phases: first, subjects learn to associate two neutral cues—for example, a bell and a light. Next, one of the cues, say the light, is associated with a biologically significant US, which leads to the development of a CR. The bell is called the preconditioned stimulus (PS) and the light is called the CS. Finally, the PS is presented alone. Despite the fact that it has never predicted the US directly, the PS will often elicit a CR, suggesting that an association was formed during preconditioning between the PS and CS. The majority of sensory preconditioning studies have been conducted in nonhuman animals (Brogden, 1939; Rizley & Rescorla, 1972). Sensory preconditioning effects have also been reported in a small number of human fear-conditioning experiments (Vansteenwegen, Crombez, Baeyens, Hermans, & Eelen, 2000; White & Davey, 1989), but human research on this topic is very limited relative to experiments using first-order Pavlovian conditioning procedures. A study utilizing sensory preconditioning procedures is presented in Chapter 5.

2.4 Associative learning models of stimulus generalization

2.4.1 Bush and Mosteller

Bush and Mosteller (1951) proposed a simple model for stimulus generalization that uses as its cornerstone an elemental theory of associative learning. Although the model is useful in terms of explaining some of the differences between similarity and discrimination, the account for stimulus generalization is tautological; generalization is argued to be a function of similarity, yet
similarity is defined by the amount of generalization. The model takes the elemental approach towards associative learning, which states that the various components that make up a CS (and are present at the time of learning) constitute a set of discrete elements. Each element can acquire some association with the US, but not all elements will become conditioned. The animal will only form associations to some element(s) on any given learning trial. Following conditioning, if the animal is presented with a new set of elements, the probability that the animal will generalize learning depends on the overlap between the similarity between the novel set and the previously conditioned set of elements. Moreover, generalization is determined by the overlap in elements that had been successfully conditioned – an overall perceptual similarity is not sufficient. Importantly, this model does not attempt to account for how organisms respond according to the perceived similarity of stimuli along the physical dimension. They state that, “any sensible measure of similarity is very much organism dependent” (Bush & Mosteller, 1951). Thus, the main limitation of this model is that it is descriptive and does not make any predictions for the shape of the stimulus generalization gradient following conditioning to a value along a particular dimension.

2.4.2 Rescorla and Wagner (1972)

A highly influential model of classical conditioning to emerge in the twentieth century was the Rescorla-Wagner model (Rescorla & Wagner, 1972). While this model does not explicitly account for stimulus generalization, modifications of the Rescorla-Wagner model have been devised to explain the process of stimulus generalization (e.g. McLaren & Mackintosh, 2002).

The basic tenet of this model is that a CS can either gain or lose associative strength on each learning trial depending on its value in predicting the US. This idea is captured by the equation:
\( \Delta V = \alpha \beta (\lambda - \Sigma V) \)

In this equation, \( V \) refers to the associative strength for the CS, thus \( \Delta V \) refers to the change in associative strength of the CS. The values \( \alpha \) and \( \beta \) range between 0 and 1 and refer to the salience of the CS and the US, respectively. The maximum value (asymptote) for the US is given by \( \lambda \), which is established by the magnitude of the US. Thus, the equation elegantly states that the change in associative strength for the CS is determined by the maximum value for the US minus the prediction for the US. If the US is not predicted then this value will be large and associative strength for the CS increases. Once the CS predicts the US with perfect accuracy the CS gains no additional associative strength. In other words, there is no new learning once the US is no longer surprising. Importantly, this model postulates that all the CSs that are present on a given learning trial can enter into association with the US as either an excitatory or inhibitory CS. This point is captured by \( \Sigma V \), which describes how CSs summate to determine associative strength. This final point is what makes the Rescorla-Wagner model an elemental model of associative learning.

Donald Blough (1975) used this model to explain how stimulus generalization can occur along physical dimensions. In line with the Rescorla-Wagner model, he postulated that the elements that constitute a CS are composed of multiple elements; but he went further in proposing that the physical elements can individually vary on a continuous scale. Thus a novel stimulus along the dimension as the CS will activate some of the stimulus elements present at conditioning, but will not activate all of these elements, causing a decrement in responding.

Evidence that the Rescorla-Wagner model could be used to explain stimulus generalization came from a study by Rescorla (1976), who demonstrated the effects of generalization on conditioned fear expression. The model allows for several predictions regarding the effects of stimulus generalization by incorporating the notion that stimuli are composed of
individual elements. For instance, by virtue of the fact that tones exist along a sensory dimension, two tones (A and B) will have certain features in common (X) that allows learning to a transfer between the stimuli. Rescorla explored this aspect of the model by exposing rats to a tone (B) prior to conditioning a novel tone (A). Since A and B share the common feature (tone, X), conditioning should be advanced to A through reinforcement of B. Through several experiments, this conclusion was verified (Rescorla, 1976).

Notably, the Rescorla-Wagner model did not initially focus on generalization between a CS and a GS. In fact, the theory failed to predict certain forms of generalization that had been known from early studies of classical conditioning studies. For instance, the model assumes that when an element that has been conditioned (A) is combined with another element that has not conditioned (B), then compound AB should elicit an equal amount of responding as A alone. This follows from the model because B has neither excitatory nor inhibitory strength, thus the associative strength of A should generalize perfectly to AB. However, Pavlov had shown half a century earlier that generalization between A and AB was not perfect; responses to AB are weaker than responses to A alone - an effect known as external inhibition (Pavlov, 1927).

2.4.3 Pearce (1987, 1994)

While the Rescorla and Wagner model of classical conditioning is elemental in nature, learning models proposed by Pearce (Pearce, 1987, 1994) are known as configural. The distinction between these two approaches is that elemental theories claim that the elements that make up a CS can enter into individual associations with the US (Rescorla & Wagner, 1972), whereas configural theories argue that these elements combine to form a compound representation that enters into a single association with the US (Pearce, 1987).

Configural models account for certain failures of the Rescorla-Wagner model, particularly in regard to explaining stimulus generalization. As mentioned, if CS A is combined
with novel CS B, there is a decrease in responding to the compound AB that is not predicted by the Rescorla-Wagner model. By relying on the concept of configural stimuli, the Pearce model is able to explain the decrement in responding from A to AB. Simply stated, generalization is determined by the similarity between the element A and the compound AB. Because AB shares a similar element, one expects some amount of generalization, but because B is present in the compound these two stimuli are not identical and, as a result, there is only partial generalization.

Pearce’s model of classical conditioning further explains stimulus generalization through the concept of a buffer of limited capacity. This buffer is the animals’ real time representation for the environment, and is always filled by the various stimuli present at any given time. When the animal is presented with a US, the collection of various stimuli serve as a single CS. Associative strength builds up to some of these stimuli through repeated CS-US learning. The stimuli that are consistently present (or attended to) will gain associative strength, similar to in the Rescorla-Wagner model. When a stimulus changes across trials, such as the property of a light or tone, a comparison is made to the representation of the CS stored in long term memory. The amount of similarity (‘S’ in equation 2 below) determines generalization. Importantly, while certain stimuli may change (i.e. a light changing from 555 nm to 590 nm), other stimuli remain constant (i.e. the surrounding context). This ensures that some amount of generalization will be maintained across trials, even if the CS itself varies to some extent. The equation for the Pearce (1987) model of generalization is intended to derive how much excitatory strength will generalize from the CS (A) to a similar stimulus (A’):

$$e_{A'} = A' S A * E_A$$

Here, $e_A$ corresponds to the excitatory strength of the generalized stimulus, $A'$, and is determined by the similarity ($S$) between A and $A'$ as well as the excitatory strength of A ($E_A$). Similarity ($S$) is a value that can range from 0 (not at all similar) to 1 (completely similar). The
important components to this model are that excitatory strength to the generalized stimulus is determined by similarity, as well as the excitatory strength that has accrued to the CS.

Notably, the Pearce (1987) model is a similarity model, and therefore does not produce monotonic generalization gradients with strong response biases characteristic of intensity dimensions (Ghirlanda, 2002)

2.4.4 McLaren and Mackintosh (2002)

McLaren and Mackintosh (2002) extended the Rescorla-Wagner model to further explain the process of stimulus generalization from an elemental position of classical conditioning. Their postulate also borrows from simple neural network models of associative learning, in which encoding of stimulus features is widely distributed across nodes that overlap to form stimulus representations. Their argument rests on three key ideas. First, the nodes that constitute stimulus elements (and, in turn, form the stimulus representation) contain overlapping gradients of activity. Therefore, a single physical stimulus can be represented by an array of activity across several nodes (or units). They suggest that stimuli are represented coarsely, meaning several nodes may be activated to varying degrees to correspond to a single representation. Second, similarity is measured by the degree to which two stimuli activate the same overlapping gradients of activity. Along a physical dimension, a CR gradient can be viewed as a “tuning curve” that shows strong activity to the CS value itself, but is also activated to some extent by closely related values. Third, organisms will not condition to all the elements presented on a given trial due to variations in their internal states as well as background “noise” from the context that is not pertinent to CS-US learning.

This postulate relies heavily on the idea that, over the course of learning trials, some elements gain associative strength and some elements lose associative strength (Rescorla & Wagner, 1972). As a result, gradients of excitation and inhibition form around particular stimulus
elements (Spence, 1937). Note that the difference between this and earlier models (Hull, 1943; Pavlov, 1927) is the reliance on an elemental model of conditioning, which emphasizes that each element of a stimulus can acquire different amounts of associative value. This elemental approach leads to interesting predictions for how animals learn to generalize behavior. For instance, while similarity is based on how many gradients of excitation overlap between the CS and another stimulus, it does not necessarily follow that more similar stimuli will evoke more generalization. This is because differential reinforcement during conditioning establishes which particular stimulus elements gain associative value. While elements common only to the CS+ gain associative value, elements common to unreinforced control stimuli (or shared between CSs) lose associative value according to the error correcting rules in the Rescorla-Wagner model. Thus, if an organism is presented with two novel stimuli that vary in similarity to the CS+, generalization is determined by those elements common to the CS+ that also helped differentiate the CS+ from a CS- (McLaren & Mackintosh, 2002). Moreover, this new stimulus may evoke even more responding than the CS+ itself (i.e. peak shift) if it contains more of those critical elements than the CS+.

In light of this and other results explained by their model, McLaren and Mackintosh (2002) argue that a configural theory (Pearce, 1987) is less tenable than an elemental theory of stimulus generalization. In particular, they argue that configural theories cannot account for the peak shift effect predicted by an elemental analysis because a configural theory does not account for how multiple elements can be arranged within a compound stimulus. For instance, the elemental view easily accommodates for the changing features of a complex stimulus (e.g. a face) by noting that each element corresponding to a stimulus feature has its own association with the US. Configural theories, on other hand, contend that the arrangement of each feature forms a wholly new stimulus compound which itself enters into association with the US. This approach
does not easily account for peak shifts when features within a compound stimulus are altered. Examples of this type of effect are presented in Chapter 3 and Chapter 4, wherein changes are made to the features of a face (compound cue) involved in the display of fearful affect (Dunsmoor, Mitroff, & LaBar, 2009; Dunsmoor, Prince, Murty, Kragel, & LaBar, 2011).

2.5 Is generalization a result of forgetting?

The passage of time is strongly linked to a flattening of the generalization gradient, i.e. loss of specificity to the CS+ (Riccio, Richardson, & Ebner, 1984). Simply put, given enough time an animal may forget the details of the CS (or context in which the CS was presented) but retain the CR. Thus, the animal may be more inclined to generalize the CR to a different stimulus or context (Riccio, Ackil, & Burchvernon, 1992). Reviews of several experiments reveal the common finding that generalization increases as the time between training and testing increases (reviewed in Riccio, et al., 1992; Riccio & Joynes, 2007; Riccio, et al., 1984).

Recent fear conditioning studies in laboratory animals have demonstrated the effect of time on stimulus generalization. A study by Biedenkapp and Rudy (2007) revealed that rats conditioned in a particular context showed enhanced generalization to new contexts as time progressed from initial conditioning. Three groups of rats were conditioned in context A and then tested in context B after 1, 7, or 15 days. Rats tested in context B after 1 day showed significantly less fear (freezing response) than in context A. Rats tested in context B after 7 or 15 days, however, showed substantial freezing behavior in context B. Thus, an increase in the time between learning and test increased the amount of generalization between the learned and novel context.

Biedenkapp and Rudy (2007) also tested for the effect of pre-exposure of the conditioned context on subsequent generalization to a novel context. Their hypothesis was that animals given enhanced pre-exposure to context A should form a better representation of the context. This
enhanced representation should be mediated by the hippocampus and should reduce forgetting and consequently thwart generalization to a novel context. A control group was pre-exposed to a separate context, X, before conditioning in context A. Results showed that following a 15 day retention interval, rats given pre-exposure to the training context showed far less generalization to novel context B than rats given pre-exposure to the control context, X. They concluded that generalization is a result of forgetting because, given enough time, the representation of the conditioned context begins to fade (Riccio, et al., 1992; Rudy & Wright-Hardesty, 2005). If the details of the representation are forgotten, then a novel context becomes capable of evoking the response that had been acquired in another context. This explanation allows for a novel interpretation of stimulus generalization which incorporates knowledge of brain mechanisms involved in the encoding and retrieval of information, as well as memory consolidation processes.

2.6 Summary

The early observation from Pavlov’s laboratory that conditioned behaviors transfer across stimuli stimulated nearly a century of theoretical and empirical research on stimulus generalization. Early debates on whether generalization was indeed an empirical phenomenon, and not merely a failure to discriminate, were largely settled by instrumental conditioning experiments in pigeons in the mid twentieth century. These studies showed that generalization gradients spanned across stimuli that could be discriminated, indicating that generalization was on some level a process and, perhaps, a cognitive act (Shepard, 1987). Associative learning models have taken this position in explaining how the representation of the CS influences generalization of excitation and inhibition.

This chapter also detailed myriad ways conditioned learning is generalized. Through various manipulations of the stimulus dimension and experimental techniques, a consistent finding is that generalization tends to diminish as physical similarity to the reinforced CS
decreases. However, if the animal has learned to discriminate between different values prior to generalization testing, then the peak of behavioral responses can sometimes shift to a value that has never been reinforced, known as the peak shift effect.

Importantly, behavior is often determined by other factors not accounted for by similarity based models. For instance, variations in stimulus intensity lead to monotonic gradients with response biases for generalized stimuli of higher intensity than the CS+. Moreover, animals frequently generalize between dissimilar objects that share a previous relationship or are associated through an intermediary.

Generalization is also linked to the mere passage of time. That is, animals will often generalize learning to new stimuli or new contexts if there is a lengthy period of time between initial learning and test. This process could simply be a result of forgetting the details surrounding the event.

The next chapters present empirical data on fear generalization in humans. These studies build off theories and research of stimulus generalization, presented in this chapter, to explore what processes determine how individuals will respond to stimuli that vary perceptually or conceptually from a learned threat.
3. Perceptual fear generalization along a dimension of emotional intensity

This chapter describes a psychophysiological experiment designed to examine the contribution of fear-relevant perceptual features on fear generalization in healthy adults. Any stimulus generalization experiment requires a stimulus dimension. In this experiment, we employed a dimension of graded emotionally expressive faces of the same identity morphed between neutral and fearful endpoints. In this way, we can assess whether fear generalization conforms to principles of intensity generalization (see Chapter 2), which has previously only been assessed along physical dimensions (e.g. increasing volume or brightness), while simultaneously exploring generalization along a perceptual gradient of similarity. Two experimental groups underwent discriminative fear conditioning between the middle value along an expression gradient (CS+) and an unreinforced CS-. The stimulus dimension and experimental design is depicted in Figure 4. In Experiment 1, the CS- was a relatively neutral face stimulus while in Experiment 2 the CS- was the most fear-intense stimulus. Keeping the CS+ constant between groups, while alternating the CS-, allows us to examine whether the perceptual features of a comparison stimulus affects subsequent intensity generalization.

The results presented in this chapter reveal that fear generalization is broadly tuned and sensitive to the amount of fear intensity of non-conditioned values, but that generalization can come under stimulus control through associative learning. These results reveal a novel form of fear generalization in humans that is based on emotional intensity value, and may have implications for understanding generalization processes in anxiety disorders characterized by heightened sensitivity to nonthreatening stimuli. The data presented in this chapter have been published as a research article in the journal *Learning & Memory* (Dunsmoor, et al., 2009).
3.1 Introduction

Fear generalization serves an adaptive function, but can become maladaptive when nonthreatening stimuli are inappropriately treated as harmful based on similarity to a known threat. For example, an individual may acquire a widespread fear of dogs after an aversive experience with a single vicious dog. In this case, recognizing that an animal is related to a feared (or fear-conditioned) animal is made possible in large part by shared physical features to the fear exemplar, such as four legs and a tail. On the other hand, fear behaviors may be sensitive to particular features associated with natural categories of threat; for example, the fear of dogs is likely associated with the threat of getting bit, and thus more attention is paid to the teeth and jaws than to fur color. Moreover, the degree to which an individual fearful of dogs responds with fear may be related to the intensity of those threatening features relative to the original threat (i.e. greater fear expression to dogs that appear more vicious). Therefore, fear generalization based on perceptual information may occur via two routes – similarity to a learned fear exemplar along nonthreatening physical dimensions or along dimensions of fear relevance. Given that fear generalization often emerges as a consequence of conditioning or observational learning, it is important to determine which characteristics of novel stimuli facilitate generalization and the extent to which generalization processes can be controlled.

A rich literature has revealed principles of generalization in nonhuman animals (see Chapter 2), but few studies of fear generalization have been conducted in humans. Moreover, the existing human studies have yet to consider the second route through which fear responses may generalize—via gradients of fear relevance. Although a wide range of neutral stimuli, such as tones or geometric figures, can acquire fear relevance through conditioning processes, other stimuli, such as threatening faces or spiders, are biologically prepared to be fear relevant (Öhman & Mineka, 2001). Compared with fear-irrelevant CSs, biologically prepared stimuli capture
attention (Öhman & Mineka, 2001), are conditioned without awareness (Öhman & Soares, 1998), increase brain activity in visual and emotional processing regions (Sabatinelli, Bradley, Fitzsimmons, & Lang, 2005), and become more resistant to extinction when paired with an aversive unconditioned stimulus (US) (Ohman, Eriksson, & Olofsson, 1975). Although the qualitative nature of the CS influences acquisition and expression of conditioned fear, it is unknown how generalization proceeds along a gradient of natural threat. For instance, human studies to date have all tested variations of a CS along physically neutral stimulus dimensions, such as tone frequency (Hovland, 1937), geometric shape (Vervliet, et al., 2006), and size (Lissek, et al., 2008).

An important stimulus feature that has been overlooked in the study of human fear generalization is intensity between the CS+ and non-conditioned generalization stimuli (non-CS). A number of generalization experiments in nonhuman animals have shown generalization to increase as a function of physical intensity – for instance from a medium volume sound that predicts the US to a loud volume sound that has never predicted the US (reviewed in Ghirlanda & Enquist, 2003). Intensity generalization often involves a shift in peak responding, such that a non-CS of greater intensity than the CS evokes a greater response than the CS itself (Ghirlanda & Enquist, 2003). Models of stimulus-intensity generalization may be particularly well suited to describe fear behaviors, since many fear-relevant stimuli (which are often involved in real-world fear learning episodes) can vary in physical or emotional intensity. For example, a soldier who experienced an explosion in a battlefield may generalize fear to other loud noises outside a battlefield situation, or an individual fearful of spiders may be more afraid around a tarantula than a daddy long-leg.

To address the role of intensity on perceptual based fear generalization, we employed an intradimensional gradient of emotionally expressive faces of the same identity morphed between
neutral and fearful endpoints. A fearful face is considered a biologically prepared stimulus that recruits sensory systems automatically for rapid motor responses (Öhman & Mineka, 2001), and detecting fearful faces may be evolutionarily selected as an adaptive response to social signals of impending danger (Dimberg & Öhman, 1996). The experimental design is presented in Figure 1.

In Experiment 1, an ambiguous face containing an intermediate fear intensity value served as CS+ and was paired with an electric shock US, while a face on the neutral end of the continuum face served as the CS- and was never paired with the US. Skin conductance responses (SCR) were recorded as a dependent measure of fear conditioning. Before and following fear conditioning, SCRs were recorded in response to non-CS face morphs of the same actor expressing several values of increasing fear intensity.
Figure 4: Perceptual generalization experimental design

(A) The conditioned and non-conditioned stimuli (S1-S5) consisted of a single identity morphed between neutral and fearful endpoints. (B) The experimental session included three phases: preconditioning, fear conditioning, and the generalization test. Fear conditioning involved discriminative learning between the S3, paired with the US (pictured as a lightning bolt) and either the most neutral face S1 (Experiment 1) or the most neutral face S5 (Experiment 2) as the unpaired control CS-. The CS+ was intermittently reinforced throughout the generalization test but never occurred following any of the other faces.
Testing generalization along an intradimensional gradient of emotional expression intensity allows for an examination of the relative contributions of fear intensity and physical similarity on the magnitude of generalized fear responses. If fear generalization is determined purely by the perceptual overlap between the CS+ and other morph values, without regard to fear intensity, then we would expect a bell-shaped generalization function with the maximum SCR centered on the reinforced (intermediate) CS+ value (S3), less responding to the directly adjacent, but most perceptually similar values, and the least amount of responding to the most distal and least perceptually similar morph values. This finding would be in line with stimulus generalization reported along fear-irrelevant dimensions (Lissek, et al., 2008) and in stimulus generalization studies using appetitive instrumental learning procedures (Guttman & Kalish, 1956). If, however, fear generalization is biased toward non-conditioned stimuli of high fear intensity, then an asymmetric generalization function should result with maximal responding to the most fear intense non-CS. This finding would suggest that fear generalization is selective to the degree of fear intensity in stimuli, similar to studies of physical intensity generalization gradients in nonhuman animals (Ghirlanda & Enquist, 2003).

We predicted that the latter effect would be observed, such that the magnitude of SCRs will disproportionately generalize to stimuli possessing a greater degree of fear intensity than the CS+ (Experiment 1). A secondary goal was to determine whether fear generalization to non-CSs can be reduced through discriminative fear learning processes. Therefore, a second group of participants was run for whom the CS- was the most fearful face (Experiment 2). In this case, we predicted that discriminative fear conditioning between an intermediate fear face CS+ and the most intense face as CS- would sharpen the generalization gradient around the reinforced CS+ value, and that responses to the most fear-intensive stimulus would decrease relative to Experiment 1. Moreover, this discriminative fear learning process may provide evidence that fear
generalization is influenced by associative learning processes and is not exclusively driven by selective sensitization to stimuli of high fear relevance (Lovibond, Siddle, & Bond, 1993). Finally, we were interested to discover whether generalization processes would yield subsequent false memory for the intensity of the CS+ in a post-experimental retrospective report. In sum, the present study has implications for understanding how fear generalization is related to the degree of fear intensity of a non-CS, the extent to which discrimination training efforts can thwart the generalization process, and how fear generalization affects stimulus recognition.

3.2 Methods

3.2.1 Participants

Seventy-one healthy adult volunteers (37 female; median age = 21; age range = 18–33 yr) provided written informed consent approved by the Duke University Institutional Review Board. An additional 12 participants were excluded from analysis due to excessively low SCRs. Participants were randomly divided into two experimental groups that differed only in regard to the unreinforced control stimulus (CS−) used during the fear-conditioning phase of the experimental session. Experiment 1 consisted of 36 participants (21 female, median age = 20) who received discriminative fear conditioning between a 55% fear intensity CS+ and an 11% fear intensity CS−. Experiment 2 consisted of 35 participants (16 female, median age = 22) who received fear conditioning with a 55% fear intensity CS+ and a 100% fear intensity CS−. Participants completed a personal history form and were excluded from analysis for self-reported history of neurologic and psychiatric illness, substance abuse, and current medication use.

3.2.2 Stimulus set

A male face morphed along a gradient from neutral-to-fearful was used as the stimulus set. This identity was taken from the Ekman pictures of facial affect (Ekman & Friesen, 1976)
and was morphed along a continuum from neutral-to-fearful using MorphMan 2000 software (STOIK) as described by (LaBar, Crupain, Voyvodic, & McCarthy, 2003). Face morphs were positioned in a full-frontal orientation and cropped to include only the face, without hair, ears, or neckline. Images were normalized for contrast and luminance and appeared on a gray rectangular background. Participants were exposed to the same face identity during all experimental phases to ensure that only fear expression and not other features related to actor identity (i.e., marking specific to an individual face such as moles or facial hair), was manipulated. Five faces along the continuum were used: 11% fear/88% neutral, 33% fear/66% neutral, 55% fear/44% neutral, 77% fear/2% neutral, and 100% fear. These stimuli are also labeled as S1, S2, S3, S4, and S5, respectively. Hence, the S3 served as the conditioned stimulus (CS+), and either the S1 (Experiment 1) or S5 (Experiment 2) served as the unpaired CS− during fear conditioning.

3.2.3 Task design and procedures

The presentation of all stimuli, recording of subjective face ratings, reaction times, and shock delivery were managed with Presentation software (Neurobehavioral Systems, Albany, CA). Prior to the start of the experimental session, the intensity of the electric shock (US) delivered to the wrist was calibrated for each participant to a level deemed “highly annoying but not painful” using an ascending staircase procedure. On each trial, participant’s rated whether or not face stimuli were expressing fear (forced choice, yes/no) as quickly and accurately as possible by pressing one of two buttons. The order of button presses was counterbalanced between participants. Participants were not informed of the CS–US contingency at any time during the experiment, but were told that some of the faces may be followed by a shock. Participants were provided a chance to practice the subjective rating task with one presentation of each of the five face stimuli prior to the start of the experiment (habituation data is not reported).
The experiment was separated into three phases that occurred in the same order for each participant: preconditioning (30 trials), fear conditioning (20 trials), and generalization testing (45 trials) (see Figure 4). Stimuli were presented for 4 s in all phases, followed by a 500-ms blank background. The blank background was then followed by a variable duration intertrial interval (ITI), which consisted of a fixation cross. The order of all face stimuli was counterbalanced between participants and pseudo-randomized such that no more than two of the same stimulus values occurred in a row. A 5–6-min break followed the first two phases of the experiment, wherein participants passively viewed a silent video of a train traveling through British Columbia (Highball Productions). Following the generalization test, participants were shown the five face stimuli used during the study and asked to identify which face had been paired with the US during the experiment. To ensure that participants would not simply point out an intermediate value along the continuum, the collection of five stimulus values was shown to the subject in a randomized, nonlinear order (i.e., not in sequence from neutral-to-fearful).

3.2.3.1 Preconditioning

Preconditioning contained six trials of each face stimulus (30 total) separated by an ITI that varied between 5, 6, and 7 s. The US was never presented during preconditioning, but participants were still instructed that some of the faces may be followed by the US. This phase was identical for Experiments 1 and 2 and allowed for a baseline measure of SCRs to the five stimulus values prior to fear conditioning.

3.2.3.2 Fear conditioning

Fear conditioning involved 10 trials of the CS+ and 10 trials of the CS− separated by an ITI that varied between 8, 9, and 10 s. The S3 (55% fear intensity) served as the CS+ in both Experiment 1 and Experiment 2 during fear conditioning and co-terminated with the US on six out of 10 trials. Importantly, in Experiment 1 the unpaired CS− was the S1 (11% fear intensity),
while in Experiment 2 the CS− was the S5 (100% fear intensity). The CS− was never paired with the US. Only the CS+ and CS− appeared during fear conditioning.

3.2.3.3 Generalization testing

Generalization testing consisted of nine trials of each stimulus (45 total) separated by an ITI that varied between 6, 7, and 8 s. The generalization test was broken into three segments, each containing three presentations of each of the five stimuli. Only a brief break (<1 min) was given between generalization test segments. The CS+ (S3) was intermittently paired with the US on 33% of generalization test trials (“steady-state” generalization test) (Honig & Urcuioli, 1981). These steady-state (“booster” trials) reinforcement procedures are intended to extend the length of time over which responding can be measured and offset the effects of extinction and habituation (Lim & Pessoa, 2008; Smith, Most, Newsome, & Zald, 2006). One unpaired US was also delivered at the start of the second and third segment of the generalization test prior to any stimulus presentation to maintain general arousal. The generalization test phase was identical for Experiments 1 and 2.

3.2.4 Psychophysiological methods and analysis

SCRs were recorded with the MP-100 system (BIOPAC systems, Goleta, CA) and sampled at 200 Hz. A conductive saline-based gel was used with Ag/AgCl electrodes that were placed on the middle phalanx of the second and third digits of the left hand. SCR analysis was carried out on AcqKnowledge software (BIOPAC systems). SCRs were considered related to stimulus presentation if the trough-to-peak response (1) occurred 1–4 s following stimulus onset, (2) lasted between 0.5 and 5.0 s, and (3) was greater than 0.02 microsiemens (μS). If these criteria were not met, the SCR was scored as zero. SCRs were hand analyzed by a trained scorer to ensure that responses were related to stimulus presentation and were not due to fluctuations occurring prior to stimulus onset, movement artifacts, or other noise in the SCR waveform as
previously described (LaBar, Cook, Torpey, & Welsh-Bohmer, 2004). Twelve participants were characterized as “nonresponders” based on a lack of measurable SCR and were therefore removed from data analysis (LaBar, et al., 2004). SCRs were square-root transformed and averaged for each participant. Mean RTs occurring prior to 100 ms or after 2000 ms from stimulus onset were not included for analysis. RTs not meeting these criteria constituted <1% of total trials. SCR and RT data were analyzed by ANOVA and polynomial trend analyses, with an α value of 0.05.

3.3 Results

Results from Experiment 1 and Experiment 2 are presented in Figure 5 and Figure 6, respectively.

3.3.1 Experiment 1: Preconditioning

An ANOVA using the five fear intensity values as repeated measures demonstrated no main effect of the amount of fear intensity on SCR magnitude, $F_{(4,140)} = 0.374, P = 0.82$. This result indicates undifferentiated responses to faces containing different levels of fear intensity prior to conditioning (Fig. 5A). A main effect of stimulus type on reaction time (RT) was revealed (Fig. 5C), $F_{(4,136)} = 19.63, P < 0.001$, as well as both linear ($P < 0.01$) and quadratic trends ($P < 0.01$). The fastest RTs were elicited by the most intense fear face (100%, S5). The slowest RTs were observed for the ambiguous stimuli, S2 and S3. Finally, participants consistently rated the S3, S4, and S5 as fear expressive, while the S1 and S2 were consistently not rated as fear expressive (Figure 5D; see also Graham, Devinsky, & LaBar, 2007). Based on these results, a categorical boundary was determined to exist between the S2 and S3 (point of subjective equality), such that faces of 55% fear intensity and greater were reliably discriminated from faces of less fear intensity that appeared neutral. Despite the fact that the S1 and S2 were categorically separate from the S3–S5, there were significant differences in RT between the S1
and S2 ($P < 0.01$), which suggests that these two within-category stimuli were perceptually discriminable. Similarly, there were significant differences in RTs for the S3, S4, and S5 as well ($Ps < 0.01$).
Figure 5: Experiment 1 Results for preconditioning and the generalization test

(A) Mean skin conductance responses (SCR) were undifferentiated among the stimulus values prior to fear conditioning, but an asymmetrical linear generalization gradient emerged during the generalization test. (B) Generalization across the three blocks of the generalization test shows a similar pattern of generalized SCRs that habituated over time. (C) Reaction times (RT) were significantly faster following fear conditioning for the S3 and S4. (D) Subjective face ratings revealed a categorical boundary between the S2 and S3. (E) Retrospective CS+ identification demonstrated a high percentage of subjects mistakenly identified the more fear-intense S4 as the CS+ post-experimentally. Error bars reflect standard error (SEM); * denotes significant differences $P < .05$; $\mu$S = microsiemens.
3.3.2 Experiment 1: Fear conditioning

SCRs were significantly greater to the CS+ (mean ± SEM: 0.72 ± 0.05) than to the CS- (0.38 ± 0.04), indicating that conditioning took place, \( t_{(35)} = 7.78, P < 0.01 \). RTs were significantly slower to the CS+ (989.73 ms ± 43.06) relative to the CS- (919.13 ms ± 38.45), \( t_{(35)} = 2.38, P = 0.02 \). Finally, participants rated the CS+ as expressing fear on 93% of trials, whereas the CS- was rated as expressing fear on <1% of trials.

3.3.3 Experiment 1: Generalization test

3.3.3.1 SCRs

Fear generalization was characterized by a linear trend, \( F_{(1, 35)} = 24.47, P < 0.01 \), while other polynomial trends were not significant \( (P > 0.22) \). Importantly, the amount of generalization to each stimulus was a function of fear intensity, and peaked at the most intense fear value, S5 (Fig. 5A). Repeated-measures ANOVA confirmed a significant main effect for fear intensity value on SCR magnitude, \( F_{(4, 140)} = 10.37, P < 0.01 \). Post-hoc Bonferroni corrected t-tests were performed and revealed significant differences between S1 and S3 \( (P < 0.005) \) and between S2 and S3 \( (P < 0.05) \), indicating a decrease in SCRs to stimuli of less fear intensity than the CS+ that were rated by participants as being neutral. The overall pattern demonstrates an asymmetrical generalization gradient around the reinforced value, consistent with the prediction of an intensity generalization gradient (Ghirlanda & Enquist, 2003) but inconsistent with a perceptual feature-based generalization gradient, which would predict a bell-shaped curve around the CS+ value. Saturated monotonic gradients occur when responses evoked by non-CS values of greater intensity than the CS+, along a dimension of increasing intensity, do not fall below the CS+ response level (Ghirlanda & Enquist, 2003). This pattern was consistent across the three blocks of the generalization test (Fig. 5B), with an overall decrease in responses, as would be expected over the course of an extended conditioning session.
3.3.3.2 Pre-to-post SCR analysis

To determine whether the stimulus generalization gradient emerged as a consequence of fear conditioning, and was not simply a function of non-associative responses to the stimulus values prior to fear conditioning, an additional analysis was conducted on the difference in SCR to all five stimulus values from preconditioning to the generalization test. For this analysis, participants’ mean response to each stimulus value during preconditioning was subtracted from their response to the same stimulus value during generalization testing. Inspection of this difference gradient (Fig. 5A) reveals a similar shape as the generalization test gradient itself. An ANOVA of the pre-to-post SCR difference revealed a main effect for stimulus type on the magnitude of the SCR, $F_{(4,140)} = 8.55$, $P < 0.001$, with a significant linear trend, $F_{(1,35)} = 30.20$, $P < 0.001$.

3.3.3.3 RT

Following fear conditioning, ANOVA revealed a main effect of stimulus type on RT (Fig. 5C), $F_{(4,136)} = 15.77$, $P < 0.001$, as well as both linear ($P < 0.01$) and quadratic trends ($P < 0.01$). The pattern of RTs during the generalization test was similar to that during preconditioning, such that the fastest RTs were elicited by the most intense fear face (100%, S5), and the slowest RTs were observed for the ambiguous stimuli (S2 and S3). There was a significant facilitation in RTs from preconditioning to the generalization test for the stimulus serving as the CS+ (S3), $t_{(34)} = 2.54$, $P < 0.05$, and the S4, $t_{(35)} = 2.22$, $P < 0.05$.

3.3.3.4 Subjective face ratings

The S1 and S2 were rated as not expressing fear on nearly all trials, whereas the S3, S4, and S5 were rated as expressing fear on the majority of trials (Fig. 5D). These results were similar to those during preconditioning indicating that fear conditioning did not have an effect on subjective face ratings to any of the stimulus values.
3.3.3.5 Retrospective CS+ identification

Following the experimental session, participants were asked to identify which stimulus had been paired with the US during the course of the experiment (Fig. 5E). Forty-five percent of participants correctly identified the S3 (55% fear intensity). Interestingly, 36% of participants falsely identified the S4 (77% fear intensity) as being paired with the US, consistent with the SCR peak shift occurring during the generalization test. This result may indicate an illusory correlation (Tomarken et al. 1989), such that the stimulus of greater fear intensity was mistakenly thought to be paired with the US for a considerable number of participants. To examine the relationship between memory accuracy for the identity of the CS+ and the extent of generalization, a subsequent ANOVA was conducted on SCRs during the generalization test using recognition accuracy as a between-subjects factor. There was no significant effect of recognition accuracy, $F(1, 30) = 0.005, P > 0.05$, indicating that the form of the generalization gradient was similar irrespective of participants’ abilities to accurately recollect the identity of the CS+.

3.3.4 Experiment 2

The goal of Experiment 2 was to determine the effect of discriminative fear conditioning between the CS+ and a CS- of greater fear intensity on the generalization gradient. If generalization operates without regard to fear intensity value, then fear conditioning with the most fear-intensive value as the unpaired control would be expected to produce a bell-shaped gradient around the CS+, or potentially shift the gradient toward the less fear-intense S2 and S1 (Hanson, 1959; Spence, 1937). Conversely, if fear generalization is driven entirely by fear intensity of the non-CS, these discriminative fear-conditioning procedures would be expected to have little effect on the generalization gradient. In this case, the generalization gradient would be nearly identical to that observed in Experiment 1. We predicted that discriminative fear conditioning in Experiment 2 would sharpen the form of the generalization gradient around the
CS+ value, and that responses to the most fear-intense stimulus would be reduced relative to Experiment 1, reflecting learning-related stimulus control.

3.3.5 Experiment 2: Preconditioning

As in Experiment 1, baseline SCRs did not significantly vary as a function of fear intensity, $F_{(4, 136)} = 1.08, P = 0.36$ (Fig. 6A). RTs were also similar to Experiment 1, such that a main effect of stimulus type on RT was found (Fig. 6C), $F_{(4, 128)} = 25.13, P < 0.001$, as well as both linear ($P < 0.01$) and quadratic trends ($P < 0.01$). The most fear-intense face evoked the fastest RT, whereas the ambiguous S2 and S3 evoked the slowest RT (Fig. 6B). Finally, the perceptual categorical boundary between the S2 and S3 was similar to Experiment 1 (Fig. 6D).
Figure 6: Experiment 2 Results for preconditioning and the generalization test

(A) Mean skin conductance responses (SCR) were undifferentiated among the stimulus values prior to fear conditioning, but a quadratic gradient that peaks before the most fear-intense stimulus emerges during the generalization test. (B) Generalization across the three blocks of the generalization test shows a similar pattern of generalized SCRs that habituated over time. (C) Reaction times (RT) were significantly faster following fear conditioning for the S1. (D) Subjective face ratings are similar to those reported in Experiment 1. (E) Retrospective CS+ identification revealed a high percentage of subjects correctly identified the S3 as the CS+ post-experimentally. Error bars reflect standard error (SEM); * denotes significant differences P < .05; µS = microsiemens.
3.3.6 Experiment 2: Fear conditioning

Significantly larger SCRs were elicited by the CS+ (mean ± SEM: 0.62 ± 0.05) as compared with the CS- (0.51 ± 0.06), indicating that conditioning took place, \( t_{(34)} = 2.42, P < 0.05 \). Participants responded with significantly slower RTs to the CS+ (1209.78 ms ± 54.77) than the CS- (891.34 ms ± 33.17), \( t_{(34)} = 9.15, P < 0.001 \). Finally, the CS+ was rated as expressing fear on 74.7% of the trials, whereas the CS- was rated as expressing fear on 99.4% of the trials. Interestingly, the endorsement rate of fear expression for the CS+ was reduced from Experiment 1 (93%).

3.3.7 Experiment 2: Generalization test

3.3.7.1 SCR analysis

In contrast to Experiment 1, fear generalization was characterized by both linear, \( F_{(1, 34)} = 12.57, P < 0.01 \), and quadratic trends, \( F_{(1, 4)} = 13.11, P < 0.01 \), with the peak of the gradient closer to the reinforced S3 value. ANOVA confirmed a main effect of stimulus type on SCR magnitude, \( F_{(4, 136)} = 9.48, P < 0.01 \). Post-hoc t-tests, corrected for multiple comparisons, showed increasing responses from S1 to S2 (\( P < 0.05 \)) and again from S2 to S3 (\( P = 0.027 \)). The increase from S2 to S3 (\( P = 0.13 \)) and from S3 to S4 (\( P = .18 \)) were not significantly different, but there was a significant decline in responding from S4 to S5 (\( P < 0.05 \)). These data show that the breadth of the intensity based generalization gradient was reduced relative to Experiment 1 largely via truncated SCRs to the most fear-intense stimulus. However, a gradient reversal was not found. Finally, the pattern of stimulus generalization was similar across the three generalization test segments (Fig. 6B).

3.3.7.2 Pre-to-post SCR analysis

A comparison of SCRs to each stimulus value from pre- to post-fear acquisition revealed a main effect for stimulus type, \( F_{(4, 136)} = 7.84, P < 0.01 \), with both linear, \( F_{(1, 34)} = 16.27, P < 0.01 \).
0.01, and quadratic, $F_{(1, 34)} = 12.29, P < 0.01$ trends. Inspection of this difference gradient (Fig. 6A) reveals a similar shape as the generalization test gradient itself. A 2 x 5 repeated-measures ANOVA was also conducted to directly compare differences in pre-to-post SCRs across Experimental groups (Experiment: 1, 2; fear intensity: S1–S5). A significant quadratic effect was found, $F_{(1, 69)} = 4.68, P < 0.05$, reflecting a sharper response gradient from Experiment 2 relative to Experiment 1.

### 3.3.7.3 RT

A main effect of stimulus type on RT was observed during the generalization test, $F_{(4, 136)} = 20.92, P < 0.001$, that was both linear ($P < 0.001$) and quadratic ($P < 0.001$). Following fear conditioning, RTs significantly increased for the least fear-intense stimulus, S1, $t_{(34)} = 2.77, P < 0.05$, as compared with preconditioning.

### 3.3.7.4 Subjective face ratings

As in Experiment 1, there was no discernible effect of fear conditioning on subjective face ratings (Fig. 6D).

### 3.3.7.5 Retrospective CS+ identification

Twenty-two out of 35 participants (62%) correctly identified the CS+ as being paired with the US during the Experiment. The S2 and S4 were falsely identified by 17% and 14% of participants, respectively, which is below chance (Fig. 6E). To examine the effect of mistaken CS+ identification on stimulus generalization, a subsequent analysis was conducted using CS+ identification accuracy as a between-subjects factor. There was no significant effect of CS+ identification on the form of the stimulus generalization gradient, $F_{(1, 32)} = 0.585, P > 0.05$, replicating Experiment 1.
3.4 Discussion

These results reveal broad generalization of the conditioned fear response to stimuli sharing similar perceptual features to a CS, but varying in the amount of fear intensity. Specifically, fear responses generalized from an ambiguously fearful face CS+ to non-CS images of the same actor displaying more or less expression of fear. In line with the prediction that generalization would increase with the intensity of non-CSs, the greatest amplitude of SCRs was recorded for stimuli that contained the highest fear intensity in Experiment 1. This finding is complemented by the fact that a number of participants mistakenly identified a more fearful stimulus as the CS+ post-experimentally. An orderly generalization gradient was also observed in Experiment 2, such that SCRs increased from stimuli containing less fear intensity than the CS+. However, the gradient was sharper overall in Experiment 2 (with a quadratic trend) and the most fear-intense stimulus evoked considerably less generalization than in Experiment 1. Further, participants in Experiment 2 successfully identified the CS+ with high frequency post-experimentally. Because Experiments 1 and 2 differed only in regard to the CS-presented during fear conditioning (Experiment 1: more fearful face, Experiment 2: less fearful face), differences in stimulus generalization between groups can be interpreted as a consequence of stimulus control gained through discrimination learning.

3.4.1 Relationship to sensory forms of intradimensional intensity generalization

The manner in which generalization increased with fear intensity in Experiment 1 is in line with prior nonhuman animal studies intensity generalization (Ghirlanda, 2002; Ghirlanda & Enquist, 2003). When animals are trained to respond to a particular stimulus within a given sensory modality (i.e., a light or noise), there is a strong bias to respond to stimuli along the same dimension that are of greater intensity than the trained CS (i.e., increases in brightness or noise
intensity) (See Chapter 2.2.3). Because there was no difference in the physical intensity of the
five stimuli along a sensory dimension (brightness, etc.), intensity was determined here by the
degree of emotional expression, a quality defined by changes in key facial-feature configurations
that constitute the display of fearful affect (Susskind et al., 2008). To our knowledge, no fear
conditioning study in humans or nonhuman animals has examined intensity generalization using a
CS- of higher physical intensity than the CS+. A nonhuman animal experiment using auditory
stimuli with appetitive reinforcement, demonstrated that linear intensity generalization gradients
could be reversed when the CS- was louder than the CS+, such that less intense non-CS sounds
elicited more generalized responses than more intense non-CS sounds (R. C. Huff, Sherman, &
Cohn, 1975). Fear generalization in the present study, however, did not show this reversed pattern
(i.e., generalization was not greater to the S1 or S2 relative to the S4 in Experiment 2). This
observation suggests that a dimension of fear intensity is not altogether equivalent to intensity
along sensory dimensions, or that conditioned fear generalizes differently from appetitive
learning. Our results further argue against a gradient interaction model for fear generalization, as
inhibition to the CS-, in summation with an excitatory gradient around the CS+, did not shift
responses to values further from the CS- (Spence, 1937). Thus, it is possible that fear
generalization was determined to some extent by sensitization to images of fear expression
(Lovibond, et al., 1993).

3.4.2 Contribution of discriminative fear learning on stimulus generalization

The difference in stimulus generalization gradients and retrospective CS+ identification
in Experiments 1 and 2 suggests that fear generalization may be influenced by the representation
of the CS+ formed during discriminative fear learning. An elemental model of stimulus
generalization may provide an account for how the representation of the CS+ was formed over
the course of fear conditioning. First, elemental models of Pavlovian conditioning have argued
that the numerous elements comprising a CS can form independent associations with the US (Rescorla & Wagner, 1972; Wagner, 2008). Over the course of learning trials, different elements can either gain or lose associative strength, depending on which elements better predict the US. In an extension of this elemental model formulated by McLaren and Mackintosh (2002), when an organism encounters a stimulus that has not been conditioned, generalization is determined by the associative strength of the elements shared with a similar conditioned stimulus (See Chapter 2.4.4). According to this view, the elements that constitute a stimulus contain overlapping gradients of activity (Blough, 1975). Through conditioning, some of these elements gain associative strength, thus increasing the excitatory gradient surrounding a particular stimulus element. Similarity is based on how many of these elements overlap between the CS+ and a non-CS. If elements common to both stimuli are those most associated with the US (through CS–US pairing), then stimulus generalization is likely to occur. Further, differential reinforcement between the CS+ and CS- establishes which particular elements gain associative value; while elements common only to the CS+ gain associative value, elements common to the CS-, or shared between CSs, lose associative value. Thus, a non-CS may evoke greater responding than even the CS+ (i.e., peak shift) if it is in greater possession of those elements that originally differentiated the CS+ from the CS-.

In terms of the present study, discriminative fear learning established that features common to the CS+ became associated with the US, whereas features common to the CS- did not become associated with the US. For Experiment 1, in which the CS- was a relatively neutral stimulus, features that differentiated the CS+ from the CS- were those associated with fear expression, while other features related to subject identity contained the most overlap and were not predictive of the US. Consequently, fear expression predicted the US, which may account for generalization to stimuli of higher fear intensity than the CS+ following fear conditioning. On the
other hand, in Experiment 2, fear expression was a poor predictor for the US, since the CS-contained more fear expression than the CS+. Consequently, features related to fear expression should have acquired very little associative strength, which may account for the decrease in SCRs to the most fear intense stimulus in this experimental group. However, while responses to the most fear-intense stimulus were reduced relative to Experiment 1, the generalization gradient was not simply the reverse of Experiment 1, as would be predicted if fear generalization were based purely on the interaction between excitatory and inhibitory elements. It is possible that conditioning an ambiguously fearful face increased the associative value of fear-relevant features in both experimental conditions, but that this propensity to generalize to fear-intense stimuli interacted with learned discrimination in Experiment 2.

3.4.3 Clinical implications

These findings have implications for certain anxiety disorders characterized by overgeneralization of fear responses to nonthreatening stimuli, such as PTSD and specific phobias. Prior fear-conditioning studies have shown that individuals with PTSD fail to limit responses to the reinforced CS+ (Grillon & Morgan, 1999). In another study, extinction to a cue that was perceptually similar to the CS+ did not result in generalization of inhibition to the originally conditioned stimulus (Vervliet, Vansteenwegen, Baeyens, Hermans, & Eelen, 2005). Vansteenwegen et al. (2007) showed that individuals fearful of spiders failed to generalize extinction of a feared spider exemplar to a novel context when provided with extinction training in a single context. However, a group of spider anxious individuals who received extinction under multiple contexts did generalize extinction of the feared stimulus (i.e., expressed less fear renewal than the group trained in one context) when it was presented in a novel context. Clearly, the mechanisms underlying fear generalization are complex and involve responses to features related both to the context and the cue-predicting threat. It is noteworthy that the present study used a
relatively ambiguous fear stimulus as the CS+ (near the point of subjective equality between emotion categories), which was sufficient to yield broad generalization gradients and false memories for the intensity of the CS+ (Experiment 1).

Finally, the limited numbers of human fear generalization studies to date have supported the idea that fear excitation is broadly tuned, whereas fear inhibition is selective (Lissek, et al., 2008; Vervliet, et al., 2005; Vervliet, Vansteenwegen, & Eelen, 2004). This capacity to broadly generalize fear may serve as an adaptive function to aid survival in an ever-changing, potentially harmful environment (Mineka, 1992). In the present study, a reduction in SCRs to the CS- resulting from discriminative fear conditioning remained specific to the CS- exemplar. These findings indicate that generalization is conserved across a wide range of perceptually similar stimuli, whereas specialization—learning when not to respond to nonthreatening stimuli—may require additional training procedures (N. C. Huff & LaBar, 2010). Thus, learning to control fear responses to nonthreatening stimuli may require extinguishing exemplars of high fear intensity that are perceptually related to a feared stimulus under multiple contexts in order to overcome a predisposed (or exacerbated) fear generalization process.

3.4.4 Conclusion

Empirical findings on stimulus generalization date back to the earliest research on classical conditioning (Pavlov, 1927). Several decades’ worth of animal research, which often incorporated instrumental conditioning procedures, further extended our knowledge on processes contributing to stimulus generalization. Combined with theoretical explanations attempting to distinguish between discrimination and generalization processes (Bush & Mosteller, 1951; Pearce, 1987; Shepard, 1987), this area of research has been active for nearly a century (Ghirlanda & Enquist, 2003). Yet, despite its theoretical relevance to understanding human fear learning and anxiety disorders, the role of generalization in the acquisition and expression of
learned fear has received little attention in human conditioning research beyond the initial case of ‘‘Little Albert’’ (Watson & Rayner, 1920). The present examination shows that fear generalization is not solely determined by the physical similarity between a CS and other stimuli. Other factors involved in fear learning and fear expression help determine the extent to which humans generalize fear, even to relatively ambiguous exemplars. These factors include a bias to generalize to non-CSs containing relatively more fear intensity than a physically similar conditioned stimulus. Importantly, this process can come under stimulus control through discriminative fear learning procedures, at least in healthy adults, suggesting that fear generalization processes are partly associative in nature. The present findings have implications for behavioral approaches to treating some anxiety disorders. Future investigations will need to address the processes involved in the maintenance of fear generalization over time, the relative contribution of associative and non-associative mechanisms, and the direct role of awareness on the expression of fear generalization in humans.
4. **Neurobehavioral mechanisms of perceptually-based fear generalization**

This chapter presents the first neuroimaging findings of fear generalization in humans, published as a research article in the journal *Neuroimage* (Dunsmoor, Prince, et al., 2011). This fMRI investigation employed the paradigm described in Chapter 3, in which participants are presented with a range of faces of the same identity morphed between neutral and fearful endpoints prior and subsequent to fear conditioning. This study followed the design of Experiment 1 from Dunsmoor et al. (2009). Many of the experimental details and background on intensity generalization are provided in the previous chapters, and will therefore not be covered in depth here. The focus of this chapter is on the neural correlates of perceptually-based fear generalization.

4.1 **Introduction**

Fear learning involves acquiring defensive behaviors to aid survival in response to environmental threats. To be adaptive, it is important that this learning is flexible such that stimuli highly similar to a learned threat are treated as potentially harmful as well. Extensive neurophysiological and brain imaging research has established several key regions involved in fear learning processes, including the amygdala, insula, cingulate gyrus, striatum, sensory cortex, and prefrontal cortex (Phelps & LeDoux, 2005). The goal of the present study is to examine neural systems contributing to the generalization of fear learning, and to link generalization-related activity to individual differences in fear expression, functional connectivity, and trait anxiety.

Neurophysiological research of fear generalization in non-human animals has focused primarily on the amygdala (e.g. Armony, et al., 1997), as this region serves a pivotal role in
forming the CS–US association and producing conditioned fear behaviors (Davis, 1992). For example, a loss of GABA release in the lateral nucleus of the amygdala has been shown to increase fear generalization (Bergado-Acosta et al., 2008; Shaban, et al., 2006). Duvarc et al. (2009) conducted a fear conditioning task in rats using two cues – a CS paired with an electric shock (CS+) and a control CS unpaired with the US (CS-) – and found large individual differences in the extent to which animals generalized fear from the CS+ to the CS−. They suggested that differences in generalization were determined in part by the bed nucleus of the stria terminals, a region closely linked with the central nucleus of the amygdala, as rats with excitotoxic lesions to this region showed a small amount of generalization to the CS−. However, the precise role of the amygdala (and other regions) in fear generalization is not clear, as generalization is often conceptualized as heightened responses to the CS−, and formal behavioral tests using graded stimuli have rarely been conducted in neuroscientific models of fear generalization. Human neuroimaging studies have shown that the amygdala, striatum, and medial prefrontal cortex flexibly respond to CSs that change in their association with an aversive US (Schiller & Delgado, 2010), but it is not clear whether these regions show graded responses to stimuli that vary in similarity from the CS+ along some featural dimension.

The present study provides a novel, systematic examination of human fear generalization using event-related fMRI with concurrent measures of psychophysiological arousal (i.e. SCR). Predictions on how the human brain mediates fear generalization were based on knowledge of the brain systems involved in acquiring and expressing learned fear. First, generalization could be mediated by regions involved in differential fear learning — that is, areas that show greater activity to a CS+ compared to a CS−. We expected that learning related regions would show enhanced activity to generalized stimuli of high emotional intensity, consistent with our prior
behavioral work showing that generalization increases as a function of the emotional intensity value of non-conditioned cues (Dunsmoor, et al., 2009).

However, it is possible that regions that show broad enhancement of activity to generalized stimuli are not directly related to the production of fear behaviors, and thus poorly reflect variability in fear learning and generalization. Therefore, to constrain interpretations of neural activity, we quantified behavioral measures of generalization by the change in SCR magnitude from pre-to-post fear conditioning, and used these residualized scores to examine brain–behavior correlates during a test of fear generalization. These individual behavioral profiles take advantage of the fact that psychophysiological responses entail considerable individual variability that may be mediated by components of the fear learning circuitry, such as the amygdala (Cheng, Knight, Smith, & Helmstetter, 2006; Knight, Nguyen, & Bandettini, 2005). Thus, we predicted that the change in fear expression following fear conditioning would be correlated with activity in limbic/paralimbic regions involved in sympathetic activation and core affective processes, such as the amygdala and insula. The amygdala is also important for enhancing the sensory representation of feared stimuli (Armony & Dolan, 2002) through reciprocal connections with sensory processing regions like the extrastriate visual cortex (Amaral, Behniea, & Kelly, 2003). We predicted enhanced connectivity between the amygdala and visual processing regions coding for the domain specific properties of the CS and related stimuli following fear conditioning. Finally, to provide a link to clinical anxiety disorders, we investigated whether amygdala connectivity related to fear generalization is correlated with individual differences in trait anxiety, in line with previous findings showing enhanced amygdala–extrastriate connectivity in phobic individuals viewing phobia relevant images (Ahs et al., 2009) and serotonin transporter (5-HTT) short allele homozygotes viewing fearful faces (Surguladze et al., 2008).
4.2 Methods

4.2.1 Participants

Twenty-five healthy adult volunteers provided written informed consent approved by the Duke University Institutional Review Board. Two participants were removed from analysis due to excessive head movement (> 3 mm in any direction), and 9 participants were not included due to a lack of SCR data, which precludes an examination of fear learning and generalization (5 participants lacked SCR data due to technical issues and 4 participants were classified as non-responders based on lack of measurable SCRs throughout the session). The analysis included 14 participants (7 female; median age 22 yrs; age range = 19-30 yrs). Participants completed the State-Trait Anxiety Inventory (STAI) (Spielberger, 1983) prior to the start of the experiment.

4.2.2 Stimulus set

The stimulus set is identical to that detailed in Chapter 3 and shown in Figure 4. The stimuli consisted of 5 morph values of the same actor identity morphed parametrically between neutral and fearful endpoints; labeled as S1-S5 for clarity.

4.2.3 Task design and procedures

The experimental paradigm is based on Experiment 1 detailed in Chapter 3 (Dunsmoor, et al., 2009). The scanning session consisted of three phases that occurred in the same order for each participant: preconditioning (3 runs), fear conditioning (2 runs), and generalization testing (3 runs). A short 5-min break followed preconditioning and generalization testing during which time participants passively viewed a silent video of a train traveling through British Columbia (Highball Productions). Each trial was 4 s in duration, during which time participants were asked to rate whether or not the face was expressing fear (forced choice: yes/no) as quickly and accurately as possible by pressing one of two buttons. The order of button presses was
counterbalanced across subjects. The intertrial interval (ITI) consisted of a white fixation cross on a black background that followed the offset of each trial. The lengths of the ITI were jittered according to an exponential distribution function. Preconditioning contained a total of 9 trials of each of the 5 morph increments (45 total trials) with an average ITI of 5 s (minimum 4 s). Fear conditioning contained a total of 16 S3 (CS+) and 16 S1 (CS−) trials (32 total) with an average ITI of 11 s (minimum of 9 s). The CS+ co-terminated with the US on 10/16 trials, whereas the CS− was never paired with the US (partial reinforcement delay conditioning procedure). The generalization test contained a total of 9 trials of each of the 5 morph increments (45 total) with an average ITI of 9 s (minimum of 5 s). The S3 was intermittently paired with the US on 6/9 trials ("steady-state" generalization test) to offset the effects of extinction over the course of an extended testing session, as routinely implemented in animal models of generalization (Honig & Urcuioli, 1981). In all phases, stimulus presentation was counterbalanced and pseudo randomized such that no more than two of the same morph increment occurred in a row. Subjects were not informed of the CS−US contingencies. Following the conclusion of the generalization test, a functional localizer task was performed to isolate cortical regions selective for processing images of faces. The localizer, based on a previously published design (J. P. Morris, Green, Marion, & McCarthy, 2008), included pseudo-randomized blocks of black and white images of faces and flowers that each contained 24 images presented for 500 ms each. Two localizer blocks were run and separated by 12 s of fixation. Data from the functional localizer were entered into a general linear model with 2 conditions of interest: faces and flowers (see below).

4.2.4 Psychophysiological methods and analysis

A long-standing issue in the study of stimulus generalization concerns measurement of the generalization gradient (Hull, 1943; Pavlov, 1927). The present analysis developed an approach adapted from the animal literature to characterize individual fear generalization profiles.
by normalizing SCRS to each morph increment as proportion of total response output (Honig & Urcuioli, 1981). This approach is particularly suitable for the present study, as it accounts for individual differences in response profiles during the preconditioning and generalization test phases. To derive this metric, preconditioning data were analyzed by dividing the sum of SCRs for each morph increment (S1–S5) by the total sum of SCRs to all morph increments during the preconditioning phase. Fear conditioning data were also normalized on the basis of response output to the CS+ and CS−. Finally, SCRs obtained during the generalization test were analyzed for each test run, as fluctuations in response patterns over time have been observed in prior behavioral stimulus generalization experiments (Dunsmoor, et al., 2009). To normalize generalization test data, the sum of responses to each stimulus type was divided by the sum of responses to all trials within each of the three generalization test runs. This yielded three generalization gradients, one for each run of the generalization test. The behavioral correlates of fear generalization were operationally defined as the difference in normalized SCRs during each generalization test run from the corresponding stimulus during preconditioning. In this way, preconditioning served as a baseline measure for each face stimulus for each participant, and allowed for an analysis of the proportional change in response patterns following fear learning. SCR data were analyzed by ANOVA and polynomial trend analyses, with an α value of 0.05 (SPSS 15.0, Chicago, IL).

4.2.5 Functional image acquisition, preprocessing, and GLM analysis

Scanning was performed on a General Electrical Signa EXCITE HD 3.0 Tesla MRI. Subjects wore ear plugs to reduce scanner noise, and head motion was minimized by using foam pads. Blood oxygenation level dependent functional images were acquired parallel to the AC-PC line using a SENSE™ spiral in sequence: acquisition matrix, 64×64; field of view, 256×256; flip-angle, 60°; 34 slices with interleaved acquisition; slice thickness, 3.8 mm with no gaps between
slices; repetition time, 2 s; echo time, 27 ms. Functional data were preprocessed using SPM 8 software (Wellcome Department of Cognitive Neurology, University College London, www.fil.ion.ucl.ac.uk) implemented in Matlab (The Mathworks Inc, Natick, MA). The first 4 functional images from each scanning run were discarded to account for magnetic equilibration effects, and remaining images were corrected for head motion using a threshold of 3 mm in any direction. Preprocessing included realignment, spatial normalization to the Montreal Neurological Institute (MNI) template using a fourth degree B-spline interpolation, and smoothing using an isotropic 8-mm3 Gaussian full width half maximum kernel. To remove low-frequency drifts, high pass temporal filtering was applied using a 128-s cutoff. Individual subject trial-related analysis was conducted using the GLM. Covariates of interest included the 5 morph increments for preconditioning (S1–S5), two morph increments for fear conditioning (CS+ and CS−), and 5 morph increments for generalization test (S1–S5). The hemodynamic response was modeled for each covariate of interest using a variable duration design that incorporated reaction time for each trial (Grinband, Wager, Lindquist, Ferrera, & Hirsch, 2008). The US was modeled as an impulse (dirac) function and, along with the 6 head motion parameters, were included as a covariate of no interest. Second-level random effects analyses were performed using one-sample t-tests in SPM8 to investigate the main effect of differential fear learning between the CS+ (S3) and CS− (S1). These contrasts included CS+ and CS− trials from the fear conditioning phase and the steady-state generalization test. A threshold value was initially set to \( p < 0.001 \), uncorrected, with an extent threshold of 5 contiguous voxels, to identify whole brain activity related to differential fear learning. For each a priori region of interest (ROI) identified from whole brain analysis, the search space for multiple comparisons was restricted to bilateral anatomical masks from the Wake Forest PickAtlas toolbox (Maldjian, Laurienti, Kraft, & Burdette, 2003) using a family wise error (FWE) correction of \( P < 0.05 \). These included separate masks for the caudate, insula,
thalamus, and anterior cingulate cortex. All reported regions survived this correction. Contrast images were then created using each generalized stimulus (S2, S4, and S5) versus the CS− (S1), and each generalized stimulus versus the CS+ (S3). For these contrasts images, the mean beta-parameters were extracted from the preconditioning and generalization test phase using an 8 mm radius sphere surrounding the peak voxel identified from independent functional contrast comparing the CS+ versus CS− (coordinates reported in Table 1). For activation in the thalamus extending into the periaqueductal gray (PAG), the peak coordinates used for the ROI analysis were selected from a meta-analytic review by Kober et al. (2008). Statistical tests conducted on the extracted parameter estimates were assessed by one-sample t-tests for each contrast and two-sample t-tests for comparing the same contrasts across the preconditioning and the generalization test phases, with an α value of 0.05 (SPSS 15.0, Chicago, IL). Data from the functional localizer were analyzed by first level contrasts of faces versus flowers. These contrast maps were analyzed in a second level random-effects analysis, yielding regions preferentially engaged by images of faces versus flowers. Search space was restricted to an anatomical mask of the fusiform gyrus (Maldjian, et al., 2003) using FWE correction of $P < 0.05$.

4.2.6 fMRI Regression analysis

Regression analyses were conducted using a second-level random effects regression model in SPM 8 to investigate brain–behavior correlations in human fear conditioning and generalization. Regression analysis of fear conditioning was conducted for the two runs of the fear conditioning phase using the difference in SCRs between the CS+ (S3) and the CS−(S1) as covariates, and brain imaging contrasts of CS+ versus CS− as the dependent variable (excluding one subject from the second run for technical problems with the SCR equipment). This analysis revealed regions positively tracking the difference in SCRs between the CS+ and CS−. Next, we conducted a regression analysis of the generalization test using the change in normalized SCRs
from pre-to-post fear conditioning to each stimulus as a covariate. Individual contrasts for S1–S5 (relative to baseline), for the three generalization test runs, were entered as the dependent variable. Mean parameter estimates from the functional ROIs identified from regression analysis were extracted and brain–behavior correlations were plotted for illustrative purposes. No outliers (defined as data points 3 standard deviations from the mean) were detected. Regions from brain–behavior analyses were initially identified for the whole brain at $P < 0.001$ uncorrected and an extent threshold of 5 contiguous voxels. All regions were then subject to FWE correction of $P < 0.05$ applied to the appropriate bilateral anatomical mask from the Wake Forest PickAtlas toolbox.

### 4.2.7 Amygdala connectivity analysis

The goal of this analysis was to examine functional connectivity between the amygdala and face-selective cortex as a function of phase (preconditioning, generalization test) and stimulus value (S1–S5). The seed region for this analysis was the left amygdala identified from the independent fear conditioning regression analysis. The face selective region was the right fusiform gyrus identified from an independent functional localizer task. A GLM was created that modeled each individual trial as a separate covariate (Rissman, Gazzaley, & D'Esposito, 2004). Correlations were computed for each subject by calculating the Pearson correlation coefficients between the mean parameter estimates from the seed region (amygdala) and fusiform gyrus. These values were converted from correlation coefficients to Z scores using the Fisher transform in order to make group inference. Inputs from the single subject level were input into a full factorial model as implemented in SPM 8, with phase (preconditioning, generalization test) and stimulus (S1–S5) as factors.
4.3 Results

4.3.1 Behavioral results

Analysis of SCRs during the preconditioning phase, using ANOVA with S1–S5 as a repeated measure, revealed no difference in SCRs as a function of stimulus intensity value, $F_{4, 52} = 0.83, P = 0.51$ (Fig. 7a). These results demonstrate that SCRs were not sensitive to stimulus intensity value prior to fear learning, consistent with research demonstrating that static facial expressions are not inherently highly arousing (Anderson, Yamaguchi, Grabski, & Lacka, 2006). Next, assessment of SCRs during the fear conditioning phase revealed that differential fear learning took place — subjects expressed greater normalized SCRs to the CS+ (mean ± SEM: 0.72 ± 0.02) than to the CS−(0.27 ± 0.02), $t (13) = 9.20, P < 0.001$. Analysis of SCRs during the generalization test phase showed a main effect of stimulus intensity, $F_{4, 52} = 9.00, P < 0.001$, with a positive linear trend across the five face exemplars ($P < 0.001$) (Fig. 7a). This asymmetric gradient is in line with generalization based on emotional intensity because the distribution of SCRs show an increasing monotonic function that saturates but does not fall below the CS+ value (Dunsmoor, et al., 2009; Ghirlanda & Enquist, 2003). The pre-to-post difference in SCR response profiles (Fig. 7b) provided the primary behavioral index of fear generalization. ANOVA of these fear generalization scores showed a main effect of stimulus intensity, $F_{4, 52} = 3.76, P = 0.009$, with a significant positive linear trend across the five face exemplars, $P = 0.01$. Whereas stimuli of less intensity than the CS+ showed an average pre-to-post decrease in SCRs, the CS+ and stimuli of high intensity showed an average increase following fear conditioning. However, behavioral responses across the emotional intensity gradient were variable overall, and these individual variations in behavior provide the basis for the neuroimaging regression analysis.
Figure 7: Behavioral results

(a) Mean normalized SCRs from preconditioning and the generalization test show that response output was undifferentiated along the neutral-to-fearful continuum during preconditioning (white bars) and shifted towards stimuli of high emotional intensity during the generalization test (black bars). (b) Difference scores, reflecting the change in response output from preconditioning to generalization test, show an average decrease in psychophysiological responses to the S1 and S2 and an average increase to the S3, S4, and S5. (c) Mean reaction times show that subjects were fastest to categorize the S4 and S5 as fearful. (d) A majority of subjects (71%) mistakenly identified the S4 as the CS+ indicating a strong illusory correlation. Error bars reflect standard error of the mean (SEM), (*) denote significant differences (p < 0.05) and (**) at p < 0.01.
On each trial, subjects rated whether or not the morphed stimulus was expressing fear. Analysis of reaction time (RT) data revealed a main effect of morph increment during preconditioning, $F_{4, 52}= 6.667, P = 0.001$, with both linear ($P = 0.01$) and quadratic ($P = 0.02$) trends (Fig. 7c). A main effect of stimulus type on RT was also observed during the generalization test, $F_{4, 52}=3.596, P = 0.01$. Post-hoc Bonferroni corrected t-tests revealed that RTs were faster for the S4 versus the S3 both before and after fear conditioning, $P < .05$. See also Supplemental Fig. 1 for subjective ratings of fear expression. At the conclusion of the experimental session, subjects were asked to identify which morphed stimulus had been paired with the US. A chi-square test revealed that a significant percentage of subjects (71%, 10 out of 14) identified the S4, $\chi^2 (4) = 23.86, P < 0.001$, from the array of stimuli (Fig. 7d). This false retrospective identification for the CS+ as a more emotionally intense stimulus confirms prior findings (Dunsmoor, et al., 2009) of a co-variation bias (Öhman & Mineka, 2001). It is important to emphasize that this bias was revealed after the generalization test, so it is unknown whether the misidentification would be similar if probed during the training itself.

4.3.2 Brain imaging results

We did not observe any brain regions that increased linearly as a function of fear intensity prior to fear conditioning. The neural correlates of differential fear learning were identified by comparing activity to the CS+ (S3) versus the CS− (S1), which revealed enhanced neural activity in the caudate, insula, and thalamus, extending into the PAG (Table 1). The reverse contrast (CS− versus CS+) revealed enhanced activity in the rostral anterior cingulate cortex (ACC) and subgenual ACC. These areas are commonly identified in human fMRI investigations of fear conditioning (Delgado, Li, Schiller, & Phelps, 2008; LaBar & Cabeza, 2006; Phelps & LeDoux, 2005). Activity related to the generalized stimuli (S2, S4, and S5) was examined by probing a priori functional ROIs identified from fear learning. First, to examine
whether differential learning activity was specific to the CS+, contrasts were created that compared activity to each generalized stimulus against the CS−. In this way, the CS− served as a reference condition (Lissek et al., 2010) for both learning (CS+ versus CS−) and generalization (S2 versus CS−; S4 versus CS−; S5 versus CS−). Activity related to these contrasts was extracted from preconditioning and the generalization test. Prior to fear conditioning, no significant effects were observed. Following fear conditioning, enhanced activity relative to the CS− was revealed for the S4 and S5 within the caudate and insula (Fig. 8a), suggesting that generalization-related activity in these regions was driven primarily by the intensity value of non-conditioned stimuli. Enhanced activity was also observed in the thalamus extending into the PAG for the S5. That these effects only emerged during the generalization test, and were not present at preconditioning, supports the idea that enhanced activity emerged as a consequence of fear learning. A similar analysis was conducted for the reverse contrast (regions that preferentially signaled the CS− relative to the CS+ during learning). No significant differences in activity to the generalized stimuli were observed in these regions during preconditioning, but enhanced activity was observed to the S2 versus the CS+ within the rostral and subgenual ACC during the generalization test, indicating that these regions exhibit a reversed intensity-based generalization gradient perhaps constituting a ‘safety’ signal (Fig. 8b).
Table 1: Fear learning-related activity

<table>
<thead>
<tr>
<th>Region</th>
<th>Hemisphere</th>
<th>MNI coordinates</th>
<th>Size (mm$^3$)</th>
<th>Peak T</th>
<th>Peak Z</th>
</tr>
</thead>
<tbody>
<tr>
<td></td>
<td></td>
<td>x   y  z</td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td><strong>CS+ &gt; CS-</strong></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>Caudate</td>
<td>Right</td>
<td>10  4  2</td>
<td>2432</td>
<td>6.89</td>
<td>5.13</td>
</tr>
<tr>
<td></td>
<td>Left</td>
<td>-10 0 14</td>
<td>640</td>
<td>4.69</td>
<td>3.96</td>
</tr>
<tr>
<td>Brainstem</td>
<td></td>
<td>2  -36  -46</td>
<td>3200</td>
<td>5.77</td>
<td>4.59</td>
</tr>
<tr>
<td>Thalamus/PAG</td>
<td></td>
<td>-6  -20  -2</td>
<td>5824</td>
<td>5.35</td>
<td>4.35</td>
</tr>
<tr>
<td>Insula</td>
<td>Right</td>
<td>34  12  6</td>
<td>1600</td>
<td>4.67</td>
<td>3.90</td>
</tr>
<tr>
<td></td>
<td>Left</td>
<td>-34 16 6</td>
<td>576</td>
<td>4.02</td>
<td>3.51</td>
</tr>
<tr>
<td>Precentral gyrus</td>
<td>Right</td>
<td>42  -8  54</td>
<td>384</td>
<td>4.27</td>
<td>3.69</td>
</tr>
<tr>
<td>Precuneus</td>
<td>Right</td>
<td>10  -48  62</td>
<td>448</td>
<td>4.02</td>
<td>3.51</td>
</tr>
<tr>
<td>Parietal Lobe</td>
<td>Left</td>
<td>-30 -32  46</td>
<td>384</td>
<td>3.82</td>
<td>3.37</td>
</tr>
<tr>
<td><strong>CS+ &gt; CS-</strong></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>Subgenual ACC</td>
<td></td>
<td>-2  32  -6</td>
<td>384</td>
<td>4.11</td>
<td>3.58</td>
</tr>
<tr>
<td>Rostral ACC</td>
<td></td>
<td>-6  44  10</td>
<td>960</td>
<td>4.11</td>
<td>3.57</td>
</tr>
</tbody>
</table>
Figure 8: Brain regions involved in differential fear learning show generalized patterns of activity.

(a) Regions of interest were identified by contrasts of CS+ versus CS−. Contrasts of each nonconditioned stimulus (S2, S4, and S5) versus the CS− revealed significant differential activity for the S4 and S5 versus the CS− (one-sample t tests) in the right caudate (x = 10, y = 4, z = 2) and right insula (x = 34, y = 12, z = 6) during the generalization test. Activity was significantly enhanced for the S5 contrast from preconditioning to the generalization test (paired-samples t test) in the right caudate. (b) Regions of interest were identified by contrasts of CS− versus CS+, and revealed activity in the subgenual (x = −2, y = 32, z = −6) and rostral (x = −6, y = 44, z = 10) ACC. Contrasts comparing non-conditioned stimuli versus the CS+ showed significant differential activity between the S2 and CS+ during the generalization test. All regions were small volume corrected with a family-wise-error p < 0.05, shown here at p < 0.001 (uncorrected) for visualization purposes. Error bars reflect SEM. For one-sample t tests, (*) denote significant differences (p < 0.05) and (**) at p < 0.01. For paired-samples t test, (/) denotes significant differences (p < 0.05).

4.3.3 Brain-behavior interactions

Analysis of brain–behavior correlations focused on the relationship between brain activity and SCRs during the fear conditioning and generalization test phases. In line with prior findings (Cheng, et al., 2006; Knight, et al., 2005), activity in the left amygdala (x = −30, y = 4, z = −26; 256mm³) was positively correlated with behavioral measures of differential learning.
during fear acquisition, $P = 0.005$. This result confirms that amygdala activity was related to measures of autonomic arousal that reflect the acquisition of conditioned fear. Based on findings from our previous behavioral study showing that the S4 in particular evoked considerable fear generalization and induced a false memory as the CS+ (Chapter 3; Dunsmoor, et al., 2009), we predicted that brain–behavior correlations in fear generalization would be revealed for this stimulus. Consistent with predictions, positive correlations were found in the left amygdala ($x = -26$, $y = 4$, $z = -18$; $1280 \text{ mm}^3$) and both right ($x = 42$, $y = 8$, $z = -6$; $6400 \text{ mm}^3$) and left ($x = -42$, $y = -12$, $z = -2$; $832 \text{ mm}^3$) insula (Fig. 9) for S4 presentations. These results show that the change in fear expression to the S4 is reflected by increases in activity within regions commonly implicated in conditioned fear learning. For comparison, no regions showed positive brain–behavior correlations for the S2, a stimulus that shared considerable perceptual overlap with the CS+ but contained less emotional intensity value. These results support the idea that emotional intensity, and not perceptual overlap, drove fear generalization. Importantly, we determined that neural responses in these regions were linked to the physiological expression of fear — when a direct S4 vs. S2 contrast was conducted irrespective of SCR levels, there were no significant results. Thus, activity in the amygdala and insula increased as a function of psychophysiological arousal to a generalized stimulus following fear conditioning. To assess the extent to which correlations identified for the S4 were selective to this stimulus, we conducted a contrast of brain–behavior correlations between the S4 and all other stimulus values, including the CS+. This analysis revealed activity in the left dorsal amygdala ($x = -26$, $y = 4$, $z = -18$; $192 \text{ mm}^3$), indicating that the amygdala showed the strongest brain–behavior effects for the S4, which was falsely identified as the CS+ by a majority of participants after the generalization test.
Figure 9: Brain–behavior interactions in fear generalization.

(a) Normalized SCR scores for the S4 illustrate the variability in response across subjects and across runs. (b) Correlations between SCR scores and brain activity were revealed in the insula and amygdala, such that increases in arousal from pre-to-post fear learning were associated with increases in brain activity in these regions.

4.3.4 Amygdala connectivity results

Finally, because the amygdala is important for modulating activity in sensory processing regions to more effectively detect and evaluate affective stimuli (Armony & Dolan, 2002;
Vuilleumier, 2005), we predicted that connectivity between the amygdala and extrastriate visual cortex would be important for supporting generalization from a learned threat to related stimuli. To examine amygdala connectivity, we used an independent functional ROI of the left amygdala identified from the fear conditioning regression analysis (x = −30, y = 4, z = −26) as a seed region for a trial-by-trial functional connectivity analysis of the preconditioning and generalization test phases (Rissman, et al., 2004). We focused on connectivity between the amygdala and a face-selective region in the fusiform gyrus (FFG) identified from the independent functional localizer task. During preconditioning, there was no difference in functional connectivity as a function of stimulus intensity value (F_{4, 52} = 2.02, \(P = 0.11\)). During the generalization test, there was a main effect of stimulus intensity on amygdala–FFG connectivity (F_{4, 52} = 3.856, \(P = 0.008\)) with a linear (\(P = 0.006\)) and quadratic (\(P = 0.036\)) trend (Fig. 10). A significant phase (preconditioning, generalization test) by stimulus interaction was observed for the CS+ versus the CS− (F_{1, 13} = 7.449, \(P = 0.017\)), indicating a learning related change in amygdala connectivity following fear conditioning. A significant phase by stimulus interaction was also observed for the S4 versus the CS− (F_{1, 13} = 7.454, \(P = 0.017\)), but not for the other generalized stimuli (\(P > 0.1\)). Prior research has also shown enhanced amygdala–fusiform connectivity in anxious individuals while viewing phobia-relevant stimuli (Ahs, et al., 2009), and in 5-HTT short allele homozygotes while viewing fearful faces (Surguladze, et al., 2008). To determine whether individual differences in trait anxiety would correlate with increases in amygdala connectivity following fear learning, we correlated the change in connectivity for the CS+ and S4 from pre- to post-conditioning with each participant's trait anxiety level. This analysis was limited to the CS+ and S4, as these were the only two stimuli to show an increase in amygdala–FFG connectivity following fear conditioning. We found that individual differences in trait anxiety were positively related to the pre-to-post fear
learning increase in amygdala connectivity for the S4 ($r (13) = 0.62, P = 0.018$), but not the CS+ ($r (13) = -0.12, P = 0.69$) (Fig. 5).
Figure 10: Learning-related and generalization-related increases in amygdala–visual cortex activity and its relation to trait anxiety.

(a) Single trial connectivity analysis, using the amygdala as a seed region, showed learning-related changes in amygdala connectivity with a face-selective region within the fusiform gyrus that was identified from an independent functional localizer task. Connectivity between these regions was undifferentiated prior to fear learning, and showed a linear increase in connectivity during the generalization test. ANOVA revealed a phase (preconditioning, generalization test) by stimulus interaction for the CS+ versus the CS−, as well as for the S4 versus the S2. (b) Increases in functional connectivity on S4 trials, from preconditioning to the generalization test, were positively related with trait anxiety scores ($r$ (13) = 0.62, $p = 0.018$). Increases in amygdala–FFG connectivity were not related with anxiety scores for the CS+ ($r$ (13) = −0.12, $p = 0.69$).
4.4 Discussion

Fear generalization is a common occurrence following a highly aversive experience that is exaggerated in some individuals. Here, subjects who underwent classical fear conditioning to a moderately intense emotional face tended to express fear to stimuli that resembled a learned threat but contained greater emotional intensity. Generalized fear expression was mirrored by a false memory of the CS+ as a more intense stimulus than it actually was, indicating a retrospective bias in estimating threat value. FMRI results confirmed several key hypotheses. First, regions involved in the acquisition of differential fear learning also showed responses to generalized stimuli as a function of emotional intensity during the generalization test. Second, activity in the amygdala and insula correlated with individual variability in physiological arousal measures related to the expression of fear generalization. Finally, functional connectivity between the amygdala and the face-selective region of the fusiform gyrus was enhanced during the generalization test to a non-conditioned stimulus of high intensity, and this increase in connectivity from pre-to-post fear learning was correlated with trait anxiety. Combined, these results provide new evidence that regions with an established role in fear learning are important for generalizing fear as well, and indicate that the neural substrates of fear generalization may be best understood on the basis of individual differences in behavior and anxiety levels.

Present findings for the striatum, insula, and thalamus/PAG extend previous knowledge concerning the role of these regions in aversive learning. For instance, prior research has shown that the striatum serves a role in learning to predict and fear an aversive stimulus (Delgado, Li, et al., 2008). The striatum has also been shown to flexibly adapt when contingencies surrounding CS–US pairing are reversed (Schiller & Delgado, 2010). The thalamus is a key region involved in aversive learning that provides sensory information to the amygdala directly and indirectly (LeDoux, 2000), and this region frequently shows enhanced responses to a CS+ in human
imaging studies of fear conditioning (LaBar & Cabeza, 2006). The periaqueductal gray supports physiological responses to aversively conditioned stimuli (LeDoux, Iwata, Cicchetti, & Reis, 1988) and is implicated as part of a core functional group of limbic regions involved in affective processes (Kober, et al., 2008). Likewise, the insula serves a role in physiological responses to affectively significant stimuli, and prior neuroimaging research has shown that uncertainty and anticipation for receiving an aversive stimulus enhances insula activity (Berns et al., 2006; Dunsmoor, Bandettini, & Knight, 2007). In all, generalized patterns of activity in the striatum, insula, and thalamus/PAG imply that areas involved in fear learning are not specific to an aversively conditioned stimulus. These regions may be important for responding to stimuli that share properties with a learned threat in order to adaptively react to potential threats from the environment.

During the generalization test, a non-conditioned stimulus of lower emotional intensity, which most closely resembled the ‘safe’ CS−, showed enhanced activation in the rostral and subgenual ACC when compared to the CS+. Prior research has shown that these regions are important for regulating emotional responses (Phelps & LeDoux, 2005; Sotres-Bayon, Bush, & LeDoux, 2004). For instance, the vmPFC is involved during the recall of learned extinction (Milad & Quirk, 2012; Phelps, Delgado, Nearing, & LeDoux, 2004), and may mediate extinction by inhibiting activity in the amygdala (Maren & Quirk, 2004). Human neuroimaging of fear conditioning has shown that during fear acquisition responses to the CS+ often decrease in the vmPFC, whereas responses to the CS− increase (Schiller & Delgado, 2010; Schiller, Levy, Niv, LeDoux, & Phelps, 2008). This pattern emerged in the present study, such that responses to the CS− were significantly enhanced relative to the CS+, suggesting that the CS− evoked activity related to regulatory processes. Given that the CS− and S2 were both associated with relative decreases in arousal following fear conditioning, it is noteworthy that the S2 was the only other
stimulus to evoke enhanced activity within these regions. This finding suggests that the regulation of fear can generalize beyond a learned safety signal to other stimuli.

The regression analysis of brain–behavior interactions yielded crucial findings on the relationship between the change in arousal following fear learning and neural activity in the amygdala and insula. Animal models of fear conditioning have consistently implicated the amygdala as the final common pathway involved in fear learning and fear expression (LeDoux, 2000). The precise role for the amygdala in fear generalization is not known, but the amygdala may initiate rapid generalized fear responses by way of direct connections with the sensory thalamus (Han, et al., 2008) as neurons in the sensory thalamus are broadly tuned (Bordi & LeDoux, 1994). Prior neuroimaging research has shown that amygdala activity is related to the production of conditioned SCRs, and does not merely track the presence of a stimulus with threat value (Cheng, et al., 2006; Knight, et al., 2005). Consistent with these previous findings, amygdala activity was related with differential SCRs (CS+ versus CS−) during the fear conditioning phase, and selectively tracked behavioral measures related to increases in arousal evoked by the S4 during the generalization test. This result extends previous fMRI findings that have demonstrated correlations in fear expression and amygdala activity to the conditioned stimulus (Cheng, et al., 2006; Knight, et al., 2005). Activity in the insula was also correlated with generalized fear expression. Previous fMRI studies have linked the insula to visceral and emotional processing in general (Phan, Wager, Taylor, & Liberzon, 2002). Theories have emerged proposing that the role of the insula is to integrate physiological information concerning internal bodily states to inform psychological awareness and decision making (Craig, 2009) which, in the context of the present study, may relate to assessing the predictive or affective value of each face exemplar.
It is important to note that the behavioral metric derived to assess fear generalization captured the change in the proportional response to each stimulus following fear conditioning, and did not reflect pure SCR magnitude to each stimulus during the generalization test (c.f. Schiller & Delgado, 2010). This distinction is critical, as this measure provides a novel way to assess how subjects change in their psychophysiological response profile from pre-to-post fear learning. This approach is in contrast to several neuroimaging investigations that have shown a relationship between the direct production of SCRs and brain activity (Critchley, Elliott, Mathias, & Dolan, 2000). The present method helps to ensure that changes in behavior are due to the intervening fear learning phase (Weinberger, 2007), either through associative learning or non-associative sensitization processes (see Dunsmoor, et al., 2009). Therefore, these findings provide a key insight into the relationship between brain activity and behavioral responses to generalized threats following an episode of fear learning.

Functional connectivity analysis revealed increased amygdala–FFG coupling during the generalization test for the CS+ and S4. The amygdala may serve a role in modulating activity in cortical regions to stimuli that have acquired affective significance, for instance through fear conditioning (Armony & Dolan, 2002). Prior research has shown that affectively salient faces preferentially engage the amygdala and FFG (Vuilleumier & Pourtois, 2007). In the present study, functional connectivity between the amygdala and FFG was undifferentiated prior to fear conditioning. The lack of differential amygdala effects during preconditioning for high versus low value fearful faces is in line with prior findings showing that static images of fearful (and angry) faces do not evoke larger responses in the amygdala than emotionally neutral faces (Fitzgerald, Angstadt, Jelson, Nathan, & Phan, 2006; LaBar, et al., 2003). Following fear conditioning, amygdala–FFG connectivity was characterized by learning-related and generalization-related effects — connectivity was enhanced to both the CS+ and the S4 following
the fear conditioning phase. These connections may facilitate fear responses by enhancing the sensory representation of stimuli related to a learned threat.

The finding that correlations between trait anxiety and increases in amygdala connectivity for the S4 is informative for the way in which amygdala connectivity contributes to stimulus processing in high anxious individuals. The association between anxiety levels and brain activity has been explored across a number of emotional processing tasks (Etkin & Wager, 2007), and high anxiety levels are frequently associated with amygdala activity evoked by negative stimuli (Bishop, 2007). The present finding is consistent with a study reporting amygdala–FFG connectivity in phobic patients viewing phobic stimuli (Ahs, et al., 2009). If amygdala connectivity serves to facilitate responses to stimuli that share properties with a feared stimulus, then this pathway might be related to overgeneralization of fears in anxiety disorders. It is interesting that connectivity was not enhanced for the S5 following fear conditioning, considering that the S5 contained the greatest degree of fear expression. This may indicate that functional connectivity is related predominately to the behavioral measures of fear conditioning (i.e., SCRs and retrospective CS+ identification), which show a bias in favor of the S4 above all other stimulus values.

These findings may be interpreted from a number of theoretical perspectives. First, a gradient-interaction theory of stimulus generalization (Spence, 1937) argues that gradients of excitation and inhibition form around the CS+ and CS−, respectively. The summation of these gradients leads to a shift in responses to a value further from the CS−, which could explain the peak-shift in behavioral and neural responses to the S4 and S5 but not the S2. However, gradient-interaction theory has not received strong support from empirical studies of stimulus generalization, as it has been repeatedly shown that inhibition and excitation do not generalize in the same manner (Ghirlanda, 2002; Rescorla, 2006). Furthermore, our previous behavioral
findings using this design argues against gradient-interaction theory as the root cause of emotion-based intensity generalization (Dunsmoor, et al., 2009). In this study, discriminatory fear learning using the most intense stimulus (S5) as the CS− resulted in a sharper generalization gradient around the CS+(S3), but did not cause a reverse gradient (i.e., greater responses to the S1 and S2 versus the S4 and S5). We also note that the BOLD signal in fMRI does not permit a direct test of inhibitory vs. excitatory gradient-interaction effects. Nonetheless, our fMRI analysis in the present study revealed that brain regions that respond selectively to the CS+ or CS− during the initial learning also generalized their response accordingly, so brain regions that support discrimination learning do make some contribution to the generalization gradient. Interestingly, the results show that effects reported for the amygdala do not peak to the CS+ value used during initial learning but instead peaks to the S4, so the amygdala's role in fear conditioning may be underestimated by using its response to the CS+ as the sole index of learning. Another interpretation comes from an elemental associative learning model of stimulus generalization (McLaren & Mackintosh, 2002). In this view, elements that predict the US accrue associative value while elements that do not predict the US lose associative value over the course of conditioning (Rescorla & Wagner, 1972). During the generalization test, similarity to the CS+ is measured by those shared elements that have gained associative value (in this case, features related to fear expression) while other perceptual elements (in this case, features related to identity) are given less weight (McLaren & Mackintosh, 2002). A peak shift can occur if a non-CS contains more of those associative elements than the CS+ itself. However, according to this model the S5 would have evoked more generalization than the CS+ and S4, as it contained the most amount of emotional intensity. Moreover, an elemental associative model does not fully accord to our previous behavioral findings which failed to show a reverse intensity based gradient (Dunsmoor, et al., 2009).
We propose that a stimulus intensity model may best explain the present results. Intensity effects are marked by a response bias (i.e., peak shift) to generalize learned behaviors towards novel stimuli that are somewhat more intense than the CS+ (Ghirlanda, 2002; Ghirlanda & Enquist, 2003). Thus, an intrinsically intense non-CS may be more likely to evoke a heightened fear response after an episode of fear learning than a stimulus that is similar to but less intense than the CS+. Overall, further brain imaging research will be needed to establish any model of generalization as neurobiologically plausible.

Interestingly, we observed a co-variation bias (or “illusory correlation”) along an emotional intensity dimension, such that the majority of subjects falsely identified a more intense face as the CS+ in a post experimental test of awareness. Previous paradigms using fear-relevant (e.g., snakes and spiders) and fear-irrelevant (e.g., flowers and mushrooms) cues have shown that subjects often mistakenly conclude that fear-relevant cues were paired with an aversive US at a higher rate, when in fact both classes of stimuli were equally paired with the US (Öhman & Mineka, 2001). This bias is more pronounced in high anxious individuals (Tomarken, Cook, & Mineka, 1989). The present result is in keeping with these previous findings, and suggests that even healthy adults are biased towards remembering details of a fear learning experience as more emotionally intense than they actually were. Alternatively, this retrospective bias in threat estimation could indicate that subjects were unable to discriminate between the CS+ and S4. However, our prior psychometric studies using these neutral-to-morphs have shown that healthy subjects can readily discriminate between morph values that are even more subtle than those chosen for the present study (Graham, et al., 2007; Thomas, De Bellis, Graham, & LaBar, 2007), and RT data from the present study reveal a clear difference in the time to classify the S3 and S4. Therefore we do not believe that the retrospective bias was indicative of a perceptual confusion during learning itself but future studies could explicitly test awareness intermittent with learning.
The relationship between generalization and discrimination has been the matter of historical
debate in the conditioned learning literature (see for example Hull, 1943; Lashley & Wade,
1946). More contemporary views on the relationship between discrimination and generalization
(e.g. Shepard, 1987) are in part informed by prevailing evidence that generalization can occur
despite the capacity to recognize the difference between exemplars (Guttman & Kalish, 1956).
That is, generalization along a dimension often follows an orderly gradient to stimuli that are both
confusable and discriminable (Pavlov, 1927). Notably, empirical research on stimulus
generalization comes predominately from studies of appetitive instrumental learning. Future
studies that focus on classically conditioned fear behaviors will be needed to fully address the
role of discriminatory processes in aversive learning, and to examine whether the present findings
extend to affectively neutral non-intensity dimensions.

A limitation of the present study is the use of a single stimulus dimension (fear intensity),
whereas generalization can occur along any dimension. In the animal literature, the use of a single
dimension is commonly used when exploring effects of intra-dimensional discrimination training
(Honig & Urcuioli, 1981). It is for this reason that we employed only a single dimension in the
present study, thus ensuring that intra-dimensional changes in behavioral and neural responses
could be attributed to the degree of emotional expression and not to other factors related to
identity. Our prior fMRI work using these face morphs has further shown that dynamic changes
in identity and emotional expression are partly dissociable in the brain (LaBar, et al., 2003). Thus,
additional research is warranted to determine how generalization is mediated along other featural
dimensions.

In conclusion, the neural and behavioral systems involved in processing and reacting to
feared stimuli are becoming increasingly well delineated across species, driven in large part by a
desire to better inform models of clinical anxiety disorders. The laboratory study of fear learning
has typically involved a systematic examination of the processes involved in acquiring, expressing, and extinguishing fears to a specific stimulus. To understand anxiety disorders marked by heightened fear responses, however, it is necessary to explore the processes involved in the generalization of fear to a wider range of stimuli, especially given that a feared stimulus can be encountered in multiple forms. Results from the present investigation demonstrate that regions involved in fear learning are not always specific to a learned threat. Moreover, individual differences in intensity-based generalization suggest a dynamic interplay between corticolimbic–autonomic coupling during fear generalization, and underscores the importance for interpreting brain activity by behavioral measures. Lastly, analysis of functional connectivity between the amygdala and FFG suggest that the amygdala may be important for modulating the sensory representation of stimuli that approximate a learned threat, and that heightened amygdala connectivity may be associated with overgeneralization of fears for individuals with heightened anxiety. Collectively, these results provide novel methodological approaches and insights into the neural basis of fear learning and generalization.
5. Conceptual similarity promotes generalization of higher order fear learning

 Whereas the previous two chapters presented studies on perceptual-based fear generalization, this chapter and the following chapter examine generalization as a function of conceptual similarity. The cognitive psychology literature on concepts is vast and contains a number of central issues and concerns (i.e., prototype versus exemplar models, organizational principles, conceptual development, etc.). These issues have been covered in a number of reviews (e.g. Barsalou, 1999; Martin, 2007b; Medin, 1989; Osherson, Wilkie, Smith, Lopez, & Shafir, 1990) and are discussed in Gregory Murphy’s exceptional tome *The Big Book of Concepts* (2002). Here, we use the term *concept* simply to refer to a mental representation of a class of stimuli, in line with the predominate view in the cognitive psychology literature on concept structure and formation. This helps distinguish perceptual similarity, as construed in the previous chapters, from conceptual similarity, as examined here and in the following chapter. Some object concepts do, by nature, tend to share metric features, but this is not always the case. For example, the study presented in this chapter uses objects that are characteristically associated but do not share physical resemblance, e.g. a spider and a spider web. In Chapter 6, objects are related through simple categorization; that is, determining that a basic level exemplar is an instance of a superordinate category.

 The distinction between perceptual and conceptual generalization may also distinguish learning & behavior processes between humans and many other animals. As stated by Lloyd Morgan in his *Introduction to Comparative Psychology* (Morgan, 1903): “[I]n comparing the psychology of man and the higher animals, the radical difference lies in the fact that man perceives particular relations among phenomena, and builds the generalized results of these
perceptions into the fabric of his conceptual thought; while animals do not perceive the relations, and have no conceptual thought, nor any knowledge – if we use this word to denote the result of such conceptual thought.” (p. 362-363). In this regard, many animal species cannot be expected to exhibit fear responses under the same circumstances as humans.

There was a small historical literature on the role of semantics in classical conditioning (Maltzman, 1977; G. H. S. Razran, 1939). In these studies, subjects would often be conditioned to a word (e.g. tobacco) and then tested for generalization to words that were either related (e.g. smoke) or unrelated (e.g. eagle). In all, the area of semantic conditioning and generalization did not garner much interest within the domain of conditioning research. One reason for this may be that words tend to make poor conditioned stimuli, particularly in fear conditioning. The studies presented below utilize a set of visual images during conditioning, and employ more contemporary procedures and measures for examining conditioning in humans than used in past studies of semantic conditioning (G. Razran, 1961). The data presented in this chapter has been published in the journal Learning & Memory (Dunsmoor, White, & LaBar, 2011).

5.1 Introduction

The idea that many fears originate from an association between neutral and aversive stimuli has continued to serve as the basis for influential accounts of human anxiety disorders (Bouton, 2002; Mineka & Zinbarg, 2006). However, the nexus of many fears defy a straightforward explanation, since following an aversive experience other stimuli that are indirectly related to the experience oftentimes become feared as well. In some cases, generalizing an acquired fear following an emotional learning episode is predicated on an association between stimuli formed before the emotional experience occurred. This form of stimulus generalization is known as sensory preconditioning (SPC) (Brogden, 1939; Gewirtz & Davis, 2000; Kimmel, 1977), and is an important process whereby the representation of a known stimulus is modified.
following a salient experience with a related stimulus. A typical SPC procedure occurs in three phases: first, subjects learn to associate two neutral cues—for example, a bell and a light. Next, one of the cues, say the light, is associated with a biologically significant unconditioned stimulus (US), which leads to the development of a conditioned response (CR). The bell is called the preconditioned stimulus (PS) and the light is called the conditioned stimulus (CS). Finally, the PS is presented alone. Despite the fact that it has never predicted the US directly, the PS will often elicit a CR, suggesting that an association was formed during preconditioning between the PS and CS. The majority of SPC studies have been conducted in nonhuman animals (Brogden, 1939; Rizley & Rescorla, 1972). SPC effects have also been reported in a small number of human fear-conditioning experiments (Vansteenwegen, et al., 2000; White & Davey, 1989), but human research on this topic is very limited relative to experiments using first-order Pavlovian conditioning procedures. Consequently, the features that promote fear generalization between previously associated stimuli are unknown.

A well-known factor important for classical conditioning (Rescorla & Furrow, 1977) and stimulus generalization overall is similarity to the CS (McLaren & Mackintosh, 2002; Pavlov, 1927). Classic experiments of stimulus generalization in animals (reviewed in Honig & Urcuioli, 1981) have revealed orderly generalization gradients of conditioned responses that track perceptual similarity to the CS along a sensory dimension (e.g. Guttman & Kalish, 1956). Prior fear-conditioning research has demonstrated perceptual-based stimulus generalization in humans as well (Dunsmoor, et al., 2009; Lissek, et al., 2008). Yet, humans frequently acquire information about a salient object that goes beyond its physical details. Thus, in many cases generalization is likely based on other regularities extracted from prior experience, such as the conceptual or semantic properties of a stimulus (Maltzman, 1977). While the ability to generalize learning between physically dissimilar objects is not unique to humans (e.g. Herrnstein, 1990; Honey &
Hall, 1989), humans are particularly adept at linking objects together on the basis of conceptual similarity, even in the absence of definable common features (Murphy, 2002). However, it is unknown whether the conceptual relationship between object representations influences generalization of higher order fear learning.

This study employed an SPC paradigm in a between-subjects design to examine whether conceptual similarity of pre-associated cues influences fear generalization. Because of our interest in models of anxiety, fear-relevant stimuli were used. The PSs—a picture of a spider and a picture of a wasp—were held constant across the three groups. In the first phase, these PSs were associated with conceptually related CSs (spider web and wasp nest, respectively), unrelated CSs (waste drums or hospital corridor, counterbalanced), or conceptually mismatched CSs (wasp nest and spider web, respectively; Figure 11). Next, one CS (CS+) was paired with an electrical shock US while the other CS (CS-) was explicitly unreinforced. Finally, the two PSs were presented alone to test for generalization of fear responses, measured here as an increase in SCRs from preconditioning to the generalization test. The effects of SPC were assessed by comparing increases in the SCR to the stimulus pre-associated with the CS+ (referred to as the PS+) versus the stimulus pre-associated with the CS- (referred to as the PS-). We hypothesized that SPC effects would be graded according to conceptual similarity, being strongest in the related group, intermediate in the unrelated group, and weakest in the mismatched group. We also hypothesized that conceptually driven SPC effects would be related to trait anxiety levels, in line with prior research demonstrating a relationship between anxiety and fear conditioning (Lissek et al., 2005).
Figure 11: Schematic of experimental design.

The PSs (spider and wasp) were held constant across the three experimental groups. In phase 1 (sensory preconditioning), subjects learned the association between the PS and a CS that was either conceptually similar, unrelated, or conceptually mismatched. In the next phase of the experiment (fear conditioning), one CS (CS+) was reinforced by an aversive electrical shock unconditioned stimulus (US) on 100% of trials, while the other CS (CS−) was never paired with the US. For the conceptually similar and mismatched conditions, the CS+ and CS− stimuli were counterbalanced between subjects as either the spider web or the wasp nest. This ensured that the PS+ and PS− were also counterbalanced between subjects as either the spider or the wasp. For the conceptually unrelated condition, the spider and wasp equally served as the PS+ (and PS−) in this condition as well. The final phase (generalization testing) immediately followed fear conditioning without a break. After three PS+ and three PS− trials, there were two CS+ trials paired with the shock. Three additional PS+ and PS− trials followed the “booster” trials. PS = preconditioned stimulus; CS = conditioned stimulus; US = unconditioned stimulus, depicted by lightning bolt.

5.2 Methods

5.2.1 Participants

Subjects were 113 healthy young adults. Data from 24 subjects was not included in the final analysis due to a lack of measurable SCR (n = 8) or a lack of differential fear conditioning
SCRs between the CS+ and CS- (n = 15). Removing these participants from analysis is justified because the study goal was to determine how fears generalized in individuals who acquired a differential fear response (i.e., evidence of fear learning was an inclusionary criterion). Data from one subject was excluded due to normalized fear generalization SCR scores that were 3 standard deviations from the group mean. The remaining 89 subjects (30 female; mean age ± SEM: 19.05 ± 0.1; age range 18–23) were randomly assigned to the conceptually similar, conceptually unrelated, or conceptually mismatched conditions. Subjects completed the Spielberger State-Trait Anxiety Inventory (STAI, Spielberger, 1983) and a questionnaire assessing attitudes toward spiders (Klorman, Weerts, Hastings, Melamed, & Lang, 1974) and wasps. The wasp questionnaire was a slightly modified version of the validated spider questionnaire. All subjects fell below the exclusionary criterion of phobia scores, set at within 1 SD from the mean of patients with a specific phobia for spiders (Fredrikson, 1983).

5.2.2 Stimulus set

The PSs included images of a spider and a wasp. For the conceptually related and mismatched groups, the CSs consisted of an image of a spider web and a wasp nest. The CSs for the conceptually unrelated group consisted of an image of a hospital corridor and waste barrel, obtained from the International Affective Picture System (IAPS: #7520, #7079) (Lang, Bradley, & Cuthbert, 2008). All images were pilot rated for valence and arousal, and matched for luminance and contrast.

5.2.3 Task design and procedures

The experimental design is depicted in Figure 1. The PSs (pictures of a spider and wasp) were held constant across the three experimental groups. In the first phase of the experiment (sensory preconditioning), subjects learned the association between the PS and a CS that was either conceptually similar, unrelated, or conceptually mismatched. Subjects were instructed to
predict which image would follow the offset of the PS as quickly and accurately as possible by pressing one of two buttons (counterbalanced between subjects). The task was intended to be simple to promote a strong association between the PS and its corresponding CS in all groups; consequently, all subjects correctly predicted the associated CS by the end of the preconditioning phase. Subjects were informed that no shocks would occur during this phase of the experiment, but were not informed of the PS–CS contingencies. In the next phase of the experiment (fear conditioning), one CS (CS+) was reinforced by an aversive electrical shock unconditioned stimulus (US) on 100% of trials, while the other CS (CS-) was never paired with the US. Subjects were informed at the start of fear conditioning that the task was to identify the image on the screen as quickly and accurately as possible by pressing one of two buttons (counterbalanced between subjects). Subjects were informed that the US would occur, but were not informed of the CS–US contingencies. For the conceptually similar and mismatched conditions, the CS+ and CS- stimuli were counterbalanced between subjects as either the spider web or the wasp nest. This ensured that the PS+ and PS- were also counterbalanced between subjects as either the spider or the wasp. For the conceptually unrelated condition, the spider and wasp equally served as the PS+ (and PS-) in this condition as well. The final phase (generalization testing) immediately followed fear conditioning without a break, and subjects continued to identify the images on the screen by pressing one of two counterbalanced buttons as quickly and accurately as possible. After three PS+ and three PS- trials, there were two CS+ trials paired with the shock. These “booster” trials were included to reduce habituation (Lim & Pessoa, 2008; Smith, et al., 2006), and data from these trials were not included in any analyses. Three additional PS+ and PS- trials followed the “booster” trials.
5.2.4 Psychophysiological analysis

Details concerning the electrical stimulation and SCR analysis are described in Chapter 3 (Dunsmoor, et al., 2009). In brief, the electrical shock US was delivered to the right wrist and managed with the BIOPAC STM200 module (BIOPAC Systems) and set to a level deemed by each subject as “highly annoying, but not painful.” SCRs were recorded from the middle phalanx of the second and third digits of the left hand using Ag/AgCl electrodes connected to the BIOPAC MP150 system (BIOPAC System, Inc.) at 200 Hz and scored with AcqKnowledge software (BIOPAC Systems, Inc.). SCRs were considered related to the stimulus presentation if the trough-to-peak response (1) occurred at least 1 sec after stimulus onset, (2) lasted between 0.5–5.0 sec, and (3) was .0.02 microsiemens (mS). A response that did not meet these criteria was scored as zero. SCRs were square-root transformed for normalization. For analysis of fear generalization, mean SCRs to the PS+ and PS- during the preconditioning phase were subtracted from mean SCRs to the PS+ and PS- during the generalization test for each subject. Statistical analysis included 3 × 2 repeated-measures ANOVA with experimental group (conceptually similar, conceptually unrelated, conceptually mismatched) as a between-subjects factor and stimulus type (CS+ and CS-, or PS+ and PS-) as a within-subjects factor. Planned two-tailed t-tests were conducted to examine a priori hypotheses and to determine differences between the stimulus types where appropriate. Results from statistical analyses were considered significant at α level of 0.05.
5.3 Results

Differential SCRs between the CS+ and CS- provided the primary index of fear conditioning (Figure 12A). Analysis of the acquisition data revealed a significant effect of stimulus \( F(1,86) = 160.63, P < 0.001 \) and trial \( F(5,430) = 16.86, P < 0.001 \), as well as a significant stimulus type (CS+, CS2) by trial (1–6) interaction \( F(5,430) = 13.68, P < 0.001 \), with a linear and quadratic trend \( P < 0.01 \). There was no main effect of experimental group \( F(2,86) = 1.51, P = 0.23 \), or any interactions involving group \( Ps > 0.05 \). As expected, post-hoc-dependent t-tests revealed that SCRs were overall greater to the CS+ than the CS- as training progressed, being significant on trials 2–6 \( Ps < 0.01 \). Overall, fear acquisition was markedly similar among the three learning groups.

The change in SCRs to the PS+ and PS- from preconditioning to the generalization test served as the primary index of SPC (Figure 12B). All trials from the preconditioning phase and the generalization test phase were included to calculate this difference score. ANOVA of the generalization test SCR data, using experimental group and stimulus type (PS+, PS-) as factors, revealed a main effect of group, \( F(2,86) = 3.19, P < 0.05 \), and a main effect of stimulus, \( F(1,86) = 11.37, P < 0.001 \), as well as an interaction between stimulus and group, \( F(2,86) = 3.17, P < 0.05 \). Planned two-tailed t-tests comparing the PS+ and PS- difference scores for each group revealed a significantly greater increase in SCRs to the PS+ compared with the PS- in the conceptually similar condition, \( t(29) = 2.85, P < 0.01 \), and the conceptually unrelated condition, \( t(28) = 2.87, P < 0.01 \). Thus, the SPC manipulation was effective in these two groups. In addition, the magnitude of fear generalization was enhanced in the conceptually similar condition, as revealed by a significantly greater increase in SCRs to the PS+ in this condition versus the conceptually unrelated condition, \( t(28) = 2.38, P < 0.05 \). Responses to the PS-, however, were not significantly different, \( P = 0.11 \). There was no difference in SCR scores to the PS+ and PS- in the conceptually
mismatched condition, $t_{(29)} = 0.09, P > 0.1$, suggesting that SPC was ineffective in this learning group. During the generalization test, subjects were asked to identify the PS+ and PS- as quickly and accurately as possible.

Analysis of the reaction time data (Figure 12C) demonstrated a main effect of group, $F_{(2,85)} = 7.40, P < 0.01$, but no effect of stimulus type and no interaction between stimulus and group. Post-hoc Bonferroni-corrected t-tests revealed that subjects in the conceptually similar condition exhibited faster RTs than subjects in the conceptually unrelated and conceptually mismatched conditions ($Ps < 0.005$).
Figure 12: Behavioral results

(A) Conditioning-related SCRs revealed markedly similar acquisition across the three experimental groups. (B) Fear generalization was calculated as the pre-to-post conditioning change in SCRs to the PS+ and PS−. The change in SCRs was significantly greater to the PS+ than the PS− in the conceptually similar and unrelated groups, but there was no difference in responses between the PS+ and PS− in the conceptually mismatched group.

Responses to the PS+ were significantly enhanced in the conceptually similar versus conceptually unrelated condition, while responses to the PS− were not different across these groups. (C) Reaction times (RT) to identify the PS+ and PS− were significantly faster in the conceptually similar group versus the other two conditions. Error bars reflect standard error of the mean (SEM). *P < 0.05, **P < 0.01. PS = preconditioned stimulus; CS = conditioned stimulus.
To explore the relationship between anxiety levels and higher order fear learning, trait anxiety levels were correlated with fear generalization to the PS+. In the conceptually similar condition, a significant correlation was found between fear generalization to the PS+ and trait anxiety levels, $r_{(28)} = 0.36, P < 0.05$ (Figure 13). In contrast, correlations with SCRs to the PS− were not significant, $r_{(28)} = 0.25, P > 0.05$, suggesting that anxiety levels were not associated with a general increase in arousal. We also examined whether the pre- to post-conditioning change in SCRs to the PS+ were related to the level of fear conditioning (calculated as the differential SCR between the CS+ and CS2), and found that these measures were not significantly related, $r_{(28)} = 0.10, P > 0.05$.

![Figure 13: Correlations between fear generalization and trait anxiety scores for subjects in the conceptually similar condition.](image)

The change in SCRs to the PS+ from pre- to post-fear conditioning were positively correlated with trait anxiety scores ($r = 0.36, P < 0.05$), whereas trait anxiety was not significantly related to the change in SCRs to the PS−. There was no relationship observed between fear generalization and trait anxiety in the conceptually unrelated or conceptually mismatched groups.
5.4 Discussion

The results of the present experiment provide novel evidence that higher order fear learning in humans is influenced by the conceptual similarity of conditional cues. SPC effects were verified in both the conceptually similar and unrelated groups (PS+, PS- responses), but the magnitude of responding to the PS+ was greater in the conceptually similar group relative to the unrelated group. Moreover, RTs were faster during the fear generalization test in the conceptually similar group relative to the other groups, and only in this group was there a positive correlation between fear generalization to the PS+ and trait anxiety. Subjects who learned the association between conceptually mismatched stimulus pairs showed only nonselective increases in SCRs during the generalization test. These findings were not attributable to group differences in initial fear learning or baseline responding. Altogether, the results support the hypothesis that conceptual relationships among stimuli modulate generalization of conditioned fear behaviors in humans.

The SCR findings are in line with the few previous investigations that have employed higher order fear learning procedures in humans (Gottfried & Dolan, 2004; Hosoba, Iwanaga, & Seiwa, 2001; Vansteenwegen, et al., 2000; White & Davey, 1989). For instance, Vansteenwegen et al. (2000) observed SPC effects using pairs of affectively neutral faces and found that generalization was reduced when the PS+ was “pre-extinguished”—that is, when the PS was presented alone following preconditioning but prior to the fear conditioning phase. Therefore, a necessary condition for SPC may be an expectation that the PS portends delivery of the CS. In accord with this suggestion, a PS and CS that are conceptually similar might serve to strengthen the inter-stimulus connection formed during preconditioning, which in turn facilitates fear generalization resulting from first-order fear learning. A similar process of stimulus belongingness has been posited to guide first-order fear conditioning in humans and nonhuman
animals (Seligman, 1970). For instance, in human fear conditioning fear-relevant CSs (e.g., snakes and spiders) appear to enter into an association with aversive electrical shocks more easily than fear-irrelevant CSs (e.g., flowers and mushrooms) (Öhman, 2009). Importantly, SPC effects were observed in the conceptually unrelated condition as well, albeit at a lower magnitude. This finding suggests that conceptual similarity does not selectively induce generalization but rather enhances generalization to pre-associated stimuli. Interestingly, SPC effects were not selective to the PS+ for the mismatched group, showing that conceptual relationships set boundary conditions for subsequent learning. The nonspecific form of fear generalization observed in the mismatched group may be attributed to two sources of generalization—sensory preconditioning and conceptual similarity—that operate in parallel on PS+ and PS- representations. In this group, responses to the PS- are likely due to the conceptual relationship with the CS+, while responses to the PS+ are likely due to higher order fear learning processes resulting from SPC.

It is important to note that the relationship between objects in the conceptually similar condition was assumed to be conceptual in nature because these object pairs (e.g., spider and spider web) are characteristically associated (Medin, 1989). However, it is also likely that these objects have frequently co-occurred in the learning history of our participants, and thus participants have been “preconditioned” prior to participation in this experiment. It is well known that experience shapes conceptual knowledge structures, so the effects observed likely include priors from both conceptual knowledge and direct learning history. If possible, future studies should attempt to dissect these influences, for instance by examining participants who have no direct prior experience with the stimuli. Because the CS+ and PS+ are assumed to be pre-associated prior to the start of the experiment, it would also be interesting to determine whether a fear generalization paradigm that did not include an explicit preconditioning phase would yield
comparable results, similar to studies that present perceptually varying stimuli following, but not prior to, acquisition training (e.g. Guttman & Kalish, 1956).

The neural mechanisms supporting SPC effects have been explored predominately in rodents. These investigations have revealed that the hippocampus and surrounding regions are important for SPC, as lesions to the hippocampus (Talk, Gandhi, & Matzel, 2002), perirhinal cortex (Nicholson & Freeman, 2000), CA1 (Port, Beggs, & Patterson, 1987), and surrounding fibers (Port & Patterson, 1984) block the effects of SPC while leaving classical conditioning intact (Port, et al., 1987; Port & Patterson, 1984). Associative learning models have implicated the hippocampal region as critical for several forms of conditioned learning that involve the representation of conditional stimuli (Gluck & Myers, 1993; Schmajuk, Lam, & Gray, 1996). SPC effects that include the transfer of fear value likely involve the amygdala as well (Gewirtz & Davis, 2000), though the precise role of the amygdala in cognitive processes that involve higher order stimulus representations remains to be elucidated. In terms of the present findings, fear generalization between related object concepts may require integration between the amygdala, hippocampus, and posterior cortical areas important for coding the perceptual and conceptual properties of objects (Martin, 2007b). Future studies will be needed to resolve the role of corticolimbic structures in human fear generalization.

The overall lack of human behavioral research on SPC is somewhat surprising, given that influential conditioning-based models of anxiety have incorporated higher order fear learning processes to explain the etiology of several anxiety disorders (Bouton, Mineka, & Barlow, 2001; Davey, 1992; Mineka & Zinbarg, 2006; Rachman, 1977). Here, we observed an effect of trait anxiety on fear generalization within the conceptually similar condition, such that subjects with high anxiety expressed a greater increase in SCRs to the PS+ following fear conditioning. The ability to acquire and express conditional fears has long served as a model for clinical anxiety
disorders (Eysenck, 1976; Watson & Rayner, 1920), and prior studies have indicated that clinically anxious subjects express broad generalization of conditioned fear responses (Lissek, et al., 2010) and fail to limit fear behaviors to conditioned stimuli (reviewed in Lissek, et al., 2005). Trait anxiety could mediate higher forms of fear learning through several processes, such as expectancy biases (Chan & Lovibond, 1996) and heightened threat appraisal (Macleod & Cohen, 1993), which could lead to the detection of generalized danger signals that are only indirectly related to the feared US. Future studies should examine whether clinical anxiety disorders are associated with a broadening of SPC effects, even in cases where the pre-associated stimuli are less conceptually related.
6. Aversive learning modulates cortical representations of object concepts

This chapter presents empirical data from a functional imaging study investigating the role of conceptual knowledge in fear learning. The goal of this study is to bridge two literatures that had never been brought together before – fear conditioning and concept formation – to address the following question: How does an aversive encounter with one member of an object category lead to widespread fear of other category members? Such “concept-based generalization” processes are theoretically linked to the development of anxiety disorders. For instance, in Specific Phobia fears are expressed categorically (e.g., arachnophobia), and in PTSD flashbacks and somatic symptoms are triggered by a host of stimuli conceptually related to a trauma. Yet, no experimental model to date has been developed to show how cortical regions that store conceptual knowledge related to object categories are modified by emotional learning experiences and become incorporated into fear processing networks in the brain.

For this study, we collaborated with Dr. Alex Martin from the Laboratory of Brain and Cognition at the National Institutes of Health – an expert on the neural correlates of object processing. The design is based on a behavioral study that has been published as a Research Article in the journal *Biological Psychology*, (Dunsmoor, Martin, & LaBar, 2012). The present fMRI investigation has been submitted for publication.

6.1 Introduction

One way emotional experiences shape our lives is by changing how we represent and respond to objects we already know. For instance, an aversive encounter with a pit bull may lead to fears and avoidance behaviors that encompass dogs in general. In this way, emotional experiences with exemplars of a particular object category may rely on existing conceptual knowledge to draw inferences about potential affective qualities of related objects through the
process of inductive reasoning (Murphy, 2002). If an aversive experience with a category exemplar modulates representations of related object concepts, this experience-dependent change in representational architecture could provide a potential mechanism for category-based fear generalization. Such a mechanism could contribute to an understanding of Specific Phobias, for which fears are typically expressed categorically.

Pavlovian fear conditioning provides an experimental procedure particularly suited for investigating how an episode of fear learning promotes generalization to related object concepts. Fear conditioning studies in rodents provide evidence for neural mechanisms involved in experience-dependent modulations in stimulus representations. Specifically, sensory information concerning the conditioned stimulus (CS) and aversive unconditioned stimulus (US) converges in the basolateral amygdala, leading to strengthened CS-US associations, production of conditioned fear responses (CR) (Maren, 2001; Pape & Paré, 2010), and plasticity in unimodal sensory cortices (LeDoux, 2000; Letzkus et al., 2011; Pape & Paré, 2010; Weinberger, 2004). For instance, auditory fear conditioning leads to learning-induced retuning of neurons in auditory cortex towards the CS frequency, increased areas of representation for the CS frequency in auditory cortex, and behavioral fear generalization to similar tones (Weinberger, 2004, 2007).

Functional magnetic resonance imaging (fMRI) studies in humans confirm activity changes in the amygdala and sensory areas during conditioning with odors (Li, Howard, Parrish, & Gottfried, 2008), tones (Knight, et al., 2005), and visual images (Lim, Padmala, & Pessoa, 2008). Furthermore, perceptually-based fear generalization in humans occurs through changes in functional connectivity between the amygdala and visual cortical areas (Dunsmoor, Prince, et al., 2011). Taken together, these investigations suggest that fear conditioning modulates activity in and connectivity between the amygdala and cortical areas coding for the sensory qualities of the CS.
Although prior evidence provides a mechanism to explain how perceptually-based information in sensory cortices is modulated through fear conditioning, whether aversive experiences modulate neural representations and organization of conceptually-based information is unknown. Candidate neural regions include those representing information about object concepts, such as the semantic categories of animals and tools (Patterson, Nestor, & Rogers, 2007; Tranel, Damasio, & Damasio, 1997). Neuroimaging investigations reveal category selectivity for these superordinate category domains along posterior occipitotemporal cortex (Beauchamp, Lee, Haxby, & Martin, 2002; Chao & Martin, 2000; Ishai, Ungerleider, & Haxby, 2000; Simmons & Martin, 2011). Whereas images of animals reliably activate lateral fusiform gyrus (FFG), posterior superior temporal sulcus, and inferior occipital regions, images of tools activate medial FFG, posterior middle temporal gyrus, and middle occipital cortex (Martin, 2007b). The organization of category-selective regions may be related to intrinsic connectivity between unique brain networks (Mahon & Caramazza, 2011; Mahon et al., 2007). For instance, amygdala-lateral FFG connectivity has been proposed as a key network involved in the representation of animate objects (Martin, 2007a), whereas connectivity between motor cortex and medial FFG is important for representing graspable objects like tools (Mahon & Caramazza, 2011). In support for this hypothesis, separated patterns of functional connectivity during rest have been identified between regions implicated in social- versus tool-related conceptual tasks (Simmons & Martin, 2011). Finally, multivariate approaches including representational similarity analyses (RSA) provide an emerging tool for probing architecture of category representations, such as those distinguishing animate and inanimate objects in temporal cortex (Kriegeskorte, Mur, & Bandettini, 2008; Kriegeskorte et al., 2008). Thus, well-established separations in the neural organization of category knowledge provide a novel framework to investigate whether instances of fear learning to category-specific exemplars modify categorical representations in
occipitotemporal cortex and associated networks, thereby enabling individuals to express fear to a range of conceptually-related objects.

The present study developed a novel trial-unique conditioning paradigm using event-related functional magnetic resonance imaging (fMRI) to investigate whether aversive learning modulates representations of object concepts in the human brain. Taking advantage of well-established organizational separations in representations for animate and inanimate objects, we utilized animals and tools as conditioned stimuli. We predicted that aversive learning to category exemplars modifies representations in occipitotemporal cortex, as well areas implicated in simpler forms of conditioning, such as the amygdala. Moreover, we predicted that multivariate patterns of activity in object-selective cortex would reveal enhanced representational similarity among different category members, which may be integral to facilitate fear generalization between physically distinct objects. Consistent with an animacy hypothesis of fear learning, we also predicted a dissociation in domain-specific representational networks such that learning to fear a category of animate (as opposed to inanimate) objects selectively utilizes amygdala-cortical connectivity during learning, and is further reflected in resting state connectivity after learning (Simmons & Martin, 2011).

Finally, we sought to determine whether object typicality is used as a learning mechanism to generalize fear. Object categories contain graded structures that can be organized around members that are regarded as more typical than other members (Rosch & Mervis, 1975). Typicality effects play a well-known role in behavioral category-based induction (Murphy, 2002; Osherson, et al., 1990; Rips, 1975); however it is unknown what role typicality plays in fear learning. Although brain imaging investigations of typicality effects are scant, a candidate region involved in generalizing from this type of information is the hippocampus. The hippocampus evinces sparse responses to different object categories (Quiroga, Kreiman, Koch, & Fried, 2008).
and is implicated broadly in the generalization of learning (Gluck & Myers, 1993; Shohamy & Wagner, 2008). We therefore predicted that the hippocampus would signal object typicality early in training, when generalizing between different members from a feared category is made on the basis of their representativeness.

In this fMRI experiment, participants learned to fear different objects from a category domain while a set of objects from another category domain were used as an unpaired control condition (Fig. 14a). For half of the participants, a subset of images of animals was paired with delivery of an electrical shock US whereas all tools were presented without reinforcement (i.e., animal CS+, tool CS-). The other half of participants received the reverse contingencies (i.e., tool CS+, animal CS-). Thus, each participant viewed the same images, but individualized learning histories determined which category attained threat-relevance.

Another important feature of this task is that individual category exemplars were only presented once. This differs markedly from standard conditioning approaches, in that participants must generalize learning beyond each specific trial-unique CS-US instance to successfully acquire conditioned fear. Thus, participants were expected to rely on category-based inductive reasoning to deduce whether or not each CS presents a threat, rather than through repeated exposure.
Figure 14: Fear conditioning procedure and behavioral results.

(a) Subjects were presented with 80 unique exemplars of animals (40) and tools (40) for 6 s each. Exemplars were never repeated during the conditioning session. Half the objects from one category (CS+) co-terminated with a 6-ms electrical shock unconditioned stimulus (US), whereas objects from the other category were never reinforced (CS-). In this example, tools served as the CS+ and animals served as the CS- (category assignment was counterbalanced across subjects). (b) Expectancy ratings showed that subjects learned the CS-US contingencies. Skin conductance responses (SCRs) revealed differentially greater responses to the category exemplars predictive of shock versus the safe category exemplars. Neither shock expectancy ratings nor SCRs were influenced by which object category (animals or tools) predicted shock. Dashed line depicts chance level of US expectancy. A+ = images of animals predicted shock; T+ = images of tools predicted shock; A- = images of animals were safe; T- = images of tools were safe; µS = microSiemens; Error bars represent ± SEM; ** = p < .01, two-tailed t-tests.
6.2 Methods

6.2.1 Participants

Thirty-four right-handed healthy adults provided written informed consent in accordance with the Duke University Institutional Review Board guidelines. Data from one subject was excluded due to technical problems with the MRI scanner. The final analysis included 33 participants (16 females; age range = 18 to 37; median age = 23 yrs). Subjects were assigned either to the group in which animals predicted shock and tools were safe (A+/T-, n = 17) or to the group in which tools predicted shock and animals were safe (T+/A-, n = 16).

6.2.2 Stimulus set

Stimuli consisted of 80 images of tools (N = 40) and animals (N = 40) presented on a white background. Images were obtained from the website www.lifeonwhite.com and from publicly available resources on the internet. These stimuli were used in a previously published behavioral study on fear generalization (Dunsmoor, et al., 2012). Notably, unlike faces and houses (two categories frequently contrasted in human neuroimaging studies (e.g. Kanwisher, McDermott, & Chun, 1997)) general knowledge about animals and tools includes information about numerous visually distinct typical and atypical instances (Rosch, 1975). The exemplars were chosen on the basis of published category norms (e.g. Van Overschelde, Rawson, & Dunlosky, 2004) to ensure a range of highly typical (e.g. hammer and dog) and less typical (e.g., planer and armadillo) items. We also collected individual participant ratings of object typicality 24 hours following fear conditioning, which demonstrated a similar range of atypical and typical exemplars among both tools and animals. We avoided the use of highly threat-relevant images such as knives and snakes so as to mitigate the potential arousal bias evoked by these objects (Öhman & Mineka, 2001). The presentation of stimuli was controlled using Presentation Software (Neurobehavioral Systems, Albany, CA).
A 6-ms electrical shock applied to the right wrist served as the unconditioned stimulus. Shock administration was controlled using the STM-100 and STM-200 modules connected to the MP-150 BIOPAC system (BIOPAC systems, Goleta, CA). The level of the shock was calibrated for each subject individually prior to the start of the experiment using an ascending staircase procedure so as to reach a level deemed “annoying but not painful” [see Dunsmoor et al. (2009) for similar procedures]. Following calibration, subjects were asked to rate the intensity of the shock on a scale from 0 (“not at all unpleasant”) to 10 (“extremely unpleasant”). The average shock intensity rating was 5.56 (SD = 1.04), and there was no difference in shock intensity ratings across groups.

6.2.3 Task design and procedures

The scanning session contained four phases that occurred in the same order for each subject: animal-tool functional localizer, preconditioning resting state, fear conditioning, and post-conditioning resting state. The animal-tool localizer was performed prior to fear conditioning so as to isolate cortical regions selectively evoked by images of animals versus tools, and vice versa. The localizer consisted of 12 alternating blocks of animals, tools, and phase-scrambled images. Blocks consisted of 10 images presented for 750 ms, with each image followed by a 250-ms blank screen. Blocks were separated by an 11-s blank screen containing a fixation cross. The order of the blocks was counterbalanced across subjects. To ensure that subjects were attending to the images, one block of each condition contained a catch trial in which a hash symbol (#) was overlaid on an image. Subjects were asked to press a button on the joystick to indicate they had seen the hash symbol. Images presented during the localizer were not reused in any other experimental phase. Data from the functional localizer were entered into a general linear model using the events animals, tools, and phase-scrambled images (see fMRI analysis).
Prior to the start of fear conditioning, subjects were given unscanned practice trials with random objects (unrelated to the task stimulus set) to get accustomed to the use of the MRI compatible joystick and shock expectancy rating procedure. No shocks were delivered during the practice trials. The fear conditioning session was a slow event-related fMRI design conducted over four runs of equal length (6 min). Each run contained 20 trials (10 animals and 10 tools) presented in a pseudorandomized order such that no more than two objects from the same category occurred in a row (see Fig 14a). We used eight different stimulus presentation orders to counterbalance the presentation of animal and tool exemplars across participants. Each trial was 6 s in duration, during which time subjects used the MRI compatible joystick to rate expectancy from 0 (“sure the shock will not occur”) to 10 (“sure the shock will occur”) by moving a cursor along a rating bar that appeared below the CS. Expectancy was calculated as the final location of the rating bar at trial offset. A 10-12 s blank screen (average = 11 s) containing a white fixation cross on a blank background followed the offset of each trial. For each participant, one object category (e.g., animals) was designated the CS+, and 50% of exemplars from this category co-terminated with delivery of the shock US (partial reinforcement delay conditioning procedures). The other object category (e.g., tools) served as the CS-, and none of its exemplars were reinforced with a shock US. Category assignment was counterbalanced across subjects (animals CS+: n = 17; tools CS+: n =16). The choice of which CS+ items were paired (or unpaired) with shock was random and counterbalanced between subjects. Every trial during the object localizer and fear conditioning session contained a distinct basic-level exemplar with a different name that was never repeated (for example, there were not two different images of elephants). Subjects were not instructed on the CS-US contingencies and were not informed that images would be presented only once during learning.
Twenty-four hours after the fMRI fear conditioning session, subjects returned to a different laboratory setting outside the MRI facility where they conducted a follow-up recognition memory test for the CS+ and CS- items (not reported here) and rated each animal and tool picture for typicality to the superordinate category. Typicality ratings were made for each item on a 10-point scale (0 = highly atypical, 10 = highly typical). Typicality ratings were Z-scored and these values were used as a parametric regressor for the analysis of brain regions modulated by object typicality.

6.2.4 Psychophysiological methods and analysis

Psychophysiological recording was controlled with the MP-150 BIOPAC system (BIOPAC systems, Goleta, CA) and grounded through the resonance frequency filter panel and shielded from magnetic interference. MRI-compatible SCR electrodes were placed on the hypothenar eminence of the palmar surface of the left hand. SCR analysis was carried out using AcqKnowledge software (BIOPAC systems) using procedures previously described (Dunsmoor, et al., 2012). An SCR was scored as a response if the trough-to-peak response occurred 1-4 s following stimulus onset, lasted between 0.5 and 5.0 s, and was greater than 0.02 microsiemens. A trial that did not meet these criteria was scored as a zero. SCRs were square-root transformed for normalization prior to statistical analysis. Three subjects were excluded from the SCR analysis due to an overall lack of measurable SCR data throughout the entire conditioning session. Statistical analysis of the behavioral data was assessed by repeated-measures ANOVA, with CS type (CS+, CS-) and group category (animals CS+/tool CS-, tools CS+/animals CS+) as factors, as well as post-hoc t-tests, considered significant at $P < 0.05$ (SPSS 15.0, Chicago, IL). Tonic skin conductance levels (SCL) were acquired throughout the pre- and post-conditioning resting state scans. SCLs were averaged over each of these two resting state time courses (6 min
each) and square-root normalized. Subjects’ mean baseline SCL was subtracted from their post-conditioning SCL to assess changes in tonic arousal from pre-to-post fear conditioning.

**6.2.5 Functional image acquisition, preprocessing, and GLM analysis**

Whole brain functional imaging was conducted on a General Electric Signa EXCITE HD 3.0 Tesla MRI scanner. Blood oxygenation level-dependent functional images were acquired parallel to the AC-PC line using a SENSE™ spiral in sequence: acquisition matrix, 64 x 64; field of view, 256 x 256; flip-angle, 60°; 34 slices with interleaved acquisition; slice thickness, 3.8 mm with no gaps between slices; in-plane resolution = 3.75 mm x 3.75 mm; repetition time, 2 s; echo time, 27 ms.

Preprocessing was conducted using SPM8 (Wellcome Trust Center, www.fil.ion.ucl.ac.uk) implemented in MATLAB (The Mathworks Inc, Natick MA). Images were spatially normalized into Montreal Neurological Institute (MNI) space, voxel size was resampled to 2 mm x 2 mm x 2 mm, and smoothed using an isotropic 8-mm³ Gaussian full width half maximum kernel. Functional images were co-registered to each subject’s high resolution T1-weighted structural scan. The first four functional images were removed from each scanning run to account for magnetic equilibration, and the remaining images were corrected for head motion using a center-of-mass movement threshold of 3 mm. Padding was inserted around the subject’s head to minimize head motion.

Statistical analysis of the preprocessed data was conducted using the general linear model in SPM8 with trial regressors for the onset and duration of each event convolved with the canonical hemodynamic response function. Head motion parameters and the shock US were included as covariates of no interest. A high-pass filter of 128 s was applied to account for low-frequency drifts. Statistical thresholds for whole-brain analyses were set at $P < .001$, with a minimum extent threshold of 60 voxels. This spatial extent for multiple comparisons provides a
cluster correction of $P < .05$, as derived using the REST AlphaSim utility (www.restfmri.net; toolkit V 1.3), which computes the alpha level using 1,000 Monte Carlo simulations to verify that activations of this cluster size were unlikely to have occurred due to chance. Based on extensive literature on the role of the amygdala in fear conditioning (Phelps & LeDoux, 2005) and the hippocampus on learning and generalization (Shohamy & Wagner, 2008), we included these as a priori regions of interest for small volume correction analysis. Thresholds for small volume correction analyses were conducted using bilateral anatomical masks from the Wake Forest PickAtlas toolbox (Maldjian, et al., 2003) and a threshold of FWE $P < .05$ with 10 contiguous voxels. Brain activations are displayed on the average anatomical image from 33 subjects. Labels for the anatomical regions provided by Talairach Client (Lancaster et al., 2000) based on peak coordinates in MNI space, converted to Talairach space using GingerALE 2.0 (Laird et al., 2010).

6.2.5.1 Independent animal-tool localizer

The animal-tool localizer task included separate regressors for blocks of animals, tools, and phase-scrambled objects. Single subject contrast images for each event versus baseline were taken to group-level random-effects analysis examining the contrast of animals versus tools (and tools versus animals). Activations resulting from the animal versus tool contrasts were inclusively masked by intact objects (animals + tools) > phase scrambled images at $P < 0.05$, uncorrected (see Table 2). A 4-mm radius sphere was drawn around the peak voxels from regions identified from the object localizer as selective to animals or tools. These ROIs were used in subsequent analyses.

6.2.5.2 Aversive learning

To examine effects of aversive learning in category-selective regions, mean beta-parameters were extracted from the six ROIs identified from the independent object localizer. Statistical tests conducted on extracted parameter estimates included a repeated-measures
ANOVA with CS type (CS+, CS-) and group (animals CS+/tool CS-, tools CS+/animals CS-) as factors, considered significant at \( P < .05 \). Whole-brain group-level fMRI analysis of aversive learning was analyzed using a 2 x 2 ANOVA with the first-level estimates of CS+ versus baseline and CS- versus baseline. The factors included CS type (CS+, CS-) and group (animals CS+/tools CS-, tools CS+/animals CS+). Main effects of CS type were assessed through a conjunction analysis of the contrast CS+ versus CS- for both groups.

6.2.5.3 **Representational similarity analysis**

To assess how aversive learning modulates the neural representation of animals and tools, we performed a representational similarity analysis (Kriegeskorte, Mur, & Bandettini, 2008). This multivariate approach examines the similarity of neural response patterns between different combinations of stimuli. This approach has been used to show that patterns of responses in inferotemporal cortex are more similar between pairs of objects from within the same category (e.g. different animate objects) than between objects from different categories (e.g. between animate and inanimate objects). To conduct this analysis, we first constructed a general linear model with each regressor containing a single trial. Preprocessing did not include spatial smoothing in order to retain information at fine spatial scales. The whole-brain model produced a single estimate of neural activity for every object. We performed separate analyses, sampling activity from the object localizer mask (animals + tools > scrambled objects) and bilateral anatomical amygdala ROIs from the AAL atlas, yielding an 80 object by 4599 (object localizer) or 468 (amygdala) voxel activity pattern for each subject. Representational dissimilarity matrices (RDM) were computed from these patterns by calculating 1-r (Pearson correlation coefficient) across all object selective voxels for every possible pairing of objects, producing a symmetric 80 by 80 RDM.
As the patterns of activity for a given stimulus are used to compute the similarity to multiple other objects within (and between) categories, conventional comparisons between correlation coefficients (which assumes independence of samples) could not be performed. Therefore we used bootstrap resampling (using the `bootstrap` function in MATLAB with 10,000 iterations) to estimate the mean similarity patterns within and between categories for each subject (see also Kriegeskorte (2008)). By computing a bootstrap statistic for each category, we were able to examine whether within-category dissimilarity for CS+ items was different from within-category dissimilarity for CS- items and between-category dissimilarities for each group using two-sample t-tests.

6.2.5.4 Psychophysiological interaction (PPI) analysis

We conducted a PPI analysis (Friston et al., 1997) implemented in SPM 8 to examine patterns of functional connectivity during the experimental task. The purpose of the PPI analysis is to demonstrate functional coupling across the brain with a source region and an experimental parameter. We used this analysis to examine whether task-based connectivity between the amygdala and category-selective regions was selectively enhanced in conditions of interest (CS+ versus CS-). The source region was provided by the amygdala ROI identified from the main effect of fear conditioning (CS+ > CS-). The representative time course was extracted from voxels in the amygdala mask - which included voxels in both left and right amygdala - using the first eigenvariate calculated from singular value decomposition. The time course from the source region, the psychological context (CS+ > CS-), and the interaction term between the time series and psychological context (PPI) were included in a general linear model. The PPI analysis reveals brain regions exhibiting stronger functional coupling with the amygdala during aversive learning, while accounting for the main effect of conditions by including the psychological and physiological time course into the GLM. We targeted category-selective ROIs identified in the
independent localizer for analysis of the interaction term to assess connectivity with the amygdala during aversive learning across learning groups.

6.2.5.5 Resting state connectivity

Immediately before and immediately after fear conditioning, subjects were instructed to lay still with their eyes open during a 6-min resting state scan. One subject from the A+/T- group did not complete the post-conditioning resting state due to time constraints and is thus not included in the resting state analysis. We used a targeted approach to examine correlations for pairs of *a priori* ROIs before and after aversive learning. Time courses were extracted from the amygdala (identified from the fear conditioning analysis contrast CS+ > CS-) and category-selective regions that dissociated animals from tools during the independent localizer using methods described above for the PPI analysis. For each subject, the time course from these regions was fit using a general linear model incorporating head motion and the mean signal over the course of the run. A 128-s filter was applied to remove low-frequency drift. Pearson correlation coefficients were calculated for each subject during baseline (pre conditioning) and post conditioning rest scans for amygdala-category ROI pairs. Correlation coefficients were Fisher Z-transformed and baseline values were subtracted from the post conditioning values to gain a measure of the change in correlations between regions from before to after fear conditioning. One-sample t-tests were used to determine significant differences in the pre-to-post change in connectivity.

6.2.5.6 Parametric modulation analysis of object typicality

To examine whether any brain regions tracked object typicality during fear conditioning, we used typicality ratings (obtained for each subject 24 hrs after the imaging session) for each CS+ image as a trial-by-trial parametric regressor. Typicality ratings were Z-scored and used in the parametric modulation analysis. Three subjects were not included in this analysis due to a lack
of variability in typicality ratings (SD of raw typicality scores < 2). First-level model estimation for the parametric modulation analysis was conducted using the GLM, with the events CS+ and its parametric modulation, CS- and its parametric modulation, head motion parameters, and US. Group level analysis included ANOVA with the parametric modulation of the CS+ across the four scanning runs as factors. The analysis focused on regions showing a linear decrease in CS+ modulation over the four runs of fear conditioning [contrast weights; 2 1 -1 -2].

Lastly, we interrogated the role of the hippocampus further by incorporating the hippocampal region identified from the parametric modulation analysis as a seed region in a PPI analysis using procedures described above. The purpose of this analysis was to examine regions exhibiting task-based connectivity with the hippocampus during early aversive learning (when the hippocampus exhibited typicality effects) but not late aversive learning (when the typicality effect subsided). The interaction term between the hippocampus time series and experimental parameter (CS+ versus CS-) was entered into a group level model that included run (1-4) as a factor. We examined whether any regions exhibited a linear decrease in connectivity across the entire conditioning session [contrast weight: 2 1 -1 -2].

6.3 Results

6.3.1 Behavioral results

Shock expectancy (Fig. 14b) was greater to the CS+ than the CS-, $F_{1, 31} = 334.54, P < 0.001$. There was no effect of group ($P = .24$) and no interaction between CS type and group, $P = .39$. Note that expectancy on CS+ trials did not reach ceiling, which is likely due to the use of a partial reinforcement schedule. In this way, subjects developed knowledge that a given superordinate category posed a threat, but they could not be certain that the shock would occur on a given trial. Analysis of SCRs (Fig. 4c) likewise revealed a main effect of CS type, $F_{1, 28} = 33.66, P < 0.001$, but no effect of group ($P = .56$) and no interaction with group, $P = .89$. There
was no correlation between SCRs and shock expectancy. Finally, tonic skin conductance levels (SCL) were acquired during 6 min resting state scans immediately before and after fear conditioning. SCLs significantly increased from before to after conditioning ($P < .001$) with no difference between groups ($P > .5$), providing further evidence that the aversive learning episode had a lasting impact on emotional arousal. Together, these results demonstrate that subjective and autonomic contingency learning differentiated as a function of CS type, but were unaffected by which category (animals or tools) was reinforced, replicating our behavioral findings using this novel paradigm (Dunsmoor, et al., 2012)

6.3.2 Brain imaging results

6.3.2.1 Animal-tool functional localizer

Prior to fear conditioning, object-selective regions along occipital-temporal cortex were localized by contrasting images of animals, tools, and phase-scrambled objects. Whole-brain analysis revealed two clusters of activation that were selective to images of animals: the right lateral fusiform gyrus (FFG) and right inferior occipital gyrus extending into the posterior superior temporal sulcus (pSTS) (See Table 2). Four clusters of activation were selective to images of tools: bilateral medial FFG/parahippocampal gyrus, and bilateral middle occipital gyrus. These regions are consistent with prior fMRI studies examining category-specificity for animate versus manmade objects (Martin, 2007b). Based on these localized activations from independent \textit{a priori} occipital-temporal regions, we created regions of interest (ROI) for use in further analyses.
Table 2: Animal-tool localizer

Regions identified from the preconditioning object localizer as preferential for images of animals or preferential for images of tools. Identified at $P < .001$, cluster corrected $P < .05$.

<table>
<thead>
<tr>
<th>Region</th>
<th>Hemisphere</th>
<th>MNI coordinates</th>
<th>Size (voxels)</th>
<th>Peak T</th>
<th>Peak Z</th>
</tr>
</thead>
<tbody>
<tr>
<td></td>
<td></td>
<td>$x$  $y$  $z$</td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>Animals greater than tools</td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>Inferior occipital gyrus</td>
<td>Right</td>
<td>46  -78  -2</td>
<td>1052</td>
<td>7.84</td>
<td>6.87</td>
</tr>
<tr>
<td>Middle temporal gyrus</td>
<td>Right</td>
<td>42  -64  8</td>
<td>4.52</td>
<td>4.29</td>
<td></td>
</tr>
<tr>
<td>Fusiform gyrus</td>
<td>Right</td>
<td>36  -52  -20</td>
<td>143</td>
<td>6.28</td>
<td>5.73</td>
</tr>
<tr>
<td>Tools greater than animals</td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>Middle occipital gyrus</td>
<td>Left</td>
<td>-28  -78  22</td>
<td>589</td>
<td>5.56</td>
<td>5.16</td>
</tr>
<tr>
<td>Fusiform gyrus</td>
<td>Left</td>
<td>-24  -46  -16</td>
<td>452</td>
<td>5.42</td>
<td>5.05</td>
</tr>
<tr>
<td>Middle temporal gyrus</td>
<td>Left</td>
<td>-44  -58  -4</td>
<td>4.49</td>
<td>4.26</td>
<td></td>
</tr>
<tr>
<td>Middle temporal gyrus</td>
<td>Left</td>
<td>-36  -54  -2</td>
<td>4.21</td>
<td>4.02</td>
<td></td>
</tr>
<tr>
<td>Fusiform gyrus</td>
<td>Right</td>
<td>28  -48  -12</td>
<td>70</td>
<td>4.62</td>
<td>4.38</td>
</tr>
<tr>
<td>Middle occipital gyrus</td>
<td>Right</td>
<td>28  -80  18</td>
<td>226</td>
<td>4.52</td>
<td>4.29</td>
</tr>
</tbody>
</table>
6.3.2.2 Aversive learning modulates activity in posterior category-selective cortical regions

A primary question in this investigation was whether aversive learning enhances activity in category-selective cortex as a function of which category (animals or tools) was threatening or safe. To investigate this question, we probed neural activity during learning using the functional ROIs defined by the independent object localizer. As expected, analysis of the mean parameter estimates extracted from these ROIs revealed interactions between CS type and group, with the sole exception of the tool-selective region of the right middle occipital gyrus (Fig 15bc and Table 3). Post-hoc t-tests showed that these interactions were driven by greater activity to images from the CS+ (versus the CS-) category within the ROI preferential to the CS+ category. In the tool-selective ROIs, (Fig 15c), greater activity to the CS+ versus the CS- was only observed in the group fear conditioned to tools. Within animal-selective ROIs, greater activity to CS+ than CS- was observed in the group fear conditioned to animals. Moreover, activity to animal CS+’s in this group was greater than activity evoked by both animal CS-’s and tool CS+’s in the opposite group (Fig 15b). Altogether, these findings provide evidence that fear learning enhanced activity within category-selective cortex as a function of the reinforcement contingency.
Table 3: Statistical analysis of aversive learning-related activity extracted from category-selective ROIs

Note: statistical tests were conducted on the mean parameter estimates from category-selective regions identified using an independent localizer prior to fear conditioning. See Figure 15 for brain imaging results. FFG = fusiform gyrus, L = left, R = right.

<table>
<thead>
<tr>
<th>Brain Region</th>
<th>ANOVA</th>
<th>Post-hoc t-tests</th>
</tr>
</thead>
<tbody>
<tr>
<td></td>
<td>CS by category interaction</td>
<td>Main effect CS</td>
</tr>
<tr>
<td>R Inferior occipital gyrus</td>
<td>*P &lt; .001</td>
<td>P = .3</td>
</tr>
<tr>
<td>R Lateral FFG</td>
<td>*P &lt; .001</td>
<td>*P = .04</td>
</tr>
</tbody>
</table>

ROIs identified as selective to animals

<table>
<thead>
<tr>
<th>Brain Region</th>
<th>ANOVA</th>
<th>Post-hoc t-tests</th>
</tr>
</thead>
<tbody>
<tr>
<td></td>
<td>CS by category interaction</td>
<td>Main effect CS</td>
</tr>
<tr>
<td>L Middle occipital gyrus</td>
<td>*P &lt; .001</td>
<td>*P = .01</td>
</tr>
<tr>
<td>L medial FFG</td>
<td>*P &lt; .001</td>
<td>P = .45</td>
</tr>
<tr>
<td>R medial FFG</td>
<td>*P &lt; .001</td>
<td>*P = .005</td>
</tr>
<tr>
<td>R middle occipital gyrus</td>
<td>P = .10</td>
<td>*P &lt; .001</td>
</tr>
</tbody>
</table>
Category-selective regions were independently identified prior to aversive learning through a functional localizer. The right lateral fusiform gyrus and right inferior occipital gyrus/posterior superior temporal sulcus showed selectively to images of animals, whereas bilateral medial fusiform and bilateral middle occipital gyrus showed selectively to images of tools. These category-selective regions of interest were then used to interrogate neural activity during aversive learning. (b) Within animal-selective regions, activity related to aversive learning was enhanced for subjects who learned to fear animals versus subjects who learned to fear tools. (c) Within tool-selective regions, activity related to aversive learning was enhanced for subjects who learned to fear tools versus subjects who learned to fear animals. pSTS = posterior superior temporal sulcus; ROIs = regions of interest; A+ = images of animals predicted shock; T+ = images of tools predicted shock; A- = images of animals were safe; T- = images of tools were safe; Error bars represent ± SEM; * = P < .05 and ** = P < .01, two-tailed t-tests. Activations displayed on subject averaged anatomical image at P < .001, cluster corrected P < .05.
6.3.2.3 Effects of aversive learning on representational similarity in object-selective cortex

In addition to the univariate fMRI analysis reviewed above, we examined activity in object selective cortex using a multivariate representational similarity analysis (Kriegeskorte, Mur, & Bandettini, 2008; Kriegeskorte, Mur, Ruff, et al., 2008). In this approach, the response patterns in a set of voxels are extracted and their similarity structure is compared across stimulus categories or experimental manipulations. For this analysis, we examined patterns of neural activity within bilateral occipital-temporal cortex identified as preferentially signaling intact objects (conjunction of animals and tools > phase scrambled objects) during the object localizer. This analysis showed a clear delineation between animate and inanimate objects (Fig 16a), in line with previous findings using this method (Kriegeskorte, Mur, Ruff, et al., 2008).
Figure 16: Representational similarity analysis (RSA)

(a) The similarity of activity patterns across voxels in object-selective occipitotemporal cortex reflects a category boundary between animals or tools during aversive learning in groups fear conditioned to animals or tools, respectively. Results are presented in a representational dissimilarity matrix (RDM), which illustrates the correlation between activity patterns evoked by two different stimuli as ‘1-r’ (Pearson correlation coefficient) (Kriegeskorte, Mur, Ruff, et al., 2008). The images used in this analysis are arranged by alphabetical order within category, with each cell depicting the similarity in response patterns between two stimuli. For display, the dissimilarity metric is presented in percentiles. (b) The average dissimilarity amongst images of animals, tools, and between animals and tools, was quantified within each subject group using bootstrap resampling. The dissimilarity amongst objects from the CS+ category was reduced for each group relative to objects from within the CS- category. Error bars represent ± SEM; ** = P < .01, two-tailed t-tests.
Importantly, by quantifying these similarity structures we are able to show how aversive learning alters these representational structures separately for each subject group as a function of which object category (animals or tools) was reinforced. We used a bootstrap resampling approach to quantify the mean dissimilarity of multivariate patterns from within the CS+ and CS- category for each subject group. This included within-category comparisons (i.e., animals-to-animals, and tools-to-tools) and the between category comparison (animals-to-tools) computed separately for the two groups (Fig 16b). Within the animal CS+ group, paired t-tests showed that the representational structure for animals-to-animals was more similar than for tools-to-tools (CS-) or animals-to-tools ($P < .01$). In contrast, the group fear conditioned to tools exhibited more similarity among tool-to-tool exemplars (CS+) than animal-to-animal (CS-) and animal-to-tool exemplars ($P < .01$). A similar increase in within-category similarity for CS+ versus CS- items was observed in the amygdala (Fig. 17). These results indicate that aversive learning enhances the representational similarity among categorically-related exemplars, which may be integral to facilitate generalization between physically distinct objects that signal a salient outcome.
136

Figure 17: Representational similarity analysis for the amygdala

(a) The similarity of activity patterns across voxels in the anatomical mask for the amygdala (defined by the AAL atlas) reflects a category boundary for objects from within the CS+ category. That is, dissimilarity amongst different images of animals was reduced for subjects fear conditioned to animals (CS+), whereas the opposite pattern was seen for images of tools for subjects fear conditioned to tools. For display, the dissimilarity metric is presented in percentiles. The images in the representational dissimilarity matrices are arranged by alphabetical order within each category, with each cell depicting the similarity in response patterns between two stimuli (see Fig 4 for more detail on the particular images). (b) The dissimilarity amongst objects from the CS+ category was quantified using bootstrap resampling in order to statistically evaluate the relationship between dissimilarity amongst categories (animals-to-animals; tools-to-tools; animals-to-tools) for each learning group. Error bars represent ± SEM; ** = P < .01, two-tailed t-tests

6.3.2.4 Role of the medial temporal lobe and other brain regions

Group-level fMRI analyses consisted of a 2 x 2 ANOVA with the factors CS type and group category. Given a strong a priori interest in the role of the amygdala in fear conditioning (LaBar, Gatenby, Gore, LeDoux, & Phelps, 1998; Phelps & LeDoux, 2005), and the role of the hippocampus in generalization of learning (Gluck & Myers, 1993; Kumaran, Summerfield, Hassabis, & Maguire, 2009; Shohamy & Wagner, 2008), we conducted an ROI analysis for these areas using a small volume correction with a family-wise error of P < .05 on the contrast for the main effect of CS type (CS+ > CS-). A main effect of CS type (CS+ > CS-) was observed bilaterally in the amygdala (Fig 18a). Learning-related effects were significant in both groups, and there was no effect of group and no interaction between CS type and group (Ps > .1). As the amygdala’s response profile typically habituates over the course of traditional fear conditioning
procedures (Buchel, Morris, Dolan, & Friston, 1998), we further interrogated its activity from voxels identified from the ROI analysis across the four scanning runs. Learning-related amygdala activity showed a steady decrease over time as responses to the CS+ diminished (Fig 18b), implicating a time-sensitive role in the initial acquisition of conditioned fear. A main effect of CS type (CS+ > CS-) was also observed in bilateral hippocampus, and there was no effect of group and no interaction between CS type and group (Ps > .1). Similar to the amygdala, learning-related activity in the hippocampus showed a decrease across conditioning runs due to diminished responses to the CS+. These findings indicate that medial temporal lobe regions were engaged in learning to fear a category of objects that predicted shock but that differences in CS+ versus CS- activations were strongest in early learning, consistent with the existing literature using more traditional fear conditioning procedures.

Whole-brain univariate analyses complemented the ROI approach to reveal other regions exhibiting effects of generalized aversive learning. Enhanced activity to the CS+ versus CS- was observed in bilateral insula, inferior frontal gyrus, dorsolateral prefrontal cortex, thalamus, and anterior cingulate cortex (ACC) (Fig. 18c and Table 4). The reverse contrast (CS- > CS+), highlighting areas preferentially sensitive to safety signals, revealed activations in ventromedial prefrontal cortex and posterior cingulate cortex. In all, the collection of areas identified as showing a main effect of aversive learning have been commonly identified in traditional human neuroimaging studies that employ a single CS+ and CS- exemplar (Delgado, Li, et al., 2008; Etkin, Egner, & Kalisch, 2011; LaBar & Cabeza, 2006; Phelps & LeDoux, 2005; Sehlmeyer, et al., 2009). The present results thus extend the role of these structures to learning about fear associations to trial-unique exemplars that span a superordinate category.
Table 4: Activation during concept-based fear learning

Regions exhibiting greater activity on reinforced (CS+) versus unreinforced (CS-) trials. Identified at $P < .001$ (cluster corrected $P < .05$) during fear conditioning across all subjects ($n = 33$) * $P < .05$, small volume corrected.

<table>
<thead>
<tr>
<th>Region</th>
<th>Hemisphere</th>
<th>MNI coordinates</th>
<th>Size (voxels)</th>
<th>Peak T</th>
<th>Peak Z</th>
</tr>
</thead>
<tbody>
<tr>
<td></td>
<td></td>
<td>$x$</td>
<td>$y$</td>
<td>$z$</td>
<td></td>
</tr>
<tr>
<td>CS+ &gt; CS-</td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>Insula</td>
<td>Right</td>
<td>32</td>
<td>20</td>
<td>-6</td>
<td>8253</td>
</tr>
<tr>
<td>Midbrain</td>
<td>Mid</td>
<td>4</td>
<td>-26</td>
<td>-6</td>
<td></td>
</tr>
<tr>
<td>Inferior frontal gyrus</td>
<td>Left</td>
<td>-34</td>
<td>20</td>
<td>-2</td>
<td></td>
</tr>
<tr>
<td>Medial frontal gyrus</td>
<td>Mid</td>
<td>2</td>
<td>20</td>
<td>44</td>
<td>4568</td>
</tr>
<tr>
<td>Cingulate gyrus</td>
<td>Mid</td>
<td>4</td>
<td>26</td>
<td>30</td>
<td></td>
</tr>
<tr>
<td>Precordentral gyrus</td>
<td>Right</td>
<td>36</td>
<td>2</td>
<td>32</td>
<td></td>
</tr>
<tr>
<td>Precuneus</td>
<td>Right</td>
<td>18</td>
<td>-66</td>
<td>30</td>
<td>1438</td>
</tr>
<tr>
<td>Cuneus</td>
<td>Right</td>
<td>12</td>
<td>-64</td>
<td>38</td>
<td></td>
</tr>
<tr>
<td>Precuneus</td>
<td>Right</td>
<td>4</td>
<td>-62</td>
<td>62</td>
<td></td>
</tr>
<tr>
<td>Posterior cingulate</td>
<td>Mid</td>
<td>0</td>
<td>-30</td>
<td>20</td>
<td>384</td>
</tr>
<tr>
<td>Angular gyrus</td>
<td>Left</td>
<td>-28</td>
<td>-56</td>
<td>44</td>
<td>1187</td>
</tr>
<tr>
<td>Precuneus</td>
<td>Left</td>
<td>-14</td>
<td>-66</td>
<td>36</td>
<td></td>
</tr>
<tr>
<td>Precuneus</td>
<td>Left</td>
<td>-16</td>
<td>-62</td>
<td>48</td>
<td></td>
</tr>
<tr>
<td>Region</td>
<td>Side</td>
<td>X</td>
<td>Y</td>
<td>Z</td>
<td>T</td>
</tr>
<tr>
<td>-------------------------</td>
<td>-------</td>
<td>------</td>
<td>------</td>
<td>------</td>
<td>-----</td>
</tr>
<tr>
<td>Cerebellum</td>
<td>Left</td>
<td>18</td>
<td>-72</td>
<td>-10</td>
<td>328</td>
</tr>
<tr>
<td>Middle frontal gyrus</td>
<td>Left</td>
<td>-40</td>
<td>42</td>
<td>12</td>
<td>335</td>
</tr>
<tr>
<td>Middle frontal gyrus</td>
<td>Left</td>
<td>-40</td>
<td>48</td>
<td>6</td>
<td></td>
</tr>
<tr>
<td>Amygdala*</td>
<td>Left</td>
<td>-20</td>
<td>0</td>
<td>-12</td>
<td>58</td>
</tr>
<tr>
<td>Amygdala*</td>
<td>Right</td>
<td>16</td>
<td>-2</td>
<td>-12</td>
<td>13</td>
</tr>
<tr>
<td>Hippocampus*</td>
<td>Right</td>
<td>20</td>
<td>-30</td>
<td>-4</td>
<td>22</td>
</tr>
<tr>
<td>Hippocampus*</td>
<td>Left</td>
<td>-32</td>
<td>-12</td>
<td>-12</td>
<td>11</td>
</tr>
</tbody>
</table>

**CS- > CS+**

<table>
<thead>
<tr>
<th>Region</th>
<th>Side</th>
<th>X</th>
<th>Y</th>
<th>Z</th>
<th>T</th>
<th>Z</th>
</tr>
</thead>
<tbody>
<tr>
<td>Anterior cingulate</td>
<td>Mid</td>
<td>0</td>
<td>52</td>
<td>-14</td>
<td>685</td>
<td>4.70</td>
</tr>
<tr>
<td>Medial frontal gyrus</td>
<td>Mid</td>
<td>-8</td>
<td>42</td>
<td>-18</td>
<td></td>
<td>4.66</td>
</tr>
<tr>
<td>Medial frontal gyrus</td>
<td>Mid</td>
<td>-8</td>
<td>60</td>
<td>8</td>
<td></td>
<td>4.43</td>
</tr>
<tr>
<td>Precuneus</td>
<td>Mid</td>
<td>-6</td>
<td>-48</td>
<td>32</td>
<td>218</td>
<td>4.63</td>
</tr>
<tr>
<td>Middle frontal gyrus</td>
<td>Left</td>
<td>-20</td>
<td>26</td>
<td>38</td>
<td>135</td>
<td>4.26</td>
</tr>
<tr>
<td>Middle frontal gyrus</td>
<td>Left</td>
<td>-16</td>
<td>34</td>
<td>40</td>
<td></td>
<td>3.60</td>
</tr>
<tr>
<td>Middle temporal gyrus</td>
<td>Left</td>
<td>-42</td>
<td>-72</td>
<td>36</td>
<td>67</td>
<td>4.05</td>
</tr>
</tbody>
</table>
Figure 18: Amygdala and whole-brain activity during aversive learning

(a) Activity in the amygdala was identified from a small volume corrected region of interest (family-wise error < .05). Differential fear-conditioned activity as a function of stimulus type (CS+, CS-) was observed in bilateral amygdala in subjects who learned to fear animals and in subjects who learned to fear tools. (b) In both left and right amygdala, differential fear-conditioned activity diminished over time as responses to the CS+ habituated. (c) Whole-brain analysis (cluster corrected < .05) revealed greater activity to the CS+ versus CS- in bilateral insula, anterior cingulate cortex, and dorsolateral PFC, whereas the CS- elicited greater activity in the ventromedial PFC (see Table 4 for full list of regions).
6.3.2.5 Testing the animacy hypothesis: Learning to fear animals is characterized by enhanced functional connectivity between the amygdala and lateral fusiform gyrus

It has been proposed that patterns of intrinsic connectivity between the amygdala and lateral occipital-temporal regions provide a key network for processing animate objects (Mahon & Caramazza, 2011). We therefore examined functional connectivity between the amygdala and category-selective ROIs during aversive learning using task-based connectivity measures and during rest both before and following aversive learning. Task-based connectivity was examined using a psychophysiological interaction (PPI) analysis (Friston, et al., 1997), with the amygdala as the source region and category-selective ROIs as targets. The PPI analysis revealed significant connectivity between the amygdala and lateral FFG in those subjects fear conditioned to animals ($t_{16}=2.52$, $P=.02$) (Fig 19a). Amygdala-lateral FFG connectivity was enhanced in the group fear conditioned to animals relative to the group fear conditioned to tools ($t_{15}=2.27$, $P=.03$), whereas the latter group did not exhibit either positive or negative coupling between these regions ($t_{15}=-.96$, $P=.35$). In contrast, connectivity with the amygdala was not evidenced with tool-selective medial FFG, or other tool-selective ROIs, in either learning group. This finding suggests that learning to fear animals selectively tuned task-based connectivity between the amygdala and animal-selective lateral FFG.

In order to examine whether aversive learning selectively modulates resting-state connectivity in domain-specific networks (Simmons & Martin, 2011), resting state scans were acquired immediately before and after aversive learning. We focused this analysis on the difference in resting state correlations between the amygdala and category-selective cortex from pre- to post-learning across groups. Functional connectivity between the amygdala and lateral FFG was selectively enhanced pre-to-post conditioning in the group fear conditioned to animals ($t$
$t_{15} = 3.87, P = .001$) but not for the group fear conditioned to tools ($t_{15} = .67, P = .51$) (Fig 19b).

For the group fear conditioned to animals, this significant increase in amygdala-lateral FFG connectivity was positively correlated with the change in tonic SCLs ($r_{12} = .54, P = .04$) (Fig 19c), suggesting a relationship between increases in domain-specific resting state networks and peripheral measures of arousal following an emotional episode. In contrast, neither group showed enhancement in connectivity between the amygdala and other category-selective ROIs. In all, both task-based and resting-state functional connectivity analyses indicate selective functional enhancement of the amygdala’s connectivity with animal-selective processing regions, which supports a special mechanism for the acquisition of fear to animate objects as postulated by the animacy hypothesis.

Figure 19: Amygdala-lateral FFG functional connectivity at learning and at rest.

(a) Results from the psychophysiological interaction (PPI) analysis show that enhanced amygdala-lateral fusiform gyrus (animal-selective ROI) connectivity during fear conditioning was selective to the group fear conditioned to animals. (b) During 6-min rest scans, connectivity between the amygdala and lateral fusiform increased significantly from before to after fear conditioning in the group fear conditioned to animals. (c) In the group conditioned to animals, increases in amygdala-lateral FFG connectivity were positively correlated with resting-state increases in tonic skin conductance levels from pre- to post-conditioning. Regions were defined on the basis of functional activity during learning (amygdala; CS+ > CS-) and from an independent category localizer (fusiform gyrus; see Methods). A+ = images of animals predicted shock; T+ = images of tools predicted shock; A- = images of animals were safe; T- = images of tools were safe; Error bars represent ± SEM; * = $P < .05$ and ** = $P < .01$, two-tailed t-tests.
6.3.2.6 Hippocampal signaling of object typicality contributes to category-based fear learning early in training through interactions with the amygdala

To test our prediction that typicality effects apply to early training trials, we used individual subjects’ typicality ratings in a trial-by-trial parametric modulation analysis to examine whether brain activity was modulated by the typicality of feared category exemplars. In line with our prediction, a linear decrease in modulation by object typicality over the course of the four learning runs revealed selective activation in the left hippocampus during early but not late training trials (Fig 20a). Based on prior research showing generalization of learning is supported by interactions between the hippocampus and midbrain (Shohamy & Wagner, 2008), and hippocampus and ventromedial prefrontal cortex (Kumaran, et al., 2009), we incorporated the hippocampus identified from the typicality analysis as a seed region in a PPI analysis to interrogate connectivity with other conditioning-related regions during early training. Results revealed a temporally-graded pattern of connectivity with the left amygdala (Fig 20b) such that task-based connectivity between the hippocampus and amygdala was strong during early training trials but then dissipated. This spatiotemporal pattern implicates one neurobehavioral mechanism through which inductive-based fear learning operates -- typical category exemplars are preferentially signaled by the hippocampus early in the learning process, which interacts with the amygdala to generalize fear to other basic-level exemplars, thus leading to enhanced representations in object-selective cortices and widespread behavioral expression of fear to the superordinate category.
Figure 20: Activity in the hippocampus is modulated by object typicality during early aversive learning.

(a) A schematic of an analysis using subjects’ own typicality measures in a parametric modulation of fear conditioning-related brain activity. (b) During the early phase of aversive learning, neural activity in the left hippocampus was modulated by the typicality of objects from the feared category. (c) A psychophysiological interaction (PPI) analysis using the left hippocampus as a seed region showed enhanced connectivity with the left amygdala during early aversive learning. This typicality effect dissipated over time, as subjects learned to abstract fear learning to all members of the superordinate category.
6.4 Discussion

Although fear conditioning is traditionally considered an evolutionarily conserved system mediating only simple forms of learned behaviors, the present results reveal how fear can generalize to more abstract, complex representations of object categories. Utilizing object concepts as stimuli during aversive learning yielded three primary findings. First, conjoint activation in brain areas associated with emotional processing and representation of object concepts supports fear generalization based on sparse, trial-unique reinforcement of select members of a superordinate category. Representational category structures were modulated by aversive learning, such that activity patterns in object-selective cortex were more similar for different members from the feared versus safe category. Second, we found evidence for domain-specific neural pathways in learning to fear animate (as opposed to inanimate) objects, providing novel evidence for the specialized role of animacy in aversive learning models. Finally, the hippocampus was modulated by typicality of feared exemplars and exhibited a time-delimited coupling with the amygdala that was strongest early in acquisition training. Taken together, these results reveal advanced conceptual abilities contribute to the pursuit of fear learning, and, importantly, emulate characteristics of real-world fear acquisition that may serve as a more ecologically-valid model of clinical anxiety disorders characterized by widespread fear.

Prior human fMRI and non-human electrophysiological experiments have shown modulation in subcortical and unimodal sensory cortex as a result of aversive learning (Pape & Paré, 2010). This modulatory effect helps ensure that, due to association with an aversive event, the CS acquires emotional significance and is treated appropriately the next time it is encountered. Yet, prior neurophysiological studies almost universally incorporate a single CS instance into the fear conditioning design, and thus it is unclear whether neural modulation of a CS leads to widespread fear of related stimuli (but see (Weinberger, 2004)). Moreover, it has
remained unclear whether aversive learning modulates activation in cortical regions implicated in broadly coding object categories, despite the fact that these are the types of stimuli most often encountered in real-world learning episodes and contribute to clinical presentation in anxiety disorders such as Specific Phobia. Our results show activity in category-selective ROIs along posterior occipitotemporal cortex separated as a function whether participants acquired fear of animals or tools. These different patterns of activity, identified from a univariate fMRI analysis, are especially noteworthy given that participants relied on inductive reasoning to assess threat-value of each instance. Thus, while participants had no direct knowledge for whether a given exemplar predicted the US, the results suggest that knowledge of the affective significance of the category enhanced the representation of basic-level members. Moreover, activations observed in other regions traditionally implicated in conditioned learning (e.g., insula, ACC, and brainstem structures) extend the role of these structures beyond fear acquisition to a single, repeated CS. These findings complement a rich behavioral literature on stimulus generalization in classical conditioning (Pavlov, 1927), and extend neuroimaging findings on fear generalization beyond stimulus sets that vary parametrically along a unimodal perceptual dimension (e.g., faces (Dunsmoor, Prince, et al., 2011) or shapes (Greenberg, Carlson, Cha, Hajcak, & Mujica-Parodi, 2011)).

The conclusion that aversive learning modulates cortical representations of object concepts is further bolstered by the multivariate representational similarity analysis. Across participants in both learning groups, this analysis distinguished between patterns of activity elicited by animals and tools along occipitotemporal cortex, as previously demonstrated (Kriegeskorte, Mur, & Bandettini, 2008; Kriegeskorte, Mur, Ruff, et al., 2008). Importantly, we discovered that representational structure of these categories was functionally altered in an experience-dependent fashion. We specifically found enhanced representational similarity
between images of animals in subjects fear conditioned to animals, and analogous effects between images of tools in subjects fear conditioned to tools. Put another way, aversive reinforcement selectively enhanced similarity structure of an object category representation that acquired fear value. A reorganization of similarity structures as a result of learning may provide a means to support broad stimulus generalization between physically dissimilar items that portend the same significant outcome.

Evidence for domain-specific neural pathways supporting fear acquisition to animate objects fits with theoretical and empirical work concerned with whether particular classes of CSs serve as prepared stimuli during aversive learning (Öhman & Mineka, 2001). In human fear conditioning, a selective association has been proposed between objects serving as evolutionarily significant throughout mammalian development, including particular animals and conspecifics (Öhman & Mineka, 2001). Notably, individuals more often develop phobias for these fear-relevant stimuli (e.g. snakes) than recently created manmade objects (e.g. weapons) (Seligman, 1971). The important role of animacy is also supported by human neuroimaging data on biological motion (Allison, Puce, & McCarthy, 2000; Beauchamp, et al., 2002), and face processing (Adolphs, 2009), as well as electrophysiological studies showing a category-selective response to images of animals in the right amygdala (Mormann et al., 2011). Here, learning to fear animals was characterized by enhanced amygdala-lateral FFG coupling during fear acquisition, and during rest immediately following acquisition. These results are in line with the hypothesis that animacy is represented by intrinsic cortical connectivity with the amygdala because it was evolutionarily more important to integrate emotional information into the representation of animals and conspecifics than for artifacts (Martin, 2007a) (Mahon & Caramazza, 2011).
Finally, we discovered a novel typicality effect during early aversive learning. The hippocampus, a region commonly implicated in generalization processes (Gluck & Myers, 1993; Kumaran, et al., 2009; Shohamy & Wagner, 2008), exhibited a time-delimited modulation by object typicality of feared exemplars and a time-sensitive functional coupling with the amygdala, which may be important for the spread of fear learning across category exemplars on the basis of representativeness. Typicality effects play a well-known role in category-based induction (Murphy, 2002; Osherson, et al., 1990; Rips, 1975), but the role of typicality in fear learning is unknown. Typicality may be relevant to processing threat, as some exemplars are regarded as more prototypically dangerous than others (e.g. a Pit bull versus a Chihuahua). Prior fMRI investigations have demonstrated hippocampal-midbrain and hippocampal-ventromedial PFC connectivity during generalization of non-affective information (Shohamy & Wagner, 2008) (Kumaran, et al., 2009). Here, the hippocampus exhibited connectivity with the amygdala, differentiating its role from prior generalization investigations and suggesting a unique role in higher-order fear learning.

As we show evidence that conceptual systems can be utilized during fear acquisition, another intriguing question is whether these systems can be recruited during fear extinction, when the CS no longer predicts danger (Milad & Quirk, 2012). For instance, a prediction based on the category induction literature would propose that extinction to more typical (Rips, 1975) or diverse (Osherson, et al., 1990) category members would lead to better generalization of extinction. Extinction serves an important role in conditioning-based models of human anxiety (Milad & Quirk, 2012). Thus, future work extending the present approach to extinction phenomena may have broad implications for models of anxiety disorders marked by persistent fear responses to stimuli or situations that resemble a known threat, such as posttraumatic stress disorder (Foa, Steketee, & Rothbaum, 1989).
In conclusion, our findings show that learning to fear an object category modulates activity in and representational architecture of category-selective cortex and the amygdala. For subjects fear conditioned with animate objects, this modulation may be provided by a special mechanism through connectivity between the amygdala and lateral occipitotemporal cortex. Finally, typicality effects during early training impact category-based fear learning through hippocampal signaling and functional interactions with the amygdala. Once the relevant superordinate category representation is activated, this role for the hippocampus decreases in importance as participants begin generalizing fear to non-reinforced category exemplars, resulting in enhanced local representations in category-selective cortices. In sum, these findings demonstrate myriad ways humans learn to fear stimuli from the environment, and provide potential avenues for understanding how conceptual systems are recruited in the overgeneralization of fear characteristic of clinical anxiety disorders.
7. Synthesis and conclusions

The series of experiments in this dissertation are aimed towards gaining a better understanding of fear generalization processes in humans. Chapter 1 covered research on human and non-human animal fear conditioning, an area that has been highly active for several decades. Likewise, there is a rich and historical literature on stimulus generalization, covered in Chapter 2, which dates back to the earliest experiments on classical conditioning from Pavlov’s laboratory.

Surprisingly, the overlap between these two fields - fear conditioning and stimulus generalization - has only rarely overlapped over the past century. The research presented in this dissertation helps synthesize these related fields by showing how fear conditioning processes interact with myriad principles of generalization to affect how individuals transfer conditioned fear behaviors between perceptually and conceptually related stimuli.

This concluding chapter discusses how integrating theoretical and empirical work on generalization may help advance models of clinical anxiety disorders, for which fear generalization is a pervasive component. First, a brief overview of conditioning-based models of anxiety is presented with a focus on how the concept of generalization has been incorporated. Next, we provide an overview for how the multiple factors contributing to stimulus generalization support the generalization of fear. Finally, we conclude by proposing future directions.

7.1 Fear generalization in human anxiety disorders

Stimulus generalization is involved to some degree in most (and perhaps all) mental disorders involving excessive fear and anxiety. The focus here is limited to Specific Phobia and PTSD, as most of the conditioning-based anxiety literature has centered on these disorders.
7.1.1 Short history of conditioning-based models of anxiety

The earliest behavioral studies of fear conditioning in humans demonstrated that laboratory based conditioning might be a useful analog to investigate psychophysiological reactions to emotional experiences. Early behaviorist researchers, such as James B. Watson, used these conditioning principles as a basis to refute theories of anxiety grounded in Freudian psychology, which placed the origin of neuroses on various complexes and developmental issues while minimizing the role of direct stimulus learning. Because fear responses elicited in the laboratory resembled those observed in phobias (e.g. “Little Albert’s” fear of rats), early conditioning research supported the notion that anxiety was a reflection of stimulus-response (S-R) learning. The process of S-R learning was widely cast as the explanatory mechanism behind the acquisition of pathological behaviors (Pavlov, 1927; Watson & Rayner, 1920).

Theories of how mental illness arises through S-R learning fell out of favor in the mid to late twentieth century, due in part to the shift away from behaviorism towards cognitive models of learning and behavior. One of the chief criticisms leveled at behaviorism was that it was too simplistic to explain the etiology of fear and anxiety disorders. For instance, behaviorism made little account for differences between humans and other animals in their ability to condition and considered all classes of conditioned stimuli more or less equally (Seligman, 1970). Moreover, S-R models could not readily account for why some individuals are more susceptible to acquire a fear disorder following a conditioning experience than others. Also, individuals frequently develop fears to stimuli or situations to which they have never been directly exposed, seemingly circumventing S-R learning altogether (Rachman, 1977).

Models of classical conditioning evolved in the twentieth century and led to new insights into human fear learning. One of the monumental shifts away from behaviorism towards cognitive models was the conception that conditioned learning involves forming mental
representations of conditioned and unconditioned stimuli (S-S learning). In contrast to the black-box approach of S-R theories, these more cognitively-oriented learning models described how contingencies surrounding CS-US pairing influence the acquisition of learned behaviors. For instance, conditioning will only occur to stimuli that provide predictive value for the US, and learning is a result of experiencing the difference between what a CS predicts and what actually occurs (e.g. Rescorla & Wagner, 1972). Advances in cognitive models of associative learning have had beneficial applications to understanding fear disorders from the perspective of behavior theory (Mineka & Zinbarg, 2006).

7.1.2 Overgeneralization of fear in Specific Phobia and PTSD

7.1.2.1 Specific Phobia

Specific Phobia is the mental disorder most parsimoniously described by models of conditioning (Davey, 1992; Merckelbach, deJong, Muris, & vandenHout, 1996; Öhman & Mineka, 2001). Phobias include an irrational and persistent fear of certain objects or situations with the strong motivation to avoid the objects or situations – oftentimes to an excessive degree that may interfere with daily life (DSM IV). Early conditioning models simply conceptualized phobias as the result of strong conditioned fear responses to those neutral stimuli that have become associated with an aversive reaction (Mowrer, 1939; Pavlov, 1927; Watson & Rayner, 1920). Contemporary views on the etiology of phobias are still grounded in the notion that many phobias arise through direct aversive experiences, but have also incorporated the idea that fears can arise through higher-order conditioning, observation, or social communication (Mineka & Zinbarg, 2006).

Foa et al. (1989) have argued that phobias are unique among anxiety disorders insofar as the feared stimulus is generally well defined. Consequently, they argue, phobias are marked by less stimulus generalization than PTSD. For instance, whereas arachnophobia is limited to a fear of
spiders or other arachnids, PTSD triggers may include a wide range of cues or situations that are only abstractly associated with the traumatic event (see the following section).

Although fear generalization in phobias is mostly limited to cues associated with the phobia, it is important to note that this can still encompass a host of diverse cues or situations. Take for example the fear of flying. While the fear itself is related to flying, an aerophobic can be placed in a state of high anxiety by the mere thought of flying, or cues and locations associated with flying. Individuals with a specific phobia also have a bias to detect a wide range of information associated with their phobia. For example, an emotional version of the Stroop task has been used to show that phobic subjects have slowed reaction times to phobia-related words (e.g., the word “web” for spider phobics) relative to non-phobic related words (Watts, McKenna, Sharrock, & Trezise, 1986). Studies of co-variation biases in phobic subjects likewise demonstrate a cognitive bias to overestimate the frequency with which an aversive US follows phobic-related versus phobic-irrelevant stimuli (De Jong, Merckelbach, Arntz, & Nijman, 1992).

In sum, individuals with phobias tend to generalize fear beyond the phobic stimulus itself towards a range of associated objects or situations. Given that most complex real world objects can assume multiple forms, it is not surprising that phobics attend to a variety of cues that bear some resemblance to a perceived threat. For instance, someone with a fear of spiders may overreact to the sudden appearance of any small bug or avoid a range of locations where spiders typically occupy (e.g. a dank basement). Put another way, the persistent fear characteristic of a specific phobia is not necessarily bound to a precise object or situation.

7.1.2.2 PTSD

PTSD is characterized by severe psychological distress following an intense emotional experience in which one is confronted with death or injury. The symptoms of PTSD are marked by re-experiencing of the traumatic event, avoidance of stimuli or situations reminiscent of the
traumatic event, and hyper-arousal (American Psychiatric Association 1994). A common thread among the diagnostic criteria for PTSD is that cues or situations that are not themselves directly relevant to the primary trauma possess the capacity to elicit fear, whether through re-experiencing the event through reminder cues, avoiding stimuli that may act as reminders, or displaying vigilance to innocuous reminder cues. Thus fear generalization is an important factor that helps characterize PTSD (Brewin, 2001; Foa, et al., 1989; Foa, Zinbarg, & Rothbaum, 1992).

Foa et al. (1992) note that the commonness of generalized fear and arousal in PTSD may be related to the unpredictability associated with the traumatic event. This explanation fits well with animal models, in which an unpredictable aversive US leads to similar generalized fear behaviors as observed in individuals with PTSD. Controllability is therefore proposed as a countervailing force that helps “inhibit generalization of fear to the context” (Foa, et al., 1992).

Empirical data provides evidence for enhanced fear generalization to safety cues (CS-) in patients with PTSD (Grillon & Morgan, 1999). While laboratory studies typically employ perceptually related simple sensory stimuli, many of the reminder cues involved in evoking generalized fear in PTSD are symbolic in nature (Keane, Marshall, & Taft, 2006). For instance, mementoes or anniversaries of a trauma often provide reminders that can provoke anxiety. Although generalization in anxiety disorders frequently occurs across conceptual domains, conditioning studies in PTSD have so far utilized only perceptual dimensions.

7.2 What factors promote fear generalization?

A central theme in these studies has been to examine the source of fear generalization. Chapter 2 discussed multiple factors known to affect stimulus generalization. Notably, the preponderance of stimulus generalization experiments are based in the appetitive domain (e.g. key pecking or salivary responses in anticipation of a food reward). In the series of fear generalization experiments presented in the above chapters, we asked what perceptual and
conceptual factors influence generalization of fear, specifically. Fear learning systems, covered in Chapter 1, encompass a well delineated neural circuitry and set of behaviors related to avoidance and increases in sympathetic arousal that are in many ways distinct from behaviors associated with appetitive learning. Understanding the source of generalization may be important to provide a more robust understanding for how generalization processes integrate with fear learning systems.

7.2.1 Physical similarity and fear generalization

Theoretical and empirical studies of stimulus generalization have by far tended to rely upon a well-defined physical dimension to examine the transfer of learning between objects. In this way, the relationship between generalized responses and the similarity between the stimulus values can be quantified. Unimodal sensory dimensions (e.g. wavelength of light) are traditionally used because these dimensions can be divided into ordered units (e.g. nanometers) and behavior can be plotted as a function of similarity to the CS+ or CS- in terms of these units.

Physical dimensions have been crucial to define the nature of the generalization gradient and resolve questions on the relationship between generalization and discrimination (e.g. the Lashley-Hull debate). Recent human fear generalization studies have used simple unimodal geometric dimensions to plot gradients of generalized conditioned responses (Greenberg, et al., 2011; Lissek, et al., 2008). The elegant paradigm of Lissek et al. (2008) has been extended to clinical populations revealing broader gradients in patients with panic disorder versus healthy controls (Lissek, et al., 2010).

Yet simple physical dimensions pose some important problems for understanding the source of fear and fear generalization in humans. First, most real world fears do not involve simple sensory stimuli like tones, lights, or geometric shapes. Rather, fear involves complex stimuli composed of multiple features that can be encountered in multiple forms and contexts.
Thus, the ecological validity of these designs is tenuous. Another problem is that there is not a direct correspondence between the physical dimension and our perception of the physical dimension. Therefore, why should behavior conform one-to-one to the perceptual dimension? This point was appreciated to some extent during early theoretical debates on stimulus generalization (Hull, 1943; Lashley & Wade, 1946). For instance, it is well known that tones separated by an octave sound more similar, which leads to irregular gradients of behavioral generalization along a sound frequency domain (Blackwell & Schlosberg, 1943). To address issues concerning the correspondence between physical and psychological similarity, Roger Shepard pioneered the use of multidimensional scaling (MDS) to determine perceived similarity between objects. However, to date there is no empirical evidence that behavioral gradients of conditioned fear responses track subjective similarity judgments between stimulus values arrived at by MDS techniques. Also, because MDS is a similarity model, it does not account for intensity effects such as monotonic gradients and response biases (Ghirlanda 2002). In sum, physical similarity alone is not sufficient to explain fear generalization.

### 7.2.2 Intensity dimensions and fear generalization

Experiments presented in Chapters 3 and 4 combined physical similarity and emotional intensity to gauge fear generalization. These experiments showed generalization between faces that varied in the expression of fear was driven predominately by emotional intensity and not mere perceptual overlap.

Intensity models may be particularly well suited to describe fear learning, as intensity is an important dimension relevant to fear behaviors. That is, more intense stimuli act as better CSs and USs in models of conditioning and the intensity (or salience) of a CS may be directly related to how well it enters association with the US (Rescorla & Wagner, 1972). In addition, more intense USs lead to broader generalization than less intense USs (Baldi, et al., 2004). Intensity is
also an important feature within certain CS classes. For example, it could be argued that fear-relevant CSs like snakes and spiders are qualitatively more emotionally intense than fear-irrelevant CSs like flowers and mushrooms. Moreover, particular members of an object category may be considered more emotionally intense than other members, and these members may evoke relatively more fear generalization. For example, a widespread fear of dogs may not necessarily be uniformly distributed across all members from the category. Instead, fear may be limited to exemplars that are considered as more typical threatening, e.g. pitbulls and dobermans versus chihuahuas and poodles.

To the best of our knowledge, intensity generalization has not received attention in the conditioning-based anxiety literature. Intensity may be an important factor in the etiology and maintenance of overgeneralized fear. Thus future models of pathological fear and anxiety should consider intensity dimensions as an explanatory construct.

### 7.2.3 Higher-order forms of fear conditioning and fear generalization

Generalization is not always a product of physical similarity. Oftentimes it is predicated on a relationship between stimuli established prior to the conditioning experience. In mediated generalization (also known as acquired equivalence), different stimuli associated with a common stimulus are treated as more similar. Shohamy & Wagner (2008) revealed that coupling between the hippocampus and midbrain provided a neurobiological source of generalization during an non-affective acquired equivalence task. Unfortunately, studies of mediated generalization within the domain of fear learning are limited (Honey and Hall 1989).

A related concept to mediated generalization in the classical conditioning domain is sensory preconditioning. Chapter 5 presented data on a sensory preconditioning experiment in human fear conditioning and showed that conceptual similarity between conditional cues
facilitates the transfer of learned fear. Thus, conceptual similarity and higher order forms of learning interact to provide a source of fear generalization in humans.

Unfortunately systematic studies of higher-order forms of conditioned fear in humans are scant. This is surprising, given that these processes frequently provide a source of fear generalization in real-world situations, and has been included in influential conditioning based models of clinical anxiety (Davey, 1992; Foa, et al., 1989; Mineka & Zinbarg, 2006). For example, Davey (1992) provided a scenario in which a man on a bus witnesses a stranger suffer a fatal heart attack. Sometime later, this man’s father dies from a heart attack. Surprisingly, this man now experiences anxiety when riding the bus. In this scenario, the association between the bus and a heart attack was formed quite unintentionally. Yet, this latent association provided the link necessary for anxiety to transfer from the emotional event (heart attack) to a seemingly unrelated situation (riding the bus).

**7.2.4 Category-based fear generalization**

While similarity is in many cases determined by the physical overlap between object features, in many circumstances determining similarity is more complex and abstract. An extremely rich literature on categorization and conceptual processing has delved into the top of how animals (in particular, humans) utilize conceptual systems to form category representations, generalize knowledge between different category members, and generate new category exemplars (Murphy, 2002). This ability to recognize objects in the environment as a member of a category allows us to utilize our preexisting knowledge in flexible ways, but is so ubiquitous that we often fail to appreciate its complexity.

Theories of how humans process and represent categories and concepts are ideally suited to describe situations in which fear is generalized between category exemplars or conceptually related stimuli or situations. In Specific Phobia, for example, fear is generally expressed
categorically; e.g., a fear of spiders typically involves all arachnids, despite variations across species. Moreover, many phobias involve widespread generalization between cues that don’t necessarily contain any physical resemblance. In blood-injection phobia, for example, an individual may be anxious around a host of objects (e.g. needles), locations (e.g. a doctor’s office), or situations (e.g. a blood drive at work) that share an underlying conceptual relationship but little or no physical attributes. Similar processes may be involved in PTSD (e.g. avoidance of myriad cues or situations reminiscent of the trauma), panic disorder (e.g. avoidance of all enclosed spaces that may include elevators, MRI machines, traffic, or lines in the grocery store), and obsessive compulsive disorder (e.g. repeatedly checking items associated with danger, including ovens and unlocked doors).

The principles of categorization may be critical to understand generalization that occurs in real-world situations involving complex stimuli and situations. The experiment presented in Chapter 6 showed that healthy volunteers readily formed the association between an object category (animals or tools) and an electrical shock US, and exhibited differentially greater SCRs to the feared versus safe category. Brain activity revealed enhanced activity in category-selective brain regions, suggesting that learning to fear a category of objects modulates activity in regions putatively coding for the properties of the stimulus class. These findings have implications for understanding why it is that aversive experiences with involving specific instances of a category leads to widespread fear of related concepts.

**7.2.5 Future directions**

Emerging techniques in brain imaging and neurophysiology are allowing neuroscientists to examine areas of learning and behavior that were identified over a century ago, like classical conditioning. Behavioral neuroscience has extensively employed conditioning procedures to examine questions of learning, memory, and emotion. While stimulus generalization per se has
not received a large amount interest, renewed focus on this historical topic has started to appear. With the increasing interest in Pavlovian fear conditioning models (LeDoux, 1996), new research on the neuroscience of fear generalization has started to increase as well. A wealth of theories and behavioral research on stimulus generalization provides a rich framework to explore future directions using neuroscience-based approaches.

Future research should address the conditions that limit the overgeneralization of fear learning, and promote the generalization of extinction learning. This research can draw from emerging evidence on the genetics underlying fear conditioning (Mahan & Ressler, 2012), which may also help to reveal what determines individual differences in the breadth of fear generalization. Overgeneralization of fear may be related to a lack of stimulus control driven in part by a failure to appreciate the differences between a learned threat (in the laboratory, this involves the CS+ that actually does portend the US) and similar but innocuous cues (i.e., the CS- or other stimulus values). In this way, individuals may benefit from the opportunity to enhance this discrimination either through extended discrimination training or some other type of perceptual learning (Goldstone, 1998). The undergeneralization of extinction (i.e. the return of fear) is a problem well known in conditioning that has clear clinical relevance (Bouton, 2002). A question for future research is whether the principles of generalization (covered in Chapter 2) can be used to promote generalization of extinction learning, and whether these methods could plausibly optimize treatment.

Additionally, future research should begin to integrate accounts of generalization from the episodic memory literature (Kumaran, 2012; Zeithamova, Schlichting, & Preston, 2012) with fear conditioning models. Recent imaging experiments on inferential reasoning have revealed a unique role for the hippocampus in generalization between experiences. Although these studies are traditionally based outside the affective domain, many of these tasks are suitable for
conditioning paradigms. Hypothetically, these literatures could be integrated to explore higher order generalization processes that promote generalization of fear behaviors.

Finally, we suggest that the fear generalization paradigm may provide a useful technique to address questions of how information is stored and organized in the human brain. Pavlovian fear conditioning serves a multifaceted role in behavioral neuroscience that extends well beyond the study of “fear.” This is largely due to the fact that fear conditioning provides a rapid, cheap, and efficient method to elicit long-lasting effects in brain regions known to be involved in learning and memory, such as the amygdala and hippocampus (Maren, 2008). Notably, fear conditioning also alters brain circuitry outside areas putatively involved in simple forms of conditioning, as shown in Chapter 6. In this way, fear is a powerful method to examine how information is modulated by experience and, by extension, how this information is stored and organized in the brain. In other words, experimenters can take advantage of the fact that the brain is tuned to detect and respond not only to direct signals of danger, but generalized signals as well. In this way, fear generalization itself could be used to gauge how connections are made and information is stored in the brain. Thus, basic research using the fear generalization paradigm may prove a simple but powerful method to examine questions that extend beyond the affective domain.
References


Graham, R., Devinsky, O., & LaBar, K. S. (2007). Quantifying deficits in the perception of fear and anger in morphed facial expressions after bilateral amygdala damage. *Neuropsychologia, 45*(1), 42-54.


Biography

Joseph Dunsmoor was born in Washington D.C. in June of 1981. He received an undergraduate degree in Psychology from James Madison University in the spring of 2004. Following graduation, he worked as a training fellow at the National Institute of Mental Health under the supervision of Dr. Peter Bandettini and Dr. David Knight. He started graduate school at Duke University in the fall of 2007 under the supervision of Dr. Kevin LaBar. He begins a postdoctoral fellowship at NYU in the fall of 2012 in the laboratory of Dr. Elizabeth Phelps.

PEER REVIEWED RESEARCH PAPERS


BOOK CHAPTERS & COMMENTARIES


HONORS & FELLOWSHIPS

2011 American Psychological Association Dissertation Research Award
2010-2012 Individual NIH Predoctoral NRSA - F31 MH090682
2009-2012 Conference Travel Fellowship, Duke University
2009 Mentoring fellowship, Duke University
2009 Student Colloquium Committee Member, Duke University
2005 Postbaccalaureate Intramural Research Training Award (IRTA), National Institutes of Health