BEHAVIORAL AND NEURAL MECHANISMS OF SPONTANEOUS GENERALIZATION

by

STEFANIA RENÉ ASHBY

A DISSERTATION

Presented to the Department of Psychology
and the Division of Graduate Studies of the University of Oregon
in partial fulfillment of the requirements
for the degree of
Doctor of Philosophy
June 2021

DISSERTATION APPROVAL PAGE

Student: Stefania René Ashby

Title: Behavioral and Neural Mechanisms of Spontaneous Generalization

This dissertation has been accepted and approved in partial fulfillment of the requirements for the Doctor of Philosophy degree in the Psychology Department by:

Dasa Zeithamova	Chairperson
Brice Kuhl	Core Member
Ben Hutchinson	Core Member
Nicole Giuliani	Institutional Representative

and

Andrew Karduna	Interim Vice Provost for Graduate Studies

Original approval signatures are on file with the University of Oregon Division of Graduate Studies.

Degree awarded June 2021

DISSERTATION ABSTRACT

Stefania René Ashby

Doctor of Philosophy

Department of Psychology

June 2021

Title: Behavioral and Neural Mechanisms of Spontaneous Generalization

Memory generalization is the process by which we extract commonalities across our individual experiences to form new knowledge that can guide future decisions. Studies examining generalization have traditionally employed tasks, like category learning, that emphasize learning categorical information via extraction of commonalities among stimuli. Generalization is then explicitly assessed via transfer of category knowledge to new examples. Separately, memory for individual experiences, or memory specificity, has been studied through episodic memory tasks that emphasize differences between stimuli. However, real-world experience rarely puts us in situations where learning goals prioritize specificity or generalization at the expense of the other. Rather, circumstances often require us to extract the commonalities across our experiences while also maintaining memory for the specific details. Thus, the goal of the dissertation was to evaluate the behavioral and neural mechanisms that support spontaneous memory generalization during learning that emphasizes memory specificity. Using a novel, paired associates learning task where blended faces were paired with full-name labels, we provided an opportunity for participants to form category knowledge based on shared surname labels. Unlike traditional category learning tasks, learning goals in the current task explicitly required participants to differentiate all faces, even those with shared

family membership. Across 3 studies, using behavioral measures of perceived similarity and neural pattern analyses during encoding, we found that the mere presence of a shared label produced behavioral and neural evidence for category-biased representations during learning. Notably, neural evidence for category-biased representations extended beyond hypothesized memory generalization regions to include widespread aspects of the brain including higher-order visual cortex. Further, we found evidence that the hippocampus may support generalization and specificity simultaneously via differential connections with other hypothesized memory generalization and specificity regions. Together, our results inform our understanding of current theories of memory generalization by demonstrating conditions under which memory generalization proceeds spontaneously during learning.

This dissertation includes previously published and unpublished co-authored material.

CURRICULUM VITAE

NAME OF AUTHOR: Stefania René Ashby

GRADUATE AND UNDERGRADUATE SCHOOLS ATTENDED:

University of Oregon, Eugene
Brigham Young University, Provo

DEGREES AWARDED:

Doctor of Philosophy, Psychology, 2021, University of Oregon
Master of Science, Psychology, 2016, University of Oregon
Bachelor of Science, Psychology, 2011, Brigham Young University

AREAS OF SPECIAL INTEREST:

Cognitive Neuroscience

PROFESSIONAL EXPERIENCE:

Graduate Research & Teaching Assistant, University of Oregon,
September 2015 – June 2021

Staff Research Associate I, University of California, Davis,
June 2012 – August 2015

Undergraduate Research Assistant, Brigham Young University,
January 2011 – January 2012

GRANTS, AWARDS, AND HONORS:

National Science Foundation GRFP Honorable Mention, Functional Neural Networks Underlying the Testing Effect, University of Oregon, 2016

Phi Kappa Phi National Honor Society, Brigham Young University, 2011

Golden Key National Honor Society, Brigham Young University, 2010

Psi Chi International Honor Society, Brigham Young University, 2010

Phi Eta Sigma National Honor Society, Brigham Young University, 2009

PUBLICATIONS:

Ashby, S.R. Chaloupka, B., & Zeithamova, D. (in prep). Category learning induces true category biases in perception that are dissociable from strategic judgment bias.

Ashby, S.R. & Zeithamova, D. (in prep). Category-biased neural representations form spontaneously during learning that emphasizes memory for specific instances. *Journal of Neuroscience.*

Bowman, C.R., **Ashby, S.R.** & Zeithamova, D. (submitted). Aging effects on instructed and non-instructed category learning.

Ashby, S.R. & Zeithamova, D. (submitted). The role of test and restudy in the retention of briefly encountered facts.

Ashby, S.R., Bowman, C.R., & Zeithamova, D. (2020). Perceived similarity ratings predict generalization success after traditional category learning and a new paired-associate learning task. *Psychonomic Bulletin and Review, 27*(4), 791-800, doi: 10.3758/s13423-020-01754-3

Rosenthal, A., Mayo, D., Tully, L.M., Patel, P.K., **Ashby, S.R.**, Titone, M., Meyer, M., Carter, C.S., & Niendam, T.A. (2020). Contributions of childhood trauma and atypical development to poor clinical course in recent onset psychosis. *Early Intervention in Psychiatry*, 1-7, https://doi.org/10.1111/eip.12931

Garcia, P.L., **Ashby, S.R.**, Patel, P.K., Pierce, K.M., Meyer, M., Rosenthal, A., Titone, M., Carter, C.S., & Niendam, T.A. (2019). Clinical and neurodevelopmental correlates of aggression in early psychosis. *Schizophrenia Research, 212,* 171-176.

Niendam, T.A., Ray, K.L, Losif, A.M., Lesh, T.A., **Ashby, S.R.**, Patel, P.K., Smuncy, J., Ferrer, E., Solomon, M., Ragland, J.D., & Carter, C.S. (2018). Association of age at onset and longitudinal course of prefrontal function in youth with schizophrenia. *JAMA Psychiatry, 75*(12), 1252-1260, doi:10.1001/jamapsychiatry.2018.2538

Kirwan, C.B., **Ashby, S.R**., & Nash, M.I. (2014). Remembering and imagining differentially engage the hippocampus: A multivariate fMRI investigation. *Cognitive Neuroscience, 5*(3-4), 177-185, doi: 10.1080/17588928.2014.933203

ACKNOWLEDGMENTS

Throughout my degree program I have received a great deal of support and assistance. I first wish to express sincere appreciation to my advisor and mentor Dr. Dasa Zeithamova, whose expertise was invaluable in formulating the research questions and methodology of my projects in the lab. Your feedback and guidance throughout grad school pushed me to grow both professionally and personally during my academic journey.

In addition, special thanks are due to all the staff at the Lewis Center for Neuroimaging for providing facilities, equipment, training, and support integral to collection of the data included in this project. I greatly appreciate their assistance over the years and the opportunity to collect imaging data on my own providing me with more breadth of training.

I would also like to thank Dr. Caitlin Bowman and all my colleagues in the Brain and Memory Lab for their guidance, assistance, and emotional support in the form of long conversations and happy distractions from the stress of graduate school. Being surrounded by a friendly support system in the lab made all the difference in completing my degree.

Lastly, I would like to thank my parents for their encouragement, wise counsel, and sympathetic ear. I could not have overcome the struggles of the last six years without their unfailing support. My degree is a direct reflection of the amazing people I have the privilege to call family.

To Grandpa C. whose support made all this possible.

Thank you. Love you.

Until we meet again.

TABLE OF CONTENTS

Chapter Page

LIST OF FIGURES

LIST OF TABLES

Table Page

CHAPTER I

INTRODUCTION

Memory allows us to store the individual details of our daily experiences. However, our memory is not a mirror reflecting a detailed and perfect recall of past events. Rather it is a flexible, reconstructive process that also supports the extraction of common details across our individual experiences. G*eneralizing* memory information across all our prior experiences is adaptive and allows us to determine the best course of action when placed in novel situations. For example, a child may take several swimming lessons over the course of a summer and store individual memories for each lesson. However, details from the individual memories pertaining to water safety and various strokes can be combined across lessons and thus guide the child's behavior and decisions at the inaugural family beach visit the following summer. Although the child has never set foot on a beach before, generalizable aspects of prior experiences can be combined and are helpful in guiding decisions for safely and successfully swimming in this new environment. Though memory generalization is widely studied across many disciplines—decision making, perception, psychology, neuroscience—the mechanisms which allow generalization to proceed from our individual experiences and inform decisions in novel situations remains an actively explored topic in the literature.

How does the brain represent memories to retain specific information while also representing generalizable knowledge? Traditionally, a multiple memory systems view has suggested that memory generalization is supported by disparate neural substrates from those supporting memories for specific information. The hippocampus has a well-

known role in supporting detailed episodic memory (Scoville & Milner, 1957; Squire & Zola, 1998) and serves as a key region for reducing memory interference between similar experiences through pattern separation processes (for review see Yassa & Stark, 2011). In contrast, other memory systems such as the striatum (Poldrack & Foerde, 2008; Poldrack & Packard, 2003) or cortex (McClelland, McNaughton, & O'Reilly, 1995; O'Reilly & Norman, 2002) learn slowly and thus only represent statistical regularities that are likely to generalize across experiences. While the multiple memory systems view is intuitive and well supported, more recent work has indicated that there may be other ways the brain supports generalization. The hippocampus may also contribute to memory generalization that is rapid and based on a small number of experiences.

Hippocampal-based generalization has been studied using multiple paradigms. Episodic inference tasks contain various learning experiences that share common elements. Participants are instructed to learn details of the individual episodes but are also tested on whether they can infer new knowledge by linking common information across individual experiences. Many studies across various domains of episodic inference find hippocampal involvement (Ryan et al., 2016; Schlichting, Mumford, & Preston, 2015; Shohamy & Wagner, 2008; Zeithamova, Dominick, & Preston, 2012; Zeithamova & Preston, 2010) and interactions between the hippocampus and putative memory generalization cortical regions like the ventromedial prefrontal cortex (Bunsey & Elchenbaum, 1996; DeVito, Lykken, Kanter, & Eichenbaum, 2010; Schlichting et al., 2015) supporting these inference judgments. More recently additional support for hippocampal-based generalization has come from studies of category learning. Category learning paradigms typically involve presenting individuals with explicit instructions to

learn the category structure of a set of stimuli. After category learning, successful transfer of category knowledge to new, never-studied stimuli is evaluated as memory generalization performance. These category learning studies have demonstrated evidence for abstracted category representations in the anterior portions of the hippocampus as well as the ventromedial prefrontal cortex (Bowman, Iwashita, & Zeithamova, 2020; Bowman & Zeithamova, 2018) and within the middle temporal gyrus (Bowman & Zeithamova, 2018) another region known to support semantic gist memory (Dennis, Kim, & Cabeza, 2008; Turney & Dennis, 2017).

Using these paradigms, we have learned much about rapid hippocampal-based generalization. However, our knowledge comes primarily from laboratory tasks where learning explicitly emphasizes memory generalization. Real world experience suggests that learning conditions are often less explicit and multiple learning goals may simultaneously be at play. For example, a child encountering multiple dogs at the park may both remember the individual dogs as well as form an overall generalized "dog" representation that can be applied to identify a new animal the next time a furry, four-legged creature is encountered. Thus, it is unlikely in a real-world context that generalized representations only form when learning goals emphasize generalization. Instead, memory generalization may proceed more spontaneously under conditions that emphasize learning episodic details of our experiences. Thus, the primary goal of the dissertation is to determine the behavioral and neural mechanisms that support the spontaneous formation of generalized memory representations under learning conditions that do not emphasize generalization.

Competing Theories of How Memory Generalization Proceeds from Learning

A vast wealth of research has explored memory generalization (F. G. Ashby, Alfonso-Reese, Turken, & Waldron, 1998; F. G. Ashby & Maddox, 2005, 2011; Zeithamova & Bowman, 2020; Zeithamova, Schlichting, & Preston, 2012) and ultimately several theories have emerged regarding how memory generalization proceeds from learning. One proposal of memory generalization postulates a *flexible retrieval* hypothesis where generalization occurs "on-demand" at retrieval when task demands require individuals to make a generalization judgment (Squire, 1992; Teyler & DiScenna, 1986; Winocur, Moscovitch, & Sekeres, 2007). Thus, during learning individual memories are stored and it is not till individuals are prompted to make a generalization judgment (e.g. categorize new examples or infer a relationship between two associated episodes) that generalization occurs. Alternatively, another proposal of memory generalization postulates an *integrative encoding* hypothesis where generalization occurs *prior* to retrieval or situations that create generalization task-demands (Shohamy & Wagner, 2008; Zeithamova, Schlichting, et al., 2012). Therefore, during learning the commonalities across experiences are integrated simultaneously.

On-demand Generalization Through Flexible Retrieval

Several areas of research have supported a flexible retrieval account of memory generalization. Exemplar theories of category learning postulate that a generalized memory representation is unnecessary for individuals to make a generalization judgment. Instead, specific memory traces representing each learned item are stored at encoding and these traces are sufficient for informing generalization at retrieval (Hintzman, 1984; Kruschke, 1992; Nosofsky, 1988). When probed to make a generalization judgement the

individual stored examples are retrieved and compared with one another "on-the-fly" to inform the generalization decision. Formal mathematical models for exemplar models of category learning have been applied to fMRI to elucidate the representations underlying concept learning. Using this modeling approach exemplar model correlates have been found in lateral occipital and posterior parietal cortices during retrieval (Mack, Preston, & Love, 2013) indicating that these regions support generalization by representing each exemplar separately.

Studies of episodic inference provide converging evidence for how separate memory representations for individual experiences can be used to inform generalization. In associative inference paradigms, pairs of items with overlapping associations are learned (e.g. AB and BC pairs) and generalization is tested by examining performance on the indirectly learned association (e.g. AC pairs; for examples see Zeithamova & Preston, 2010; Zeithamova, Schlichting, et al., 2012). Retrieval-based accounts of inference assert that hippocampal pattern-separated episodes are stored for every experience during encoding; even experiences that are related to one another are orthogonalized. Thus, generalization is a result of reactivation of multiple related memory traces (AB and BC pairs) that are recombined dynamically or "on-the-fly" during retrieval (Kumaran, 2012) in response to a generalization task-demand. Consistent with retrieval-based accounts, Banino, Koster, Hassabis, and Kumaran (2016) found that better memory for directly studied items (AB and BC pairs) predicted better generalization performance and computational modeling showed that a retrieval-based account of generalization best fit the data.

Memory Integration During Encoding

Others argue that people may also link information across distinct episodes during encoding, prior to retrieval or situations that create generalization task-demands. Prototype theories of category learning postulate that memory generalization is supported by a generalized category representation created by averaging features abstracted across category exemplars (Posner & Keele, 1968; Reed, 1972). Thus, memory generalization is facilitated by comparing new, incoming information against the category prototype representation which is then dynamically updated over the course of learning. Fitting a formal mathematical model for prototype category learning, Minda and Smith (2001) found evidence that a prototype account fit the data best indicating that learning category information during encoding is a function of how similar individual items are to a given prototype representation.

Other episodic inference work supports an integrative encoding account. For example, in studies of associative inference, it is also possible that when overlap between items is encountered (e.g. studying a BC pair after already learning AB) memory for the overlapping pair (AB) is reactivated by hippocampal pattern completion processes and combined with the BC representation. Thus, an integrated ABC representation during encoding is constructed (for review see Zeithamova, Schlichting, et al., 2012). Consistent with the integrative encoding hypothesis, Shohamy and Wagner (2008) found that the degree of neural activation increases within the hippocampus across encoding is associated with performance on subsequent memory generalization tests. Further, Schlichting et al. (2015) found evidence for integrated memory representations in anterior hippocampus and posterior medial prefrontal cortex *prior* to an explicit inference test.

Other support for integrative encoding comes from observations of impaired cognition in amnesic patient populations which have shown impaired recognition memory but spared ability to classify and make generalization judgments (Knowlton & Squire, 1993). Because amnesiacs cannot use a flexible retrieval process due to impaired memory specificity for individual events, their intact ability to generalize may indicate that generalizable information was learned independently from their ability to learn the episodic details at encoding. Together this evidence would suggest that generalized memory representations can be constructed spontaneously during learning.

Generalization May Proceed from Learning Via Both Mechanisms

Flexible retrieval and integrative encoding may not be mutually exclusive mechanisms of how memory generalization proceeds from learning. Instead it is plausible that both integrative encoding and flexible retrieval processes are at play and disparate findings across domains of study reflect the widely different circumstances surrounding individual task parameters and learning goals resulting in dominance of one or the other mechanism (for more discussion see review by Zeithamova & Bowman, 2020). One possibility is that separate memory representations may be flexibly linked on-demand at retrieval, but this process may result in constructing integrated memory representations. This is consistent with work in associative inference that shows individuals have significantly more false memories for directly studied pairs (AB and BC) *after* but not before successful AC inference trials are tested (Carpenter & Schacter, 2017, 2018). Further, patterns of neural activity in the anterior hippocampus during AB retrieval after successful inference trials (AC test) are more similar to neural activity patterns in the overlapping BC trials (Carpenter, Thakral, Preston, & Schacter, 2021). Thus, memories

for the directly learned associations must have been formed at encoding and only after task-demands required a generalization decision (integration of the AC items at retrieval) did a generalized memory representation form that interfered with the ability to recall the direct associations.

Alternatively, both types of representations may co-exist. Representations for item- or episode-specific information contain the detailed memories of our individual experiences while generalized memory representations contain information combined across our varied experiences and support more conceptual memory information. If learning proceeds via both mechanisms generalized memory representations may be both measurable during retrieval as well as during encoding alongside our memories for individual experiences. There is emerging evidence that the hippocampus may support the simultaneous construction of both types of representations via a division of labor along the long axis of the hippocampal body (Poppenk, Evensmoen, Moscovitch, & Nadel, 2013). Animal work has shown that information is represented in the hippocampus at multiple levels of spatial specificity via an anterior-posterior gradient in receptive field size along the long axis of the hippocampus. Kjelstrup et al. (2008) found larger receptive fields in the ventral (analogous to human anterior) hippocampus and smaller receptive fields in the dorsal (analogous to human posterior) hippocampus. Similar findings in the human hippocampus by Brunec et al. (2018) showed more overlap of spatial representational patterns in the anterior hippocampus compared to the posterior hippocampus. Thus, the anterior portions of the hippocampus may be capable of supporting learning of generalizable information while maintenance of episodic details may be supported by the posterior hippocampus. Alternatively, other work posits that

support for both representations simultaneously could also be made through differential pathways (monosynaptic vs. trisynaptic) within hippocampal subfields (Schapiro, Turk-Browne, Botvinick, & Norman, 2017; Schlichting, Zeithamova, & Preston, 2014; Zeithamova, Manthuruthil, & Preston, 2016). Importantly, the hippocampus appears to be a structure that is capable of supporting information across multiple levels of specificity in service of both episodic memory and memory generalization.

Can Memory Generalization Proceed Spontaneously?

While existing work has provided many new insights into the possible mechanisms of generalization, the majority of studies have focused on memory generalization as it proceeds from explicit instructions. For example, in traditional category learning tasks participants are often explicitly instructed that there is a category structure. The category structure is learned via feedback-based learning where a category label is guessed, and corrective feedback is received (F. G. Ashby & Maddox, 2005). However, it is clear from other work that explicit awareness of relationships between learned items is not necessary for memory generalization. Shohamy and Wagner (2008) found that overlap between some elements across episodes induced integration of that information into a generalized representation even though there was no explicit awareness amongst participants regarding the relationships. Examples of incidental generalization can also be seen in episodic inference. Transitive inference tasks—where hierarchical relationships between items are learned (A > B, B > C, C > D) and then unlearned relationships are tested (B ? D)—typically withhold details that the individual associations being learned together form a hierarchical relationship (Heckers, Zalesak, Weiss, Ditman, & Titone, 2004; Ryan et al., 2016; Zalesak & Heckers, 2009). Yet,

activity in the hippocampus during encoding predicts inference performance even after controlling for how well individuals learned the directly studied associations (Heckers et al., 2004), suggesting that the hierarchical knowledge is being spontaneously formed.

Studies of incidental category learning also show that even without explicit instruction to categorize, it is still possible for individuals to acquire generalized category knowledge (Aizenstein et al., 2000; Bozoki, Grossman, & Smith, 2006; Gabay, Dick, Zevin, & Holt, 2015; Kéri, Kálmán, Kelemen, Benedek, & Janka, 2001; Love, 2002; Reber, Gitelman, Parrish, & Mesulam, 2003; Wattenmaker, 1993). In these tasks, category information is present but not emphasized during learning as participants are typically distracted with a cover task at encoding. For example, Aizenstein and colleagues (2000) used black and white dot pattern stimuli that changed to one of three colors (red, blue, or yellow). Participants were only instructed to make a corresponding button press when the dot colors changed. However, unbeknownst to participants, which color the dots would become was determined by the spatial pattern of the dots. Their results showed that after learning, even though participants had no conscious awareness of there being an underlying category structure, they were more accurate at classifying never-studied distortions of the three dot pattern category prototypes than they were at classifying never-studied distortions of prototypes that were not learned during the incidental training.

Other incidental category learning tasks have used only a single category paradigm often referred to as "A/not-A" learning (F. G. Ashby & Maddox, 2005; F. G. Ashby & O'Brien, 2005). In these tasks, participants learn examples of a single category during a study phase. Following this phase, they are then informed that all the items they

learned were members of a single category and then asked to make a judgment as to whether new items are also members of the same category they learned previously. Together, previous findings make it clear that individuals are capable of learning category information that can be generalized to new situations incidentally and under conditions which do not promote awareness of the generalizable information. Thus, it is conceivable that information may be spontaneously generalized during encoding even without explicit awareness.

Category Bias in Perception as a Means to Measure Spontaneous Generalization

Category learning may induce conditions which allow us to uniquely measure memory generalization outside of the confines of an explicit generalization task. Acquiring category knowledge has been shown to bias our perception. Leveraging these perceptual biases during learning of category information may allow for a more incidental measure of generalization under conditions when generalization task demands are greatly minimized. Early work showing this influence comes from studies examining categorical perception. Categorical perception effects are best defined as an ability to better differentiate stimuli when they belong to different categories than when they belong to the same category (Goldstone, 1994a). For example, in prior work examining speech categories, individuals were better able to distinguish speech sounds from one another when they were from different phonemic structure categories than when they both had similar phonemic structures (Liberman, Harris, Hoffman, & Griffith, 1957). Other work examined categorical perception effects using color stimuli that varied on a graded scale from green to blue. Gilbert, Regier, Kay, and Ivry (2006) found that participants were slower to make a discrimination judgment between two colors that were

within the same color category than they were at making judgments between colors across the category boundary.

Additional work has expanded these categorical perception findings to determine if categorical perception could be induced by learning new category structures. Beale and Keil (1995) found participants were better able to discriminate morphed face stimuli when they straddled a learned category boundary than when they were learned to be within the same category. Further, Folstein, Palmeri, and Gauthier (2013) demonstrated the same result while also controlling for within and between category similarity of the stimuli and showed that participants were better able to discriminate along a category-relevant dimension that was diagnostic of category membership rather than along a category-irrelevant dimension that did not coincide with the learned category structure. Utilizing a traditional feedback-based category learning task, Livingston, Andrews, and Harnad (1998) presented artificial stimuli that varied on two dimensions (e.g. artificial microorganisms that varied across category boundary according to shape and length of artificial cilia projections) and participants made subjective ratings of perceived similarity for pairs of stimuli. They found participants rated across-category pairs as being less similar to one another than pairs within either category (between-category expansion or acquired distinctiveness). Thus, category learning induced a perceptual category bias that was primarily driven by an *expansion effect* where items between learned category boundaries were "pushed apart" in perceptual space.

In addition to this category biased expansion effect in perceptual similarity space for items learned to be between category boundaries, other studies have found perceptual changes for items learned to be within categories. Oftentimes, category learning causes

within-category items be perceived as more similar to one another leading to a category biased *compression effect*. Gureckis and Goldstone (2008) found items within a learned category to be less discriminable after learning while other studies that objectively measured changes in perceived similarity found participants rated items within learned categories as more perceptually similar to one another after compared to before learning (Goldstone, Lippa, & Shiffrin, 2001; Kurtz, 1996; Livingston et al., 1998). Goldstone, Lippa, and Schriffin (2001) found both compression and expansion effects in perceptual similarity ratings after learning.

Together these findings suggest that tracking perceptual category biases after learning may provide a unique way to measure learning related category knowledge. Moreover, detecting a category bias in perceived similarity ratings after learning but prior to an explicit generalization task may be a good index of generalizable category knowledge that is present prior to a task that explicitly demands a generalization judgement. A category learning task that presents category-relevant information but emphasizes encoding of detailed episodic memory while also controlling for within and between-category similarity amongst stimuli would be an excellent way to explore the questions presented.

Goal and Structure of the Dissertation

The primary goal of the dissertation is to determine the behavioral and neural mechanisms that support the spontaneous formation of generalized memory representations under learning conditions that do not emphasize generalization. We developed a novel category learning paradigm that emphasized learning of unique, detailed episodic information while also providing generalizable information in the form

of an underlying category structure. In addition to theoretical discussions in Chapters 1 and 5, we addressed this question in three empirical studies described in Chapters 2-4.

In the empirical chapters we leveraged measures of perceived similarity changes following category learning, memory performance for transferring category labels to never-studied stimuli, and neural measures of category-biased information during encoding to examine generalization. In Chapter 2, we introduce the novel category learning paradigm to assess the extent to which signatures of category knowledge are present in a task that emphasizes memory for stimulus-specific information. The goal of Chapter 2 was to determine to what degree a category-bias in perception following learning may be a useful behavioral index of memory generalization. We found that individuals were able to successfully generalize to never-studied examples although learning goals emphasized specificity and individuation of studied stimuli. The degree of category bias after learning predicted subsequent generalization performance providing evidence for generalization immediately after learning and prior to retrieval.

In Chapter 3, we utilized fMRI during encoding to determine whether evidence for category-biased information is present in neural representations during encoding and prior to retrieval. We found patterns of activity that were biased towards category-relevant information across widespread aspects of the cortex including some regions hypothesized to support memory generalization. Lastly, in Chapter 4 we examined intrinsic background connectivity between the hippocampus and putative generalization and specificity regions to explore how the hippocampus is able to simultaneously support memory generalization while maintaining specificity. We found differential connections between anterior and posterior portions of the hippocampus with putative generalization

and specificity regions providing further evidence for a hippocampal mechanism supporting spontaneous generalization during encoding.

CHAPTER II

PERCEIVED SIMILARITY RATINGS PREDICT GENERALIZATION SUCCESS AFTER TRADITIONAL CATEGORY LEARNING AND A NEW PAIRED-ASSOCIATE LEARNING TASK

From Ashby, S.R., Bowman, C.R., & Zeithamova, D. (2020). Perceived similarity ratings predict generalization success after traditional category learning and a new paired-associate learning task. *Psychonomic Bulletin and Review, 27*(4), 791-800, doi: 10.3758/s13423-020-01754-3

Categorization helps us organize information from the world around us into meaningful clusters relevant to behavior. A hallmark of category knowledge is the ability to categorize new instances (memory generalization), allowing us to use our prior experiences to guide decisions in novel situations (Knowlton & Squire, 1993; Nosofsky & Zaki, 1998; Poldrack et al., 2001; Reber, Stark, & Squire, 1998). Category knowledge also results in biases in perception, which can manifest as increased perceived similarity of items within a category, decreased perceived similarity of items from different categories, or a combination of both (Beale & Keil, 1995; Goldstone, 1994a; Goldstone et al., 2001; Kurtz, 1996; Livingston et al., 1998). These perceptual biases are often thought to reflect stretching of the perceptual space along the category-relevant dimensions and/or shrinking along the category-irrelevant dimension, resulting from shifts of attention to the relevant features (Goldstone & Steyvers, 2001; Kruschke, 1996; Medin & Schaffer, 1978; Nosofsky, 1991; Nosofsky, 1986). While a category bias on peception can emerge relatively quickly following category learning, it remains unknown to what degree it reflects the quality of category knowledge and relates to subsequent categorization and generalization performance. If category learning results in changes of

perceptual space and persistent attentional shifts to category-relevant features, the degree of category bias on perception should be a good indicator of the quality of category knowledge. On the other hand, if good learners more accurately encode all information—which may allow them to better determine which information is category relevant and which irrelevant—then the degree of category bias may not be a good predictor of category knowledge. Thus, one goal of the current study was to measure both category bias in perception and generalization in a single study to determine to what degree category bias in perception following category learning can be used as a measure of generalizable category knowledge by predicting performance on unstudied items.

Most categorization studies explicitly instruct participants to learn categories. Several studies have also compared categorization tasks that focus on contrast across categories and commonalities within categories to identification tasks that focus on learning stimulus-specific information (Nosofsky, 1986; Shepard & Chang, 1963; Shepard, Hovland, & Jenkins, 1961). However, in the real world, category information can be available alongside information about specific items or individuals, without an explicit goal to form category knowledge. For example, when attending a wedding and meeting many new individuals, one's objective is to remember individual people and learn their unique names. Yet, some guests may share last names, providing an opportunity to also extract categorical structure across individuals. Past work has shown that category knowledge can be extracted without explicit instruction (Aizenstein et al., 2000; Bozoki et al., 2006; Gabay et al., 2015; Kéri et al., 2001; Love, 2002; Reber et al., 2003; Wattenmaker, 1993). However, how category learning proceeds when category information is available, but instructions emphasize learning of specific information is

rarely addressed. While some show that categorization performance can be predicted from performance on identification tasks that emphasize discrimination of individual items (Nosofsky, 1986), others have found that learning and generalizing concept information is more challenging when learning is focused on discrimination of individual stimuli (Soto & Wasserman, 2010). Thus, in Experiment 2, our goal was to assess signatures of category knowledge – generalization and category bias on perception – in a task that emphasizes memory for stimulus-specific information and more closely resembles an episodic paired-associate learning task than a traditional category-learning task.

In the current paper, we assessed (a) category bias on perception, (b) category generalization success, and (c) their relationship after traditional category learning (Experiment 1) and after a novel task where category information was available but instructions emphasized stimulus-specific information (Experiment 2). Participants were shown faces that belonged to three categories (families), designated by a family name. Face stimuli were created as blends of never-seen "parent" faces, resulting in increased physical similarity between faces that shared a parent. Some physically similar faces were members of the same family while others were members of different families, allowing us to dissociate the effect of category membership from physical similarity. In Experiment 1, faces were encountered in the context of a traditional feedback-based category learning task, emphasizing similarities among faces belonging to the same family and how they contrast with faces belonging to different families. In Experiment 2, faces were encoded through observational, face-full name paired-associate learning. While family names were identical to Experiment 1, with each family name shared across

several faces, first names were unique for each face, requiring participants to remember individual faces and differentiate faces within each family. Perceived similarity ratings were collected immediately before and after learning to test for the emergence of category bias in perception. We also tested participants' ability to generalize family names to new face-blend stimuli. The category bias in perceived similarity ratings after learning was related to subsequent generalization success in order to determine the extent to which category bias in perception reflects the quality of category knowledge.

The current design allowed us to also address additional questions regarding the nature of category bias in perception. First, what drives category bias in perception has been variable across studies. Some studies have shown *between-category expansion* or *acquired distinctinctiveness,* where items across a learned category boundary become more discriminable (Beale & Keil, 1995; Folstein et al., 2013; Goldstone, 1994a; Gureckis & Goldstone, 2008; Wallraven, Bülthoff, Waterkamp, van Dam, & Gaißert, 2014) and are perceived as more dissimilar after category learning (Goldstone et al., 2001). Category bias can also manifest as *within-category compression* or *acquired equivalence*, where items within a learned category become less discriminable (Gureckis & Goldstone, 2008; Soto, 2019) and are perceived as more similar after category learning (Goldstone et al., 2001; Kurtz, 1996; Livingston et al., 1998). As relatively few studies show both compression and expansion effects following category learning (but see Goldstone et al., 2001; Gureckis & Goldstone, 2008), we were interested to what degree both expansion and compression effects can be observed after category learning of the face-blend stimuli with equated within-category and between-category physical similarity. Furthermore, the aforementioned studies on learning-related category bias

have focused on traditional category learning. Thus, the degree to which within-category compression and between-category expansion can be observed after learning that emphasizes memory for stimulus-specific information remains unknown.

Finally, using perceived similarity to probe category knowledge in Experiment 2 can help us link research on the emergence of conceptual knowledge to another area of generalization research: episodic inference. Episodic inference refers to the ability to integrate information across distinct experiences that share content to infer new information (e.g. infering that two people are likely a couple after seeing each of them with the same child on different occasions). Whether people spontaneously integrate memories of related events as they are encoded (Cai et al., 2016; Gershman, Schapiro, Hupbach, & Norman, 2013; Schlichting et al., 2015; Shohamy & Wagner, 2008; Zeithamova, Dominick, et al., 2012) or whether links between related memories are formed in response to generalization demands (Banino et al., 2016; Carpenter & Schacter, 2017, 2018) remains debated. Here, observing evidence for the formation of a category representation under conditions that minimize generalization demands – such as observing category bias in perceived similarity ratings after learning but before the explicit generalization test —would suggest that participants may extract category information and form category representations spontaneously.

Method

Participants

Healthy participants—N = 39 in Experiment 1 and N = 43 in Experiment 2—were recruited from the University of Oregon community via the university SONA research system and received course credit for their participation. Except for the learning phase,

all procedures were identical across experiments and will be presented together. All participants provided written informed consent, and experimental procedures were approved by Research Compliance Services at the University of Oregon. From Experiment 1, four participants were excluded due to chance performance (accuracy $\leq$ .33) in categorizing the training faces. From Experiment 2, participants were excluded for failing to make responses on more than 25% of categorization trials (n = 3) and incomplete data (n = 1). After exclusions, analyses were carried out with the remaining 35 participants for Experiment 1 (M_{age} = 20.43, SD_{age} = 2.58, 18-32 years, 21 females) and 39 participants for Experiment 2 (M_{age} = 19.26, SD_{age} = 1.13, 18-23 years, 21 females). These sample sizes provide 80% power for detecting medium size effects ($d \geq 0.5$) using planned one-sample and paired t-tests and strong ($r \geq .5$) correlations, as determined in G-Power (Faul, Erdfelder, Buchner, & Lang, 2009; Faul, Erdfelder, Lang, & Buchner, 2007).

Stimuli

Stimuli were grayscale images of blended faces constructed by morphing two unaltered face images together using FantaMorph Version 5 by Abrosoft. We used blended faces because it allowed us to maintain realistic-looking stimuli while also controlling for within- and across-category physical similarity. Faces were also convenient for creating the face-name learning task in Experiment 2 that was intuitive for the participants and yielded the right level of difficulty as verified through a pilot study. Prior work has shown that category effects differ based on whether morphed faces are constructed from parents within one race versus across two races (Levin & Angelone, 2002). Thus, we restricted all parent faces to be Caucasian to ensure that the resulting

face-blend stimuli were comparably similar to all other faces with a shared parent. Additionally, all parent faces were of a single gender (male) to ensure that face-blends maintained a realistic appearance. Parent faces were compiled over several years from multiple sources, including the Dallas Face Database (O'Toole et al., 2005), CVL Face Database provided by the Computer Vision Laboratory, University of Ljubljana, Slovenia (Peer, 1999), and Google Image Search. Faces were selected primarily based on whether they would blend well with other faces (e.g., visibility of both ears, no facial hair, etc.) but were not formally equated for features such as attractiveness or memorability.

The stimulus structure is presented in Figure 2.1. For each participant, three category-relevant parent faces and three category-irrelevant parent faces were randomly selected from a total set of twenty faces. Each of the three category-relevant parent faces were individually morphed with each of the three category-irrelevant parent faces with equal weight given to each parent face (50/50 blend). The resulting nine blended faces were then used as training stimuli. Faces that shared a category-relevant parent shared a family name (belonged to the same category). Faces that shared a category-irrelevant parent belonged to different families. As faces sharing any parent (category-relevant or category-irrelevant) shared physical traits, physical similarity alone was not diagnostic of category membership. Because of the blending procedure used, an equal number of category-relevant and category-irrelevant parent faces were selected to provide equal exposure to the relevant and irrelevant category features. With an uneven number of relevant vs. irrelevant parent faces (e.g. two relevant parent faces blended with multiple irrelevant parent faces to create family members), unsupervised learning could take place,

making the features of the relevant parent faces more prominent through increased exposure instead of being category-learning driven. We chose a three-way category structure, which provided nine blended faces to learn and therefore 36 pairwise similarity rating comparisons. We determined that the three-way structure provided the best balance of a reasonable number of training stimuli to learn but still provided adequate pairwise comparisons for similarity ratings. Generalization stimuli were new faces created by

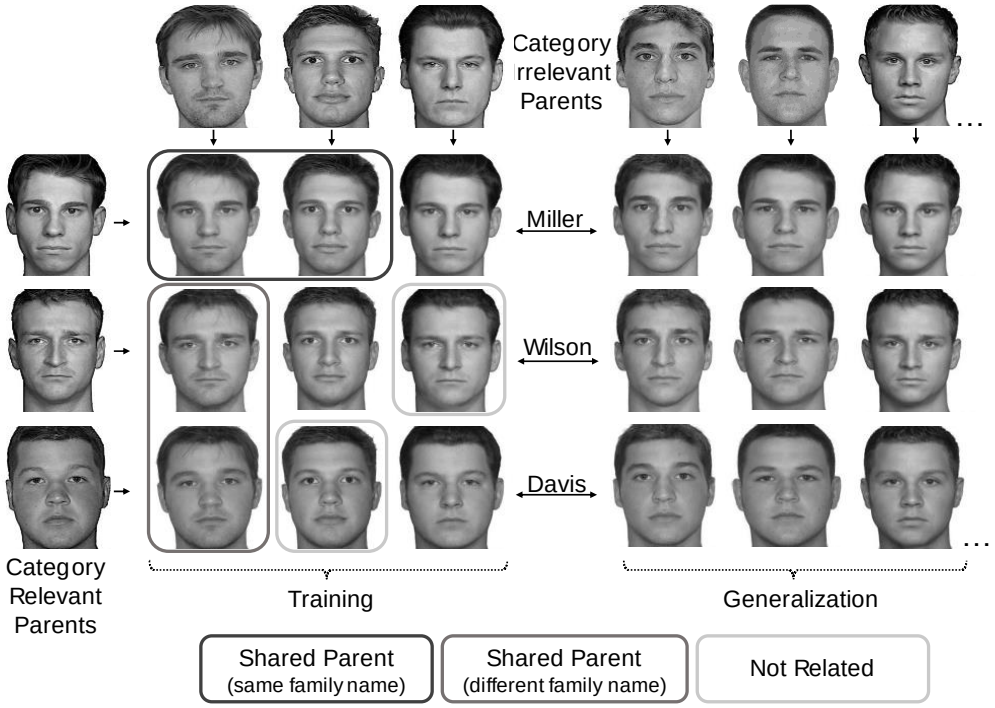

Figure 2.1. Example face-blend stimuli. Parent faces on the leftmost side are designated "category relevant parents" as these parents determined family membership—Miller, Wilson, or Davis—during learning and generalization. Parent faces across the top are designated "category-irrelevant parents" as these parents introduced physical similarity across families but did not determine categories. Three category-irrelevant parents were used for learning. The rightmost three category-irrelevant parents are a subset of new faces used for generalization. Parent faces were never viewed by participants, only the resulting blended faces. The face blending procedure produced pairs of faces that shared a category-relevant parent and belonged to the same family (shared parent - same family name; example indicated with dark grey box), pairs of faces that shared a category-irrelevant parent and belonged to different families (shared parent- different family name; example indicated with medium grey box). Non-adjacent pairs did not share a parent and were not related (example indicated with light grey boxes).

blending category-relevant parent faces with fourteen remaining parent faces not used for creation of the training faces.

Procedure

Both experiments consisted of the following phases: passive viewing, pre-learning similarity ratings, learning (different in each Experiment), passive viewing, post-learning similarity ratings, and category generalization. Additionally, Experiment 2 included cued-recall of face-name associations before the category generalization phase. Self-paced breaks separated the phases.

Passive viewing. To familiarize participants with the stimuli and give them an idea of the degree of similarity between all faces before collecting perceived similarity ratings, participants first viewed each of the nine training stimuli individually, once in a random order without any labels and without making any responses. Face-blends were shown for 3s with a 1s inter-stimulus-interval (ISI). Passive viewing of the face-blends immediately before the pre- and post-learning similarity rating phases was also included as a pilot of a future neuroimaging experiment. No responses were collected during viewing.

Pre-learning similarity ratings. To validate that participants were sensitive to the similarity structure among faces introduced by the blending process and to obtain baseline similarity ratings, participants rated the subjective similarity of pairs of faces to be used during the learning phase. All possible 36 pairwise comparisons of the 9 training faces were presented and participants rated the similarity of the two faces on a scale from one to six (1 = two faces appeared very dissimilar, 6 = two faces appeared very similar). Face pairs and the similarity rating scale were displayed for 5s with a 1s ISI. Face pairs

were then binned into three conditions for analyses depending on whether they 1) shared a parent and a family name, 2) shared a parent face but did not share a family name, or 3) did not share a parent face (see example pairs in Figure 2.1).

Learning phase.

Experiment 1: Feedback-based category learning. On each trial, a training face was presented on the screen along with family names (Miller, Wilson, Davis) as response options. Participants were instructed to indicate family membership via a button press and received corrective feedback after each trial. Each face was viewed simultaneously with the family name response options on the screen for 4s, received corrective feedback for 1s, and trials were separated by a 1s ISI. Each face was presented 16 times total, evenly split across 2 blocks.

Experiment 2: Observational learning of face—full name associations. To test the robustness of category learning outside of a traditional categorization task, Experiment 2 provided an opportunity to form associations between faces from the same families in the context of a face-full name associative learning task. On each trial, participants studied a face-name pair that was presented on-screen for 2s and then made a prospective memory judgement for 2s on a scale from one to four (1 = definitely will not remember, 4 = definitely will remember). Trials were separated by a 4s ISI and participants viewed each face-name pair twelve times, evenly split across 3 blocks. Prospective memory judgments were included to facilitate participant engagement with the observational learning task and were not considered further. Family names were identical to Experiment 1 and shared across faces whereas first names were unique to each face. While the inclusion of face-specific first names required participants to differentiate individual faces, the inclusion of

the shared family names provided an opportunity to form links between related faces. We designed the task to determine to what degree experiences that overlap in content (here, last name) tend to affect perception and be related in memory, bridging traditional category research with research on generalization through episodic inference (Schlichting & Preston, 2015; Zeithamova, Dominick, et al., 2012). However, we subsequently discovered similarities between our task and a study by Medin, Dewey, and Murphy, (1983). In Medin et al. (1983), participants also learned first and shared last names of faces but under a feedback-based categorization paradigm rather than a paired-associate observational paradigm. Because our task did not employ feedback-based learning, participants were not provided with cues as to the number of first names or surnames. The fact that family names were repeated across faces or that there was a category structure among faces was not explicitly emphasized to participants. This allowed us to see if we could replicate results from Experiment 1 under very different conditions, in a task that does not resemble traditional category learning and where category information is present but not emphasized.

Post-learning similarity ratings. Perceived similarity ratings were repeated after the learning phase with the same timing as pre-learning ratings. Of main interest was a potential category bias in perceived similarity, i.e., whether faces that shared a parent would be rated as more similar when they had the same family name than when they had different family names.

Cued recall of face-name associations. Experiment 2 included a self-paced cued-recall task of face-name associations. Participants viewed each training face individually on a computer screen and handwrote the full name of each face on a sheet of

paper. Participants advanced the trials at their own pace but were not able to skip faces or go back and look at faces already named. Participants were encouraged to make their best guess as to the first and family names of each face even if they were not confident in their memory.

Generalization phase. As the last phase of both Experiments, category knowledge was tested directly using categorization of old and new faces. In addition to the nine training faces, participants categorized 42 never-seen faces, consisting of 14 new blends of each of the three category-relevant parent faces. Participants were asked to select via button press the family name for each face, which were presented individually for 4s, from the three options (Miller, Wilson, Davis) presented on the screen. Trials were separated by an 8s ISI. No feedback was provided, and participants were encouraged to make their best guess when unsure of family membership.

Results

Learning Phase

Experiment 1: Feedback-based category learning. Overall percent correct across training was 76% (SD = 14%), which was well above chance (33% for three categories; one-sample $t(34) = 17.66$, $p < .001$, $d = 3.01$). Categorization accuracy improved across training, from 66% in the first half to 85% in the second half ($t(34) = 9.72$, $p < .001$, $d = 1.63$), demonstrating learning over time.

Experiment 2: Observational learning of full name—face associations. Observational learning provided no measure of accuracy from the learning phase. Therefore, in Experiment 2 a cued-recall task was included to assess how well

participants learned the face-full name pairs. Participants recalled on average 52% of first names and 65% of family names.

Similarity Ratings

We compared mean face similarity ratings in each pair-type (shared parent-same family name, shared parent-different family name, not related) using repeated-measures ANOVA. Analyses were performed separately in each phase (pre-learning, post-learning). We also assessed learning-related rating changes by comparing ratings across phases. For all ANOVAs, a Greenhouse-Geisser correction for degrees of freedom (denoted as *GG*) was used wherever Mauchly's test indicated a violation of the assumption of sphericity.

Experiment 1. Pre-learning ratings (Fig. 2.2A) demonstrated that participants were sensitive to the physical similarity structure introduced with the face-blending procedure. A one-way, repeated measures ANOVA showed a significant effect of pair type ($F(2, 68) = 58.74$, $p < .001$, $\eta_p^2 = .63$), driven by lower perceived similarity for faces that did not share a parent compared to those that shared a parent (with or without shared family name, both $t > 9.17$, $p < .001$, $d > 1.50$). Faces that shared a parent were perceived as equally similar to one another irrespective of whether they also shared the same—not yet presented—family name ($t(34) = -0.17$, $p = .87$, $d = 0.03$).

Post-learning ratings (Fig. 2.2B) revealed a category bias on perceived similarity: pairs of faces sharing a parent and family name were perceived as significantly more similar than faces that shared a parent but not a family name ($M_{diff} = 0.72$, $SD_{diff} = 1.41$, $t(34) = 3.02$, $p = .005$, $d = 0.51$). Faces that shared a parent remained rated as more similar than unrelated faces (both $t > 6.85$, $p < .001$, $d > 1.15$).

To further test the effect of learning, we conducted a 2 x 3 (timepoint [pre-learning, post-learning] x pair-type [shared parent-same family name, shared parent-different family name, not related]) repeated-measures ANOVA. There was no main effect of timepoint ($F(1, 34) = 0.04$, $p = .85$, $\eta_p^2 = .001$). There was a significant main effect of pair-type ($F(1.63, 55.38) = 61.21$, $p < .001$, $\eta_p^2 = .64$, *GG*), and a significant interaction between timepoint and pair-type ($F(1.64, 55.88) = 11.85$, $p < .001$, $\eta_p^2 = .25$, *GG*). Follow-up pre-post comparisons within each pair-type (Fig. 2.2C) revealed that this interaction was driven by both a significant *increase* in similarity ratings for faces sharing a parent and a family name ($t(34) = 3.02$, $p = .005$, $d = 0.51$) and a significant *decrease* in similarity ratings for faces only sharing a parent but not a family name ($t(34) = -2.33$, $p = .026$, $d = -0.39$). There was no significant change in similarity ratings for faces that did not share a parent ($t(34) = -0.18$, $p = .86$, $d = -0.03$).

Experiment 2. As in Experiment 1, participants were sensitive to the face similarity structure. Pre-learning similarity ratings (Fig. 2.2E) differed significantly among pair types ($F(1.46, 55.47) = 72.22$, $p < .001$, $\eta_p^2 = .655$, *GG*), driven by lower perceived similarity of faces that did not share a parent compared to faces that shared a parent (with and without shared family names, both $t > 10.65$, $p < .001$, $d > 1.70$). For faces that shared a parent, ratings did not significantly differ when face pairs had the same or different—not yet presented—family names ($t(38) = 1.82$, $p = .077$, $d = 0.29$). A category bias was found in post-learning ratings (Fig. 2.2F) with pairs of faces sharing a

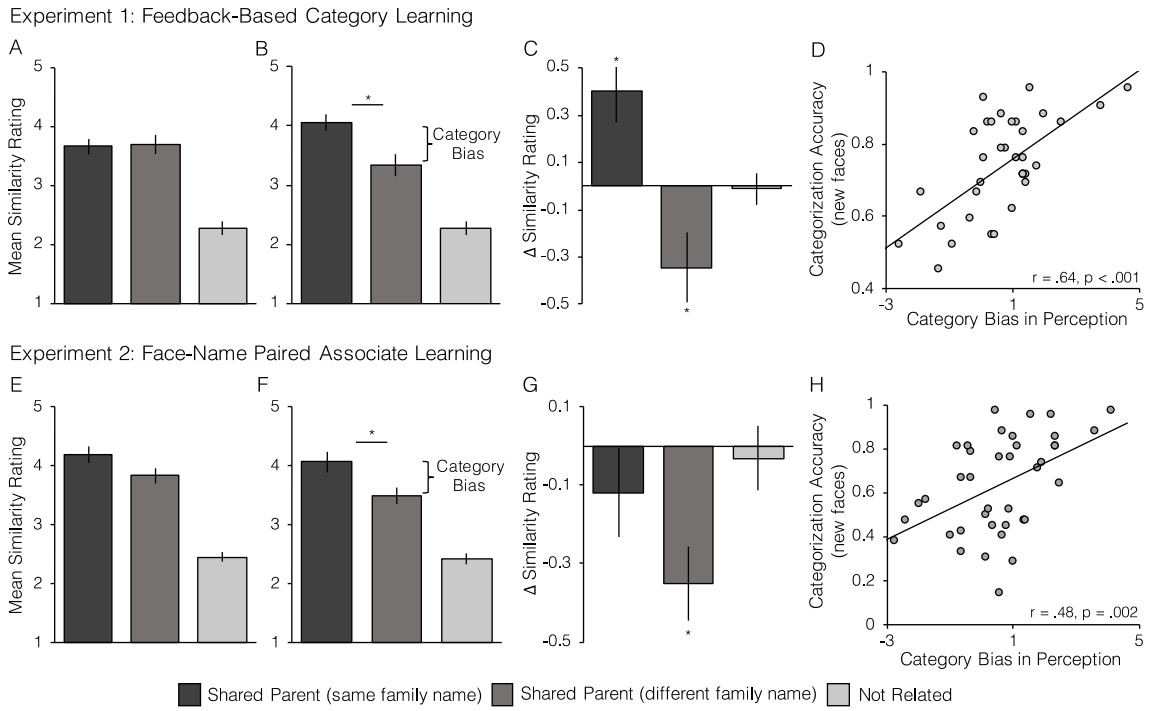

Figure 2.2. Behavioral results for traditional category and paired associate learning. Top panel are results from the traditional category learning experiment. Bottom panel (shaded grey) are results from the face-name paired associate learning experiment. ***A & E.*** Average similarity ratings for faces that share a parent and family name, faces that only share a parent, and faces that don't share any parents before learning. ***B & F.*** Average similarity ratings for the same pairwise comparisons after learning. Asterisk represents a significant ($p < .05$) difference in post-learning similarity ratings for faces that belong to the same family vs. faces that share physical similarity but belong to different families (i.e. a category bias in perception). ***C & G.*** Changes in similarity ratings from pre- to post-learning. Asterisk denotes significant ($p < .05$) increases and decreases in perceived similarity for faces. ***D & H.*** Positive relationship between indirect (category bias in perception) and direct (categorization accuracy for new faces) measures of memory generalization.

parent and family name perceived as significantly more similar than faces that shared a parent but not a family name ($M_{diff} = 0.58$, $SD_{diff} = 1.52$; $t(38) = 2.39$, $p = .022$, $d = 0.38$).

Testing the effect of learning, the 2 x 3 (timepoint x pair-type) repeated-measures ANOVA revealed a significant main effect of timepoint ($F(1, 38) = 5.20$, $p = .028$, $\eta_p^2 = .120$), with overall similarity ratings being lower post-learning than pre-learning ($M_{pre} = 3.49$, $SD_{pre} = 0.51$; $M_{post} = 3.33$, $SD_{post} = 0.59$; $t(38) = -2.28$, $p = .028$, $d = 0.37$). There was also a significant main effect of pair-type ($F(1.28, 48.60) = 60.42$, $p < .001$, $\eta_p^2 = .614$, *GG*), and a significant interaction between timepoint and pair-type ($F(1.67, 63.37) =$

4.21, p = .03, η_p^2 = .10, *GG*). Follow-up pre-post comparisons within each pair-type (Fig. 2.2G) revealed that the interaction was driven by a significant *decrease* in similarity ratings for faces sharing a parent but not a family name (t(38) = -3.71, p = .001, d = -0.59), but there were no significant changes in similarity ratings for other pair-types (both t < -1.04, p > .30, d < -0.18). Thus, changes in perceived similarity were affected by category membership in both experiments.

Although not significant (p = .077), we noted a numerical tendency towards a category bias in pre-learning similarity ratings. Parent faces were randomly selected for each participant to serve as category-relevant or category-irrelevant parents, but some of the category-relevant parent faces may have been more salient, leading to a numerically greater pre-learning similarity rating. Thus, we tested whether the post-learning category bias on perceived similarity was reliably greater than pre-learning bias. A 2 x 2 (timepoint [pre-learning, post-learning] x pair-type [shared parent-same family name, shared parent-different family name]) repeated-measures ANOVA showed only a marginal interaction between timepoint and condition (F(1, 38) = 2.87, p = .098, η_p^2 = .07). We thus controlled for pre-learning similarity rating differences in subsequent analyses that assessed the relationship of post-learning ratings and generalization performance.

Category Generalization

Experiment 1. Participants correctly categorized 85% of training faces (SD = 17%) and 74% of new faces (SD = 13%), which was well above chance (.33 for three categories; both one-sample t(34) > 18.12, p < .001, d > 3.06). A paired-samples t-test showed higher categorization accuracy for the training faces than for the new faces (t(34)

= 5.48, $p < .001$, $d = 0.93$). We next tested whether the category bias on perceived similarity ratings (an indirect measure of category knowledge) was related to subsequent generalization success. A Pearson's correlation showed a significant positive relationship between the category bias on perceived similarity ratings and generalization accuracy ($r(33) = .64$, $p < .001$; Fig. 2.2D). The category bias on perceived similarity in the post-learning phase was a significant predictor of subsequent generalization performance even when pre-learning similarity ratings were considered (multiple regression: pre-learning differences in perceived similarity $\beta = .30$, $t(34) = 1.80$, $p = .08$; post-learning category bias $\beta = .46$, $t(34) = 2.75$, $p = .01$).

Experiment 2. Participants correctly categorized 70% of training faces (SD = 23%) and 64% of new faces (SD = 22%), which was well above chance (.33 for three categories; both one-sample $t(38) > 8.65$, $p < .001$, $d > 1.38$). A paired-samples t-test showed higher categorization accuracy for the training faces than for new faces ($t(38) = 2.12$, $p = .04$, $d = 0.34$). The post-learning category bias on perceived similarity ratings was significantly correlated with generalization accuracy (Pearson's $r(37) = .48$, $p = .002$; Fig. 2.2H). Further, the category bias was a significant predictor of subsequent generalization performance even when pre-learning similarity ratings were controlled for (multiple regression: pre-learning category bias $\beta = -.22$, $t(38) = -0.86$, $p = .40$; post-learning category bias $\beta = .66$, $t(38) = 2.57$, $p = .01$).

Discussion

The current study investigated category learning using measures of perceived similarity and category generalization across two experiments. Face-blend stimuli were used to control physical similarity within and across categories (families). Experiment 1

was a traditional feedback-based category-learning task, with three family names serving as category labels. In Experiment 2, the shared family name category label was encountered in the context of a face-full name paired-associate learning task, where first names were unique for each face. Participants were able to successfully apply category labels to new faces in both experiments, demonstrating that category information can be extracted in support of generalization even when task goals do not emphasize learning categories at encoding. Past work of incidental category learning has shown that individuals can extract category structures when not instructed using patterns of physical similarity as category cues (Aizenstein et al., 2000; Love, 2002; Reber et al., 2003; Wattenmaker, 1993). We extend these prior findings by showing that category structure can also be extracted when category membership is dissociable from physical similarity and further when individuals are actively learning information that differentiates individual items *within* the same category.

Learning-related changes in perceived similarity ratings were observed in both experiments. In both cases, following learning, participants rated faces sharing a category label as more similar than equally physically similar faces that did not share a category label. These results extend prior studies finding changes in perceived similarity as a result of explicit category learning (Goldstone, 1994b, 1994a; Livingston et al., 1998) to a novel task that exposes participants to a category label but requires individuation of stimuli within a category. Observing category bias after the face-name paired-associate learning also indicates that the mere presence of a shared piece of information can bias perception even outside the context of a traditional category-learning task.

The current results also indicate that similarity ratings provide a useful tool to index category knowledge while minimizing explicit generalization demands. In both experiments, category bias in similarity ratings observed after learning predicted subsequent generalization of category information to new examples. To our knowledge, this is the first study relating the strength of a perceptual category bias to the quality of learned category information (as measured by generalization success). The finding that good category generalizers were those who showed the greatest distortion in perceptual representations (rather than those with representations better aligned with physical similarity) is consistent with the view that category bias in perception results from learning-related attentional shifts and differential weighting of perceptual features based on their category relevance (Goldstone & Steyvers, 2001; Kruschke, 1996; Medin & Schaffer, 1978; Nosofsky, 1991; Nosofsky, 1986). Our findings tie together research on categorical perception and concept generalization, and newly indicate that perceived similarity ratings reflect the quality of new category knowledge robustly across two distinct tasks involving category learning.

Interestingly, while perceptual biases occurred in both experiments, they took different forms. In Experiment 1, similarity ratings for faces within a family increased while similarity ratings for faces that were physically similar but belonged to different families decreased. These results provide a new example of a category structure in which both within-category compression and between-category expansion are observed after traditional feedback-based category learning (Gurekis & Goldstone, 2008; Goldstone, Lippa & Shiffrin, 2001), and aligns well with the task demands of treating some stimuli as distinct and some as equivalent. Based on prior work on attentional shifts after

category learning (Goldstone & Steyvers, 2001; Kruschke, 1996; Nosofsky, 1991), this result indicates that participants both focused more strongly on features that differentiate between categories (features of the relevant parent faces) and decreased attention to features that do not differentiate between categories (features of the irrelevant parent faces that affected physical similarity of faces but not family membership).

In contrast, the changes in perceived similarity after the face-name paired-associate learning in Experiment 2 were primarily driven by decreased similarity for faces that were physically similar but belonged to different families. We did not observe increases in perceived similarity ratings for faces belonging to the same family. While more difficult category structures are thought to trigger within-category compression (Pothos & Reppa, 2014), this does not explain differences observed here as category structure was the same across experiments and category learning was easier rather than more difficult in Experiment 1, where compression was observed. Rather, we suspect that learning goals at encoding drove the differences in the pattern of category bias between experiments. Although it is not possible to rule-out a contribution from other factors, such as feedback-based vs. observational learning, the goal of learning a full name for each face (including the unique first names) in the paired-associates task was likely a key factor. It required participants to look for differences between *all* faces, even faces within the same family, in order to differentiate between categories as well as between "brothers" within the same family. That meant that all features remained relevant for task goals in Experiment 2, as the features of category-irrelevant parent faces were important for discriminating two members of the same family, such as differentiating Brad Miller

from Ryan Miller. Thus, participants could not simply ignore the category-irrelevant features as they could in Experiment 1.

Notably, the category bias was measured *after* learning but *before* the explicit generalization test, meaning that the category bias was present prior to explicit generalization demands. Yet, the presence of a shared piece of information (same last name) was sufficient to affect how faces became represented, even in Experiment 2 where no features were irrelevant for the task at hand. This finding is consistent with the notion that people spontaneously link related episodes into an integrated representation at encoding (Shohamy & Wagner, 2008; Zeithamova, Dominick, & Preston, 2012) rather than in response to explicit generalization demands (Banino et al., 2016; Carpenter & Schacter, 2017, 2018). As a strategic decision to rate faces with the same last name as more similar can contribute to biases in similarity ratings (Goldstone, 1994b; Goldstone et al., 2001), we cannot definitively attribute our findings to spontaneous integration during learning. However, our results *do* indicate that evidence for the formation of category knowledge can be demonstrated even when generalization task demands are greatly minimized, and outside of a traditional category learning task. The nature of the resulting category representations—such as whether they are exemplar-based (Hintzman, 1986; Medin & Schaffer, 1978), prototype-based (Homa, Cross, Cornell, Goldman, & Shwartz, 1973; Posner & Keele, 1968), or cluster-based (Love & Medin, 1998)—cannot be resolved in the current study as any model of category learning that postulates learning-related attentional shifts would predict the emergence of perceptual biases.

In summary, we build on long lines of research on category learning (*for reviews see* Ashby & Maddox, 2011; Seger, 2008) and categorical perception (*for reviews see*

Goldstone & Hendrickson, 2010; Harnad, 2006) by demonstrating that category bias in perception reflects the quality of learned category knowledge. We further extend prior work beyond traditional category learning, to demonstrate perceptual biases and successful generalization even after learning that emphasizes individuation of category members, with the specific pattern of learning-related perceptual shifts reflecting goals during learning. Lastly, relating our results to hypotheses generated from studies of episodic inference, our data align with the notion that individuals may spontaneously link related information at encoding, prior to explicit demands to generalize.

Open Practices

None of the experiments discussed in the current report were preregistered. Data and materials for all experiments are freely available in the *Blended-Face Similarity Ratings and Categorization Tasks* repository on the Open Science Framework (https://osf.io/e8htb).

CHAPTER III

CATEGORY-BIASED NEURAL REPRESENTATIONS FORM SPONTANEOUSLY DURING LEARNING THAT EMPHASIZES MEMORY FOR SPECIFIC INSTANCES

This chapter contains unpublished co-authored material. The graduate student is the primary author of this chapter with input from her adviser Dasa Zeithamova (second author). The graduate student contributions to this chapter include task design, data collection, all data processing and analyses, figure creation, initial drafting of the manuscript, and incorporation of edits based on feedback from the second author.

The ability to link details across our varied experiences and organize them into meaningful clusters of information that can be readily applied in new situations is an important aspect of memory. The organization of memory in service of generalization to new situations has been often studied using category learning paradigms. In traditional category learning tasks, individuals explicitly learn to categorize a set of stimuli and then memory generalization performance is measured through successful transfer of category knowledge to new, never-studied examples. Oftentimes, category learning involves learning which stimulus features are category-relevant (determining category membership) and which features are irrelevant for category membership (Goldstone & Steyvers, 2001; Medin & Schaffer, 1978). Attending to category-relevant information while discarding category-irrelevant information has been shown to bias perception after learning such that items within the same category are perceived as more similar while items across categories are perceived as less similar to one another (S. R. Ashby, Bowman, & Zeithamova, 2020; Beale & Keil, 1995; Goldstone, 1994a; Goldstone et al., 2001; Kurtz, 1996).

Traditional category learning approaches have been fruitful for understanding how memory generalization proceeds when task goals emphasize learning generalizable information. However, there are other situations where information about category membership is present, but task goals instead require differentiation of individual members of a category from one another. We recently examined whether individuals would extract category information and display a perceptual category bias when stimuli shared generalizable information but task instructions emphasized memory for individual stimuli (Ashby, Bowman, & Zeithamova, 2020). Using a paired associate learning task, participants learned face-full name associations for face-blend stimuli. Face-blends were created by morphing together never-studied "parent" faces resulting in increased physical similarity for faces that shared a parent. Some faces that were physically similar were then assigned a shared family name (belonged to the same category) while other faces that were physically similar had different family names, allowing us to dissociate the effects of physical similarity from category membership. Each blended face stimulus was also paired with a unique first name and the instructions emphasized learning a full name for each face. After the paired-associate learning, participants showed a category-bias in perceptual similarity ratings of the face-blends, where faces with the same last name were rated as more similar than faces that were physically equally similar but had different family names. This indicated that category-relevant information is still extracted even when task goals at encoding emphasize learning of individual items. Further, we found that the category-bias in perception measure predicted performance on a subsequent categorization test of never-studied face-blends, indicating that category-bias in perception may be a good index of the extent of category learning.

The category-bias in perception was measurable after learning, but prior to an explicit generalization test, providing behavioral evidence that category information was extracted in service of generalization prior to the explicit demand to generalize, and even when task goals directed individuals to learn face-specific information. While this data suggests that participants formed category representations spontaneously during encoding, we cannot rule out the possibility that the act of making similarity judgments after category learning carries inherent strategic cues to rate same-category items as more similar to one another. Thus, in the current study, we set out to utilize neural evidence to determine whether representations that form at encoding already reflect category information. Whether related memories are linked in service of memory generalization spontaneously during encoding or in response to task demands at retrieval is an active area of debate within the literature (for a review see Zeithamova & Bowman, 2020). Some argue that individual memories are stored at encoding and memories are only related to one another on-the-fly at retrieval in response to generalization demands (Banino et al., 2016; Carpenter & Schacter, 2017, 2018). Others argue that overlap between events leads to reactivation of prior related memories during learning, resulting in the spontaneous formation of an integrated memory that links related experience as they are encoded (Cai et al., 2016; Gershman, Schapiro, Hupbach, & Norman, 2013; Shohamy & Wagner, 2008). Thus, the first goal of this study was to test whether related faces are spontaneously linked to extract category knowledge *before* any explicit generalization demands. To achieve this goal, we measured neural representations of individual face stimuli using functional MRI and pattern information analyses to test for

the presence of category information and category bias in neural representations during encoding of face-full name associations.

A second question we had was where in the brain category-biased representations may spontaneously form. Several regions have been identified to support organization of related memories in service of generalization. The ventromedial prefrontal cortex (VMPFC) has been shown to integrate new information while taking into account prior memories (van Kesteren et al., 2013). Learning-related interactions between the anterior hippocampus (AHIP) and the VMPFC have also been shown to support integration across memories (Schlichting et al., 2015; Zeithamova, Dominick, et al., 2012), and abstract category representations in AHIP and VMPFC support the transfer of concept information to new examples (Kumaran, Summerfield, Hassabis, & Maguire, 2009). Additionally, portions of the lateral temporal cortices, in particular the middle temporal gyrus (MTG), have been implicated in semantic memory (Mummery et al., 2000) and gist representations (Dennis et al., 2008; Turney & Dennis, 2017). VMPFC, MTG and AHIP have been also shown to represent abstract category knowledge in an explicit categorization task where all stimulus features were category-relevant and jointly determined category membership (Bowman, Iwashita, et al., 2020; Bowman & Zeithamova, 2018). However, it is unknown whether these regions also would also reflect a representational shift of individual stimuli based on feature relevance to align with their category membership and in a paradigm where explicit task goals require attending to category-irrelevant information for successful learning of full names for individual stimuli.

Alternatively, as behavioral category bias in perception is thought to be driven by attentional shifts towards category-relevant and away from category-irrelevant features (Nosofsky, 1986, 1991), neural reflections of such category biases may not be localized to putative generalization regions. In one study that investigated how neural representations align with learned attentional bias during categorization, Mack, Preston, and Love (2013) found relatively widespread evidence for attention-biased neural representations after category learning across the brain, including lateral occipital cortex, posterior parietal cortex, and lateral prefrontal regions. Recruitment of prefrontal regions has been reported in other studies of category learning (Nosofsky, Little, & James, 2012; Seger et al., 2000) and left dorsolateral prefrontal activity was found in individuals that showed a larger degree of category knowledge (Seger et al., 2000). It has been proposed that greater dorsolateral activity may reflect attentional processes that guide examining features and making decisions as to whether or not features are category diagnostic (Seger et al., 2000). Furthermore, attention is known to have widespread effect on neural processing across the brain, from high-level cognitive regions to sensory cortices (Hämäläinen, Hiltunen, & Titievskaja, 2002; Kanwisher & Wojciulik, 2000; Olson, 2001). Thus, if the presence of a category label during face-name learning results in a spontaneous attentional shift towards category-relevant features, then we may observe category-biased neural representations widespread across the cortex.

Method

Participants

Forty-four healthy participants were recruited from the University of Oregon and surrounding community via the university SONA research system and community fliers.

Participants received monetary compensation for their participation ($10/hr outside the scanner and $20/hr inside the scanner). All participants provided written informed consent, were right-handed, native English speakers, and were screened for neurological conditions and medications known to affect brain function. Experimental procedures were approved by Research Compliance Services at the University of Oregon. Four participants were excluded from analyses: two for movement in excess of 1.5mm frame-wise displacement within a run, one due to operator error resulting in poor data quality, and one for having an undisclosed migraine disorder and subsequent migraine headache in the middle of the scanning session. The remaining sample of 40 participants (22 female, 18 male; age 18-30 years; M_{age} = 21.33, SD_{age} = 2.92) are reported in all analyses.

Stimuli

Stimuli were grayscale images of blended faces that we previously developed and made publicly available (OSF Repository: https://osf.io/e8htb/; see also Ashby, Bowman & Zeithamova, 2020). The stimulus set comprises of a pool of 20 face photographs (so called "parent" faces, never shown to the participants in our study) and all 190 pairwise computer blends of those 20 parent faces.

Training stimuli. To create the training blended faces, 6 parent faces were randomly chosen for each participant, three of them assigned as category-relevant and three assigned as category-irrelevant. Each of the three category-relevant parent faces were individually morphed with each of the three category-irrelevant parent faces, with equal weight given to each parent face (50/50 blend; see Figure 3.1). The resultant nine face-blends were then used as stimuli in the learning task, with faces sharing a parent face being physically more similar than faces that did not share a parent face. Faces that

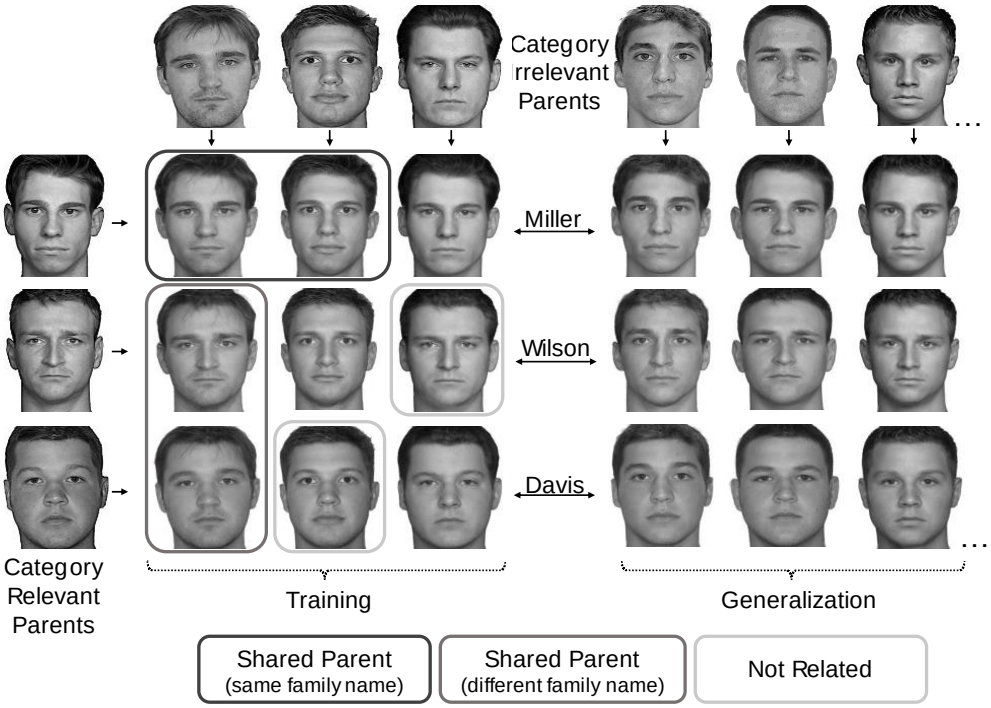

Figure 3.1[*]. Structure of face-blend stimuli. Parent faces on the leftmost side are designated "category relevant parents" as these parents determined family membership—Miller, Wilson, or Davis—during learning, recognition, and generalization. Parent faces across the top are designated "category-irrelevant parents" as these parents introduced physical similarity across families but did not determine categories. Three category-irrelevant parents were used for learning. The rightmost three category-irrelevant parents are a subset of new faces used for generalization. Parent faces were never viewed by participants, only the resulting blended faces. The face blending procedure produced pairs of faces that shared a category-relevant parent and belonged to the same family (shared parent - same family name; example indicated with dark grey box), pairs of faces that shared a category-irrelevant parent and belonged to different families (shared parent- different family name; example indicated with medium grey box). Non-adjacent pairs did not share a parent and were not related (example indicated with light grey boxes).

[*]*Figure is adapted with permission from Ashby, Bowman, & Zeithamova 2020.*

shared a category-relevant parent also shared a family name (belonged to the same category) while faces that shared a category-irrelevant parent had different family names. Thus, using blended-faces provided us with realistic-looking face stimuli while allowing us to control within- and between-category similarity. Because pilot data indicated that some parent faces were more distinct and thus more prominent in the resulting blend while other faces were more average and thus less prominent in the resulting blend, we took two additional steps not implemented in our prior work to better equate pre-learning perceived similarity of the face-blends that shared a parent within and between

categories. First, we limited the pool of possible parent faces for the creation of the training stimulus set to 10 faces (from the full set of 20) that were of intermediate distinctiveness based upon an item analysis of pre-learning similarity rating data that we collected through pilot testing and previously published studies (see Ashby et al., 2020; Bowman, Ashby, & Zeithamova 2021). Second, we implemented a yoking procedure between subjects so that two participants were assigned the same parent faces with reversed category-relevant and category-irrelevant parent designation. This ensured that if one parent face happened to have more salient features, it would be equally frequently assigned as a category-relevant parent or a category-irrelevant parent.

Test stimuli. In addition to the nine training stimuli, 52 new face-blend stimuli were created for subsequent old/new recognition test and a surprise generalization test. To create new test stimuli, the three category-relevant parent faces were blended with 14 new parent faces (all parent faces not used for training stimuli) resulting in 14 new face-blends per category.

Experimental Design

The experiment consisted of the following phases (Figure 3.2): initial exposure (passive viewing), pre-learning similarity ratings, observational learning of face-full name associations (scanned), post-learning similarity ratings, cued-recall of face-name associations, old/new recognition test (scanned), and category generalization (scanned). Only the fMRI data from the observational learning phase were analyzed for the purpose of the current paper, testing for the formation of category-biased neural representations when task goals emphasize face-specific information.

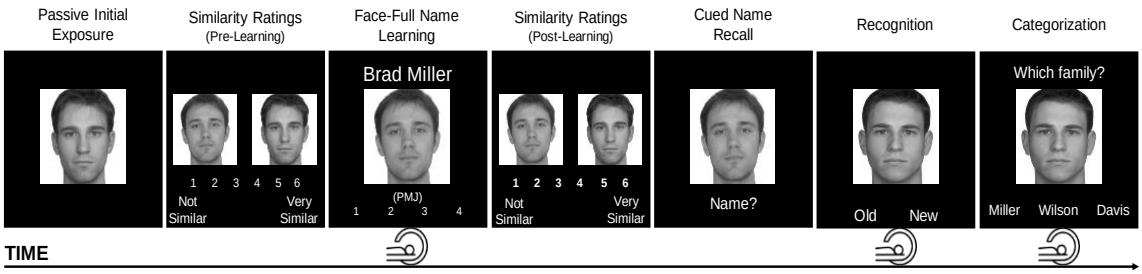

Figure 3.2. Full Imaging Procedure. Participants passively viewed the 9 training faces and rated the subjective similarity of all 36 pairwise comparisons of the training faces prior to entering the scanner. Face-full name learning was scanned and completed in four runs. Anatomical scans were collected during post-learning similarity ratings to minimize time spent in scanner. Cued name recall was completed with participants communicating their answers to researchers verbally through the scanner intercom system. The recognition phase was scanned and consisted of 51 trials (9 old and 42 new faces) split into three runs. The categorization phase was also scanned and used the same faces as the recognition phase and was also split into three runs.

Passive viewing. Prior to entering the scanner, participants first passively viewed each of the nine training stimuli individually, once in random order without any labels and without making any responses. Face-blends were shown for 3s with a 1s inter-stimulus-interval (ISI). This was done to familiarize participants with the stimuli, minimize novelty effects during the learning phase, and provide participants with an estimate of the degree of similarity between all faces prior to collecting the pre-learning perceptual similarity ratings.

Pre-learning similarity ratings. Prior to entering the scanner, participants rated the subjective similarity of all pairs of training faces. This allowed us to verify that participants were sensitive to the inherent similarity structure among faces introduced by the blending procedure. All possible 36 pairwise comparisons of the 9 training faces were presented and participants rated the subjective similarity of the two faces on a scale from one to six (1 = the two faces appeared very dissimilar, 6 = the two faces appeared very similar). The face pairs and the rating scale were presented simultaneously for 5s with a 1s ISI. For subsequent analyses face pairs were binned into three conditions depending on whether they 1) shared a parent and a family name, 2) shared a parent but did not share a

family name, or 3) did not share a parent (see example pairs in Figure 3.1). Because there are 9 pairs of faces that share a relevant parent, 9 pairs of faces that share an irrelevant parent, and 18 pairs of faces that do not share a parent, we presented the 9+9 pairs of faces with shared parents twice, with counterbalanced left-right position of the two faces.

Observational learning of face-full name associations (scanned). Participants were next placed in the MRI machine and scanned during learning of the face-full name associations across 4 training runs. During learning, participants studied a face-full name pair for 3s and then made a prospective memory judgement on a scale from one to four (1 = definitely will not remember, 4 = definitely will remember) for 2s. Prospective memory judgments were included to encourage participant engagement with the observational task and were not considered further. All trials were separated by a 3s ISI. Each face-full name pair was studied 3 times per run for a total of 12 exposures across all of learning. Family names (Miller, Wilson, Davis) were shared across faces that shared a category-relevant parent face. Nine unique first names (Brad, John, Paul, Steve, Tyler, Andy, Ryan, Kyle, Eric) were randomly assigned to each face. This structure allowed for participants to differentiate individual faces, even within the same family, while also providing an opportunity to form links between related faces in service of memory generalization. Participants were instructed to learn each individual's full name and repetition of family names across faces or the presence of any category structure was not explicitly emphasized to participants.

Post-learning similarity ratings. Post-learning perceived similarity ratings were collected in the scanner while anatomical data was collected (see fMRI data acquisition

below). Timing and presentation of face-pairs was identical to the pre-learning similarity rating procedure, in a new random order.

Cued recall of face-name associations. To assess learning success, participants completed cued-recall of the face-full name associations. During this recall phase, participants viewed each of the nine training faces individually for as much time as needed while still lying in the scanner. Participants were instructed to vocalize the first and last name of each face and the researcher, listening through the scanner intercom system, recorded their responses on a piece of paper. Trials were advanced by the researcher at the request of the participant. Participants were encouraged to make their best guess as to the full names of the faces even if they were not confident in their memory.

Recognition (scanned). An old/new recognition test was also used as another learning performance metric for the individual faces. In addition to the nine training faces, participants were exposed to 42 never-seen faces that consisted of the 14 new blends of each of the three category-relevant parent faces. Participants were asked to select via button press whether or not the face presented was old—meaning it was a face they had already studied while in the scanner—or new. No feedback was given. The 51 trials were split into 3 runs of 17 trials each (each run contained 14 new and 3 old faces) and each trial was presented for 4s with an 8s ISI. Imaging data from the recognition phase were not considered further in the current report.

Generalization (scanned). Lastly, category knowledge was directly tested using categorization of old (training) and new face blends. New face blends were the same as those used in the recognition phase. Participants were asked to select via button press the

family name for each face from the three options (Miller, Wilson, Davis) presented on the screen. No feedback was provided. The 51 trials were split into 3 runs of 17 trials each (14 new and 3 old faces) with 4s trials and an 8s ISI. Imaging data from the categorization phase were not considered further in the current report.

fMRI Data Acquisition

Imaging data was collected using a 3T Siemens MAGNETOM Skyra scanner at the University of Oregon Lewis Center for Neuroimaging using a 32-channel head coil. Foam padding was used around the head to minimize head motion. The scanning session started with a localizer SCOUT sequence followed by four functional runs of the learning task, and three functional runs each of the recognition and generalization tasks using a multiband gradient echo pulse sequence [TR = 2000 ms; TE = 26 ms; flip angle = 90 º; matrix size = 100 x 100; 72 contiguous slices oriented 15º off the anterior commissure-posterior commissure line to reduce prefrontal signal dropout; interleaved acquisition; FOV = 200 mm; voxel size = 2.0 x 2.0 x 2.0 mm; Generalized Autocalibrating Partially Parallel Acquisition (GRAPPA) factor = 2]. For each task run, 110 volumes were collected for the learning task and 104 volumes each for the recognition and categorization tasks. Only data from the learning phase are presented here. A standard high-resolution T1-weighted MPRAGE anatomical image [TR = 2500 ms; TE = 3.43 ms; TI = 1100 ms; flip angle = 7º; matrix size = 256 x 256; 176 contiguous slices; FOV = 256 mm; slice thickness = 1 mm; voxel size = 1.0 x 1.0 x 1.0 mm; GRAPPA factor = 2] and a custom anatomical T2 coronal image [TR = 13,520 ms; TE = 88 ms; flip angle = 150º; matrix size = 512 x 512; 65 contiguous slices oriented perpendicularly to the main axis of the hippocampal body; interleaved acquisition; FOV = 220 mm; voxel size = 0.4 x 0.4 x

2 mm; GRAPPA factor = 2) were collected to facilitate anatomical localization of the neural signals.

Preprocessing and Single-Trial Modeling

Raw dicom images were converted to Nifti format using MRIcron's (https://www.nitr.org/projects/mricron) dcm2nii function. Functional, behavioral and anatomical data were organized in the Brain Imaging Data Structure (BIDS) format for public dissemination on OpenNeuro (*forthcoming*). Functional images were entered into a single-trial fMRI Expert Analysis Tool (FEAT) model from FSL Version 6 (www.fmrib.ox.ac.uk/fsl). First the functional images were skull stripped using the Brain Extraction Tool (BET) and corrected for within-run motion using MCFLIRT by realigning all volumes to the middle volume. Next, we applied high-pass temporal filtering (60s) and minimal spatial smoothing using a 2mm FWHM Gaussian Kernel. No slice timing correction was applied.

Individual trials were modeled using the GLM including nuisance regressors representing the six, standard motion regressors for rotational and translational motion. A regressor for the individual trial onset times for the training was included in each model and events were modeled with durations of 3s (the period of time the face-name pair was on the screen prior to the prospective memory judgment). This was then convolved with the hemodynamic response function as implemented in FSL (gamma function: phase = 0s, SD = 3s, mean lag time = 6s) resulting in beta weight estimations for each individual trial, for each functional run of the training task. We next concatenated the resultant beta images for each trial across time creating a single betaseries image for each of the four functional runs. Across-run realignment was then applied to the betaseries images for

each run using Advanced Normalization Tools (ANTs; http://stnava.github.io/ANTs/) with the first volume of the fourth run of the training task used as the reference volume. The first volumes of all other task runs were registered to the reference volume and the resulting transformation was applied to the concatenated betaseries images. Lastly, we concatenated all the realigned betaseries images across runs for pattern analyses.

Regions of Interest (ROIs)

Three regions of interest (ROIs) were selected for their hypothesized roles in memory generalization. We selected the VMPFC because of its well established role in supporting memory integration (Schlichting et al., 2014; Zeithamova & Bowman, 2020; Zeithamova, Dominick, et al., 2012), MTG because of its role in semantic and gist memory (Dennis et al., 2008) and our recent findings of its role in category learning (Bowman & Zeithamova, 2018), and the anterior portion of the hippocampus (AHIP) given recent proposals that AHIP (ventral hippocampus in rodents) may be uniquely involved in forming coarser, generalized representations (*for review see* Poppenk, Evensmoen, Moscovitch, & Nadel, 2013).

Three additional ROIs were included as control regions. Because the face-blend stimuli share physical similarity both within and across category boundaries, we chose two visual ROIs that we expected would be sensitive to the physical similarity between face-blends but perhaps not the learned category structure: lateral occipital cortex (LO) and the posterior fusiform gyrus (PFUS). We also explored the posterior hippocampus (PHIP) to test for anterior-posterior dissociation within the hippocampus.

ROIs were defined in each individual participant's native space using the cortical parcellation and subcortical segmentation routines from Freesurfer version 6

(https://surfer.nmr.mgh.harvard.edu/) of the T1-Weighted MPRAGE anatomical image. Bilateral masks for each ROI were created by collapsing together across hemispheres. The VMPFC ROI was defined as the Freesurfer medial orbitofrontal cortex label. To obtain separate AHIP and PHIP regions, we divided the Freesurfer hippocampal ROI at the middle slice. In the event that there were an odd number of hippocampal slices for a given participant, the middle slice was assigned to the posterior hippocampus (PFUS) and not included in the AHIP definition for that participant. All ROI functional analyses were conducted in native space of each participant.

Statistical Analysis

Memory performance for faces and names. To index participants' memory for the individual faces and their names that participants encountered during the paired-associates task, we recorded the proportion of first names and the proportion of last names correctly recalled during the cued recall test. Additionally, we used a measure of corrected hit rate (hits – false alarms) from the recognition task to determine how well participants were able to identify the individual faces encountered during learning. Recognition performance was evaluated using a one-sample t-test comparing corrected hit rate against zero.

Categorization performance. Generalization performance was measured as the accuracy (percent correct) for categorizing new face blends during the surprise categorization task. We also recorded percent correct categorization of the training faces. One-sample t-tests compared categorization performance against chance performance (33.3% for three categories), separately for training faces and for new stimuli. A paired

sample t-test was used to compare categorization performance for the training faces against categorization performance for the new faces.

Similarity ratings. Of main interest from the similarity ratings task was the category bias in perception (similarity ratings for two faces that shared parent and family name – two faces that shared parent but had different family names) from the post-learning similarity ratings. First, we examined perceptual similarity ratings separately for the pre- and post-learning phases. Within each phase we compared mean similarity ratings for faces in each pair-type (shared parent-same family name, shared parent-different family name, not related) using repeated-measures ANOVA. To examine learning-related changes we also compared across phases using a 2x3 (timepoint [pre-learning, post-learning] x pair-type [shared parent-same family name, shared parent-different family name, not related] repeated measures ANOVA. A Greenhouse-Geisser correction for degrees of freedom (denoted as GG) was used wherever Mauchly's test indicated there was a violation of the assumption of sphericity in the data.

Lastly, to determine whether the category-bias in similarity ratings predicts generalization performance (see also Ashby et al., 2020) we used a Person correlation to examine the relationship between the indirect and direct measures of generalization. To confirm that individual differences in pre-learning similarity ratings did not account for this relationship we also used a multiple regression including both the pre- and post-learning category biases in the model as predictors of generalization success.

fMRI classification of category-relevant and category-irrelevant information. Our first approach was to use multi-voxel pattern analysis (MVPA) classification analysis within each a-priori ROI to test to what degree it is possible to decode the category-

relevant and the category-irrelevant parent structure among the training stimuli. Each face-blend seen during category learning contained features shared with other face-blend stimuli with whom it shared the same parent, whether a category-relevant or a category-irrelevant parent. However, it belonged to the same family category only with faces with whom it shares the same category-relevant parent. Thus, we reasoned that if both the category-relevant and category-irrelevant information are decodable in a given region, that may indicate that the region is sensitive to the physical similarity shared between stimuli. In contrast, if a classifier can decode the category-relevant but not the category-irrelevant information in a region, the region primarily represents category information rather than physical similarity. Finally, a classifier may be able to decode both types of information but perform better when decoding category-relevant information compared to category-irrelevant information. This also would indicate category-biased representations during learning.

We predicted that classifier accuracy would be greater for category-relevant compared to category-irrelevant information in regions known to support memory generalization. Further, we predicted above-chance classification of both category-relevant and category-irrelevant information in visual control regions as they should be sensitive to the physical similarity of the faces regardless of the learned category information. Critically, this classification would test whether that category representations are spontaneously formed even when a task emphasizes individuation of individual exemplars and when a category label is present but not emphasized.

To test these predictions, we used PyMVPA (www.pyvmpa.org; see also Hanke et al., 2009) and trained two separate classifiers, one to classify the category-relevant parent

faces and one to classify the category-irrelevant parent faces among the nine training faces. We used a support vector machine (SVM) classifier and a leave-one-run-out cross-validation procedure across all 4 blocks of learning. Classifier success was tested to see if performance was greater than theoretical chance performance (33.3% for three categories) using one-tailed, one-sample t-tests for category-relevant and category-irrelevant classification within each ROI. Differences in classification accuracy for category-relevant vs. category-irrelevant information was examined using paired-samples t-tests within each ROI. All t-tests were corrected for multiple comparisons using the Bonferroni correction.

Neural pattern similarity representations of category information. Our second approach was to use representational similarity analysis (RSA) to directly test for the existence of a category bias in neural representations. Since pairs of faces that share a category-relevant parent and pairs that share a category-irrelevant parent are equated for physical similarity, greater neural pattern similarity for pairs that share a category-relevant parent would demonstrate that learning altered neural representations to reflect a category bias. As with the MVPA approach, we predicted generalization regions, but not necessarily the visual control regions, would demonstrate this neural category bias. To test for the category-biased representations, we first measured the degree of neural pattern similarity using a Pearson correlation within each ROI for all pairs of trials that (a) shared a parent and also shared a family name and (b) shared a parent and had different family names. The resulting R-values were Fisher z-transformed to conform to normality and permit statistical analyses. For each participant and ROI, we then calculated the category bias in neural pattern similarity by subtracting the mean pattern

similarity for the two types of pairs (Shared Parent Same Family Name Similarity – Shared Parent Different Family Name Similarity), and dividing the difference by their variability to quantify the category bias in terms of normalized distance Cohen's D. The pattern of results remains the same when raw (not normalized) similarity differences are used. Category biases in neural representations for each hypothesized generalization ROI were then tested against zero using one-tailed, one-sample t tests to assess if there was greater neural pattern similarity for faces that shared parents and were within the same family compared to faces that shared parents but were from different families.

Searchlight classification of category-relevant and category-irrelevant information. Because the anatomical ROI approach may be insufficient by either including uninformative voxels or excluding informative voxels (Kriegeskorte & Bandettini, 2007; Kriegeskorte, Goebel, & Bandettini, 2006), and because we were interested in how any potential category representations may be distributed across the brain, we also conducted a MVPA searchlight analysis to classify the category-relevant and category-irrelevant parent faces across the entire brain. This allowed for a data-driven approach to discover where in the brain, outside the *a priori* ROIs, category-biased representations may form during learning. The searchlight analysis was completed using a 3mm sphere which then was iteratively swept across the entire brain using PyMVPA producing separate searchlight accuracy maps for category-relevant and category-irrelevant decoding for each subject. Individual subject searchlight maps were then normalized to the standard MNI template space using ANTs. Transformations to standard space were calculated between each subject's reference volume (run 4 of

training) and the standard template and then applied to the searchlight maps for category-relevant and category-irrelevant classification.

Next, individual subject maps in standard space were merged into two 4D maps (one for category-relevant and one for category-irrelevant) and smoothed (Gaussian Kernel: 4mm) in preparation for group-level statistics. In order to compute one-sample t-tests on the merged images to statistically test which regions in the brain represented category-relevant and category-irrelevant information, we first subtracted theoretical chance performance from each merged image (1/3) and then masked the images with the standard MNI template whole-brain mask. Lastly, we used FSL Randomise with Threshold-Free Cluster Enhancement (TFCE) to perform two one-sample t-tests using the category-relevant and category-irrelevant merged, smoothed, and masked images. The resultant t-stat images were then thresholded using the cluster-corrected p-value image to produce maps with only statistically significant clusters. In the event that the TFCE procedure produced a large significant cluster that spanned many functional regions and extended across lobes, we applied an additional voxel-wise threshold (T = 3.5) in order to separate the larger cluster for better characterization of the functional regions evoked.

Searchlight neural pattern similarity representations of category information. We also tested for category-biased neural representations across the entire brain by running an RSA searchlight analysis. As with the MVPA approach, we used a 3mm sphere to iteratively sweep across the entire brain comparing pattern similarity between face stimuli using PyMVPA. The subtraction described in the ROI analysis above were also carried out to produce a searchlight map of category representations for each subject. Searchlight maps were next normalized to the standard MNI template space

using ANTs and transformations were calculated as outlined above in the MVPA searchlight. Next, the individual subject maps in standard space were merged into a single 4D map, smoothed with a 4mm Gaussian Kernel, and masked with a standard MNI template whole-brain mask. Statistical analysis was performed using one-sample t-tests to test against zero and TFCE with FSL Randomise. As with the MVPA classifier searchlight, the resultant t-stat images were then thresholded using the cluster-corrected p-value image to produce maps with only statistically significant clusters representing a neural category bias.

Results

Behavioral

Memory for faces and names. We first examined recall accuracy from the cued-recall task to assess how well participants stayed on task and learned the first and family names during the observational paired-associates learning. On average, participants were able to recall 58% of first names and 65% of family names, similar to our prior behavioral study (52% of first names, 65% of family names, see Ashby et al., 2020). Next, we examined performance for identifying individual faces during the recognition phase as a secondary measure of learning success. We examined performance for identifying faces during the recognition phase as either old or new using a corrected hit rate (hits – false alarms) to account for unequal exposure to old (n = 9) and new training faces (n = 42). We found evidence for good recognition as the average corrected hit rate for participants was 79.5% (SD = 17%) which was well above zero ($t(39) = 29.19$, $p < .001$, $d = 4.616$). The hit rate was 89.1% (SD = 11.1%) and the false alarm rate was 9.6% (SD = 11.2%).

Categorization performance. Next, we examined performance for learning the category-relevant information by assessing categorization accuracy during the surprise categorization task. We examined accuracy separately for the training faces and the never-learned faces. During categorization, participants correctly categorized 69% (SD = 21%) of the old faces that were learned during the observational training and 62% (SD = 18%) of the new faces that were never viewed during learning, which is well above chance (both $t(39) > 10.00$, $p < .001$, $d > 1.58$). A paired-samples t-test showed lower categorization accuracy for the new faces than for the training faces ($t(39) = -3.19$, $p = .003$, $d = .505$). The successful categorization of the new faces into the appropriate family categories indicates that category information extracted during learning was successfully generalized.

Similarity ratings. For our indirect measure of memory generalization, we examined perceptual similarity ratings separately for pre- and post-learning phases. Pre-learning similarity ratings confirmed that participants were sensitive to the similarity structure among stimuli, introduced by the blending procedure (Figure 3.3a). We found a significant main effect of pair type ($F(2, 78) = 96.18$, $p < .001$, $\eta_p^2 = .71$), driven by lower similarity ratings for faces that did not share a parent compared to those that shared a parent (both $t(39) \geq 12.705$, $p < .001$, $d \geq 2.010$). Faces that shared a parent were rated equally similar to one another regardless of whether or not they shared the same family name (which had yet to be presented to participants; $t(39) = -.566$, $p = .574$, $d = .09$).

Post-learning similarity ratings (Figure 3.3b) also differed by pair type ($F(1.67, 65.13) = 91.93$, $p < .001$, $\eta_p^2 = .702$, *GG*), again driven by higher ratings for pairs of faces that shared a parent (category-relevant or irrelevant) compared to faces that did not share

a parent (both $t(39) \geq 12.664$, $p < .001$, $d \geq 2.002$). In contrast to our previous study, we did not find evidence for a category bias in post-learning ratings, as ratings remained comparable between pairs of faces that shared a relevant parent and those that shared an irrelevant parent ($t(39) = 0.211$, $p = .834$, $d = .033$).

A 2x3 (timepoint [pre-learning, post-learning] x pair-type [shared parent-same family name, shared parent-different family name, not related] repeated measures ANOVA showed a significant main effect of timepoint ($F(1,39) = 5.890$, $p = .020$, $\eta_p^2 = .131$) driven by greater perceived similarity ratings before learning compared to after (Figure 3.3c; $t(39) = 2.406$, $p = .020$, $d = .38$) and a significant main effect of pair-type ($F(1.739, 67.806) = 110.575$, $p < .001$, $\eta_p^2 = .739$, *GG*) where faces that shared a parent were rated as more similar than unrelated faces (both $t(39) \geq 14.127$, $p < .001$, $d \geq 2.234$). The interaction between timepoint and pair-type was not significant ($F(2, 78) = .740$, $p = .480$, $\eta_p^2 = .019$).

Although the overall effect of the category bias in post-learning similarity ratings was not significant, in our prior work we found a post-learning category bias in perception that predicted generalization performance (Ashby et al., 2020). Thus, we wanted to examine whether individual differences in the category-bias were still related to performance on the generalization task. We predicted that we would replicate our result from our previous behavioral study finding a positive relationship between the post-learning category-bias in perception and generalization. As predicted, Pearson correlation showed a significant relationship such that larger post-learning category biases in perception were associated with better generalization performance during the categorization task (Figure 3.3d; $r(39) = .57$, $p < .001$). Further, the category bias on

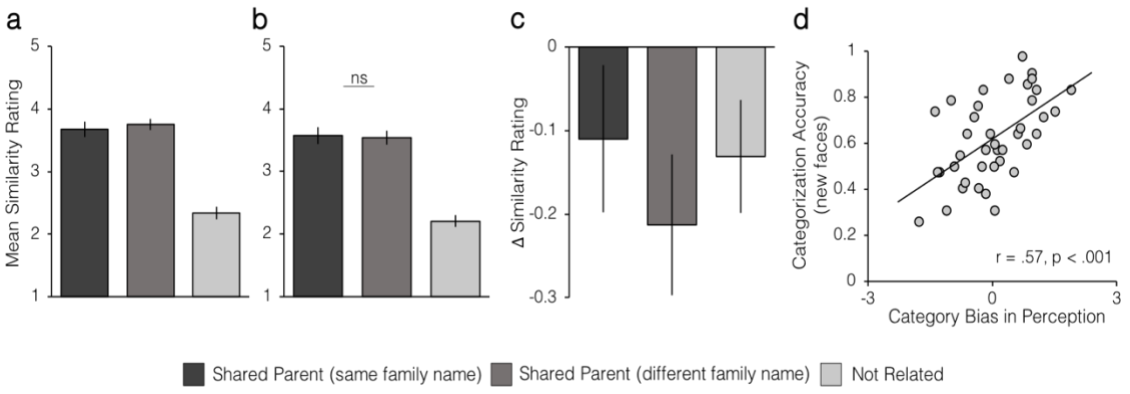

Figure 3.3. Behavioral Category Bias. **A.** Average similarity ratings for faces that share a parent and family name, faces that only share a parent, and faces that don't share any parents before learning. **B.** Average similarity ratings for the same pairwise comparisons after learning. No significant category bias in perception was found averaged across subjects. **C.** Changes in similarity ratings from pre- to post-learning. An overall significant decrease in perceived similarity for faces. **D.** Positive relationship between indirect (category bias in perception) and direct (categorization accuracy for new faces) measures of memory generalization.

perceived similarity post-learning remained a significant predictor of subsequent generalization performance even when pre-learning similarity ratings were considered (multiple regression: pre-learning category bias $b = .137$, $t(39) = 0.69$, $p = .49$; post-learning category bias $b = .47$, $t(39) = 2.36$, $p = .024$). These results successfully replicate our previous work providing further evidence that a learning-evoked category bias in perceptual similarity ratings may be a useful indirect measure of memory generalization when task-related demands to generalize are minimized.

Region of Interest Analyses

Classification of category-relevant and category-irrelevant visual information.

Each face-blend that was viewed during learning contained features that were both category-relevant and category-irrelevant. Our first goal was to determine whether category-biased neural representations are detectable during learning. We predicted that we would find category-biased neural representations that extended beyond the physical similarity of the stimuli by showing neural pattern classification for

category-relevant information to a larger degree than category-irrelevant information. Our second goal was to determine if category-biased neural representations during learning are uniquely represented in putative generalization regions (VMPFC, MTG, and AHIP). We predicted that category-biased representations would be measurable in putative generalization regions but not in control regions.

To test this hypothesis, we first examined classifier performance within putative generalization regions. MVPA classifier performance for decoding category-relevant and category-irrelevant information during learning in each of the a-priori ROIs is presented in Figure 3.4a (left side). Significance for all t-tests was determined by a Bonferroni adjusted alpha level of p = .0167 (α = .05 divided by 3 regions). One-sample t-tests compared classifier accuracy for generalization regions against chance performance (33.3% for 3 categories) revealing significant decoding of category-relevant information in MTG ($t(39) = 3.95$, $p < .001$, $d = .57$), which remained significant after correcting for multiple comparisons

No other generalization regions significantly decoded category-relevant information and none of the three regions decoded category-irrelevant information. To evaluate whether MTG had greater classification accuracy for category-relevant vs. category-irrelevant information, we followed up with a paired-samples t-test to compare decoding accuracies across conditions. We found better decoding performance within MTG for category-relevant information than category-irrelevant information ($t(39) = 2.31$, $p = .013$, $d = .37$, one-tailed) indicating a neural category bias within MTG during learning. Thus, our pattern classification results provide compelling evidence for a neural category bias in MTG, but we did not see significant evidence for category-biased

representations in the other hypothesized generalization regions—although VMPFC

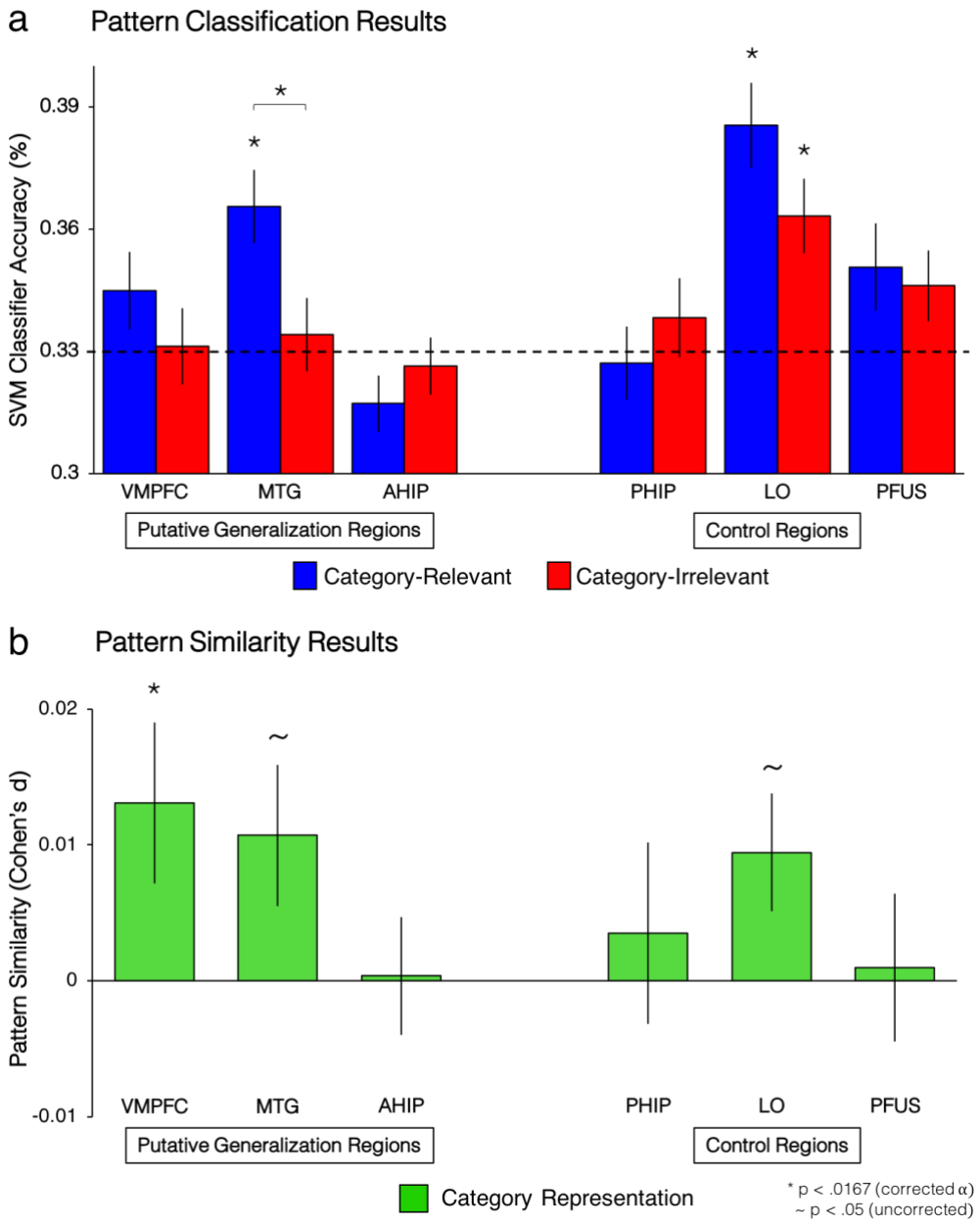

Figure 3.4. Pattern classification and pattern similarity analyses within six a-priori regions of interest. **A.** Mean classifier accuracies across all of paired-associates learning for category-relevant (blue) and category-irrelevant (red) parent face decoding. **B.** Pattern similarity—depicted as effect sizes—for category representations (green) across all of the paired-associates learning. Error bars represent the across-subject SEM. Stars indicate regions where pattern classification was significantly greater than chance (.333 for three categories) and survived Bonferroni correction for multiple comparisons ($p < .0167$). Tildes indicate regions where pattern similarity was significantly greater than zero uncorrected but did not survive correction for multiple comparisons.

followed the same pattern numerically.

Next, we examined classifier performance within the control regions (Figure 3.4a, right side). We predicted that visual control regions (LO, PFUS) would be sensitive to the physical similarity of the faces rather than the learned category information and thus would classify both category-relevant and category-irrelevant information to a similar degree. As PHIP has been shown to be involved in episodic memory we did not have any specific predictions for the patterns of activity we may see in this region during a category-learning task with specificity goals. One-sample t-tests compared classifier accuracy for the control regions against chance performance revealing significant decoding of category-relevant information in LO ($t(39) = 5.31$, $p < .001$, $d = .79$) that survived correction for multiple comparisons. As predicted we also found significant decoding of category-irrelevant information in LO ($t(39) = 3.64$, $p < .001$, $d = .52$), indicating that visual cortex was sensitive to the physical similarity of the faces. While the classification of category-relevant information in LO was greater than category-irrelevant information, the difference did not reach significance ($t(39) = 1.73$, one-tailed uncorrected $p = .046$, two-tailed corrected $p > 0.05$).

Neural pattern similarity representations of category information. Our second approach to testing for category-bias in neural representations was to leverage RSA, determining if neural activity show greater similarity for pairs of faces that shared a category-relevant parent face than pairs of faces that shared a category-irrelevant parent face. We predicted that generalization regions would show significant category representations indicating a learning-driven neural category bias.

We first examined evidence for category bias in neural representations in the hypothesized generalization regions (Figure 3.4b, left side). One-sample t-tests compared differences in pattern similarity against zero revealing significant category representations in VMPFC ($t(39) = 2.21$, $p = .0165$, $d = .35$, one-tailed) and MTG ($t(39) = 2.06$, $p = .023$, $d = .33$, one-tailed). Category representations in VMPFC survived correction for multiple comparisons while category representations in MTG did not ($p > .0167$). Next, we examined category representations in the control regions (Figure 3.4b, right side) and found a significant category representation in LO ($t(39) = 2.18$, $p = .0175$, $d = .34$) which also did not survive correction for multiple comparisons but remained marginal. Overall, two of the hypothesized generalization regions, as well as LO, showed some evidence of category bias in neural representations of individual faces.

Whole-Brain Searchlight Analyses

The ROI-based classification analyses indicated that learning-related category information is measurable during encoding in MTG and LO. The ROI-based pattern similarity analyses further suggest that a neural category-bias may be measurable during learning in VMPFC, MTG, and LO during encoding; however, only representations in VMPFC remained significant once corrected for multiple comparisons. To further test to what degree any potential category-biased representations are unique to hypothesized generalization regions or rather wide-spread across the brain, we conducted a whole-brain searchlight to allow for a more data-driven approach to find regions which may carry learning-related category information during encoding.

Searchlight classification of category-relevant and category-irrelevant information. Whole-brain searchlight maps for decoding of category-relevant and

category-irrelevant information across the learning phase are presented in Figure 3.5a.

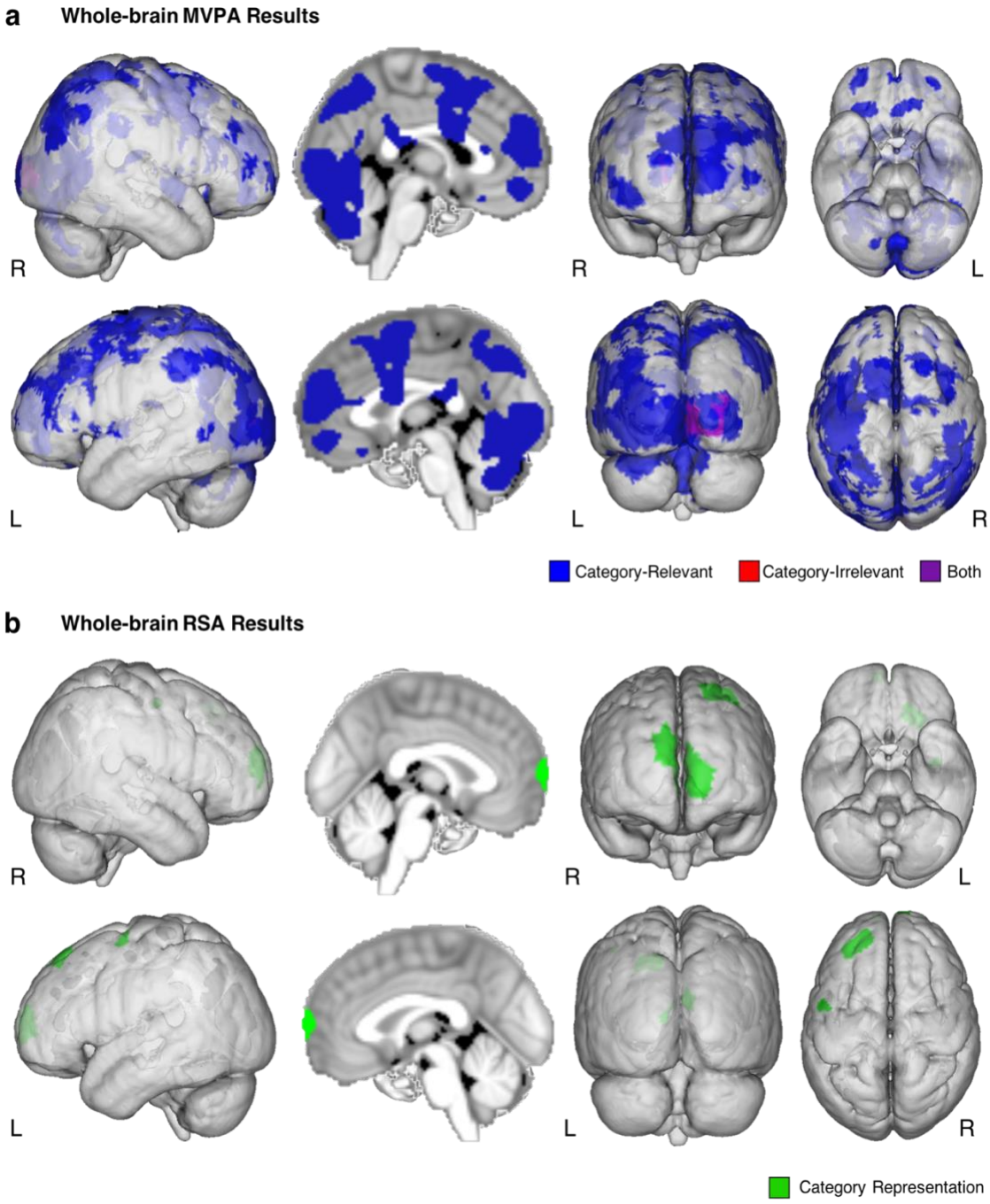

Figure 3.5. Whole-brain Searchlight Results. **A.** MVPA searchlight maps for category-relevant (blue) and category-irrelevant (red) decoding across all four runs of learning. Category-irrelevant decoding in LO largely overlapped with decoding for category-relevant information (purple). **B.** RSA searchlight map for category representations (shared parent same family name – shared parent different family name). Animations fully displaying the pattern of results across the entire brain are available on OSF for both MVPA and RSA searchlight results.

For category-relevant classification, using Threshold-Free Cluster Enhancement in FSL randomize yielded a single large cluster that survived cluster correction but encompassed many functional regions (Multi-Regional Cluster peak: MNI -42, 26, 4; t = 6.29; 43,497 voxels). To better characterize the large cluster, we applied an additional voxel-wise correction (t = 3.5) to the already thresholded map in order to parse the cluster into definable functional regions (Table 3.1). Notably, regions that classified category-

Table 3.1. Learning Phase Searchlight MVPA Results. Peak coordinates are reported separately for regions significantly classifying category-relevant and category-irrelevant information.

				Peak Coordinate		
Region	**Hemisphere**	**Cluster Size**	**T-statistic**	**X**	**Y**	**Z**
Category-Relevant						
Multi-Regional Cluster	L + R	43497	6.31	-42	26	4
Early Visual Cortex	*L+R*	*1517*	*5.26*	*18*	*-74*	*4*
Paracingulate Gyrus	*L+R*	*1196*	*4.98*	*-10*	*50*	*24*
Superior Parietal Lobule/ Precentral + Postcentral Gyrus	*L*	*1065*	*5.05*	*-30*	*-50*	*54*
Superior Parietal Lobule + Angular Gyrus	*R*	*747*	*5.10*	*38*	*-48*	*48*
Inferior + Middle Frontal Gyrus	*L*	*698*	*6.31*	*-42*	*26*	*4*
Superior Parietal Lobule + Angular Gyrus	*L*	*424*	*4.64*	*-36*	*-52*	*38*
Superior Frontal Gyrus	*L*	*379*	*5.53*	*-12*	*8*	*64*
Precuneous + Cuneal Cortex	*L*	*158*	*3.94*	*-6*	*-80*	*38*
Superior Lateral Occipital Cortex	*R*	*122*	*4.49*	*48*	*-64*	*34*
Middle Frontal Gyrus	*R*	*118*	*5.46*	*30*	*14*	*34*
Postcentral Gyrus	*R*	*114*	*4.68*	*40*	*-28*	*60*
Precuneous Cortex	*R*	*109*	*4.15*	*22*	*-62*	*26*
Occipital Fusiform Gyrus	*L*	*90*	*4.23*	*-26*	*-84*	*-24*
Middle Frontal Gyrus	*R*	*87*	*4.26*	*36*	*10*	*52*
Inferior Frontal Gyrus (pars opercularis)	*L*	*66*	*4.39*	*-48*	*8*	*20*
Frontal Pole	*R*	*60*	*4.41*	*32*	*52*	*-6*
Anterior Supramarginal Gyrus	*L*	*47*	*4.37*	*-58*	*-38*	*36*
Superior Lateral Occipital Cortex	*L*	*27*	*3.86*	*-28*	*-84*	*30*
Anterior Cingulate Gyrus	*R*	*24*	*3.81*	*2*	*4*	*26*
Inferior Lateral Occipital Cortex	*L*	*20*	*3.80*	*-44*	*-76*	*0*
Caudate Nucleus	*L*	*12*	*3.67*	*-8*	*4*	*6*
Category-Irrelevant						
Lateral Occipital Cortex	R	315	5.57	16	-94	-4

Cluster size is the number of voxels; peak coordinate is given in MNI space. L = left; R = Right. Sub-clusters were obtained by applying additional voxel-wise thresholding (t = 3.5) and are identified in italics.

relevant information were widespread and distributed across large portions of the frontal lobes, parietal lobes, occipital lobes and the midline. In contrast, MVPA searchlight for category-irrelevant classification yielded only a single small cluster fully confined to LO (no additional thresholding applied). This cluster almost entirely overlapped with classification of category-relevant information in the visual cortex (see purple in Figure 3.5a).

Searchlight neural pattern similarity representations of category information. Next we used the whole-brain searchlight approach to perform the RSA analysis and look for a neural category bias (shared parent-same family name > shared parent-different family) across the entire brain. Whole-brain searchlight maps depicting category-biased representations across all of the learning phase are presented in Figure 3.5b. Category-representations survived cluster correction in bilateral frontal pole, left superior frontal gyrus + middle frontal gyrus, and the left precentral gyrus (Table 3.2).

The pattern of results across searchlight analyses demonstrates that during learning many regions spontaneously form category-biased neural representations even though task-demands at encoding emphasized specificity. The category-irrelevant information is important for the explicit task goals of remembering the full name for each specific face, but despite this our results indicate that neural representations are biased

Table 3.2. Searchlight RSA Results. Peak coordinates are reported for regions with a significant category-biased neural representation during learning.

				Peak Coordinate		
Region	**Hemisphere**	**Cluster Size**	**T-statistic**	**X**	**Y**	**Z**
Frontal Pole	L + R	631	4.04	8	66	18
Superior Frontal Gyrus + Middle Frontal Gyrus	L	129	3.49	-18	42	52
Precentral Gyrus	L	19	3.89	-44	-4	58

Cluster size is the number of voxels; peak coordinate is given in MNI space. L = left; R = Right.

towards category-relevant information. Further, category-relevant information as well as neural category-bias seem to be relatively widespread across the brain and are not unique to our hypothesized generalization regions.

Discussion

Prior work has indicated that category learning induces a perceptual category-bias where items within categories are perceived as more similar to one another than items across category boundaries. A category-bias in perception has also been demonstrated under other task conditions that vary from those of the traditional category learning paradigm extending these findings to a task where category-irrelevant features are also important for explicit task goals (S. R. Ashby et al., 2020). To directly test whether neural category-biased representations are formed spontaneously during learning we scanned individuals using fMRI as they completed an observational paired-associates learning task that required maintenance of both category-relevant and category-irrelevant information. Participants learned face-full name associations using facial stimuli that were blended to maintain physical similarity both within and across family category boundaries. Ratings of perceptual similarity were collected both before and immediately after learning and a subsequent categorization task that included never-studied face-blends was administered to measure memory generalization. Although the category bias in similarity ratings did not reach significance across the group, we replicated our prior work that showed that individual differences in similarity ratings category-bias predicted performance on a subsequent generalization task. Pattern information analyses of fMRI data revealed evidence for significant or marginal category-biased neural representations during learning in putative generalization regions (middle temporal gyrus and

ventromedial prefrontal cortex) that served as regions of interest. Unexpectedly, we found evidence for both category-relevant and irrelevant information in lateral occipital cortex with numerically greater evidence for category-relevant information. Furthermore, whole-brain searchlight analyses showed evidence for category-relevant information widely distributed across the brain. Together, our results indicate that category information is measurable during learning (even under task conditions that emphasize learning face-specific information) demonstrating that category-biased neural representations form spontaneously during encoding and are not merely the product of generalization task-goals at retrieval.

Category-bias in behavioral ratings predicts subsequent generalization performance

We found an overall decrease in similarity ratings after learning which replicated our prior findings (S. R. Ashby et al., 2020). The paired-associates task required individuals to pay attention to both the category-relevant and the category-irrelevant features as task goals at encoding required participants to discriminate not only between families but also between "brothers" within the same family. The overall expansion effect in similarity ratings after learning in this task indicates that category learning may utilize feature weighting where more attentional resources are allocated to features that support the learning goals of the task at hand (Nosofsky, 1991). In our prior study we found that though task goals were to learn individual identities, merely including the shared family name category label was sufficient to elicit a category bias in perceptual similarity ratings. Here, we did not find an overall category-bias in post-learning similarity ratings across all subjects. However, we did replicate our prior finding that individual differences in the strength of the category-bias in post-learning similarity ratings predicts subsequent

generalization of category information to new instances, even when controlling for pre-existing perceptual similarity biases. This is consistent with traditional category-learning work that has theorized that a category bias in perception is due to an attentional shift to items and features that are learned to be relevant to the learned category (Goldstone & Steyvers, 2001; Kruschke, 1996; Nosofsky, 1991). Together with our prior work (S. R. Ashby et al., 2020), we show a novel evidence for this effect, where category-irrelevant information was still relevant to task goals at encoding and the mere presence of the shared family label was sufficient to induce a category bias in some individuals which allowed them to generalize the category label to never-studied faces during the surprise categorization task. Individuals who generalized information well showed the largest distortion in their perceptual representations of the face-blend stimuli even though face-blends were controlled for physical similarity within and across category boundaries.

Category-biased neural representations are measurable during encoding

Whether related events are linked on-the-fly at retrieval in response to generalization demands (Banino et al., 2016; Carpenter & Schacter, 2017, 2018) or whether they are spontaneously linked during encoding (Shohamy & Wagner, 2008; Zeithamova, Dominick, et al., 2012) remains a hotly debated discussion in the literature. Our prior work provided preliminary behavioral evidence that a category bias in perceptual similarity ratings may be a good indicator of the degree of available generalizable category knowledge *prior* to explicit generalization task demands (Ashby et al., 2020). Thus, category-biased information prior to retrieval may indicate that generalization may occur spontaneously during learning. While measuring the perceptual category-bias after learning greatly minimized task-related demands to make

generalization decisions, it was not possible to rule out that probing similarity judgments may have induced a strategic generalization demand to rate shared family faces as more similar to one another. The current study allowed us to more definitively determine whether category-biased representations are formed during encoding by observing neural evidence for a category-bias during learning, in the absence of explicit task demands.

Among the hypothesized generalization regions, we found the most robust evidence for category-biased neural representations in the middle temporal gyrus (MTG). Category-relevant information was decodable in neural patterns of activity during learning to a greater extent than category-irrelevant information. Studies of semantic gist memory have found the MTG to be involved in generalization processes by evaluating incoming information in light of existing schema representations (Turney & Dennis, 2017; Webb, Turney, & Dennis, 2016). Deng, Booth, Chou, Ding, and Peng (2008) found learning-related increases in MTG activation when processing semantically related transfer items but not for trained stimuli which may be reflective of accessing semantic information when integrating new information with existing knowledge. Our findings are consistent with this prior work indicating that the MTG is sensitive to the generalizable category-relevant information during learning and may contribute to updating the category representation during learning. We also found modest evidence for category-biased information in the ventromedial prefrontal cortex (VMPFC) which survived correction for multiple comparisons and is consistent with other work which show evidence for abstracted memory representations in VMPFC (Bowman & Zeithamova, 2018; Kumaran et al., 2009).

Unexpectedly we did not see evidence for category-biased information in anterior hippocampus (AHIP). As past work has found evidence for abstract, category representations in AHIP during categorization tasks or traditional category-learning paradigms (Bowman, Iwashita, et al., 2020; Bowman & Zeithamova, 2018), we speculate that the lack of category-biased representations within hippocampus in the current study may reflect disparate task goals. The learning goals of the present study required individuals to encode individual face-name pairs and avoid interference between items within the same category. Thus, the hippocampus in the current task may require more resources allotted to pattern separation processes (Yassa & Stark, 2011) in order to reduce interference in light of task goals.

We also found classification for category-relevant information during learning that was not unique to our theorized memory generalization regions. Instead we found evidence for category-relevant information more widespread across the brain. Classification for category-relevant information also involved regions theorized to maintain working memory in light of task goals (caudate nucleus) and bias attention towards category-relevant information (inferior frontal gyrus). As the caudate nucleus (CN) is consistently activated during learning tasks in animals (Fernandez-Ruiz, Wang, Aigner, & Mishkin, 2001; Teng, Stefanacci, Squire, & Zola, 2000) and in studies of human category learning (Poldrack, Prabhakaran, Seger, & Gabrieli, 1999; Seger & Cincotta, 2005) finding category-relevant information in this region is consistent with past work. It's been posited that CN activity may be modulated by working memory load (Poldrack et al., 1999) and more recent work has found evidence for stronger CN activity when encountering new overlapping stimuli (Brown & Stern, 2014). Activity in inferior

frontal gyrus (IFG) has been shown in several experiments of semantic memory (Thompson-Schill, D'Esposito, Aguirre, & Farah, 1997; A. D. Wagner, Pare-Blagoev, Clark, & Poldrack, 2001) and learning-related activation increases in IFG have been observed during processing of never-studied "transfer" items in a semantic learning task (Deng et al., 2008). Together, these results provide evidence for category-biased shifts in attentional processes during learning. CN representations for the category-relevant information during learning in the current study may be reflective of the working memory resources needed in the current task to maintain the category-relevant features while updating the appropriate category-representation in light of new face-blends encountered during learning. Additionally, IFG may support category learning by actively evaluating the importance of incoming information during learning. This is in line with suggestions that the IFG may work to evaluate semantic representations in light of the task at hand (Gabrieli, Poldrack, & Desmond, 1998; Gold & Buckner, 2002; Poldrack et al., 1999; A. D. Wagner et al., 2001) and may serve as a key region for biasing attention towards category-relevant information (Mack et al., 2013).

Category-biased neural representations may reflect attentional allocation to category-relevant information

Theories of category learning postulate that a key part of learning is an allocation of attentional resources away from category-irrelevant information to category-relevant features. This results in a stretching and shrinking of perceptual space where items within a category are perceived as more similar to one another and are more difficult to discriminate (Goldstone et al., 2001; Gureckis & Goldstone, 2008; Kurtz, 1996; Livingston et al., 1998; Soto, 2019), while items from different categories become less

similar to one another (Goldstone et al., 2001) and are easier to discriminate (Beale & Keil, 1995; Folstein, Palmeri, & Gauthier, 2013; Goldstone, 1994a; Gureckis & Goldstone, 2008). Here we see that category-relevant information can be decoded across large portions of the brain including regions theorized to maintain working memory in light of task goals and bias attention towards category-relevant information. While information for physically similar faces that did not align with category membership was decodable in the brain, the extent of brain involvement in representing this category-irrelevant information was small and largely overlapped with representations for category-relevant information.

Neural pattern similarity analyses indicated several regions including ventromedial prefrontal cortex (VMPFC), and to a marginal degree the lateral occipital cortex (LO) and MTG, that represented faces that shared a parent and family name as more similar to one another than faces that shared a parent but differed in their family name. Although we hypothesized that VMPFC would reflect a category bias during learning, our finding of category-biased representations in this region are tenuous as are the marginal findings in LO and MTG. This finding is consistent with prior work by Mack et al., (2013) that showed learned attention-weighted neural similarity patterns during category learning are widespread across cortex and include visual cortices, but contrasts with other work that has found abstract category representations predominately driven by these regions (Bowman, Iwashita, et al., 2020; Zeithamova & Bowman, 2020; Zeithamova, Maddox, & Schnyer, 2008). We speculate that these differences may reflect the differences in the attentional shifts required by distinctive category structures. In the category learning paradigms that previously found strong category representations in

VMPFC and MTG, all features of the stimuli were equally relevant for categorization and learning involved primarily linking together category labels. In contrast, the category structure in the current study was more similar to that of Mack et al. (2013) which required learning both which features are relevant and irrelevant to determine category membership. Thus, the neural category bias may reflect an attentional shift to category-relevant features, which employs a large extent of the brain rather than being specific to regions implicating in generalization and memory integration. Here we extend these findings to a task where the category-irrelevant information cannot merely be ignored and instead is important to maintain to accomplish the task goals that require individuation of all faces including those within the same family category. Though the current study provides evidence for widespread category-biased neural representations during learning it is important to note that the current study cannot distinguish the style of category-biased representations formed during learning. Whether the category representations formed during learning are abstract, generalized representations of the families as would be predicted by prototype theory (Posner & Keele, 1968) or whether they are individual, specific representations for face-family name associations as predicted by exemplar theory (Kruschke, 1992; Nosofsky, 1986) is unknown because both theories would predict similar attentional shifts in perceptual space.

Summary

The current findings build off our prior behavioral work showing category-biased perceptual effects after learning and extends those findings to demonstrate category-biased neural representations during learning. Critically we found category-biased neural representations throughout the cortex during a learning task that contained category

information but emphasized differentiating stimuli both within and between category boundaries. Thus, our findings demonstrate that neural category representations can form spontaneously during learning in the absence of explicit generalization task demands. We also extend prior findings of attention-weighted representations widely distributed across the cortex during retrieval to a learning task that requires attention to both category-relevant and category-irrelevant information.

CHAPTER IV

HIPPOCAMPAL INTERACTIONS WITH CORTICAL MEMORY REGIONS DURING SPONTANEOUS GENERALIZATION

The hippocampus has long been known to support detailed episodic memory (Scoville & Milner, 1957), but recent work has also implicated the hippocampus as an important structure for memory generalization (Bowman & Zeithamova, 2018; Shohamy & Wagner, 2008; Zeithamova, Dominick, et al., 2012). How the hippocampus is able to support both processes is not well understood and is currently an emerging area of interest within the literature (Berens & Bird, 2017; Schapiro et al., 2017). One proposal calls for a division of labor expressed along the long-axis of the hippocampus with the posterior portion supporting memory specificity and the anterior portion supporting memory generalization (Brunec et al., 2018; Collin, Milivojevic, & Doeller, 2015; Poppenk et al., 2013). Recent work from our lab (see Frank, Bowman, & Zeithamova, 2019) found evidence for anterior/posterior dissociations showing differential intrinsic connections between posterior hippocampus and known specificity regions and between anterior hippocampus and known generalization regions. Furthermore, individual differences in hippocampal connectivity with the ventromedial prefrontal cortex was associated with individual differences in memory generalization performance. Although differential connections between anterior/posterior hippocampus and several cortical regions persisted across multiple task phases of the experiment including resting state, connectivity between inferior frontal gyrus and the hippocampus was less stable indicating this connection may be driven more by task engagement. Thus, the extent to

which differential anterior/posterior hippocampal connections with putative generalization and specificity regions reflects stable connections or differential engagement depending on task goals is less understood. In the current study, we sought to test the differential anterior/posterior hippocampal connections in the context of a novel paradigm where task goals emphasize memory specificity, but individuals also spontaneously generalize information during learning. Additionally, because memory generalization occurs spontaneously and is not the explicit goal of the novel task, we tested whether individual differences in hippocampal connectivity are associated with behavioral measures of memory generalization under these circumstances.

Division of Labor Within the Hippocampus

Long-axis specialization of the hippocampus has been found in various domains. In rodent work, receptive field size varies along the hippocampal axis with the smallest fields, representing more fine-grained detailed information, residing in in the dorsal hippocampus (analogous to the human posterior hippocampus), and larger receptive fields, representing more course-grained information, residing in the ventral hippocampus (analogous to human anterior hippocampus; see Poppenk et al., 2013). Additional work examining spatial representations in the hippocampus have found greater posterior compared to anterior activity for detailed representations of individual features and exact locations (Doeller, King, & Burgess, 2008; Hassabis et al., 2009; Nadel, Hoscheidt, & Ryan, 2013) and greater anterior compared to posterior activity for representations of more relative locations (Ekstrom, Copara, Isham, Wang, & Yonelinas, 2011; Morgan, MacEvoy, Aguirre, & Epstein, 2011).

In humans, studies of associative inference where pairs of items are encoded that share an overlapping element (AB, BC pairs), representations for the individual AB and BC pairs remained individualized in the posterior hippocampus while there was evidence in anterior hippocampus for an integrated ABC representation (Schlichting et al., 2015). Along the same lines, more recent work examining category learning found generalized concept representations in the anterior hippocampus but not the posterior hippocampus both during category learning (Bowman, Iwashita, et al., 2020) and generalization of category information to new examples (Bowman & Zeithamova, 2018). Thus, overwhelming evidence suggests that the hippocampus can support both processes simultaneously via a long-axis division of labor (for review see Sekeres, Winocur, & Moscovitch, 2018) with the posterior hippocampus supporting specificity and anterior hippocampus supporting generalization.

Cortical Regions Supporting Memory Generalization

In addition to the differential functions of the hippocampus supporting specificity and generalization, other cortical regions also differentially contribute to these processes. The ventromedial prefrontal cortex (VMPFC) has been shown to support the construction of schema representations (Baldassano, Hasson, & Norman, 2018; Brod, Lindenberger, Werkle-Bergner, & Shing, 2015; Ghosh, Moscovitch, Colella, & Gilbo, 2014) and to support memory integration by linking together memories during encoding (Schlichting et al., 2015; Zeithamova, Schlichting, et al., 2012) and facilitating generalization of conceptual information to never-before seen stimuli (Bowman et al., 2020; Bowman & Zeithamova, 2018; Zeithamova, Maddox, & Schnyer, 2008; for review see Zeithamova & Bowman, 2020). Further, the VMPFC and anterior hippocampus are known to interact

with one another in formation of generalized memory representations (Pajkert et al., 2017; Zeithamova, Dominick, et al., 2012), providing support to the long-axis division of labor account for hippocampal specialization. Portions of temporal cortices have also been implicated in generalization. The middle temporal gyrus (MTG) is recruited by semantic memory processes (Mummery et al., 2000; Renoult, Irish, Moscovitch, & Rugg, 2019), concept learning tasks (Bowman & Zeithamova, 2018), and is also a region that is frequently reported in investigations of "gist" representations (Dennis et al., 2008; Turney & Dennis, 2017). A study that used TMS to inhibit neural activity within the MTG induced impairment in the ability to flexibly retrieve conceptual knowledge (Davey et al., 2015) further supporting the role of MTG as a region vital for storing and manipulating conceptual knowledge in service of memory generalization.

Cortical Regions Supporting Memory Specificity

While the hippocampus has long been studied as the premiere structure for episodic memory, other cortical regions have also been implicated in representing detailed, item-specific information. Portions of lateral parietal cortex, namely angular gyrus (ANG), has been implicated in preventing interference between similar memories in service of specificity (Hutchinson, Uncapher, & Wagner, 2009; Kuhl & Chun, 2014; Xiao et al., 2017), and studies of exemplar models of categorization (which rely on each item encountered being stored as an individual, unique representation) also show exemplar correlates within lateral parietal cortices (Mack et al., 2013). Other work utilizing TMS disruption of activity within ANG found impairments with retrieval of concepts at a more specific level. For example, when presented with a learned image of a dog on the screen participants had difficulty retrieving the specific verbal label

corresponding to the breed “Corgi”, but not the more general category membership “Animal” (Davey et al., 2015). The inferior frontal gyrus (IFG) is another region that promotes specificity by supporting autobiographical retrieval (Greenberg et al., 2005) and resolving interference between related events (Bowman & Dennis, 2016; Kuhl, Dudukovic, Kahn, & Wagner, 2007; Stramaccia, Penolazzi, Altoè, & Galfano, 2017) to preserve specificity.

Prior Study that Identified an Anterior/Posterior Dissociation in Functional Connectivity to Memory Specificity and Generalization Regions

Recent work by Frank, Bowman & Zeithamova (2019) set out to explore how the hippocampus may interact with these cortical generalization and specificity regions in order to further support the dual-role hypothesis of the hippocampus. In their study, they tracked the intrinsic functional connectivity between the anterior and posterior hippocampus and putative generalization (VMPFC, MTG) and specificity (ANG, IFG) regions. Participants completed a traditional feedback-based category learning paradigm outside the scanner. After learning, participants were scanned during three task phases: resting state, passive viewing, and a concept generalization task. As predicted, low frequency fluctuations in specificity regions (ANG, IFG) was more strongly coupled with low frequency fluctuations in posterior compared to anterior hippocampus. Low frequency fluctuations in VMPFC was more strongly coupled with low frequency fluctuations in anterior compared to posterior hippocampus while evidence for coupling between MTG and anterior hippocampus was not reliable. Notably, these couplings remained fairly stable across the three different task phases although connectivity with IFG did increase during phases that involved stimulus presentation (i.e. greater functional

connectivity between hippocampus and IFG during each task phase compared to rest) indicating that interactions between hippocampus and IFG may be affected by task engagement.

These findings are the first of their kind in the literature to show differential relationships between anterior and posterior hippocampus and cortical regions supporting generalization and specificity. Although their results remained stable across various phases of their experiment in most regions, it's still unknown whether these findings would replicate in a completely different task during encoding rather than after learning has already taken place. If these results truly reflect intrinsic connections between the hippocampus and cortical regions, we would predict that these findings would be replicable in a drastically different paradigm.

The Current Study

In the current study, we sought to replicate and extend the findings by Frank and colleagues (2019). Our primary goal was to determine the stability of differences in hippocampal-cortical connectivity along the long-axis of the hippocampus by examining intrinsic connectivity during a specificity-focused paired associates learning task. We also explored whether individual differences in hippocampal-cortical connectivity was associated with behavioral measures of memory generalization during learning that elicits spontaneous generalization. During fMRI, participants completed the same observational, face-full names paired associates learning as was described in Chapter's 2 and 3 of the dissertation. To measure intrinsic connections between regions we utilized the same measures of background connectivity as implemented by Frank and colleagues (2019) by removing the trial-by-trial signal due to task-related fluctuations and measuring the

remaining “background” fluctuations that are thought to be indicative of the intrinsic connections between regions (Van Dijk et al., 2010). Given the prior study’s findings that background connectivity was stable across levels of engagement (i.e. differential anterior/posterior connectivity with generalization regions was not greater during the generalization task) and other studies that have also found background connectivity measures to show similar information as resting-state connectivity analyses (Frank, Preston, & Zeithamova, 2019; Gratton et al., 2018; Touroutoglou, Andreano, Barrett, & Dickerson, 2015), we hypothesized that the differential connectivity effects uncovered by Frank et al. (2019) are stable and reflect intrinsic connections which will be replicable under different task conditions and with very different stimuli. Specifically, we predicted that we would find the posterior hippocampus to be more functionally connected to hypothesized specificity regions (IFG, ANG) and the anterior hippocampus to be more functionally connected to hypothesized generalization regions (VMPFC, MTG) during encoding.

In Chapter 3 of the dissertation we found evidence for category-biased neural representations in cortical visual regions, namely the lateral occipital cortex (LO) and posterior fusiform gyrus (PFUS). This is consistent with other literature that has found hippocampal connectivity with the visual cortex. Learning-related connectivity changes between the hippocampus and fusiform gyrus has been shown in tasks that utilize facial stimuli (Bokde et al., 2006; Takashima et al., 2009; I. C. Wagner, Rütgen, & Lamm, 2020) and increased connectivity between the hippocampus and fusiform gyrus during sleep has been shown to benefit subsequent learning for face-location associations (van Dongen, Takashima, Barth, & Fernández, 2011). Hippocampal connectivity with the

lateral occipital cortex following learning has been associated with better retrieval performance (Tambini, Ketz, & Davachi, 2010), and enhanced connectivity between the anterior hippocampus and high-level visual cortex has been shown to predict individual differences in memory for high-reward associations (Murty, Tompary, Adcock, & Davachi, 2017). Thus, we predicted that we would find the hippocampus to be functionally connected to higher order visual cortex (PFUS, LO) during encoding.

Lastly, as Frank et al. (2019) found preliminary evidence indicating that connectivity between anterior hippocampus and VMPFC predicts generalization performance during categorization, we were interested in whether the same evidence would be seen during encoding that elicits spontaneous generalization. Although an examination of individual differences predicting behavior require larger sample sizes to be adequately powered, we reasoned that an exploratory approach to the data may be informative when interpreted cautiously and in conjunction with prior findings. We predicted that individual differences in VMPFC-anterior hippocampal connectivity would be associated with performance on behavioral measures of memory generalization during encoding.

Method

Participants

Participants were collected as part of the project presented in Ashby and Zeithamova (in prep) and discussed in Chapter 3 of the dissertation. Forty-four participants were recruited from the University of Oregon community, gave written informed consent, and scanned at the Lewis Center for Neuroimaging on the university campus. Four participants were excluded for excess motion (two participants), scanner

operator error (one participant), and an undisclosed neurological condition (one participant). Thus, analyses included the remaining forty participants (22 female, 18 male; age 18-30 years; M_{age} = 21.33, SD_{age} = 2.92). All research activities were approved by the University of Oregon Research Compliance Services.

Procedure and fMRI Data Acquisition

Participants completed the same experimental procedure and fMRI scanning was completed using the same acquisition procedures previously described (Ashby & Zeithamova, in prep; see also Chapter 3 of the dissertation).

Regions of Interest (ROIs)

Regions of Interest were defined in each individual participant's native space using both the cortical and subcortical segmentation routines from Freesurfer version 6 (https://surfer.nmr.mgh.harvard.edu/) of the T1-Weighted MPRAGE anatomical image. Bilateral masks for each ROI were created by collapsing together across hemispheres.

Given recent work that has suggested a division of labor along the long axis of the hippocampus, with posterior hippocampus (PHIP) supporting memory specificity and anterior hippocampus (AHIP) supporting memory generalization, we examined these regions separately. Anterior and posterior hippocampal ROIs were defined by dividing the Freesurfer hippocampal ROI at the middle slice. In the event that there were an odd number of hippocampal slices for a given participant, the middle slice was assigned to the posterior hippocampus.

Two regions of interest (ROIs) were selected for their hypothesized roles in memory specificity: inferior frontal gyrus (IFG) and angular gyrus (ANG). The IFG ROI was obtained by combining the three IFG subregions—Freesurfer labels: pars

opercularis, pars orbitalis, and pars triangularis—while the ANG ROI was defined using the 2009 Freesurfer parcellation. Two additional ROIs were selected for their hypothesized roles in memory generalization: ventromedial prefrontal cortex (VMPFC) and middle temporal gyrus (MTG). The VMPFC ROI was defined as the Freesurfer medial orbitofrontal cortex label (MOFC). Lastly, because our prior work showed that category information was represented in higher-order visual cortex (Chapter 3), we also included two additional Visual ROIs: lateral occipital cortex (LO), posterior fusiform gyrus (PFUS).

fMRI Preprocessing

Raw dicom images were converted to Nifti format using MRIcron's (https://www.nitr.org/projects/mricron) dcm2nii function. Functional, behavioral and anatomical data were organized in the Brain Imaging Data Structure (BIDS) format for public dissemination on OpenNeuro (forthcoming). First, using FSL Version 6 (www.fmrib.ox.ac.uk/fsl), functional images were skull stripped using the Brain Extraction Tool (BET) and corrected for motion within each scanner run using FLIRT to realign all images within a run to the middle volume. Across-run realignment was then applied to functional images for each run using Advanced Normalization Tools (ANTs; http://stnava.github.io/ANTs/) with the first volume of the first run of the training task used as the reference volume. The first volumes of all other task runs were registered to the reference volume and the resulting transformation was applied to the remaining functional runs. The registered functional data was next passed into an FSL FEAT model to apply a high-pass temporal filter (60s) with minimal spatial smoothing (2mm FWHM Gaussian Kernel).

According to past work examining functional connectivity (see Murphy, Birn, & Bandettini, 2013; Power, Barnes, Snyder, Schlaggar, & Petersen, 2012), connectivity measures can be artificially inflated by noisy data. To better control for physiological confounds we extracted the timeseries signal for cerebrospinal fluid (csf), white matter, (wm), and whole-brain signal (wb). Next, to control for motion artifacts we also calculated the framewise-displacement (FD) and the global signal change (DVARS) for each functional scan. These were all used as nuisance regressors when calculating connectivity (see below) and also used to determine if individual volumes needed to be scrubbed from analyses and excluded. Volumes were excluded from analyses if either the FD was greater than 0.5mm or if DVARS was over 0.5%. Additionally, when volumes were flagged for exclusion, we also scrubbed the volume before and after the flagged motion event. Our scrubbing procedure resulted in removal of an average of .65% of volumes from analysis.

Calculating Background Connectivity

In order to measure background connectivity, we filtered out any task-based activity (i.e. mutual responses to stimulus onset) that could drive coactivation between regions that may not actually be functionally connected (Frank, Bowman, et al., 2019; Norman-Haignere, McCarthy, Chun, & Turk-Browne, 2012; Tambini, Rimmele, Phelps, & Davachi, 2017). We used a low-pass filtering approach by setting the low-pass filter below the frequency of the task to remove task-related signals. Low-pass filtering was accomplished by applying a Gaussian linear (10s) bandpass filter to remove functional activity that was cycling faster than the task-driven frequency (8s trials). The 10s filter was chosen by examining the power spectrum of the lateral occipital cortex from a

handful of subjects (see Figure 4.1 for a representative subject) and setting a conservative threshold that we felt was appropriate to remove the task-related frequencies. Volume scrubbing as described above was completed after low-pass filtering and timeseries was extracted from the low-pass filtered data for each ROI.

To examine connectivity, we calculated partial correlations between each hippocampal ROI (AHIP, PHIP) and each cortical memory ROI (VMPFC, MTG, IFG, ANG) and each visual ROI (LO, PFUS). We controlled for motion and physiological noise by adding the six standard realignment motion parameters (rotation and translation in each X, Y, Z plane), cerebrospinal fluid, white matter, whole brain signal, plus their derivatives as nuisance regressors. Volumes that were scrubbed were removed from all regressors. The correlation coefficients were then Fisher z-transformed for analysis.

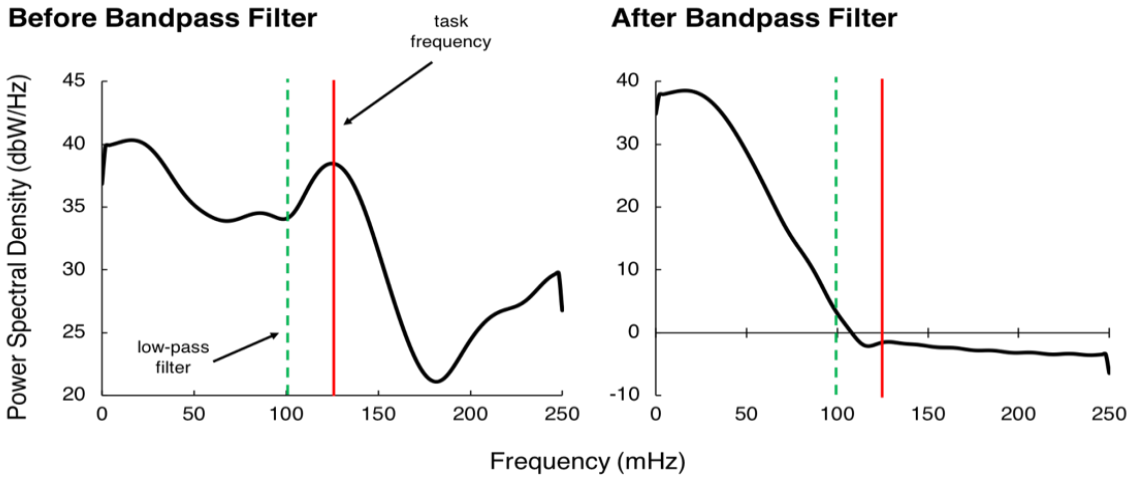

Figure 4.1. Bandpass filtering for a representative subject. Task signal from the LO before (left) and after (right) the bandpass filter was applied. The task frequency is 8s (solid red bar). To only let signal through the filter that is slower than the task frequency we set a filter below the task frequency at 10s (dashed green line). The conservative filter ensured that task-driven coactivation was removed from analyses (see right).

Results

Connectivity with Cortical Memory Regions

To determine whether functional connectivity to cortical regions is different for anterior and posterior portions of the hippocampus, we conducted a 2 [Hippocampus: anterior, posterior] x 4 [Cortical ROI: VMPFC, MTG, IFG, ANG] repeated-measures ANOVA (Figure 4.2). Of interest was a hippocampus ROI x cortical ROI interaction. We predicted that the posterior hippocampus would be more functionally connected to regions previously implicated in memory specificity (IFG, ANG) and anterior hippocampus would be more functionally connected to known generalization regions (VMPFC, MTG). As predicted, there was a significant hippocampus ROI by cortical ROI interaction ($F(1.83, 71.27) = 9.636$, $p < .001$, $\eta_p^2 = .198$, *GG*). For significant interactions, follow-up t-tests were conducted to compare the anterior and posterior hippocampus connectivity with each of the four cortical ROIs. In line with our predictions, we found that VMPFC was more functionally connected to anterior hippocampus ($t(39) = 2.52$, $p = .008$, $d = .399$, one-tailed) while ANG was more functionally connected to posterior hippocampus ($t(39) = 3.80$, $p < .001$, $d = .60$, one-tailed). Contrary to our predictions, MTG and IFG were functionally connected to the same degree with both hippocampal ROIs (both t's < 0.43, p's $> .33$ one-tailed). We found a significant main effect of Cortical ROI ($F(2.22, 86.73) = 5.128$, $p = .006$, $\eta_p^2 = .116$, *GG*) driven by significantly less functional connectivity overall between the hippocampus and ANG compared to all

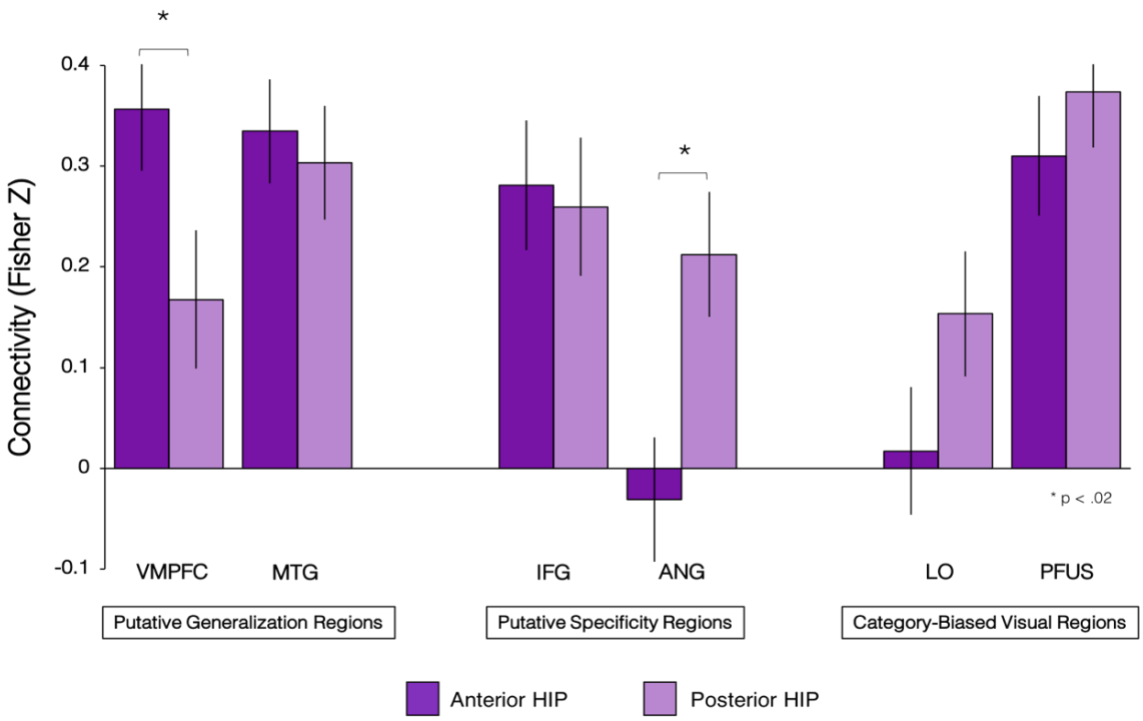

Figure 4.2. Functional Connectivity Results. Functional connectivity between anterior hippocampus (dark purple), posterior hippocampus (light purple) and the six ROIs are presented. Connectivity values are Fischer Z transformed for comparisons. Stars designate significant differences in hippocampal connectivity for VMPFC and ANG.

other cortical regions (all cortical regions compared to ANG $t > 2.32$, $p \leq .025$). Lastly, there was no significant main effect of hippocampal ROI ($F(1, 39) = 0$, $p = .994$, $\eta_p^2 = 0$).

Connectivity with Visual Regions

Past work has shown the hippocampus to be linked with perceptual regions (for review see A. C. H. Lee, Yeung, & Barense, 2012). Work in mice has found correlated spatial representations in primary visual cortex and hippocampus (Saleem, Diamanti, Fournier, Harris, & Carandini, 2018) even in the absence of visual information (Fournier, Saleem, Diamanti, Wells, & Harris, 2019), indicating that the hippocampus communicates with visual regions. In Chapter 3 of the dissertation we unexpectedly found category-relevant information represented in visual cortex. Thus, we reasoned that visual regions may be functionally connected to the hippocampus but whether there

would be anterior vs. posterior connectivity dissociations with visual regions is unknown. To explore differential connectivity between anterior and posterior hippocampus and visual control regions, we conducted a 2 [Hippocampal ROI: anterior, posterior] x 2 [Visual ROI: LO, PFUS] repeated-measures ANOVA. We found a significant main effect of visual ROI ($F(1,39) = 17.28$, $p < .001$, $\eta_p^2 = .31$) driven by larger functional connectivity between the hippocampus and PFUS ($t(39) = 4.16$, $p < .001$, $d = .657$; see Figure 3.2). There was no main effect of hippocampal ROI connectivity ($F(1, 39) = 2.42$, $p = .128$, $\eta_p^2 = .058$) nor a hippocampal ROI by visual ROI interaction ($F(1, 39) = 1.04$, $p = .314$, $\eta_p^2 = .026$) indicating that anterior and posterior hippocampus were functionally connected with visual regions to the same degree.

Connectivity-Behavior Relationships: Exploratory Analyses

Next, we wanted to examine how functional connectivity between hippocampus and putative generalization and specificity regions may be related to our behavioral measures of memory generalization and specificity. Although we do not have the power necessary to properly examine individual differences, we wanted to explore the possibility that these connectivity measures are related to our behavioral measures of memory generalization. We predicted that functional connectivity between AHIP and memory generalization regions (VMPFC, MTG) would be correlated with performance on behavioral measures of memory generalization (generalization accuracy, category bias in similarity ratings). Using a Pearson's correlation, we did not find any significant correlations between AHIP - putative generalization regions connectivity and behavioral measures of memory generalization (all $r's < 0.13$, $p's > .44$). We also predicted that functional connectivity between PHIP and memory specificity regions (IFG, ANG)

would be associated with performance on behavioral measures of memory specificity (first name recall, corrected hit rate). There was no significant correlation between PHIP - putative specificity regions connectivity and behavioral measures of memory specificity (all r's < 0.12, p's >. 38). We did not have specific predictions about how AHIP connectivity with specificity regions or PHIP connectivity with generalization regions would be associated with behavior, and we did not find any correlation with behavior for those connections (all r's < .23, p's > .14).

Chapter 3 of the dissertation showed learning-related category representations in visual cortex and prior work by other research groups indicated learning-related hippocampus-visual cortex connectivity predicting individual differences in memory (Murty et al., 2017; Takashima et al., 2009; Tambini et al., 2010). Next, we examined correlations between hippocampal connectivity with the visual regions and behavioral measures of memory generalization (generalization accuracy, category bias). We first examined connectivity relationships with generalization performance. We found a significant relationship between AHIP-PFUS connectivity and generalization accuracy ($r(39) = 0.385$, $p = .014$), where greater functional connectivity between AHIP and PFUS was associated with better performance on the generalization test (Figure 4.3a; top). Functional connectivity between AHIP and LO was marginally related to generalization performance ($r(39) = .279$, $p = .081$, see Figure 4.3b; top). Next, we examined connectivity relationships with the indirect generalization measure—the category bias in perception. We found a marginal relationship between AHIP-PFUS connectivity and the category bias in perception ($r(39) = .298$, $p = .062$, see Figure 4.3a; bottom). When examining PFUS connectivity correlations with the two generalization measures we

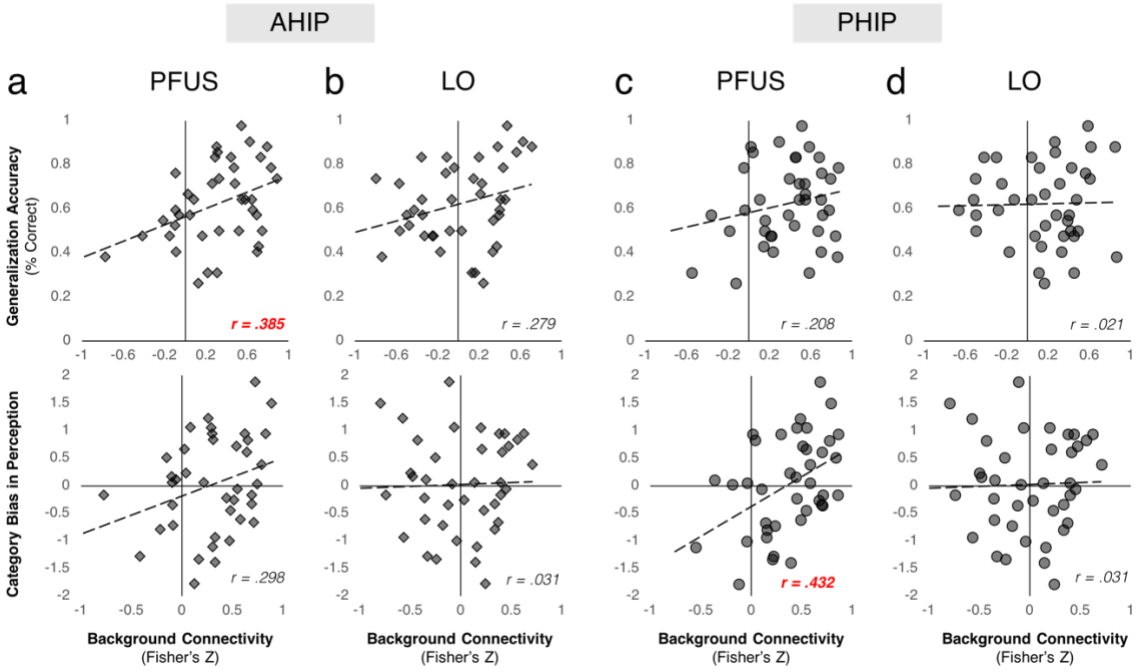

Figure 4.3. Correlations between anterior and posterior hippocampus connectivity with visual control regions and behavioral measures of memory generalization. **A.** Relationship between AHIP-PFUS connectivity and memory generalization accuracy. **B.** Relationship between AHIP-LO connectivity and memory generalization accuracy. **C.** Relationship between AHIP-PFUS connectivity and the indirect memory generalization measure—category bias in perception. **D.** Relationship between PHIP-PFUS connectivity and the category bias in perception. None of the correlations survived correction for multiple comparisons (Bonferroni corrected α = .0125, 4 comparisons).

found only a single significant relationship between PHIP-PFUS connectivity and the category bias in perception measure (r(39) = .432, p = .005; see Figure 4.3c; bottom); all other correlations were not significant (see Figure 4.3c top & Figure 4.3d top and bottom). No correlations survived correction for multiple comparisons (Bonferroni corrected α = .0125 for 4 comparisons). Although the exploratory analyses must be taken with caution due to the low powered approach and concerns with multiple comparisons, results preliminarily indicate that connections between anterior/posterior hippocampus and higher order visual cortex may be associated with how well individuals generalize information to never-before studied examples.

Discussion

The main aim of the current study was to replicate findings by Frank and colleagues (2019) showing differential hippocampal connectivity with cortical memory regions in service of memory generalization and specificity within a task that emphasizes memory specificity. We tested whether there was evidence for differential connections between the hippocampus (posterior, anterior) and cortical memory regions known to support memory specificity (ANG, IFG) and memory generalization (VMPFC, MTG). Consistent with findings from Frank and colleagues (2019), we found the ANG to be more functionally connected to the posterior hippocampus and VMPFC to be more functionally connected to the anterior hippocampus. We also did not find differential connectivity preferences between the hippocampus and the MTG. In contrast to the original work, we did not find evidence for differential hippocampal connectivity with the IFG. When exploring connectivity-behavior relationships, we did not find evidence for individual differences in hippocampal connectivity with cortical memory regions tracking individual differences in behavioral measures of specificity or generalization. Unexpectedly, we found individual differences in hippocampal connectivity with higher-level visual regions (LO/PFUS) that tracked individual differences in measures of memory generalization. Taken together, the current findings replicated findings from the original study showing differential connectivity between anterior and posterior hippocampus with ANG and VMPFC during a novel learning task that focuses on specificity but also elicits spontaneous generalization. Our findings strengthen the prior work by adding additional evidence that connections between anterior hippocampus and VMPFC and between posterior hippocampus and ANG reflect stable, intrinsic

relationships that are replicable under different task demands. Through exploratory analysis of connectivity-behavior relationships we also add new insight into the possible relationships between the hippocampus and higher order visual regions in support of spontaneous generalization.

Posterior Hippocampus Connections with Specificity Regions

Past work has implicated the ANG as a region supporting retrieval of detailed episodic memory (Johnson, Suzuki, & Rugg, 2013; Kuhl & Chun, 2014; H. Lee, Samide, Richter, & Kuhl, 2019; Richter, Cooper, Bays, & Simons, 2016; Xiao et al., 2017) and the IFG as a region primarily responsible for resolving interference between highly similar or related items (Bowman & Dennis, 2016; Kuhl et al., 2007). Recent findings by Frank, Bowman, and Zeithamova (2019) demonstrated that both ANG and IFG are more strongly connected with the posterior compared to the anterior hippocampus when collapsed across multiple experimental phases. As predicted, we replicated the finding that the ANG was more functionally connected to the posterior hippocampus during the paired associates learning. In the original study, the ANG-posterior hippocampus connectivity was stable across different task phases. Here we provide additional evidence for the stability of this finding by demonstrating the same finding it an independent dataset and during a drastically different task phase. However, in contrast to the original study, we did not find any posterior vs. anterior connectivity differences with the IFG. While Frank et al. (2019) did find evidence for greater posterior hippocampal-IFG connectivity this finding was barely significant. Further, their results also showed that hippocampal-IFG connectivity varied by task phase demonstrating less differential connectivity with increasingly more task engagement. As the current study examined background connectivity in the context of

an observational paired associates learning paradigm our finding is not at odds with the original paper. Rather, the current result provides further evidence that hippocampal-IFG connectivity may be driven by task engagement whereas connectivity between the hippocampus and the ANG did not vary across task phases and thus reflects a more stable, intrinsic connection.

Anterior Hippocampus Connections with Generalization Regions

The VMPFC and MTG have been shown to support integration of information across experiences in service of memory generalization through their involvement in schema representations (van Kesteren et al., 2013), overgeneralization resulting in false memories (Garoff-Eaton, Slotnick, & Schacter, 2006), and concept generalization (Bowman, Iwashita, et al., 2020; Bowman & Zeithamova, 2018). Because more recent work has indicated that the anterior hippocampus represents information at a course-grained scale that is advantageous for generalization (Brunec et al., 2018; Collin et al., 2015), we predicted that the anterior hippocampus would be more functionally connected to VMPFC and MTG. As predicted, we replicated the finding that the VMPFC was more functionally connected with the anterior hippocampus. This is in line with work that shows interactions between the hippocampus and VMPFC in support of integration across memories (Van Kesteren, Rijpkema, Ruiter, & Fernández, 2010; Zeithamova, Dominick, et al., 2012) but extends these findings to more specifically implicate the anterior portion of the hippocampus in this process.

Connectivity between the hippocampus and VMPFC has been shown to relate to memory generalization performance in prior work (Frank, Bowman, et al., 2019; Gerraty, Davidow, Wimmer, Kahn, & Shohamy, 2014; Kumaran et al., 2009; Van Kesteren et al.,

2010; Zeithamova et al., 2008). Frank et al. (2019) found stronger VMPFC-anterior hippocampus connectivity to be associated with worse generalization performance. This finding was counterintuitive as prior work examining connectivity-behavior relationships found that task-based connectivity was associated with stronger generalization performance (Kumaran et al., 2009; Zeithamova et al., 2008). However, the authors noted that they were not the first to find a negative relationship between VMPFC-hippocampal background connectivity and generalization performance (see also Gerraty, Davidow, Wimmer, Kahn, & Shohamy, 2014; Van Kesteren et al., 2010) reasoning that lower levels of baseline or post-encoding background connectivity may indicate that information has already been successfully integrated.

To test this hypothesis, we also examined whether low-frequency fluctuations between the VMPFC and hippocampus were associated with measures of memory generalization. The current study included two measures of memory generalization: 1) a direct measure of memory generalization as performance on an explicit generalization test, and 2) an indirect measure of memory generalization as a category bias in perceptual similarity ratings after learning. In contrast with the original study, we found that neither measure of memory generalization was significantly associated with the strength of VMPFC-anterior or VMPFC-posterior hippocampal connectivity. Although the current study utilized a larger sample size than the original, we acknowledge that the sample size of the current study is still not optimal for examining individual differences. Whether our disparate findings reflect a true null finding, are due to task differences, or are due to an underpowered ability to measure individual differences in the current data cannot be determined. Future work that is specifically designed with studying these individual

differences is needed in order to determine the nature of the relationship between VMPFC-hippocampal connectivity and behavior.

No Differential Connectivity Preferences Between the Hippocampus and MTG

Given the wealth of prior work that has implicated the MTG in studies that examine memory integration (Bonnici et al., 2012; Takashima et al., 2009; Tompary & Davachi, 2017), we predicted that the anterior hippocampus would be more functionally connected to MTG than the posterior hippocampus. However, we did not find evidence for differential connectivity between the hippocampus and MTG in the current study. Though there was a numerical difference between anterior and posterior hippocampus connectivity with MTG, this difference did not approach reaching significance which is consistent with Frank et al. (2019)'s marginal evidence for anterior hippocampus-MTG connectivity. Although the prior study did not find any evidence for changes in hippocampal-MTG connectivity across task phases, more recent work has indicated that the MTG may communicate with multiple systems to support memory integration. Ren et al. (2020) observed that greater functional connectivity between MTG and the hippocampus was associated with the ability to construct new concepts while greater connectivity between MTG and executive control regions was associated with breaking down the boundaries of old concepts. As such, the mechanism through which the MTG supports concepts and generalization may be through its interaction with multiple neural systems of which the hippocampus is just one. Thus, measures of background connectivity between anterior hippocampus and MTG may not be the best indicator of the relationship between these two regions.

Individual Differences in Hippocampal Connectivity with Cortical Visual Regions Tracks Generalization Performance

In Chapter 3 we found that category information was represented in higher-order visual cortex. Therefore, we included two additional visual regions in our analyses (LO, PFUS) that were not part of the original study. Although we did not find any significant anterior vs. posterior hippocampal connectivity differences with either visual region, we examined whether there were any connectivity links with behavioral measures of memory generalization. We found a significant positive relationship between performance on the generalization task and PFUS-anterior hippocampal connectivity, with a similar pattern for PFUS-posterior hippocampal connectivity that did not reach significance. We also found a significant positive relationship between the category bias in perceptual similarity measure of memory generalization and PFUS-posterior hippocampus connectivity, with a similar pattern for PFUS-anterior hippocampal connectivity that did not reach significance. We did not find any significant relationships between behavior and LO-hippocampal connectivity.

Our findings for hippocampal-visual cortex connectivity are consistent with past research demonstrating task-related changes in functional connectivity between the hippocampus and higher order visual cortex. Increased connectivity between the hippocampus and fusiform face area has been demonstrated during encoding of face information (Rajah, McIntosh, & Grady, 1999) and during imagination processes involving construction of new, never-encountered scenes (Zeidman, Mullally, & Maguire, 2015). Findings by Zeidman et al. (2015) are particularly relevant to the present study as imagination of never-encountered scenes rely on an integrative process that may be similar

to the process utilized for successful memory generalization. Increased hippocampal connectivity with visual cortex has also been implicated in subsequent memory effects where greater connectivity between the hippocampus and visual cortex at encoding is associated with better memory at retrieval (Ranganath, Heller, Cohen, Brozinsky, & Rissman, 2005) and disruption of connectivity between the hippocampus and higher order visual cortex is evident in elderly individuals diagnosed with Alzheimer's Disease (Wang et al., 2006). Together, past work indicates that communication between the hippocampus and higher order visual cortex is integral for successful memory. Our current findings extend the importance of communication between these regions to a new aspect of memory: spontaneous memory generalization.

Conclusions

The current study provides additional support for theories of functional dissociations along the long axis of the hippocampus by demonstrating connections between the anterior hippocampus and a key generalization region (VMPFC) and between the posterior hippocampus and a key specificity region (ANG). Replication of these findings under learning conditions that emphasize memory specificity but also elicit spontaneous generalization bolster previous work further confirming that these connections reflect stable, intrinsic communication networks between regions. Further, unexpected exploratory findings for hippocampal-higher order visual cortex connectivity relationships with increased memory generalization provide preliminary evidence that the hippocampus interacts with visual cortex to support spontaneous generalization during learning.

CHAPTER V

GENERAL DISCUSSION

The goal of the dissertation was to evaluate the behavioral and neural mechanisms that support spontaneous generalization during learning that emphasizes memory specificity. Often memory generalization has been studied under learning conditions that either explicitly prompt generalization or under conditions where generalization proceeds more incidentally to the task at hand. However, our real-world observations often highlight circumstances in which it may be beneficial to remember the details of our individual experiences as well as the commonalities across experiences simultaneously. To our knowledge there is no research in the literature that has determined whether memory generalization proceeds during learning that emphasizes maintaining specificity. Thus, we developed a novel, observational, paired associates learning task where a shared label provided an opportunity to form categorical knowledge but learning goals explicitly required participants to differentiate all stimuli, even those with shared labels.

Integrated Summary of Results

In our behavioral testing (Chapter 2) we found evidence for a category bias in perceived similarity ratings indicating items learned to be within a category were perceived as more similar to one another than equally physically similar faces from different families. The category bias in perception predicted performance on an explicit generalization task that required applying learned category labels to never-studied stimuli. Critically, the category-bias in perception was measurable immediately after learning and *prior* to the categorization task that had explicit generalization demands. Thus, we reasoned that a category bias in perception may be a good behavioral index of

memory generalization that occurs during encoding rather than in response to task-demands to generalize. However, because collecting perceived similarity ratings may itself carry a minimal task-demand to rate same-category items as more similar to one another, the extent to which the category-bias reflected bias acquired during learning itself could not be definitively determined.

To better determine whether the category bias observed in Chapter 2 reflected real category-biased changes that occurred during learning, we examined neural biases during learning using functional MRI (Chapter 3). We replicated our previous finding that individual differences in category-bias in perceptual similarity ratings predicted memory generalization performance. Overall, during learning we found evidence for widespread category-biased neural representations throughout the cortex. This included some regions implicated in prior work as important structures for memory generalization, but also other regions, including higher-order visual cortex. Results indicated that although both category-relevant and category-irrelevant information was pertinent to task goals during encoding, representations were overwhelmingly biased towards category-relevant information. Our findings are consistent with past work that indicates category learning may induce attentional shifts towards category-relevant information (Goldstone & Steyvers, 2001; Kruschke, 1996; Medin & Schaffer, 1978; Nosofsky, 1991; Nosofsky, 1986) but extend these findings to a task where category-irrelevant information is still relevant for the explicit task goals.

Our behavioral findings that individuals were able to both remember the individual stimuli encountered during training as well as form generalizable category knowledge may indicate that memory for specific details is maintained alongside

generalizable category information. Given the well-known hippocampal role in supporting memory for specific details (Scoville & Milner, 1957; Squire & Zola, 1998), and more recent evidence that implicates the anterior hippocampus in memory generalization (Bowman, Iwashita, et al., 2020; Bowman & Zeithamova, 2018; Kumaran et al., 2009), we explored whether the hippocampus could support both processes via an anterior/posterior division of labor (Chapter 4). We examined background connectivity during the paired associates learning task to determine whether anterior hippocampus is more functionally connected with putative memory generalization regions and whether posterior hippocampus is more functionally connected with putative memory specificity regions. Consistent with prior findings by Frank, Bowman, and Zeithamova (2019), we found functional dissociations along the long axis of the hippocampus and extend these findings to a task that emphasizes memory specificity but elicits spontaneous generalization. Although hippocampal connectivity with these putative specificity and generalization regions did not track task behavioral performance, preliminary exploratory findings demonstrated that greater connectivity between the hippocampus and higher-order visual cortex was associated with increased memory generalization. Our findings indicate that there are differential intrinsic connections between the hippocampus and key cortical generalization and specificity regions which may guide spontaneous generalization during encoding. The hippocampus may also interact with higher-order visual cortex to support spontaneous generalization during learning.

Together, our results provide evidence that spontaneous generalization may occur during learning even when task-demands during encoding require differentiation of all stimuli. Category-biased neural representations, which are also reflected in category-

biased perceptual similarity ratings, spontaneously form during encoding. While widespread cortex reflects category-biased neural representations, the hippocampus may also be at play by providing simultaneous communication to specificity and generalization networks during learning.

Category Learning Biases Attention to Category-Relevant Information Even When Task Goals Emphasize Specificity

Attention has long been assumed to guide successful category learning. Exemplar models of category learning posit that categorization involves comparing the similarity of previously learned items with new incoming information to determine category membership (Medin & Schaffer, 1978; Nosofsky, 1986). Alternatively, prototype models posit that categorization involves comparing the similarity of new incoming information to a "prototypical" category representation created by extracting the central tendency across all learned category exemplars (Homa et al., 1973; Posner & Keele, 1968). While the hypothesized mechanisms underlying category learning in these two models are quite different (comparisons to all previously learned items vs. comparisons to an abstract representation) one thing they do share in common is their prediction of attentional shifts to category-relevant information during learning. Specifically, attentional shifts serve to "compress" and "expand" perceptual space. When items are learned to belong to the same category more attention is allocated towards features of stimuli that would help determine category membership (i.e. category-relevant information). Thus, items within a category become less discriminable from one another (Gureckis & Goldstone, 2008) and are perceived as more similar to one another after learning (Goldstone et al., 2001; Kurtz, 1996; Livingston et al., 1998). Additionally, when items are learned to belong to

different categories, less attention is allocated to the features that distinguish categories resulting in these items being more discriminable (Beale & Keil, 1995; Folstein et al., 2013) and perceived as less similar to one another after learning (Livingston et al., 1998).

Exemplar models of category learning predict that individuals should be best at categorizing old items and new items that are closest to the old exemplars (Nosofsky, 1987; Zaki, Nosofsky, Stanton, & Cohen, 2003). Thus, shifting attention to category-relevant features would guide in determining how similar new items are to the already-stored memory representations of each old item. Prior work has found evidence for attention-biased exemplar representations in the brain (Mack et al., 2013). These attention-biased representations were found in lateral occipital cortex, inferior parietal cortex, inferior frontal gyrus, and insular cortex. Consistent with this prior work, we found evidence for category-biased information widespread across the cortex during learning including higher-level visual cortices indicating that category-relevant information was prioritized. On the other hand, Mack and colleagues (2013) also examined an exemplar model without selective attention that instead was derived from the physical similarity of the training stimuli and found only a single region in primary visual cortex tracked this information. Our finding of only a single region in lateral occipital cortex that represented category-irrelevant information is also consistent with this prior work.

Prototype models of category learning predict that individuals should be best at categorizing the never-studied category prototypes themselves (even better than categorizing learned exemplars) and that performance should suffer as exemplars share less features with the prototype (Minda & Smith, 2001). Thus, shifting attention to

category-relevant features guides in determining how similar new items are to the stored category prototype. Prior work has found evidence for abstract category representations in the anterior hippocampus as well as established memory generalization regions like the ventromedial prefrontal cortex and middle temporal gyrus (Bowman, Iwashita, et al., 2020; Bowman & Zeithamova, 2018). Consistent with this work we found evidence for category-biased information in ventromedial prefrontal cortex and middle temporal gyrus during learning and demonstrated intrinsic connections between anterior hippocampus and middle temporal gyrus. Although abstract prototype representations in prior work were found to be unique to hypothesized generalization regions, we found evidence for category-biased information more widespread across the brain. We hypothesize that these differences may be due to the structure of the categories learned and differences in task goals. In prior studies (Bowman, Iwashita, et al., 2020; Bowman & Zeithamova, 2018), all features of the stimuli were equally relevant for determining category while in the novel paired associates task used here participants had to learn both the relevant and irrelevant category features in order to categorize as well as tell all stimuli apart from one another. Thus, in the current task attentional shifts to complete task goals may have recruited a larger extent of the brain than was required in these prior studies.

In the current experiments it is clear that related experiences are already linked in some manner at encoding resulting in behavioral and neural category-biases. However, because both exemplar and prototype models predict the same attentional shifts towards category-relevant information, it is not possible to determine whether we can consider the representations we observed during encoding as truly "generalized" memory representations. However, given recent findings by Bowman, Iwashita and Zeithamova

(2020) that show evidence for both prototype and exemplar representations formed in parallel during the learning process, we speculate that individuals may store both exemplar and prototype representations during encoding. While we cannot fit formal prototype and exemplar models to the data collected in the current study, behavioral data are consistent with this idea. Individuals were able to remember the first names of the stimuli across two experiments presented in Chapters 2 and 3. Further, in Chapter 3 we found good recognition memory performance which has also been shown in another study using the same paradigm in both younger and older adults (Bowman, Ashby, & Zeithamova, 2020). This indicates that participants must have stored representations of individual faces. Additionally, across experiments participants were able to successfully categorize never-seen faces indicating that although learning emphasized specificity (and there is good behavioral evidence for successful specificity) they were still able to extract category information in service of memory generalization. Because our results demonstrated that individuals attained a good level of specificity in their memories for the individual training faces as well as being able to generalize to never-studied faces, it may indicate that under the task-demands of the current paradigm both specific and generalized memory representations are constructed during learning. Our findings extend prior knowledge for the role of attentional shifts in category learning to a new paradigm which prioritizes both category-relevant and category-irrelevant information.

Does Category Bias in Perception Reflect a True Learning-Driven Perceptual Change or a Strategic Decision to Generalize Because of Similar Labels?

Learning category information is thought to alter perception such that items within a category are viewed as more similar and/or items from different categories are

viewed as less similar to one another after learning (Beale & Keil, 1995; Goldstone, 1994a, 1994b; Goldstone et al., 2001; Rosch & Mervis, 1975). However, there has been some concern in the literature about whether these perceptual category biases after category learning reflect actual changes in perception or merely a *strategic judgment bias* to rate same-category items as more similar than between-category items (Goldstone et al., 2001). Throughout the studies presented in the dissertation, we found evidence for category-biased perceptual ratings changes that also predicted generalization success. Further, we found category-biased neural representations during learning indicating that category knowledge may be spontaneously linked at encoding rather than in response to explicit generalization task demands. However, though the category bias predicted generalization performance, we cannot fully determine whether or not a strategic decision to rate same-category faces as more similar to one another is reflected in the category-bias measure. Given the findings, a combination of perceptual changes and strategic judgment bias may be reflected in the perceived similarity ratings.

To test this idea, we pre-registered and are in the process of running a new behavioral study to determine to what degree the category bias in similarity ratings after category learning indicates a true change in perception/attention vs. a strategic decision bias. In the currently ongoing study, we tracked category bias after a traditional feedback-based category learning task with category structures learned under two conditions (Figure 5.1). For participants randomly assigned to the first condition (N = 93), category membership is in line with the physical similarity of face-blend stimuli and items in the same category share physical characteristics (as was true in the experiments presented throughout the dissertation). For participants randomly assigned to the second condition

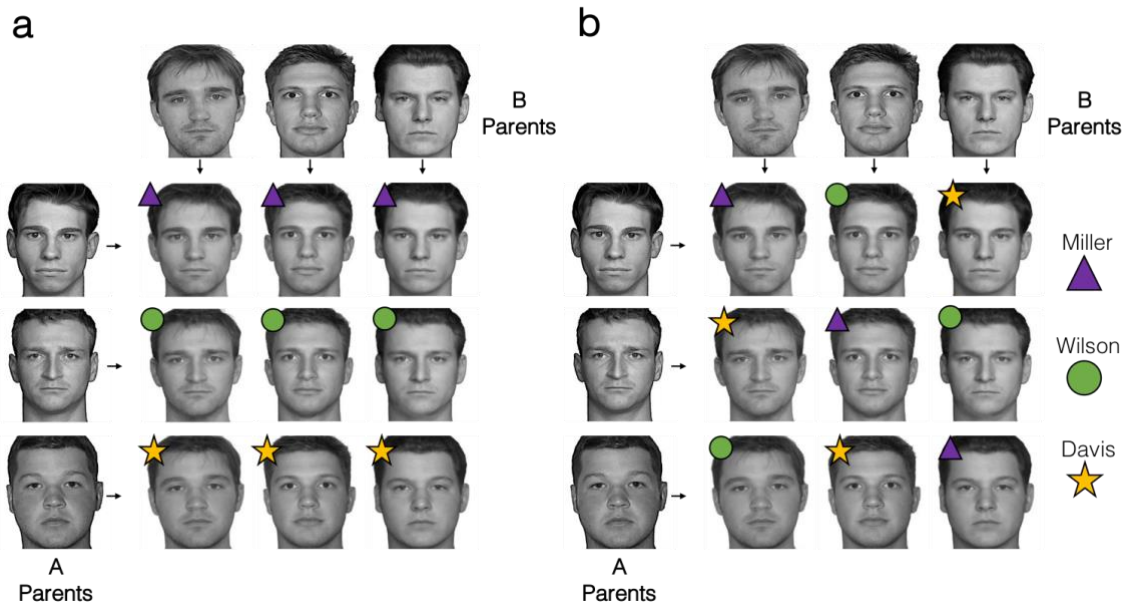

Figure 5.1. Different family category structures for two conditions. **A.** In condition 1, family assignment is determined by blending with shared 'A' parents. **B.** In condition 2, family assignment is dissociated from physical similarity excluding any shared parents.

(N = 97), category membership is completely dissociated from physical similarity and items in the same category do not share any physical characteristics (although items across category boundaries still share physical similarities).

Preliminary results from this study are presented in Figure 5.2. Unexpectedly we found strong evidence for a category bias in perception in the condition where within-category faces shared physical similarity, but no evidence for a category bias in the condition where faces did not share within-category similarities (Figure 5.2b). Notably, learning the category structure was more difficult in the non-physical similarity condition (Figure 5.2a, red line) and thus we ran a control analysis where we limited the subjects included in Condition 2 to only be the top performers (N = 34, see Figure 5.2c) to equate the two groups for learning. However, even after controlling for the degree of learning by

the end of training, a category bias in perception only emerged in the condition where faces within a category shared physical similarities (Figure 5.2d). Because a strategic judgment bias account would predict a category bias in both conditions—as shared labels should bias increased similarity regardless of physical similarity—our preliminary data

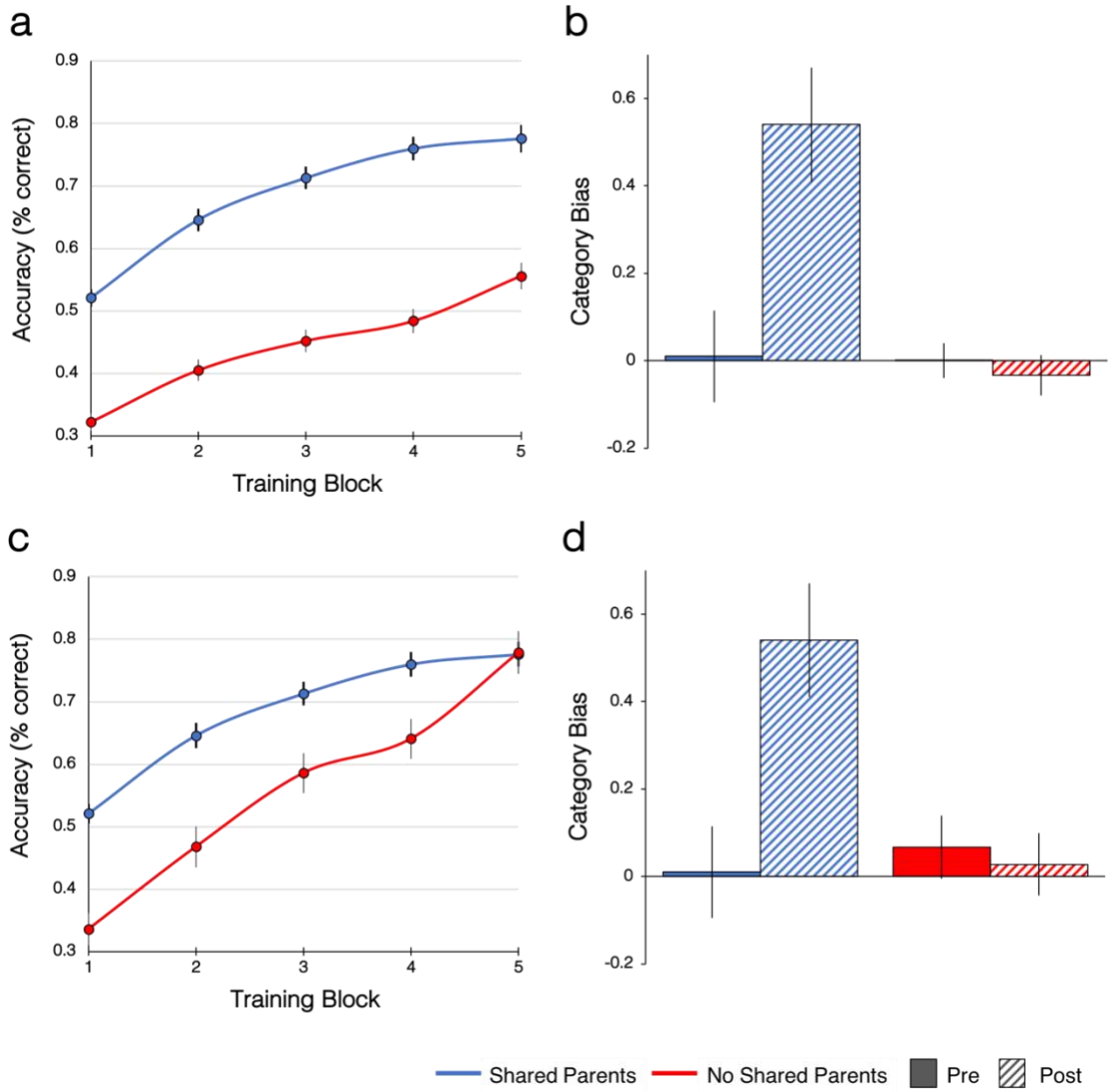

Figure 5.2. Preliminary data indicating category-bias reflects true learning-related perceptual changes. **A.** Mean accuracy across all five training blocks of the feedback-based category learning task. Learning the category structure in Condition 2 was more difficult than in Condition 1. **B.** A significant post-learning category bias is present in the condition in which within-category faces share physical similarity. **C.** Training performance across all five training blocks after subjects in Condition 2 were limited to top performers. By block five, performance between groups is equated. **D.** A significant post-learning category bias is present only in Condition 1 even when both groups are equated for learning.

indicate that the category-bias in perceived similarity ratings reflect true changes in perception following learning.

Our results are consistent with work by Goldstone and colleagues (2001) who examined differences in similarity ratings between categorized objects and neutral, uncategorized objects. They reasoned that if similarity ratings reflected an actual change in perception all objects learned to be within the same category should have similar ratings when compared to a never-studied neutral object (e.g. A and B are in the same category and E is the neutral stimulus. Similarity ratings for A/E and B/E should become more similar after learning). Alternatively, if strategic judgment bias accounted for the similarity ratings then they predicted that there should not be any greater concordance of similarity ratings between objects in the same category compared to a neutral object. Consistent with learning-induced perceptual changes they found that same-category items relative to a neutral item became more similar to one another after learning. Together, these findings provide exciting new evidence that the category-bias in perception measure collected in the current studies is an accurate reflection of real biases in perception that indicate the degree of category knowledge acquired during learning. The ability to measure the category-bias in perception allows for the detection of generalization processes under minimal task-demands and extends our ability to detect generalization in paradigms without explicit generalization tests.

The Role of the Hippocampus in Spontaneous Category Learning

The most widely known and accepted function of the hippocampus is to support encoding of detailed episodic memory (Scoville & Milner, 1957) and to reduce interference between similar experiences as they are encountered through pattern

separation (Yassa & Stark, 2011). However, more recent work has begun to uncover contributions of the hippocampus to other processes like episodic inference (Ryan et al., 2016; Schlichting et al., 2015; Shohamy & Wagner, 2008; Zeithamova, Dominick, et al., 2012; Zeithamova & Preston, 2010). It has been theorized that the hippocampus may be able to support multiple processes via a division of labor along the long axis of the hippocampal body (Brunec et al., 2018; Poppenk et al., 2013). Consistent with this theory we replicated previous findings for differential anterior/posterior hippocampal connectivity with ventromedial prefrontal cortex and angular gyrus (Frank, Bowman, et al., 2019). Specifically, we found greater functional connectivity between the anterior hippocampus and the ventromedial prefrontal cortex and greater connectivity between the posterior hippocampus and angular gyrus. In contrast with work that suggests the anterior hippocampus may play a role in memory generalization (Bowman, Iwashita, et al., 2020; Bowman & Zeithamova, 2018), we did not find any evidence across our studies for category-biased representations in anterior hippocampus.

As our study is the first to our knowledge to examine neural representations underlying category learning in a task that emphasizes specificity, we suspect that the relative lack of hippocampal involvement in representing category biased information may be due to task-demands to treat all information separately, even items within the same category. Though neural category bias was measurable throughout the cortex during learning, the hippocampus may have been recruited by our task to perform more pattern separation processes as needed for task goals to differentiate all stimuli from one another, even “brothers” within the same family.

To our surprise, we did find evidence for category-biased information in higher

order visual cortices and preliminary evidence indicating that the hippocampus may interact with the visual cortex to support memory generalization. While the hippocampus has been traditionally thought of as a region that encodes detailed memories, recent work has found the hippocampus to also be involved in perceptual discrimination (Barense et al., 2005; A. C. H. Lee & Rudebeck, 2010). A representational-hierarchical model has considered that medial temporal lobe structures like the perirhinal cortex may serve as an extension of the ventral visual stream (Saksida & Bussey, 2010; Ungerleider & Mishkin, 1982) which is involved in object identification (Goodale & Milner, 1992). Because the hippocampus is largely connected with the perirhinal cortex, it has further been suggested that the hippocampus itself may sit at the top of the hierarchy and may be important for higher-order visual processing like assessing combinations of features to allow for successful discrimination of complex stimuli (for review see Lee, Yeung, & Barense, 2012). Our findings that enhanced intrinsic connections between hippocampus and visual cortex predict better memory generalization performance are consistent with this idea. Due to the complex nature of the stimuli category structure in the current paradigm, interactions between the hippocampus and visual cortex may reflect attentional processes for determining which features are category-relevant and which are category-irrelevant. Thus, although category-biased representations were not found during learning in the hippocampus, the hippocampus may have facilitated category-biased representations in visual cortex as evidenced through intrinsic connections between regions.

Broader Implications

Our finding that inclusion of a mere category label was enough to bias representations of individual faces both perceptually and neurally are timely in light of

current social concerns regarding racism. Implicit bias is the notion that our perceptions or actions are unconsciously stereotyped to value one group above another even though we may not have conscious awareness of said bias (Amodio, 2014). A stereotype in and of itself can be considered a type of generalization about a particular group of people (Stevens & Abernethy, 2018) and thus due to underlying cultural biases we tend to generalize people by placing them into clusters based on race. Implicit bias comes in to play when we "overgeneralize" individuals or entire racial groups into additional categories (e.g. good/bad, criminal/law-abiding, truthful/liar etc.).

Prior work has postulated that negative racially driven biases develop through an associative learning process (Olsson, Ebert, Banaji, & Phelps, 2005) that proceeds much like the current experiments presented here. However, rather than merely pairing faces and names (or even just people that share racial features) race is often paired with fear and that race-fear association results in a negative bias towards other groups (Santos, Meyer-Lindenberg, & Deruelle, 2010). Thus, once an association is made, neural representations for racial categories may become biased much like the category-representations we observed in the current studies with the added complexity of fear associations.

The current results add an interesting layer to the implicit bias discussion. The fact that we found such widespread evidence for perceptual and neural category biases using faces that were held constant for in-group and out-group physical similarity is striking considering that racial biases are rooted in very salient physical differences between groups. As our preliminary data presented above indicates that perceptual evidence for category biases require at least some degree of shared physical similarity to

manifest, we speculate that representations may be biased on the account of race to a larger degree than we observed here and/or may even include more widespread cortical involvement. This is consistent with prior work that has implicated the amygdala in racial bias (Amodio, 2014) as well as other work that has found dorsolateral prefrontal cortex and anterior cingulate cortex modulation to a greater degree in individuals who measure high on behavioral measures of implicit bias (Richeson et al., 2003).

Unfortunately, other work has demonstrated that implicit bias is quite difficult to extinguish (Bouton, 1994; Sloman, 1996). Our results are consistent with this idea as we found that biased representations were measurable even when individuals were tasked with focusing on each person as an individual. Still, participants in the current study formed behavioral and neural category biased representations although category information was not pertinent to the task at hand. It has been proposed that this difficulty in extinguishing racial biases stems from consistent immersion in cultural routines (e.g. news and entertainment media consumption) that reinforce negative stereotypes (Amodio, 2014). We suspect that part of the difficulty with extinguishing implicit bias may also be due to the spontaneous nature of generalization. We found evidence that information is spontaneously linked as it is encoded and not just in response to generalization judgment demands. As individuals encounter racial biases in media, they may immediately begin to link that information with existing representations and therefore the biased representation is reactivated during encoding. Thus, biased representations are brought back to memory consistently and extinguishing already-established racial biases is an uphill battle. More research is needed to understand the best methods in which we can overcome this challenge. While memory generalization is

a useful heuristic to make quick work of organizing the world around us into meaningful clusters of information, implicit bias is a clear example of how this heuristic can be disadvantageous and harmful. Moving forward, it may be that the best way to improve implicit bias is to stop negative race associations from even occurring in the first place. Thus, a two-fold approach is needed where we can establish ways to root out our own already-established implicit biases, or at least actively work against established biases, while also creating a better world where the younger generation never creates these types of associations in the first place.

General Conclusions

Across three empirical studies described in Chapters 2-4, we investigated the behavioral and neural mechanisms that support spontaneous generalization during learning that emphasizes memory specificity. We demonstrated that the mere presence of a category label was sufficient to cause individuals to link category-relevant information in support of memory generalization even though task goals required differentiation of all stimuli and encoding of both category-relevant and category-irrelevant information. We demonstrated that representations spontaneously become more category-biased during encoding as evidenced through behavioral biases in perception and neural biases during encoding throughout the brain. Together, our results inform our understanding of theories of memory generalization by demonstrating conditions under which memory generalization may proceed spontaneously during encoding and has broader implications for our understanding of the nature of stereotyping and implicit cognitive biases.

REFERENCES CITED

Aizenstein, H., MacDonald, A., Stenger, V., Nebes, R., Larson, J., Ursu, S., & Carter, C. (2000). Complementary category learning systems identified using fMRI. *Journal of Cognitive Neuroscience*, *12*(6), 977–987.

Amodio, D. M. (2014). The neuroscience of prejudice and stereotyping. *Nature Reviews Neuroscience*, *15*(10), 670–682. https://doi.org/10.1038/nrn3800

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

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

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

Ashby, F. G., & O'Brien, J. B. (2005). Category learning and multiple memory systems. *Trends in Cognitive Sciences*, *9*(2), 83–89. https://doi.org/10.1016/j.tics.2004.12.003

Ashby, S. R., Bowman, C. R., & Zeithamova, D. (2020). Perceived similarity ratings predict generalization success after traditional category learning and a new paired-associate learning task. *Psychonomic Bulletin and Review*. https://doi.org/10.3758/s13423-020-01754-3

Baldassano, C., Hasson, U., & Norman, K. A. (2018). Representation of real-world event schemas during narrative perception. *Journal of Neuroscience*, *38*(45), 9689–9699. https://doi.org/10.1523/JNEUROSCI.0251-18.2018

Banino, A., Koster, R., Hassabis, D., & Kumaran, D. (2016). Retrieval-based model accounts for striking profile of episodic memory and generalization. *Scientific Reports*, *6*, 1–15. https://doi.org/10.1038/srep31330

Barense, M. D., Bussey, T. J., Lee, A. C. H., Rogers, T. T., Davies, R. R., Saksida, L. M., … Graham, K. S. (2005). Functional specialization in the human medial temporal lobe. *Journal of Neuroscience*, *25*(44), 10239–10246. https://doi.org/10.1523/JNEUROSCI.2704-05.2005

Beale, J. M., & Keil, F. C. (1995). Categorical effects in the perception of faces. *Cognition*, *57*, 217–239.

Berens, S. C., & Bird, C. M. (2017). The role of the hippocampus in generalizing configural relationships. *Hippocampus*, *27*(3), 223–228. https://doi.org/10.1002/hipo.22688

Bokde, A. L. W., Lopez-Bayo, P., Meindl, T., Pechler, S., Born, C., Faltraco, F., … Hampel, H. (2006). Functional connectivity of the fusiform gyrus during a face-matching task in subjects with mild cognitive impairment. *Brain*, *129*(5), 1113–1124. https://doi.org/10.1093/brain/awl051

Bonnici, H. M., Chadwick, M. J., Lutti, A., Hassabis, D., Weiskopf, N., & Maguire, E. a. (2012). Detecting representations of recent and remote autobiographical memories in vmPFC and hippocampus. *The Journal of Neuroscience : The Official Journal of the Society for Neuroscience*, *32*(47), 16982–16991. https://doi.org/10.1523/JNEUROSCI.2475-12.2012

Bouton, M. E. (1994). Conditioning, remembering, and forgetting. *Journal of Experimental Psychology: Animal Behavior Processes*, *20*(3), 219–231. https://doi.org/10.1037//0097-7403.20.3.219

Bowman, C. ., Iwashita, T., & Zeithamova, D. (2020). Tracking prototype and exemplar representations in the brain across learning. *ELife*, 1–24. https://doi.org/https://doi.org/10.7554/eLife.59360

Bowman, C. R., Ashby, S. R., & Zeithamova, D. (2020). *Age effects on category learning and their relationship to deficits in memory specificity*. *PsyArXiv*. https://doi.org/https://doi.org/10.31234/osf.io/xr3ad

Bowman, C. R., & Dennis, N. A. (2016). The neura basis of recollection rejection: Increases in hippocampal-prefrontal connectivity in absence of a shared recall-to-reject and target recollection network. *Journal of Cognitive Neuroscience*, *28*(8), 1194–1209. https://doi.org/10.1162/jocn

Bowman, C. R., & Zeithamova, D. (2018). Abstract memory representations in the ventromedial prefrontal cortex and hippocampus support concept generalization. *The Journal of Neuroscience*, *38*(10), 2811–2817. https://doi.org/10.1523/JNEUROSCI.2811-17.2018

Bozoki, A., Grossman, M., & Smith, E. E. (2006). Can patients with Alzheimer's disease learn a category implicitly? *Neuropsychologia*, *44*(5), 816–827. https://doi.org/10.1016/j.neuropsychologia.2005.08.001

Brod, G., Lindenberger, U., Werkle-Bergner, M., & Shing, Y. L. (2015). Differences in the neural signature of remembering schema-congruent and schema-incongruent events. *NeuroImage*, *117*, 358–366. https://doi.org/10.1016/j.neuroimage.2015.05.086

Brown, T. I., & Stern, C. E. (2014). Contributions of medial temporal lobe and striatal memory systems to learning and retrieving overlapping spatial memories. *Cerebral Cortex*, *24*(7), 1906–1922. https://doi.org/10.1093/cercor/bht041

Brunec, I. K., Bellana, B., Ozubko, J. D., Man, V., Robin, J., Liu, Z. X., … Moscovitch, M. (2018). Multiple scales of representation along the hippocampal anteroposterior axis in humans. *Current Biology*, *28*(13), 2129–2135. https://doi.org/10.1016/j.cub.2018.05.016

Bunsey, M., & Elchenbaum, H. (1996). Conservation of hippocampal memory function in rats and humans. *Nature*, *379*(6562), 255–257. https://doi.org/10.1038/379255a0

Cai, D. J., Aharoni, D., Shuman, T., Shobe, J., Biane, J., Song, W., … Silva, A. J. (2016). A shared neural ensemble links distinct contextual memories encoded close in time. *Nature*, *534*(7605), 115–118. https://doi.org/10.1038/nature17955

Carpenter, A. C., & Schacter, D. L. (2017). Flexible retrieval: When true inferences produce false memories. *Journal of Experimental Psychology: Learning, Memory, and Cognition*, *43*(3), 335–349.

Carpenter, A. C., & Schacter, D. L. (2018). False memories, false preferences: Flexible retrieval mechanisms supporting successful inference bias novel decisions. *Journal of Experimental Psychology: General*, *147*(7), 988–1004. https://doi.org/10.1037/xge0000391

Carpenter, A. C., Thakral, P. P., Preston, A. R., & Schacter, D. L. (2021). Reinstatement of Item-Specific Contextual Details During Retrieval Supports Recombination-Related False Memories. *NeuroImage*, 118033. https://doi.org/10.1016/j.neuroimage.2021.118033

Collin, S. H. P., Milivojevic, B., & Doeller, C. F. (2015). Memory hierarchies map onto the hippocampal long axis in humans. *Nature Neuroscience*, *18*(11), 1562–1564. https://doi.org/10.1038/nn.4138

Davey, J., Cornelissen, P. L., Thompson, H. E., Sonkusare, X. S., Hallam, G., Smallwood, J., & Jefferies, X. E. (2015). Automatic and Controlled Semantic Retrieval : TMS Reveals Distinct Contributions of Posterior Middle Temporal Gyrus and Angular Gyrus, *35*(46), 15230–15239. https://doi.org/10.1523/JNEUROSCI.4705-14.2015

Deng, Y., Booth, J. R., Chou, T. L., Ding, G. S., & Peng, D. L. (2008). Item-specific and generalization effects on brain activation when learning Chinese characters. *Neuropsychologia*, *46*(7), 1864–1876. https://doi.org/10.1016/j.neuropsychologia.2007.09.010

Dennis, N. A., Kim, H., & Cabeza, R. (2008). Age-related differences in brain activity during true and false memory retrieval. *Journal of Cognitive Neuroscience*, *20*(8), 1390–1402. https://doi.org/10.1162/jocn.2008.20096

DeVito, L. M., Lykken, C., Kanter, B. R., & Eichenbaum, H. (2010). Prefrontal cortex: Role in acquisition of overlapping associations and transitive inference. *Learning & Memory*, *17*(3), 161–167. https://doi.org/10.1101/lm.1685710

Doeller, C. F., King, J. A., & Burgess, N. (2008). Parallel striatal and hippocampal systems for landmarks and boundaries in spatial memory. *Proceedings of the National Academy of Sciences of the United States of America*, *105*(15), 5915–5920. https://doi.org/10.1073/pnas.0801489105

Ekstrom, A. D., Copara, M. S., Isham, E. A., Wang, W. chun, & Yonelinas, A. P. (2011). Dissociable networks involved in spatial and temporal order source retrieval. *NeuroImage*, *56*(3), 1803–1813. https://doi.org/10.1016/j.neuroimage.2011.02.033

Faul, F., Erdfelder, E., Buchner, A., & Lang, A.-G. (2009). Statistical power analyses using G*Power 3.1: Tests for correlation and regression analyses. *Behavior Research Methods*, *41*(4), 1149–1160. https://doi.org/10.3758/BRM.41.4.1149

Faul, F., Erdfelder, E., Lang, A. G., & Buchner, A. (2007). G*Power: A flexible statistical power analysis program for the social, behavioral, and biomedical sciences. *Behavior Research Methods*, *39*(2), 175–191. https://doi.org/10.3758/BF03193146

Fernandez-Ruiz, J., Wang, J., Aigner, T. G., & Mishkin, M. (2001). Visual habit formation in monkeys with neurotoxic lesions of the ventrocaudal neostriatum. *Proceedings of the National Academy of Sciences of the United States of America*, *98*(7), 4196–4201. https://doi.org/10.1073/pnas.061022098

Folstein, J. R., Palmeri, T. J., & Gauthier, I. (2013). Category learning increases discriminability of relevant object dimensions in visual cortex. *Cerebral Cortex*, *23*(4), 814–823. https://doi.org/10.1093/cercor/bhs067

Fournier, J., Saleem, A. B., Diamanti, E. M., Wells, M. J., & Harris, K. D. (2019). Modulation of visual cortex by hippocampal signals.

Frank, L. E., Bowman, C. R., & Zeithamova, D. (2019). Differential functional connectivity along the long axis of the hippocampus aligns with differential role in memory specificity and generalization. *Journal of Cognitive Neuroscience*, *31*(12), 1958–1975. https://doi.org/10.1162/jocn_a_01457

Frank, L. E., Preston, A. R., & Zeithamova, D. (2019). Functional connectivity between memory and reward centers across task and rest track memory sensitivity to reward. *Cognitive, Affective and Behavioral Neuroscience*, *19*(3), 503–522. https://doi.org/10.3758/s13415-019-00700-8

Gabay, Y., Dick, F. K., Zevin, J. D., & Holt, L. L. (2015). Incidental auditory category learning. *Journal of Experimental Psychology: Human Perception and Performance*, *41*(4), 1124–1138. https://doi.org/10.1037/xhp0000073

Gabrieli, J. D. E., Poldrack, R. A., & Desmond, J. E. (1998). The role of left prefrontal cortex in language and memory. *Proceedings of the National Academy of Sciences of the United States of America*, *95*(3), 906–913. https://doi.org/10.1073/pnas.95.3.906

Garoff-Eaton, R. J., Slotnick, S. D., & Schacter, D. L. (2006). Not all false memories are created equal: The neural basis of false recognition. *Cerebral Cortex*, *16*(11), 1645–1652. https://doi.org/10.1093/cercor/bhj101

Gerraty, R. T., Davidow, J. Y., Wimmer, G. E., Kahn, I., & Shohamy, D. (2014). Transfer of Learning Relates to Intrinsic Connectivity between Hippocampus, Ventromedial Prefrontal Cortex, and Large-Scale Networks. *The Journal of Neuroscience : The Official Journal of the Society for Neuroscience*, *34*(34), 11297–11303. https://doi.org/10.1523/JNEUROSCI.0185-14.2014

Gershman, S. J., Schapiro, A. C., Hupbach, A., & Norman, K. A. (2013). Neural context reinstatement predicts memory misattribution. *Journal of Neuroscience*, *33*(20), 8590–8595. https://doi.org/10.1523/JNEUROSCI.0096-13.2013

Ghosh, V. E., Moscovitch, M., Colella, B. M., & Gilbo, A. (2014). Schema representation in patients with ventromedial PFC lesions. *Journal of Neuroscience*, *34*(36), 12057–12070. https://doi.org/10.1523/JNEUROSCI.0740-14.2014

Gilbert, A. L., Regier, T., Kay, P., & Ivry, R. B. (2006). Whorf hypothesis is supported in the right visual field but not the left, *103*(2).

Gold, B. T., & Buckner, R. L. (2002). Common prefrontal regions coactivate with dissociable posterior regions during controlled semantic and phonological tasks. *Neuron*, *35*(4), 803–812. https://doi.org/10.1016/S0896-6273(02)00800-0

Goldstone, R. L. (1994a). Influences of categorization on perceptual discrimination. *Journal of Experimental Psychology: General*, *123*(2), 178–200. https://doi.org/10.1037/0096-3445.123.2.178

Goldstone, R. L. (1994b). The role of similarity in categorization: Providing a groundwork. *Cognition*, *52*, 125–157.

Goldstone, R. L., & Hendrickson, A. T. (2010). Categorical perception. *Wiley Interdisciplinary Reviews: Cognitive Science*, *1*(1), 69–78. https://doi.org/10.1002/wcs.26

Goldstone, R. L., Lippa, Y., & Shiffrin, R. M. (2001). Altering object representations through category learning. *Cognition*, *78*(1), 27–43. https://doi.org/10.1016/S0010-0277(00)00099-8

Goldstone, R. L., & Steyvers, M. (2001). The sensitization and differentiation of dimensions during category learning. *Journal of Experimental Psychology: General*, *130*(1), 116–139.

Goodale, M. A., & Milner, A. D. (1992). Separate visual pathways for perception and action. *Trends in Neuroscience*, *15*(1), 20–25. https://doi.org/https://doi.org/10.1016/0166-2236(92)90344-8

Gratton, C., Laumann, T. O., Nielsen, A. N., Greene, D. J., Gordon, E. M., Gilmore, A. W., … Petersen, S. E. (2018). Functional Brain Networks Are Dominated by Stable Group and Individual Factors, Not Cognitive or Daily Variation. *Neuron*, *98*(2), 439-452.e5. https://doi.org/10.1016/j.neuron.2018.03.035

Greenberg, D. L., Rice, H. J., Cooper, J. J., Cabeza, R., Rubin, D. C., & Labar, K. S. (2005). Co-activation of the amygdala , hippocampus and inferior frontal gyrus during autobiographical memory retrieval, *43*, 659–674. https://doi.org/10.1016/j.neuropsychologia.2004.09.002

Gureckis, T. M., & Goldstone, R. L. (2008). The Effect of the Internal Structure of Categories on Perception. *Proceedings of the 30th Annual Conference of the Cognitive Science Society*. https://doi.org/10.4314/jlt.v44i2.71793

Hämäläinen, H., Hiltunen, J., & Titievskaja, I. (2002). Activation of somatosensory cortical areas varies with attentional state: An fMRI study. *Behavioural Brain Research*, *135*(1–2), 159–165. https://doi.org/10.1016/S0166-4328(02)00145-6

Hanke, M., Halchenko, Y. O., Sederberg, P. B., Hanson, S. J., Haxby, J. V., & Pollmann, S. (2009). PyMVPA: A Python toolbox for multivariate pattern analysis of fMRI data. *Neuroinformatics*, *7*(1), 37–53. https://doi.org/10.1007/s12021-008-9041-y.PyMVPA

Harnad, S. (2006). Categorical Perception. In L. Nadel (Ed.), *Encyclopedia of Cognitive Science* (pp. 1–5). https://doi.org/10.1002/0470018860.s00490

Hassabis, D., Chu, C., Rees, G., Weiskopf, N., Molyneux, P. D., & Maguire, E. A. (2009). Decoding Neuronal Ensembles in the Human Hippocampus. *Current Biology*, *19*(7), 546–554. https://doi.org/10.1016/j.cub.2009.02.033

Heckers, S., Zalesak, M., Weiss, A. P., Ditman, T., & Titone, D. (2004). Hippocampal activation during transitive inference in humans. *Hippocampus*, *14*(2), 153–162. https://doi.org/10.1002/hipo.10189

Hintzman, D. L. (1984). MINERVA 2: A simulation model of human memory. *Behavior Research Methods, Instruments, & Computers*, *16*(2), 96–101. https://doi.org/10.3758/BF03202365

Hintzman, D. L. (1986). Schema abstraction in a multiple-trace memory model. *Psychological Review*, *93*(4), 411–428.

Homa, D., Cross, J., Cornell, D., Goldman, D., & Shwartz, S. (1973). Prototype abstraction and classification of new instances as a function of number of instances defining the prototype. *Journal of Experimental Psychology*, *101*(1), 116–122. https://doi.org/10.1037/h0035772

Hutchinson, J. B., Uncapher, M. R., & Wagner, A. D. (2009). Posterior parietal cortex and episodic retrieval: Convergent and divergent effects of attention and memory. *Learning and Memory*, *16*(6), 343–356. https://doi.org/10.1101/lm.919109

Johnson, J. D., Suzuki, M., & Rugg, M. D. (2013). Recollection, familiarity, and content-sensitivity in lateral parietal cortex: a high-resolution fMRI study. *Frontiers in Human Neuroscience*, *7*(May), 1–15. https://doi.org/10.3389/fnhum.2013.00219

Kanwisher, N., & Wojciulik, E. (2000). Visual attention: Insights from brain imaging. *Nature Reviews Neuroscience*, *1*, 91–100. Retrieved from http://www.nature.com/nrn/journal/v1/n2/abs/nrn1100_091a.html

Kéri, S., Kálmán, J., Kelemen, O., Benedek, G., & Janka, Z. (2001). Are Alzheimer's disease patients able to learn visual prototypes? *Neuropsychologia*, *39*(11), 1218–1223. https://doi.org/10.1016/S0028-3932(01)00046-X

Kjelstrup, K. B., Solstad, T., Brun, V. H., Hafting, T., Leutgeb, S., Witter, M. P., … Moser, M. B. (2008). Finite scale of spatial representation in the hippocampus. *Science*, *321*(5885), 140–143. https://doi.org/10.1126/science.1157086

Knowlton, B. J., & Squire, L. R. (1993). The learning of categories: Parallel brain systems for item memory and category knowledge. *Science*, *262*, 1747–1749. https://doi.org/10.1126/science.8259522

Kriegeskorte, N., & Bandettini, P. (2007). Analyzing for information, not activation, to exploit high-resolution fMRI. *NeuroImage*, *38*(4), 649–662. https://doi.org/10.1016/j.neuroimage.2007.02.022

Kriegeskorte, N., Goebel, R., & Bandettini, P. (2006). Information-based functional brain mapping. *Proceedings of the National Academy of Sciences of the United States of America*, *103*(10), 3863–3868. https://doi.org/10.1073/pnas.0600244103

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

Kruschke, J. K. (1996). Dimensional relevance shifts in category learning. *Connection Science*, *8*(2), 225–247. https://doi.org/10.1080/095400996116893

Kuhl, B. A., & Chun, M. M. (2014). Successful Remembering Elicits Event-Specific Activity Patterns in Lateral Parietal Cortex. *Journal of Neuroscience*, *34*(23), 8051–8060. https://doi.org/10.1523/JNEUROSCI.4328-13.2014

Kuhl, B. A., Dudukovic, N. M., Kahn, I., & Wagner, A. D. (2007). Decreased demands on cognitive control reveal the neural processing benefits of forgetting. *Nature Neuroscience*, *10*(7), 908–914. https://doi.org/10.1038/nn1918

Kumaran, D. (2012). What representations and computations underpin the contribution of the hippocampus to generalization and inference? *Frontiers in Human Neuroscience*, *6*(June), 1–11. https://doi.org/10.3389/fnhum.2012.00157

Kumaran, D., Summerfield, J. J., Hassabis, D., & Maguire, E. A. (2009). Tracking the Emergence of Conceptual Knowledge during Human Decision Making. *Neuron*, *63*(6), 889–901. https://doi.org/10.1016/j.neuron.2009.07.030

Kurtz, K. J. (1996). Category-based similarity. In G. W. Cottrell (Ed.), *Proceedings of the Eighteenth Annual Conference of the Cognitive Science Society* (p. 290).

Lee, A. C. H., & Rudebeck, S. R. (2010). Investigating the interaction between spatial perception and working memory in the human medial temporal lobe. *Journal of Cognitive Neuroscience*, *22*(12), 2823–2835. https://doi.org/10.1162/jocn.2009.21396

Lee, A. C. H., Yeung, L., & Barense, M. D. (2012). The hippocampus and visual perception. *Frontiers in Human Neuroscience*, *6*, 1–17. https://doi.org/10.3389/fnhum.2012.00091

Lee, H., Samide, R., Richter, F. R., & Kuhl, B. A. (2019). Decomposing Parietal Memory Reactivation to Predict Consequences of Remembering. *Cerebral Cortex*, *29*(8), 3305–3318. https://doi.org/10.1093/cercor/bhy200

Levin, D. T., & Angelone, B. L. (2002). Categorical perception of race. *Perception*, *31*(5), 567–578. https://doi.org/10.1068/p3315

Liberman, A. M., Harris, K. S., Hoffman, H. S., & Griffith, B. C. (1957). The discrimination of speech sounds within and across phoneme boundaries. *Journal of Experimental Psychology*, *54*(5), 358–368. https://doi.org/10.1037/h0044417

Livingston, K. R., Andrews, J. K., & Harnad, S. (1998). Categorical perception effects induced by category learning. *Journal of Experimental Psychology: Learning Memory and Cognition*, *24*(3), 732–753. https://doi.org/10.1037/0278-7393.24.3.732

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

Love, B. C., & Medin, D. L. (1998). SUSTAIN: A model of human category learning. *Proceedings of the Fifteenth National Conference on Artificial Intelligence (AAAI-98)*, 671–676.

Mack, M. L., Preston, A. R., & Love, B. C. (2013). Decoding the brain's algorithm for categorization from its neural implementation. *Current Biology, 23*(20), 2023–2027. https://doi.org/10.1016/j.cub.2013.08.035

McClelland, J. L., McNaughton, B. L., & O'Reilly, R. C. (1995). Why there are complementary learning systems in the hippocampus and neocortex: Insights from the successes and failures of connectionist models of learning and memory. *Psychological Review*, *102*(3), 419–457. https://doi.org/10.1037/0033-295X.102.3.419

Medin, D. L., Dewey, G. I., & Murphy, T. D. (1983). Relationships between item and category learning: Evidence that abstraction is not automatic. *Journal of Experimental Psychology: Learning, Memory, and Cognition*, *9*(4), 607–625. https://doi.org/10.1037/0278-7393.9.4.607

Medin, D. L., & Schaffer, M. M. (1978). Context theory of classification learning. *Psychological Review*, *85*(3), 207–238. https://doi.org/10.1037/0033-295X.85.3.207

Minda, J. P., & Smith, J. D. (2001). Prototypes in category learning: The effects of category size, category structure, and stimulus complexity. *Journal of Experimental Psychology: Learning Memory and Cognition*, *27*(3), 775–799. https://doi.org/10.1037/0278-7393.27.3.775

Morgan, L. K., MacEvoy, S. P., Aguirre, G. K., & Epstein, R. A. (2011). Distances between real-world locations are represented in the human hippocampus. *Journal of Neuroscience, 31*(4), 1238–1245. https://doi.org/10.1523/JNEUROSCI.4667-10.2011

Mummery, C. J., Patterson, K., Price, C. J., Ashburner, J., Frackowiak, R. S. J., & Hodges, J. R. (2000). A voxel-based morphometry study of semantic dementia: Relationship between temporal lobe atrophy and semantic memory. *Annals of Neurology, 47*(1), 36–45. https://doi.org/10.1002/1531-8249(200001)47:1<36::AID-ANA8>3.0.CO;2-L

Murphy, K., Birn, R. M., & Bandettini, P. A. (2013). Resting-state FMRI confounds and cleanup. *NeuroImage*, *80*, 349–359. https://doi.org/10.1016/j.neuroimage.2013.04.001.Resting-state

Murty, V. P., Tompary, A., Adcock, R. A., & Davachi, L. (2017). Selectivity in postencoding connectivity with high-level visual cortex is associated with reward-motivated memory. *Journal of Neuroscience, 37*(3), 537–545. https://doi.org/10.1523/JNEUROSCI.4032-15.2016

Nadel, L., Hoscheidt, S., & Ryan, L. R. (2013). Spatial cognition and the hippocampus: The anterior-posterior axis. *Journal of Cognitive Neuroscience, 25*(1), 22–28. https://doi.org/10.1162/jocn_a_00313

Norman-Haignere, S. V., McCarthy, G., Chun, M. M., & Turk-Browne, N. B. (2012). Category-selective background connectivity in ventral visual cortex. *Cerebral Cortex*, *22*(2), 391–402. https://doi.org/10.1093/cercor/bhr118

Nosofsky, R. M. (1986). Attention, similarity, and the identification-categorization relationship. *Journal of Experimental Psychology: General*, *115*(1), 39–57. https://doi.org/10.1037/0096-3445.115.1.39

Nosofsky, R. M. (1987). Attention and Learning Processes in the Identification and Categorization of Integral Stimuli. *Journal of Experimental Psychology: Learning, Memory, and Cognition*, *13*(1), 87–108. https://doi.org/10.1037/0278-7393.13.1.87

Nosofsky, R. M. (1988). Exemplar-Based Accounts of Relations Between Classification, Recognition, and Typicality. *Journal of Experimental Psychology: Learning, Memory, and Cognition*, *14*(4), 700–708. https://doi.org/10.1037/0278-7393.14.4.700

Nosofsky, R. M. (1991). Tests of an exemplar model for relating perceptual classification and recognition memory. *Journal of Experimental Psychology: Human Perception and Performance*, *17*(1), 3–27. https://doi.org/10.1037/0096-1523.17.1.3

Nosofsky, R. M., Little, D. R., & James, T. W. (2012). Activation in the neural network responsible for categorization and recognition reflects parameter changes. *Proceedings of the National Academy of Sciences*, *109*(1), 333–338. https://doi.org/10.1073/pnas.1111304109

Nosofsky, R. M., & Zaki, S. R. (1998). Dissociations between categorization and recognition in amnesic and normal individuals. *Psychological Science*, *9*(4), 247–255.

O'Reilly, R. C., & Norman, K. A. (2002). Hippocampal and neocortical contributions to memory: Advances in the complementary learning systems framework. *Trends in Cognitive Sciences*, *6*(12), 505–510. https://doi.org/10.1016/S1364-6613(02)02005-3

O'Toole, A. J., Harms, J., Snow, S. L., Hurst, D. R., Pappas, M. R., Ayyad, J. H., & Abdi, H. (2005). A video database of moving faces and people. *IEEE Transactions on Pattern Analysis and Machine Intelligence*, *27*(5), 812–816. https://doi.org/10.1109/TPAMI.2005.90

Olson, C. R. (2001). Object-based vision and attention in primates. *Current Opinion in Neurobiology*, *11*(2), 171–179. https://doi.org/10.1016/S0959-4388(00)00193-8

Olsson, A., Ebert, J. P., Banaji, M. R., & Phelps, E. A. (2005). Psychology: The role of social groups in the persistence of learned fear. *Science*, *309*(5735), 785–787. https://doi.org/10.1126/science.1113551

Pajkert, A., Finke, C., Shing, Y. L., Hoffmann, M., Sommer, W., Heekeren, H. R., & Ploner, C. J. (2017). Memory integration in humans with hippocampal lesions. *Hippocampus*, *27*(12), 1230–1238. https://doi.org/10.1002/hipo.22766

Peer, P. (1999). CVL Face Database. Retrieved from http://www.lrv.fri.uni-li.si/facedb.html

Poldrack, R. A., Clark, J., Paré-Blagoev, E. J., Shohamy, D., Creso Moyano, J., Myers, C., & Gluck, M. A. (2001). Interactive memory systems in the human brain. *Nature*, *414*, 546–550. https://doi.org/10.1038/35107080

Poldrack, R. A., & Foerde, K. (2008). Category learning and the memory systems debate. *Neuroscience and Biobehavioral Reviews*, *32*(2), 197–205. https://doi.org/10.1016/j.neubiorev.2007.07.007

Poldrack, R. A., & Packard, M. G. (2003). Competition among multiple memory systems: converging evidence from animal and human brain studies. *Neuropsychologia*, *41*(3), 245–251. Retrieved from http://www.ncbi.nlm.nih.gov/pubmed/12457750

Poldrack, R. A., Prabhakaran, V., Seger, C., & Gabrieli, J. D. (1999). Striatal activation during acquisition of a cognitive skill. *Neuropsychology*, *13*(4), 564–574. https://doi.org/10.1037/0894-4105.13.4.564

Poppenk, J., Evensmoen, H. R., Moscovitch, M., & Nadel, L. (2013). Long-axis specialization of the human hippocampus. *Trends in Cognitive Sciences*, *17*(5), 230–240. https://doi.org/10.1016/j.tics.2013.03.005

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

Pothos, E. M., & Reppa, I. (2014). The fickle nature of similarity change as a result of categorization. *Quarterly Journal of Experimental Psychology*, *67*(12), 2425–2438. https://doi.org/10.1080/17470218.2014.931977

Power, J. D., Barnes, K. a, Snyder, A. Z., Schlaggar, B. L., & Petersen, S. E. (2012). Spurious but systematic correlations in functional connectivity MRI networks arise from subject motion. *NeuroImage*, *59*(3), 2142–2154. https://doi.org/10.1016/j.neuroimage.2011.10.018

Rajah, M. N., McIntosh, A. R., & Grady, C. L. (1999). Frontotemporal interactions in face encoding and recognition. *Cognitive Brain Research*, *8*(3), 259–269. https://doi.org/10.1016/S0926-6410(99)00030-0

Ranganath, C., Heller, A., Cohen, M. X., Brozinsky, C. J., & Rissman, J. (2005). Functional connectivity with the hippocampus during successful memory formation. *Hippocampus*, *15*(8), 997–1005. https://doi.org/10.1002/hipo.20141

Reber, P. J., Gitelman, D. R., Parrish, T. B., & Mesulam, M. M. (2003). Dissociating explicit and implicit category knowledge with fMRI. *Journal of Cognitive Neuroscience, 15*(4), 574–583. https://doi.org/10.1162/089892903321662958

Reber, P. J., Stark, C. E. L., & Squire, L. R. (1998). Contrasting cortical activity associated with category memory and recognition memory. *Learning & Memory, 5*, 420–428. https://doi.org/10.1101/lm.5.6.420

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

Ren, J., Huang, F., Zhou, Y., Zhuang, L., Xu, J., Gao, C., … Luo, J. (2020). The function of the hippocampus and middle temporal gyrus in forming new associations and concepts during the processing of novelty and usefulness features in creative designs. *NeuroImage*, *214*(March), 116751. https://doi.org/10.1016/j.neuroimage.2020.116751

Renoult, L., Irish, M., Moscovitch, M., & Rugg, M. D. (2019). From Knowing to Remembering: The Semantic–Episodic Distinction. *Trends in Cognitive Sciences*, *23*(12), 1041–1057. https://doi.org/10.1016/j.tics.2019.09.008

Richeson, J. A., Baird, A. A., Gordon, H. L., Heatherton, T. F., Wyland, C. L., Trawalter, S., & Shelton, J. N. (2003). An fMRI investigation of the impact of interracial contact on executive function. *Nature Neuroscience*, *6*(12), 1323–1328. https://doi.org/10.1038/nn1156

Richter, F. R., Cooper, R. A., Bays, P. M., & Simons, J. S. (2016). Distinct neural mechanisms underlie the success, precision, and vividness of episodic memory. *ELife*, *5*(OCTOBER2016), 1–18. https://doi.org/10.7554/eLife.18260

Rosch, E., & Mervis, C. B. (1975). Family resemblances. *Cognitive Psychology*, *7*, 573–605. https://doi.org/10.1186/gb-2002-3-12-reports0063

Ryan, J. D., D'Angelo, M. C., Kamino, D., Ostreicher, M., Moses, S. N., & Rosenbaum, R. S. (2016). Relational learning and transitive expression in aging and amnesia. *Hippocampus*, *26*(2), 170–184. https://doi.org/10.1002/hipo.22501

Saksida, L. M., & Bussey, T. J. (2010). The representational-hierarchical view of amnesia: Translation from animal to human. *Neuropsychologia*, *48*(8), 2370–2384. https://doi.org/10.1016/j.neuropsychologia.2010.02.026

Saleem, A. B., Diamanti, E. M., Fournier, J., Harris, K. D., & Carandini, M. (2018). Coherent encoding of subjective spatial position in visual cortex and hippocampus. *Nature*, *562*(7725), 124–127. https://doi.org/10.1038/s41586-018-0516-1

Santos, A., Meyer-Lindenberg, A., & Deruelle, C. (2010). Absence of racial, but not gender, stereotyping in Williams syndrome children. *Current Biology, 20*(7), 307–308. https://doi.org/10.1016/j.cub.2010.02.009

Schapiro, A. C., Turk-Browne, N. B., Botvinick, M. M., & Norman, K. A. (2017). Complementary learning systems within the hippocampus: A neural network modelling approach to reconciling episodic memory with statistical learning. *Philosophical Transactions of the Royal Society B*, *372*(1711), 20160049. https://doi.org/10.1098/rstb.2016.0049

Schlichting, M. L., Mumford, J. A., & Preston, A. R. (2015). Learning-related representational changes reveal dissociable integration and separation signatures in the hippocampus and prefrontal cortex. *Nature Communications*, *6*, 1–10. https://doi.org/10.1038/ncomms9151

Schlichting, M. L., & Preston, A. R. (2015). Memory integration: Neural mechanisms and implications for behavior. *Current Opinion in Behavioral Sciences*, *1*, 1–8. https://doi.org/10.1016/j.cobeha.2014.07.005

Schlichting, M. L., Zeithamova, D., & Preston, A. R. (2014). CA1 subfield contributions to memory integration and inference. *Hippocampus*, *24*(10), 1248–1260. https://doi.org/10.1002/hipo.22310

Scoville, W. B., & Milner, B. (1957). Loss of recent memory after bilateral hippocampal lesions: Memory and memories-looking back and looking forward. *Journal of Neurology, Neurosurgery and Psychiatry*, *20*(11), 11–21. https://doi.org/10.1136/jnnp-2015-311092

Seger, C. A. (2008). How do the basal ganglia contribute to categorization? Their roles in generalization, response selection, and learning via feedback. *Neuroscience and Biobehavioral Reviews*, *32*(2), 265–278. https://doi.org/10.1016/j.neubiorev.2007.07.010

Seger, C. A., & Cincotta, C. M. (2005). The roles of the caudate nucleus in human classification learning. *Journal of Neuroscience*, *25*(11), 2941–2951. https://doi.org/10.1523/JNEUROSCI.3401-04.2005

Seger, C. A., Poldrack, R. A., Prabhakaran, V., Zhao, M., Glover, G. H., & Gabrieli, J. D. E. (2000). Hemispheric asymmetries and individual differences in visual concept learning as measured by functional MRI. *Neuropsychologia*, *38*(9), 1316–1324. https://doi.org/10.1016/S0028-3932(00)00014-2

Sekeres, M. J., Winocur, G., & Moscovitch, M. (2018). The hippocampus and related neocortical structures in memory transformation. *Neuroscience Letters*, *680*(August 2017), 39–53. https://doi.org/10.1016/j.neulet.2018.05.006

Shepard, R. N., & Chang, J. J. (1963). Stimulus generalization in the learning of classifications. *Journal of Experimental Psychology*, *65*(1), 94–102. https://doi.org/10.1037/h0043732

Shepard, R. N., Hovland, C. I., & Jenkins, H. M. (1961). Learning and memorization of classifications. *Psychological Monographs: General and Applied*, *75*(13), 1–42. https://doi.org/https://doi.org/10.1037/h0093825

Shohamy, D., & Wagner, A. D. (2008). Integrating memories in the human brain: Hippocampal-midbrain encoding of overlapping events. *Neuron*, *60*, 378–389. https://doi.org/10.1016/j.neuron.2008.09.023

Sloman, S. A. (1996). The empirical case for two systems of reasoning. *Psychological Bulletin*, *119*(1), 3–22.

Soto, F. A. (2019). Categorization training changes the visual representation of face identity. *Attention, Perception, and Psychophysics*, *81*(5), 1220–1227. https://doi.org/10.3758/s13414-019-01765-w

Soto, F. A., & Wasserman, E. A. (2010). Missing the Forest for the Trees: Object-discrimination Learning Blocks Categorization Learning. *Psychological Science*, *21*(10), 1510–1517. https://doi.org/10.1177/0956797610382125

Squire, L. R. (1992). Memory and the hippocampus: A synthesis from findings with rats, monkeys, and humans. *Psychological Review*, *99*(2), 195–231. https://doi.org/10.1037/0033-295X.99.3.582

Squire, L. R., & Zola, S. M. (1998). Episodic memory, semantic memory, and amnesia. *Hippocampus*, *8*(3), 205–211. https://doi.org/10.1002/(SICI)1098-1063(1998)8:3<205::AID-HIPO3>3.0.CO;2-I

Stevens, F. L., & Abernethy, A. D. (2018). Neuroscience and racism: The power of groups for overcoming implicit bias. *International Journal of Group Psychotherapy*, *68*(4), 561–584. https://doi.org/10.1080/00207284.2017.1315583

Stramaccia, D. F., Penolazzi, B., Altoè, G., & Galfano, G. (2017). Neurobiology of Learning and Memory TDCS over the right inferior frontal gyrus disrupts control of interference in memory : A retrieval-induced forgetting study. *Neurobiology of Learning and Memory*, *144*, 114–130. https://doi.org/10.1016/j.nlm.2017.07.005

Takashima, A., Nieuwenhuis, I. L. C., Jensen, O., Talamini, L. M., Rijpkema, M., & Fernández, G. (2009). Shift from hippocampal to neocortical centered retrieval network with consolidation. *Journal of Neuroscience*, *29*(32), 10087–10093. https://doi.org/10.1523/JNEUROSCI.0799-09.2009

Tambini, A., Ketz, N., & Davachi, L. (2010). Enhanced Brain Correlations during Rest Are Related to Memory for Recent Experiences. *Neuron*, *65*(2), 280–290. https://doi.org/10.1016/j.neuron.2010.01.001

Tambini, A., Rimmele, U., Phelps, E. A., & Davachi, L. (2017). Emotional brain states carry over and enhance future memory formation. *Nature Neuroscience*, *20*(2), 271–278. https://doi.org/10.1038/nn.4468

Teng, E., Stefanacci, L., Squire, L. R., & Zola, S. M. (2000). Contrasting effects on discrimination learning after hippocampal lesions and conjoint hippocampal-caudate lesions in monkeys. *Journal of Neuroscience*, *20*(10), 3853–3863. https://doi.org/10.1523/jneurosci.20-10-03853.2000

Teyler, T. J., & DiScenna, P. (1986). The hippocampal memory indexing theory. *Behavioral Neuroscience*, *100*(2), 147–154. https://doi.org/10.1037/0735-7044.100.2.147

Thompson-Schill, S. L., D'Esposito, M., Aguirre, G. K., & Farah, M. J. (1997). Role of left inferior prefrontal cortex in retrieval of semantic knowledge: A reevaluation. *Proceedings of the National Academy of Sciences of the United States of America*, *94*(26), 14792–14797. https://doi.org/10.1073/pnas.94.26.14792

Tompary, A., & Davachi, L. (2017). Consolidation Promotes the Emergence of Representational Overlap in the Hippocampus and Medial Prefrontal Cortex. *Neuron*, *96*(1), 228-241.e5. https://doi.org/10.1016/j.neuron.2017.09.005

Touroutoglou, A., Andreano, J. M., Barrett, L. F., & Dickerson, B. C. (2015). Brain network connectivity-behavioral relationships exhibit trait-like properties: Evidence from hippocampal connectivity and memory. *Hippocampus*, *25*(12), 1591–1598. https://doi.org/10.1002/hipo.22480

Turney, I. C., & Dennis, N. A. (2017). Elucidating the neural correlates of related false memories using a systematic measure of perceptual relatedness. *NeuroImage*, *146*, 940–950. https://doi.org/10.1016/j.neuroimage.2016.09.005

Ungerleider, L. G., & Mishkin, M. (1982). Two cortical visual systems. *Analysis of Visual Behavior*, 549–586.

Van Dijk, K. R. A., Hedden, T., Venkataraman, A., Evans, K. C., Lazar, S. W., & Buckner, R. L. (2010). Intrinsic functional connectivity as a tool for human connectomics: Theory, properties, and optimization. *Journal of Neurophysico*, 297–321.

van Dongen, E. V., Takashima, A., Barth, M., & Fernández, G. (2011). Functional connectivity during light sleep is correlated with memory performance for face-location associations. *NeuroImage*, *57*(1), 262–270. https://doi.org/10.1016/j.neuroimage.2011.04.019

van Kesteren, M. T. R., Beul, S. F., Takashima, A., Henson, R. N., Ruiter, D. J., & Fernández, G. (2013). Differential roles for medial prefrontal and medial temporal cortices in schema-dependent encoding: From congruent to incongruent. *Neuropsychologia*, *51*(12), 2352–2359. https://doi.org/10.1016/j.neuropsychologia.2013.05.027

Van Kesteren, M. T. R., Rijpkema, M., Ruiter, D. J., & Fernández, G. (2010). Retrieval of associative information congruent with prior knowledge is related to increased medial prefrontal activity and connectivity. *Journal of Neuroscience*, *30*(47), 15888–15894. https://doi.org/10.1523/JNEUROSCI.2674-10.2010

Wagner, A. D., Pare-Blagoev, E. J., Clark, J., & Poldrack, R. A. (2001). *Recovering meaning: Left prefrontal cortex guides controlled semantic retrieval*. *Neuron*. https://doi.org/10.1016/S0026-0495(96)90074-8

Wagner, I. C., Rütgen, M., & Lamm, C. (2020). Pattern similarity and connectivity of hippocampal-neocortical regions support empathy for pain. *Social Cognitive and Affective Neuroscience*, *15*(3), 273–284. https://doi.org/10.1093/scan/nsaa045

Wallraven, C., Bülthoff, H. H., Waterkamp, S., van Dam, L., & Gaißert, N. (2014). The eyes grasp, the hands see: Metric category knowledge transfers between vision and touch. *Psychonomic Bulletin and Review*, *21*(4), 976–985. https://doi.org/10.3758/s13423-013-0563-4

Wang, L., Zang, Y., He, Y., Liang, M., Zhang, X., Tian, L., … Li, K. (2006). Changes in hippocampal connectivity in the early stages of Alzheimer's disease: Evidence from resting state fMRI. *NeuroImage*, *31*(2), 496–504. https://doi.org/10.1016/j.neuroimage.2005.12.033

Wattenmaker, W. D. (1993). Incidental concept learning, feature frequency, and correlated properties. *Journal of Experimental Psychology: Learning, Memory, and Cognition*, *19*(1), 203–222. https://doi.org/10.1037/0278-7393.19.1.203

Webb, C. E., Turney, I. C., & Dennis, N. A. (2016). What's the gist? The influence of schemas on the neural correlates underlying true and false memories. *Neuropsychologia*, *93*(July), 61–75. https://doi.org/10.1016/j.neuropsychologia.2016.09.023

Winocur, G., Moscovitch, M., & Sekeres, M. (2007). Memory consolidation or transformation: Context manipulation and hippocampal representations of memory. *Nature Neuroscience*, *10*(5), 555–557. https://doi.org/10.1038/nn1880

Xiao, X., Dong, Q., Gao, J., Men, W., Poldrack, R. A., & Xue, G. (2017). Transformed neural pattern reinstatement during episodic memory retrieval. *Journal of Neuroscience*, *37*(11), 2986–2998. https://doi.org/10.1523/JNEUROSCI.2324-16.2017

Yassa, M. A., & Stark, C. E. L. (2011). Pattern separation in the hippocampus. *Trends in Neurosciences*, *34*(10), 515–525. https://doi.org/10.1016/j.tins.2011.06.006

Zaki, S. R., Nosofsky, R. M., Stanton, R. D., & Cohen, A. L. (2003). Prototype and Exemplar Accounts of Category Learning and Attentional Allocation: A Reassessment. *Journal of Experimental Psychology: Learning Memory and Cognition*, *29*(6), 1160–1173. https://doi.org/10.1037/0278-7393.29.6.1160

Zalesak, M., & Heckers, S. (2009). The role of the hippocampus in transitive inference. *Psychiatry Research - Neuroimaging*, *172*(1), 24–30. https://doi.org/10.1016/j.pscychresns.2008.09.008

Zeidman, P., Mullally, S. L., & Maguire, E. A. (2015). Constructing, perceiving, and maintaining scenes: Hippocampal activity and connectivity. *Cerebral Cortex*, *25*(10), 3836–3855. https://doi.org/10.1093/cercor/bhu266

Zeithamova, D., & Bowman, C. . (2020). Generalization and the hippocampus : More than one story ? *Neurobiology of Learning and Memory*, *175*, 1–10. https://doi.org/10.1016/j.nlm.2020.107317

Zeithamova, D., Dominick, A. L., & Preston, A. R. (2012). Hippocampal and ventral medial prefrontal activation during retrieval-mediated learning supports novel inference. *Neuron*, *75*(1), 168–179. https://doi.org/10.1016/j.neuron.2012.05.010

Zeithamova, D., Maddox, W. T., & Schnyer, D. M. (2008). Dissociable prototype learning systems: Evidence from brain imaging and behavior. *The Journal of Neuroscience*, *28*(49), 13194–13201. https://doi.org/10.1523/JNEUROSCI.2915-08.2008

Zeithamova, D., Manthuruthil, C., & Preston, A. R. (2016). Repetition suppression in the medial temporal lobe and midbrain is altered by event overlap. *Hippocampus*, *26*(11), 1464–1477. https://doi.org/10.1002/hipo.22622

Zeithamova, D., & Preston, A. R. (2010). Flexible memories: Differential roles for medial temporal lobe and prefrontal cortex in cross-episode binding. *Journal of Neuroscience*, *30*(44), 14676–14684. https://doi.org/10.1523/JNEUROSCI.3250-10.2010

Zeithamova, D., Schlichting, M. L., & Preston, A. R. (2012). The hippocampus and inferential reasoning: Building memories to navigate future decisions. *Frontiers in Human Neuroscience*, *6*, 1–14. https://doi.org/10.3389/fnhum.2012.00070