How Experience Shapes Category Learning

by

Troy Mitchell Houser

A dissertation accepted and approved in partial fulfillment of the

requirements for the degree of

Doctor of Philosophy

in Psychology

Dissertation Committee:

Dagmar (Dasa) Demircan, Chair

Brice Kuhl, Core Member
Deepu Murty, Core Member
Ian Greenhouse, Institutional Representative

University of Oregon

Spring 2026

©2026 Troy Mitchell Houser

This work is openly licensed via CC BY-NC-ND 4.0.



Dissertation Abstract

Troy Mitchell Houser

Doctor of Philosophy in Psychology

Title: How Experience Shapes Category Learning

Category learning and generalization of category knowledge has traditionally been studied using very similar stimuli and procedures. This work has led to some general conclusions in the field, including that learning variability enhances subsequent generalization and that people abstract out commonalities across experiences for later use. Much less attention has been paid to how external circumstances shape the category representations that we form. How flexible are the representations that we use to categorize information? What situations bias us toward remembering general or specific details? What neural substrates underly these cognitive processes? The present dissertation aims to explore 1) the degree to which people rely on shared versus stimulus-unique features to generalize following different levels of learning variability, 2) how generalization following incidental category learning compares to generalization following the more traditional feedback-based learning, and 3) the neural correlates of incidentally-acquired category representations. Across three empirical studies, I demonstrate different degrees of specificity in category representations. Moreover, I show that the ventromedial prefrontal cortex can use detailed representations of categories that may underpin incidental category learning. Together these findings provide insight into the flexibility of category representations and generalization in humans.

This dissertation includes previously published and unpublished co-authored material.

CURRICULUM VITAE

NAME OF AUTHOR: Troy Mitchell Houser

GRADUATE AND UNDERGRADUATE SCHOOLS ATTENDED:

University of Oregon, Eugene, OR
Temple University, Philadelphia, PA
Harrisburg Area Community College, Harrisburg, PA

DEGREES AWARDED:

Doctor of Philosophy, Psychology, 2026, University of Oregon
Master of Science, Psychology, 2022, University of Oregon
Bachelor of Art, Psychology, 2021, Temple University

AREAS OF SPECIAL INTEREST:

Cognitive Neuroscience

PROFESSIONAL EXPERIENCE:

Graduate Employee, University of Oregon
September 2021 – May 2026
Research Assistant, Temple University
August 2018 – August 2021

PUBLICATIONS:

Houser, T.M., Acker, R., & Zeithamova, D. (in prep). Memory precision fluctuates with theta rhythmicity in humans.

Zhou, S.*, **Houser, T.M.***, & Zeithamova, D. (in prep). Neural dedifferentiation helps memory specificity in older adults.

Houser, T.M. & Zeithamova, D. (in prep). Incidental category learning leads to exemplar representations in the ventromedial prefrontal cortex.

Houser, T.M. & Zeithamova, D. (in review). Low coherence during category learning limits both exemplar memory and concept generalization. *Journal of Experimental Psychology: Learning, Memory, & Cognition.*

Zeithamova, D.*, **Houser, T.M**.*, & Bowman, C.R. (2026). Evaluating cognitive models with permutation testing: A case study of prototype and exemplar categorization. *Advances in Methods and Practices in Psychological Science, 9*(2): 1-28.

Houser, T.M. (2025). A framework for uniting space and time in the mind and brain. *Frontiers in Cognition, 4:1611855.*

Skalaban, L.J., Neeson, A.L., **Houser, T.M.**, DuBrow, S., Davachi, L., & Murty, V.P. (2025). Goal orientation shifts attentional focus and impairs reward-motivated memory. *Learning & Memory, 32*(2), a054020.

Houser, T.M. (2025). Explore-exploit behavior in humans as a sequential sampling process. *Current Psychology,* 1-13.

Houser, T.M. (2024). A boundedly rational model for category learning. *Frontiers in Psychology, 15*(1477514).

Houser, T.M. (2024). The relationship between temporal discounting and foraging. *Current Psychology, 43*(39), 31149-31158.

Knouse, M.C., Kniffin, A.R., English, E.A., Cuadrado, W., **Houser, T.M.**, & Briand, L.A. (2024). PKMζ alters oxycodone-taking in a dose-and sex-dependent manner. *Addiction Neuroscience, 12*(100169).

Houser, T.M., Resnick, A., Zeithamova, D. (2024). Successful generalization after training to remember specific events. *Frontiers in Cognition*.

Houser, T.M.[*], Krantz, L.[*], Zeithamova, D. (2024). Retrieval-based inference in the acquired equivalence paradigm. *Frontiers in Cognition.*

Houser, T.M. (2023). How generalization relates to the exploration-exploitation tradeoff. *Frontiers in Cognition,* 2:1132766. Doi: 10.3389/fcogn.2023.1132766.

Houser, T.M., Tompary, A., & Murty, V. (2022). Agency enhances temporal order memory in an interactive exploration game. *Psychonomic Bulletin & Review, 29*(6): 2219-2228.

Shen, X., **Houser, T.**, Smith, D. V., & Murty, V. P. (2021). Machine-learning as a validated tool to characterize individual differences in free recall of naturalistic events. *Psychonomic Bulletin & Review, 30*(1): 308-316.

Houser, T. M. (2022). Spatialization of time in the entorhinal-hippocampal system. *Frontiers in Behavioral Neuroscience, 15,* Article 807197. https://doi.org/10.3389/fnbeh.2021.807197

Marshall P.J., **Houser T.M.** and Weiss S.M. (2021) The Shared Origins of Embodiment and

Development. *Front. Syst. Neurosci.* 15:726403. doi: 10.3389/fnsys.2021.726403.

*****_These authors contributed equally to this work._

TALKS:

Houser, T.M. (2023). Neural correlates of categories following incidental category learning. *Learning and Memory Nanosymposium, Society for Neuroscience, 2023*.

Houser, T.M. (2023). Predicting explore-exploit behavior from personality traits. *Oregon Council for Problem Gambling Summit.*

ABSTRACTS:

Houser, T.M., Acker, R., Zeithamova, D. (2026). Optimal encoding time windows for memory precision. *Cognitive Neuroscience Society, 2026.*

Zhou, S., **Houser, T.M.**, C.R., Bowman, Zeithamova, D. (2026). Memory for category exemplars and hippocampal connectivity in aging. *Cognitive Neuroscience Society, 2026.*

Houser, T.M., Bowman, C.R., Zeithamova, D. (2025). Hippocampal-control network functional connectivity differentially predicts memory in young and older adults. *Cognitive Neuroscience Society, 2025.*

Chaloupka, B., **Houser, T.M.**, Zeithamova, D. (2025). Representational gradient of the hippocampal long axis extends beyond memory to visual processing. *Cognitive Neuroscience Society, 2025.*

Zhou, S., **Houser, T.M.**, Bowman, C.R., Zeithamova, D. (2025). Default and control network connectivity changes predict performance in aging. *Cognitive Neuroscience Society, 2025.*

Costa, L., **Houser, T.M.**, Hutchinson, B.J., Zeithamova, D. (2025). Memory effects of event boundaries caused by spatial change. *Cognitive Neuroscience Society, 2025.*

Costa, L., **Houser, T.M.**, Zeithamova, D. (2024). Effect of event boundaries caused by spatial change. *University of Oregon Undergraduate Research Symposium, 2024*.

Sathya, A., **Houser, T.M.**, Zeithamova, D. (2024). Differences in categorization strategy between different political affiliations. *University of Oregon Undergraduate Research Symposium, 2024*.

Matheson, M., **Houser, T.M.**, Zeithamova, D. (2024). Impact of reward certainty on generalization of categories. *University of Oregon Undergraduate Research Symposium, 2024*.

Krantz, L., **Houser, T.M.**, Zeithamova, D. (2024). Retrieval vs. integration-based methods in acquired equivalence. *University of Oregon Undergraduate Research Symposium, 2024*.

Acker, R., **Houser, T.M.**, Zeithamova, D. (2024). Theta oscillations in memory. *University of Oregon Undergraduate Research Symposium, 2024*.

Houser, T.M., Bowman, C.R., Zeithamova, D. (2024). Neural dedifferentiation and reduced specific memory in aging. *Cognitive Neuroscience Society, 2024*.

Houser, T.M., Berkman, E.T., Zeithamova, D. (2024). Predicting personality from representations of uncertainty in the brain. *Social Affective Neuroscience Society, 2024*.

Krantz, L., **Houser, T.**, Zeithamova, D. (2023). Determining the underlying cognitive mechanisms behind acquired equivalence. *University of Oregon Undergraduate Research Symposium, 2023*.

Acker, R., **Houser, T.**, Zeithamova, D. (2023). Theta oscillations in memory encoding and retrieval. *University of Oregon Undergraduate Research Symposium, 2023*.

Nashawi, L., **Houser, T.**, Zeithamova, D. (2023). Predicting explore-exploit behavior from personality traits. *University of Oregon Undergraduate Research Symposium, 2023*.

Houser, T.M., Zeithamova, D. (2023). Successful generalization of conceptual knowledge after training to remember specific events. *Learning & Memory, 2023*.

Houser, T.M., Zeithamova, D. (2023). Successful generalization of conceptual knowledge after training to remember specific events. *Cognitive Neuroscience Society, 2023*.

Houser, T.M., DuBrow, S., Zeithamova, D. (2022). Exploration-exploitation tradeoff during navigation of abstract task space. *Society for Neuroscience, 2022*.

Knouse, M.C., **Houser, T.M.,** Briand, L. (2022). The role of PKMζ in oxycodone self-administration and accumbal electrophysiology. *Society for Neuroscience, 2022*.

Houser, T.M., Tompary, A., & Murty, V. (2021). Agency enhances temporal order memory in an interactive exploration game. *Psychonomics, 2021*.

Acknowledgments

I first want to extend my sincerest thank you to my advisor, Dr. Dasa Zeithamova. Her expertise in all things research was critical to all aspects of this dissertation and to any skills I have acquired along the way. Her caring nature ensured that my needs as a graduate student were taken care of. She has been an inspiration both in and out of the laboratory. I am forever grateful for everything she has taught me.

I would also like to thank my colleagues in the Brain & Memory Lab, without which none of this would have been possible (including my sanity). The graduate students have always been there for me, and the research assistants and lab managers are of the highest quality.

I thank Dr. Sarah DuBrow, who is the reason any of this happened. Without her, this journey would not have begun. Our shared excitement for temporal context will be something I cherish always.

I thank Dr. Deepu Murty, who missed me so much at Temple that he followed me out to Oregon. He has been a shining light with me at every step of this journey and a motivating force that I am forever grateful for.

I thank my best and oldest friend, Shane Schleinkofer, with whom I have faced more challenges than grad school could ever offer. Can't wait to be close-by again.

My biggest thanks goes out to my family, for which words will never be enough. I am proud to say my parents and siblings made me who I am today. My wife, Jessica, is my inspiration, my teacher, and my home—in so many ways not written in this dissertation, she is the biggest contributor to this work. My daughter Camille is simply the greatest wonder I could

have ever been gifted with. I had no idea the kind of pure and raw joy and excitement that life could generate before meeting her. She has made me a true believer in Bergsonian time.

Many others belong here, though the space needed is too much. I truly love each of you. Thank you.

For my brother Larry

Table of Contents

INTRODUCTION 17

WHAT IS A CATEGORY? 19
HOW IS A CATEGORY REPRESENTED? 22
NATURE CARVES THE MIND AT ITS JOINTS 25
Effect of variability on category learning and generalization 27
Effect of incidental encoding on category representations and generalization 31
PREVIEW OF CURRENT DISSERTATION 35
REFERENCES CITED 37

LOW EXEMPLAR COHERENCE DURING CATEGORY LEARNING LIMITS BOTH EXEMPLAR MEMORY AND CONCEPT GENERALIZATION 48

METHODS 54
Participants 54
Materials 54
Training sets. 55
Recognition and categorization test sets. 58
Procedure 60
Training. 60
Recognition test. 60
Categorization test. 61
Statistical analyses 62
Training. 62
Categorization test, unambiguous faces. 62
Categorization test, ambiguous faces. 62
Recognition test. 63
RESULTS 63
Training performance 63
Categorization performance 64
Categorization accuracy for unambiguous trials 64
Categorization of ambiguous blends 66
Recognition performance 70
Endorsement of unambiguous faces. 70
Endorsement of ambiguous faces. 71
DISCUSSION 73
Conclusion 77
REFERENCES CITED 79

SUCCESSFUL GENERALIZATION OF CONCEPTUAL KNOWLEDGE AFTER TRAINING TO REMEMBER SPECIFIC EVENTS 83

EXPERIMENT 1 86
Methods *87*
Participants. 87
Materials. 87
Stimuli. 87
Training set. 87
Training scenes. 88
Testing set. 88
Experimental design. 88
Training. 88
Associative memory test. 89
Category generalization test. 90
Statistical analysis. 91
Prospective memory judgments during training. 91
Associative memory. 91
Categorization accuracy during generalization test. 91
Categorization strategies during generalization test. 92
Model selection. 94
Results *95*
Prospective memory judgments during training. 95
Associative memory. 95
Categorization accuracy. 97
Categorization strategies. 97
Discussion *99*
EXPERIMENT 2 100
Method *100*
Participants. 100
Materials. 101
Stimuli. 101
Experimental design. 101
Statistical analysis. 102
Training. 102
Associative memory. 103
Categorization test accuracy and strategy analysis. 103
Results *103*
Training data. 103
Associative memory (paired-associate group only). 104
Categorization accuracy. 105

Categorization strategies. 107
Relationship between specific and general knowledge (Exp 1 and 2) 108
GENERAL DISCUSSION 112
Conclusion *117*
REFERENCES CITED 118

INCIDENTAL CATEGORY LEARNING LEADS TO EXEMPLAR REPRESENTATIONS IN THE VENTROMEDIAL PREFRONTAL CORTEX 125

METHODS 129
Materials *129*
Stimuli. 129
Training set. 130
Training scenes. 130
Testing set. 130
Procedure *131*
Training. 131
Associative memory test. 132
Category generalization test. 132
FMRI data acquisition. 132
Statistical analysis. 133
Prospective memory judgments during training. 133
Associative memory. 133
Categorization accuracy during generalization test. 133
Categorization strategies during generalization test. 134
Model selection. 136
fMRI preprocessing. 137
Regions of interest. 138
Model-based fMRI analysis. 138
Exploratory whole brain analysis. 140
Exploratory brain-behavior analysis. 141
RESULTS 141
Prospective memory judgments during training *141*
Associative memory *142*
Categorization accuracy during generalization test *142*
Categorization strategies *143*
ROI analysis *144*
Whole brain analysis *144*
Brain-behavior relationship *146*
DISCUSSION 148
Conclusion *153*

References cited 154
GENERAL DISCUSSION 161
How experience plus goals shape category learning 161
Does the specific tradeoff with the abstract? 164
Incidental category learning leads to exemplar representations in the ventromedial prefrontal cortex 165
Broader implications 167
Quixotic categories 167
Situating conceptual knowledge 170
Back to Don Quixote 172
Technical implications 173
Limitations and Future Directions 174
General conclusion 175
References cited 177

List of figures

Figure 1.1. Central competing theories on what a category is. 22
Figure 1.2. Visualization of how behavior matches Wittgenstein's graded category structures. 25
Figure 1.3. Example of the dot pattern prototype distortion paradigm. 29
Figure 1.4. FMRI markers of exemplar and prototype representations. 36
Figure 2.1. Training stimuli (faces in the middle boxes) were made by blending a never-shown shared parent face (left side) with a set of never-shown unique parent faces (right side). 53
Figure 2.2. Example near-prototype and ambiguous stimuli. 59
Figure 2.3. Example experimental design. 61
Figure 2.4. Categorization training accuracy. 64
Figure 2.5. Unambiguous categorization performance. 66
Figure 2.6. Categorization responses to ambiguous trials. 69
Figure 2.7. Recognition memory performance. 72
Figure 3.1. Experiment design. 90
Figure 3.2. Experiment 1 training and generalization performance. 96
Figure 3.3. Experiment 1 computational model fitting results. 98
Figure 3.4. Experiment 2 stimuli and training performance. 104
Figure 3.5. Experiment 2 test performance. 107
Figure 3.6. Experiment 2 categorization strategies. 108
Figure 3.7. Associative memory and subsequent strategy use. 110
Figure 3.8. Generalization-associative memory relationship. 111
Figure 4.1. Experimental design. 128
Figure 4.2. Regions of interest from a representative participant. 138
Figure 4.3. Summed similarity estimates and their correlations. 140
Figure 4.4. Behavioral performance. 141
Figure 4.5. Generalization strategies in behavior and brain. 143
Figure 4.6. Clusters exhibiting significant activation in response to model predictions. 145
Figure 4.7. Visualization of clusters significantly tracking dissimilarity to the prototype. 146
Figure 4.8. Brain-behavior relationship. 147

List of tables

TABLE 2.1. EXAMPLE STRUCTURE OF TEST STIMULI. 57
TABLE 2.2. MIXED ANOVA PREDICTING THE PROPORTION OF RESPONSES CONSISTENT WITH SHREAD VS. UNIQUE FEATURE RELIANCE. 69
TABLE 4.1. CLUSTERS SIGNIFICANTLY TRACKING NEGATIVE PROTOTYPE MODEL PREDICTIONS. .. 145

Chapter 1

Introduction

As the father of a toddler, I am constantly answering the question "What is this?" And, when I am not, I am actively telling her how to label items. As she acquires more and more mappings from things in the world to labels, she communicates more efficiently. Even beyond communication, she seems to move through the world in a more organized, structured manner, as if the labels she learns enables filing away of experiences for systematic use in the future. Most awe-inspiring is watching her apply previously acquired knowledge to a novel situation. For example, she will frequently point at never-before-seen animals and label them correctly. It is as if, for example, when she learned how to buckle her car-seat straps, she learned how to buckle *all* car-seat straps. This ability to generalize from previously acquired knowledge is prominent in all aspects of her life. Indeed, it is prominent in all aspects of everybody's life. They might seem like trivial instances, but this is only because generalization is so ubiquitous, it gets overlooked. It is no small feat to act adaptively in a situation you have never encountered before. Moreover, generalization (and categorization) has particularly deep ties with the study of cognition. A pioneering cognitive scientist, Roger Shepard, once argued that generalization should be the *first principle of psychology* (Shepard, 1987).

What makes generalization the first principle of psychology? Why not attention or perception or any of the other core phenomena in all the introductory psychology textbooks? Imagine entering a grocery store in a foreign nation and knowing that the milk was near the cheese. In doing so, you had to generalize the previously acquired knowledge that milk and cheese are typically found nearby in grocery stores to the present situation. The adaptive nature of generalization is clear from this example but the reason Shepard elevated generalization above

all other psychological phenomena is because it is literally *always* occurring. In the extreme, one can say that every moment in time is a novel phenomenon, yet we transition through time, carrying our histories with us wherever we go, acting as if the *now* is the same as the *just-before*, unaware of the remarkable cognitive and neurobiological processes required to facilitate such behavior.

We will not go so far as Shepard in elevating generalization beyond all other psychological phenomena, for, as the above examples illustrate, generalization does not act alone. For my daughter to accurately label a never-before-seen squirrel as 'squirrel', she first needs to have some kind of understanding of what a squirrel is. To infer the layout of a grocery store, we need category knowledge of 'grocery stores'. Thus, generalization and categorization go hand-in-hand in facilitating adaptive and efficient behavior in the world.

A category is a group of items or events similar enough to be treated equivalently for some goal. Categorization is an indispensable survival capacity. It is how our ancestors could differentiate mushrooms from toadstools and garden snakes from rattlesnakes (Smith, 2014). For vervet monkeys, specific calls for different predators are critical for their survival (P. C. Lee, 1990). Social categorization is how people form bonds, find mates, and can shape how one remembers others (Flade et al., 2019; Hewstone et al., 1991; Maner et al., 2012). A more provocative account proposes that categorization is *the* fundamental computational principle guiding all brain function (Barrett & Miller, 2025). This account argues that the brain's overarching goal is to reduce the infinite dimensional world that we live in to fewer, digestible chunks that enable our living within it. Indeed, even at the earliest stages of sensory processing in the brain, signals are compressed, removing redundancies and noise, in turn enhancing signal quality and decreasing processing costs (Babadi & Sompolinsky, 2014; Barranca et al., 2014;

Ganguli & Sompolinsky, 2012; Singer et al., 2018). At a broader level, we can imagine that people compress multiple high dimensional ensembles of metabolic, olfactory, visual, haptic, and auditory information into a single, low-dimensional category for 'threat'. This ability to transcend detailed physical differences is what in turn allows us to generalize from the past to the present (Bernardi et al., 2020). Moreover, by treating the many as one, categorization supports metabolic efficiency, thereby satisfying a key biological constraint for life and evolution, which is energy optimization (Barrett & Miller, 2025; Sterling, 2012, 2018; Sterling & Laughlin, 2015).

What is a category?

The utility, ubiquity, and connection to generalization clearly motivates the need for researching categorization; however, categorization itself is not so clearly defined as one might think at first blush. For starters, debates about what a category even is have raged for thousands of years with few signs of an end in sight. Aristotle suggested that a category is defined by a set of features that all category members share and only category members share. For example, according to this logic, a dog might be a dog if and only if it has four legs. Here, four legs is a feature that is shared by all dogs and the combination of all dog features is unique to dogs. This argument derived from Plato's theory of forms, which argued that mathematical entities, such as a triangle, had a set of necessary and sufficient conditions that made them what they are. Thus, a triangle is a triangle if and only if it is two-dimensional, a polygon, has three sides, three vertices, and three internal angles.

Ludwig Wittgenstein was a philosopher in the mid-twentieth century who rejected the Aristotelian premise on the grounds of it being a grammatical illusion—something that only holds true in the abstract and not in the real world. For example, a triangle, for which Aristotle's definition holds true, is not something observed in the real world, but a dog, which is observed in

the real world, has no set of necessary and sufficient conditions that make it a dog. Wittgenstein diagnosed the purported error as a "craving for generality" (Wittgenstein, 1953). He famously used the concept of a *game* as an example. If we try to list the necessary features of a game, we may come up with a list such as competition, rules, skill, and fun. Competition precludes games like solitaire, rules preclude many children's games, skill precludes any games of chance, and fun can preclude professional sports. In brief, no single feature is shared by all games. However, competition, rules, skill, and fun *do* exist in many, perhaps a majority of, games. Thus, Wittgenstein derived a new framework that relied on the notion of overlapping similarities rather than necessary conditions. For Wittgenstein, categories are defined by a network of overlapping similarities, such as family resemblances.

A family often has overlapping physical features but not every family member has every overlapping feature. For example, two of three siblings might have brown hair but the third has blonde hair. However, the blonde-haired sibling may share the brown eye color with one other sibling. Thus, we can say that *on average* siblings have more shared than individually unique features. The notion that *on average* members of a category are similar is the key component of Wittgenstein's proposal, for it entails that category membership is graded rather than all-or-none. In other words, whereas things in Aristotle's world either are or are not members of a given category, things in Wittgenstein's world are more or less members of a given category (Figure 1). Thus, every item can be a more or less typical example of a category. A big reason that most of category research assumes Wittgenstein's categories are more accurate than Aristotle's is because people's behavior matches a graded category structure. We will unpack this next.

If we define a family as a category (e.g., the Miller family) and Aristotle is correct, then all Millers are equally Miller and everybody else is equally non-Miller. As such, we would

expect that a Miller trusts all other Millers equally and distrusts all non-Millers equally. If instead Wittgenstein is correct, a Miller's level of trust should be proportional to Miller family resemblance. Consistent with Wittgenstein's proposal, likeability, cooperation, presumed kinship, and trust of a person scale with their resemblance to the rater/subject (DeBruine, 2002; DeBruine et al., 2008; Ewing et al., 2015, 2019; Giang et al., 2012). In a landmark study, Posner and Keele (1968) had participants group an array of dots into Category A or not Category A, where the array could be more or less similar to an arbitrarily defined perfect Category A array, called a prototype. For example, imagine a random array of nine dots that serves as the prototype. Now if we shifted the location of one of these dots, keeping the same location of the other eight, then this new array would be a typical member of Category A because it is so similar to the prototype for Category A (Posner et al., 1967). If we shifted the location of 7 of these dots, the new array would be an atypical example. After training participants to learn which arrays counted as Category A and which ones were not Category A, Posner and Keele (1968) tested people's ability to categorize novel arrays (i.e., generalize their Category A knowledge). Exactly as would be expected assuming graded category knowledge, people were more accurate at categorizing typical arrays and performance worsened with decreasing typicality (Posner & Keele, 1968). Direct ratings of typicality converge with this finding. For example, speed of processing is faster when asked to verify category membership of a typical category member (e.g., "Is a robin a bird?") than when asked to verify category membership of an atypical category member (e.g., "Is an ostrich a bird?") (Rosch & Mervis, 1975). Figure 2 illustrates the main findings of generalization gradients in categorization performance and speed of processing.

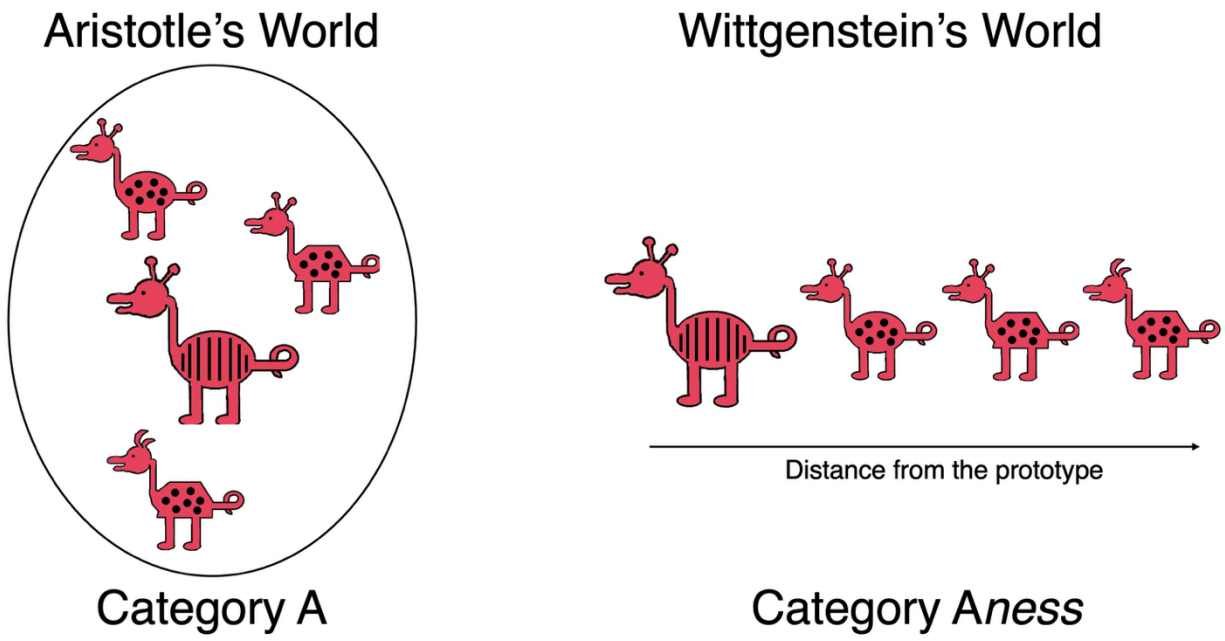

Figure 1.1. Central competing theories on what a category is.

Aristotle defined category boundaries as an all-or-none phenomenon, such that all things inside a category equally denote the overarching category. Wittgenstein defined category boundaries as graded, such that each thing has different levels of how much they denote the overarching category. It is thus perhaps useful to imagine that Aristotle's things in the world either are or are not Category A while Wittgenstein's things in the world all have different levels of Aness.

How is a category represented?

With the research of Roger Shepard, Michael Posner, and Eleanor Rosch (among others), the question around which the study of categorization revolved shifted from "What is a category?" to "How is a category represented?" What kind of representations could explain the typicality effects observed by these pioneering scientists? Posner and Keele (1968) suggested a process of abstraction, where people build up a representation of a category that is different from the instances actually observed. In other words, people's behavior is biased toward more typical examples of a category because they are literally representing the category as a never-before-seen category average (or some other central tendency). Thus, if one has encountered five trees

in their life, their representation of 'tree' is the set of overlapping features from across all five encounters. This central tendency is the prototype of a category. Upon encountering a sixth tree, prototype theory assumes that the person will compare the sixth tree to the tree prototype to determine whether or not it is in fact a tree.

The utility of prototypes is easy to see. They subsume swaths of data in smaller chunks, which enables fast and efficient processing of information. Prototypes thus likely enable people to do things like respond to danger, problem solve, and "think on their feet". They likely save space in the brain as well. On the other hand, this account of category representations assumes that people discard substantial amounts of potentially relevant information, which seems contrary to our experience. To illustrate, I have both a gist-like understanding of what a rollercoaster is and a vivid, detailed memory of the first time I went on a rollercoaster. What if the brain did not discard these details?

Prototype theory argues that people build up abstract representations across multiple experiences, retaining only the overlapping features for subsequent generalization. This account beautifully explains typicality effects in categorization studies; however, it entails something important that can change the interpretation of the underlying mechanism. Specifically, the representation of a prototype is built from individual instances, but the implicit assumption is that individual instances are not stored or at least rapidly forgotten, though nothing about this account depends on the timing of the combination of instances. In short, the exact same behavior can then occur if individual instances are stored and only combined *in the moment*. The shift in interpretations centers not on what either account predicts of performance, but what they predict is represented in the mind and brain. Prototype theory says a category is denoted with a single summary representation whereas the alternative just mentioned assumes all category members

are represented and flexibly combined when making category judgments. In the above example, all five trees will have been stored in memory and individually compared to the sixth. The sixth tree's similarity to each of the previous five is then summed to obtain a single estimate of how likely it is that the sixth tree is in fact a tree. The summation at the time of retrieval means the alternative can also explain typicality effects, for idiosyncratic details get washed out via summation. The alternative account is known as exemplar theory, where individual category members are known as exemplars (Nosofsky, 1986).

Considerable progress has been made with the developments of prototype and exemplar theories. Decades of investigation into their computational characteristics, behavioral predictions, and neural underpinnings have yielded several key discoveries that have advanced the field rapidly. These include category learning optimizing attention allocation (Kim & Rehder, 2011; Nosofsky, 1986, 1987; Rehder & Hoffman, 2005a, 2005b), neural substrate overlap between category generalization and episodic memory (Bowman & Zeithamova, 2018; Zeithamova et al., 2012), and prototype-like representations in single cells (Daume et al., 2024; Quiroga, 2012; Quiroga et al., 2005; Rey et al., 2020). One area of research into categorization that remains relatively neglected, however, is how experience shapes our category representations. This open question requires another subtle shift in the focus of categorization research. That is, whereas early scientific study shifted the overarching question from "What is a category?" to "How is a category represented?", they nevertheless both assume some true underlying nature of a category. In one domain, the nature is objective and in another it is subjective. Instead of assuming category structure or category representations a priori, this dissertation will focus on how the structure of our experience shapes what kind of representations we form.

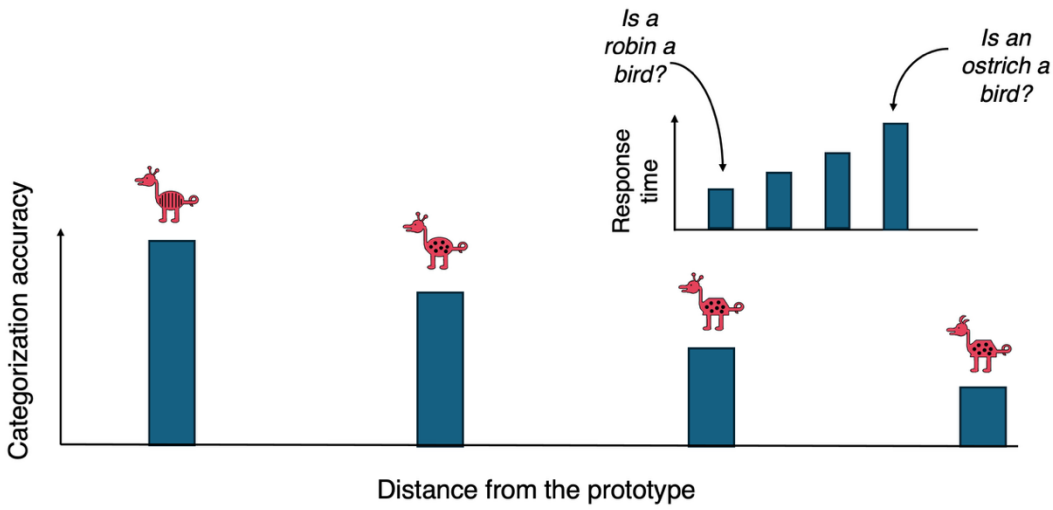

Figure 1.2. Visualization of how behavior matches Wittgenstein's graded category structures.

Accuracy in categorizing stimuli is best when categorizing prototypes themselves and gradually decreases with increasing dissimilarity to the prototype. This gradient of performance is flipped for response times, where responses are faster when judging more typical items.

Nature carves the mind at its joints

Plato popularized the notion of *carving nature at its joints*, referring to the mind's ability to abstract out perfect forms from experience (i.e., see the triangle in the world or the essence of a dog). This dissertation approaches the same topic from the opposite end: Rather than investigating how the mind carves up nature into digestible bits of information, I tested across three empirical studies how the world can flexibly carve out shapes (categories) in the mind.

There are many ways that the world can shape categories in the mind and brain. Most of these ways can be distilled to two approaches: the category structure during category learning and the manner in which people learn categories. Category structure can refer to numerous distinct ways in which the world, or experience, shapes category learning. To illustrate, the seminal work of Shepard et al. (1961) shows that the number of features required to attend to for categorization accuracy dictates the difficulty of learning category structures. Rule-plus-

exception learning refers to learning category structures where stimuli mostly adhere to a rule (e.g., all blue stimuli belong to Category A) but several exceptions exist (e.g., blue stimuli *with stripes* belong to Category B). This structure typically leads to categorization strategies that mix internalization of the rule and memorization of the individual stimuli that are exceptions (Nosofsky, Palmeri, et al., 1994). More complex rule-based structures, such as exclusive-OR (XOR; a Boolean logic expression denoting that an argument is true if its inputs are not alike and otherwise false), conjunctive, and disjunctive ones all shape how easy categories are to learn (Feldman, 2000, 2003, 2006, 2023; Rabi et al., 2015). Moreover, the structure of a category shapes what brain regions underpin category learning. For example, rule-based learning engages the prefrontal cortex while similarity-based learning engages the medial temporal lobe (Badre et al., 2010; Davis et al., 2014; JD et al., 2001; Kahnt et al., 2012; Mack et al., 2013; Mansouri et al., 2020; Paniukov & Davis, 2018; Reinert et al., 2021; Zeithamova et al., 2019). Other cognitive functions are differentially engaged depending on category structure as well. For example, rule-based category structures are thought to require working memory (Rabi & Minda, 2014), while similarity-based structures require reinforcement learning (F. G. Ashby et al., 1998).

The other prominent way that the world shapes category learning is through the manner of category learning. That is, *how* two people learn the exact same category structure can differ and this difference in how can lead to different category representations and generalization strategies. The overwhelming majority of categorization research uses classification, or supervised, approaches, which involves people guessing the category membership of a stimulus and being provided corrective feedback. Another popular approach, however, is inference-based category learning, which involves providing people with the category label associated with a given stimulus and prompting them to infer the value of a missing feature (Yamauchi &

Markman, 2000). Inference learning has been shown to facilitate more attention to non-diagnostic features than classification learning (Chin-Parker & Ross, 2004) and several studies show that these two approaches lead to fundamentally different category representations (Markman & Ross, 2003; Yamauchi et al., 2002; Yamauchi & Markman, 1998). Contrastive learning, blocked versus interleaved training, A/not-A, and sorting constitute additional styles of category learning that have been used to study representational differences in behavior, the brain, development, and across species (Chin-Parker & Cantelon, 2017; Minda et al., 2008, 2024; Posner & Keele, 1968; Rabi & Minda, 2016; Rosch & Mervis, 1975; Smith et al., 2004).

While the opportunities to study how the world shapes category representations in the mind are vast, the current dissertation zeroed in on two. Firstly, I manipulated the structure of stimuli by comparing groups of people trained with differentially variable stimulus sets. Second, I manipulated the manner in which people learned categories by using a novel incidental, paired-associates learning task. These two areas of focus are expanded upon in turn next.

Effect of variability on category learning and generalization

Saying the structure of experience shapes category representations assumes that the same person with a different experience can represent the same category differently. For example, experiences can be more or less variable around a category center. When choosing books for my daughter to read, I am often faced with the choice between reading her something similar to her favorite (*Green Eggs and Ham*) or something drastically different so that she broadens her knowledge. Is one option better than the other if the goal is for her to generalize reading to other books? Interestingly, Posner & Keele (1968) found that people trained with Category A exemplars (dot arrays; see Figure 3) that were more atypical of Category A generalized more accurately to novel exemplars than people trained with more typical exemplars. This finding was

subsequently replicated (Peterson et al., 1973). More variable experiences provide a better understanding of the breadth of a category (Cohen et al., 2001; Hahn et al., 2005). If one has experience with a three-legged dog, encountering a new dog with a missing leg in the future may reduce potential confusion about whether or not it is a dog. On the flip side, only reading Dr. Seuss likely restricts one's ability to read other material. The notion that more variable experiences promote better generalization is an idea that now cuts across several disciplines, informing educational, machine learning, and sports training recommendations (Raviv et al., 2022). Importantly, this notion is motivated by substantial empirical evidence. Infants are able to generalize certain concepts only after being presented with multiple exemplars (Vukatana et al., 2015). Bean bag tosses are more accurate following practice from three versus one location (Kerr & Booth, 1978). Face recognition is enhanced when shown four versus one different face image (Dukes & Bevan, 1967). People are better able to solve statistical problems after training with four compared to one example (Quilici & Mayer, 1996). Although there is a great deal of nuance and domain-specific details, the general conclusion arrived at in many of these studies is that variability slows learning but enhances generalization.

With respect to categorization studies in particular, the point that variability slows learning is often overlooked. This is an important detail because category learning studies comparing generalization following more or less variable experiences train participants to a learning criterion (Peterson et al., 1973; Posner & Keele, 1968). For example, before the generalization test, participants may be required to achieve 90% accuracy on the past 10 training trials. People trained with more variable instances take much longer to reach the learning criterion, meaning they perform many more training trials than people trained with less variable instances. Hintzman (1986) noted why this is problematic: when retrieving exemplar memory

traces of more variable instances, averaging across so many traces effectively washes out the noise (i.e., variability). This improved quality signal then boosts generalization. In other words, it may not be that variability itself enhances generalization but instead that an increased number of memory traces does. Indeed, when equating training duration, Bowman and Zeithamova (2020) found that people trained with typical exemplars generalized better than people trained with atypical exemplars. This was later replicated with different stimuli and category learning paradigms (Bowman & Zeithamova, 2023, 2025a; Hu & Nosofsky, 2024).

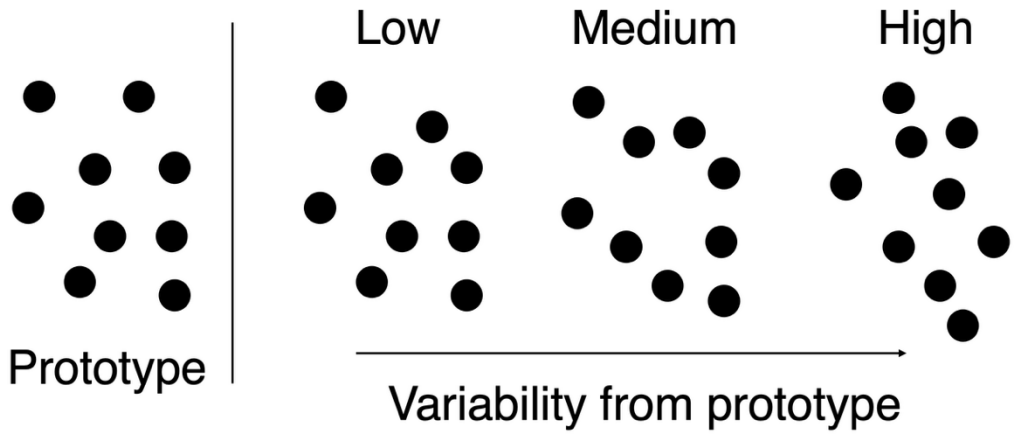

Figure 1.3. Example of the dot pattern prototype distortion paradigm.

A random array of dots in a two-dimensional grid serves as an arbitrarily defined prototype of Category A (left). To construct exemplars with more or less variability about the prototype, prototype dots can be jittered, such that each dot is randomly displaced by a small amount. For simplicity, here I simply jittered one, four, and eight dots to create the low, medium, and high variability exemplars, respectively. Thus, what Posner and Keele (1968) found is that training with high variability exemplars leads to better generalization later on than if one trained with low variability exemplars.

The studies on the variability of experience tend to emphasize a coherent set of training instances, or instances typical of some central tendency, as opposed to noise. Noise denotes *random* variability, which entails that the variance is not diagnostic of category membership. To illustrate, imagine learning to separate red and blue flowers into two categories. All the red flowers share the same color and all the blue flowers share the same color. Thus, color is a

diagnostic feature of the two coherent sets of flowers. However, one flower from each category may have thorns and all other flowers do not. In this case, the thorns are a non-diagnostic feature because having or not having thorns does not help one discriminate red versus blue flowers. As such, the appearance (or non-appearance) of thorns can effectively be treated as noise. Dealing with noise in our environment, learning to filter out irrelevant information, is a central feature of cognition and is also researched across several domains (Barretto-García et al., 2023; Findling et al., 2019; Findling & Wyart, 2021; J. K. Lee et al., 2023; Li et al., 2017; Prat-Carrabin & Woodford, 2022; Salvador et al., 2022), though it is not the only form of variability we encounter (Bowman & Zeithamova, 2025a; Schapiro et al., 2018; Tandoc et al., 2024).

Category exemplars can be atypical, or have few shared features with other exemplars of the same category, while retaining unique features that are also category diagnostic (because having exemplar-unique features necessarily means that those features are also unique to whatever category the exemplar belongs to). Thus, what happens when shared *and* unique features are informative? Will people attend to both? Will people rely on one versus the other when generalizing? In Chapter 2 of the current dissertation, I set out to answer these open questions using a behavioral paradigm carefully tailored to dissociating learning, recognition, and generalization differences between people trained with more or less typical exemplars. Though there was nuance in the findings, I found that training with more typical exemplars enhances category learning, recognition, and generalization when training duration is equated. I also found that, regardless of one's training exemplar typicality, people rely on shared features when generalizing, even when unique features are also category diagnostic. This begs the question: Does every structure to experience promote abstraction? Are we incapable of using the fine-grained details of an experience for subsequent on-the-fly use? To test this, in Chapters 3

and 4, I evaluated signatures of category learning and generalization following a novel behavioral paradigm specifically designed to foster exemplar representations.

Effect of incidental encoding on category representations and generalization

Exemplar theory originated from Medin and Schaffer's (1978) *context model*, which stated that the psychological similarity between two stimuli is context-dependent. For example, we can imagine that ignoring a flower's thorns in the above example will lead one to think two similarly red flowers—one with thorns and one without—are highly similar. This is because, in the current context, color is relevant and thorns are irrelevant. However, perhaps one must now trim the thorns, requiring one to group the flowers into 'thorns' and 'no thorns' piles, in which case the same two flowers may appear dissimilar. In other words, similarity depends on context. Nosofsky (1986, 1987) introduced a generalized version (GCM) that captures context dependency via selective attention weights that can systematically modify the shape of psychological space in which exemplars are embedded. This means that psychological space gets stretched along attended dimensions or features and shrunk along less attended ones. The exemplar model can be used to predict category judgments of novel stimuli by mapping the stimuli to psychological space and computing its similarity to every exemplar within the relevant category. Ergo, GCM tells us similarity depends on context because context guides attention to context-dependent relevant features.

The ability of GCM to capture human categorization was so successful that the debate between exemplar and prototype theory seemed settled (Kruschke, 1992; McKinley & Nosofsky, 1995; Nosofsky, 1986, 1987, 1991, 1992; Nosofsky, Gluck, et al., 1994; Nosofsky & Kruschke, 1992; Nosofsky & Zaki, 1998, 2002; Zaki & Nosofsky, 2001). However, some caveats began to

emerge later. First, the stimulus sets used in the studies verifying the GCM were almost always comprised of low dimensional exemplars (i.e., stimuli with roughly 3-4 features). This is noteworthy because low dimensionality will a priori be easier to memorize and recall in detail but is not the only type of stimuli we encounter in the world. In fact, low dimensional information is rare. When using higher dimensional stimuli, people build up prototypes and use prototypes to categorize novel stimuli (Minda & Smith, 2001, 2002; Smith & Minda, 1998, 2000, 2002). A more technical issue with the initial body of work emphasizing exemplar theory is that the GCM is inherently more capable of capturing *any* pattern of behavioral category judgments than the computational prototype model. This inherent flexibility means that the GCM may not be capturing the true phenomenon of categorization but instead is using its (over-) parameterization to arbitrarily fit the data better. Again, when controlling for this issue, the prototype model is observed to characterize people's categorization behavior better than GCM (Minda & Smith, 2001, 2002; Smith & Minda, 1998, 2000, 2002).

Furthermore, substantial work has now investigated the neural correlates of category representations and the dissociation between prototype and exemplar substrates maps onto that found in spatial navigation and declarative memory. Specifically, place cells in the hippocampus, which respond to specific spatial locations one occupies (O'Keefe & Dostrovsky, 1971), can respond to more or less area. That is, their tuning widths can be narrow, broad, and anywhere in between, corresponding to their spatial coverage of the environment. The longitudinal axis of the hippocampus has been found to be comprised of place cells with tuning widths that scale with their location along this axis (Poppenk et al., 2013; Strange et al., 2014). Thus, place cells in the anterior hippocampus have broad tuning widths, responding to larger portions of space, whereas place cells in the posterior hippocampus have narrow tuning widths (Figure 4). Other work has

similarly revealed that the posterior hippocampus tracks fine-grained details while the anterior hippocampus is responsible for generalization (Brunec et al., 2018; Brunec & Momennejad, 2022; Collin et al., 2015; Poppenk & Moscovitch, 2011; Schlichting et al., 2015). Moreover, the ventromedial prefrontal cortex (vmPFC) was shown to track episodic generalization (Zeithamova et al., 2012). The hippocampus and vmPFC have now repeatedly been implicated in categorization as well. Prototypes have consistently been observed in the anterior hippocampus and vmPFC (Bowman et al., 2020; Bowman & Zeithamova, 2018, 2025b; Y. Chen et al., 2026), while exemplars have been observed in a more distributed network, with marginal evidence for their existence in the posterior hippocampus (Bowman et al., 2020).

Together, the tides have turned yet again, this time away from exemplar theory towards prototypes under more naturalistic conditions when stimuli are high dimensional and model fitting is conducted with a permutation-based approach (Zeithamova et al., 2026). Nevertheless, this shift from earlier work emphasizing exemplars and more recent work emphasizing prototypes does hint at the notion that category representations are highly flexible and can change under different circumstances. An overwhelming majority of categorization research structures category learning the same way (Kurtz, 2015; Markman & Ross, 2003): people are shown instances, prompted to guess which of two categories it belongs in, and subsequently provided corrective feedback. Kurtz (2007, 2015) has argued that this form of training primarily emphasizes discrimination. That is, by repeatedly choosing between categories, one does not really need to learn what makes a category a category but rather the line in stimulus space that divides the two categories. Assuming this is what happens, details and exemplars get blurred out of any representation as people are simply representing what generally falls on either side of the line. Alternatively, the presence of category labels likely implicitly motivates people to abstract

out shared features from across multiple exemplars, for they are thus being told at the outset that all that matters are features relevant for defining the labels. Both of these possibilities likely bias people toward forming prototype representations and abstracting out shared features. However, if the structure of experience shapes the representations we form, then we should expect different category representations to emerge following a different style of category learning. In Chapters 3 and 4, I focused on structuring experience such that exemplar representations will form.

Given that the traditional category learning paradigm seems to implicitly motivate abstraction, I designed a novel category learning paradigm specifically to motivate exemplar memorization. It relies on two key components: incidental learning and the absence of category labels. Specifically, people were shown exemplars atop a scene background, where the background could be a forest or a mountain. They were told to remember the unique exemplar-scene association. Importantly, scenes aligned with exemplar similarity structure, such that exemplars paired with mountains were more similar to each other than exemplars paired with forests and vice versa. In Chapter 3, I directly compared generalization following this paradigm to generalization following the more traditional feedback-based category learning.

Recently, there has been growing interest in learning categories incidentally (Clapper & Bower, 1994, 2002; Doya, 1999; Gabay et al., 2015; Gabay, Madlansacay, et al., 2023; Gabay, Roark, et al., 2023; Gabay & Holt, 2015; Lim et al., 2019; Love, 2002, 2003; Roark et al., 2022; Tricomi et al., 2006; Unger & Sloutsky, 2022), given that incidental learning is much more prominent in the real world (Clapper, 2006; Clapper & Bower, 1994, 2002; Unger & Sloutsky, 2022). Incidental learning occurs without the explicit goal of grouping information into distinct categories. Outside of the laboratory, people are rarely instructed to search for category-diagnostic features or receive immediate feedback. Instead, incidental category learning suggests

that people simply detect correlations between features in their environment (Billman & Knutson, 1996; Gabay et al., 2015; Lim et al., 2019), such that overlapping similarity gets encoded in the brain. Without a goal or label to guide attention, it was predicted that incidental learning would facilitate broader attention across exemplar features, leading to better exemplar memorization and an exemplar-based generalization strategy. In Chapter 4 of this dissertation, I additionally investigated the neural correlates of category representations following incidental category learning using functional magnetic resonance imaging (fMRI).

Preview of current dissertation

To preview the general findings from this collection of empirical studies, I found that category representations and generalization strategy are both quite flexible in healthy adults. In the first experiment, I observed evidence for abstraction despite high variability in training information, but I also observed that people can rely on specific, unique details when abstraction fails them, suggesting that both forms of category representations are learned simultaneously to differing degrees depending on training variability. In the second and third experiments, I found that incidental category learning through paired associates' observation can facilitate greater reliance on specific information when generalizing, which is distinct from previous work in the lab showing greater reliance on general information after learning categories via feedback-based training. The third experiment additionally presents evidence of exemplar representations in the brain when people generalized incidentally learned category information, providing novel insight into the flexibility of the brain to use different representations to meet task demands. This body of research has important implications for a thorough understanding of category learning, category representations, and generalization of category knowledge.

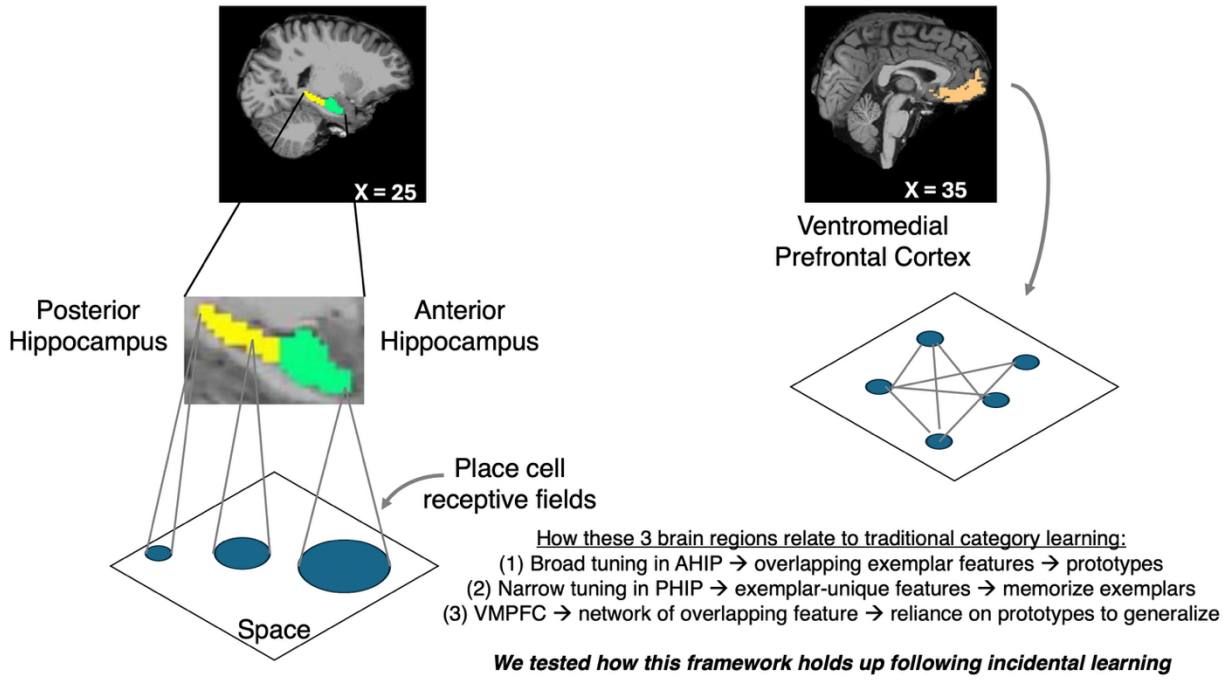

Figure 1.4. FMRI markers of exemplar and prototype representations.

References cited

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). https://doi.org/10.1037/0033-295X.105.3.442

Babadi, B., & Sompolinsky, H. (2014). Sparseness and Expansion in Sensory Representations. *Neuron*, *83*(5). https://doi.org/10.1016/j.neuron.2014.07.035

Badre, D., Kayser, A. S., & D'Esposito, M. (2010). Frontal cortex and the discovery of abstract action rules. *Neuron*, *66*(2). https://doi.org/10.1016/j.neuron.2010.03.025

Barranca, V. J., Kovačič, G., Zhou, D., & Cai, D. (2014). Sparsity and Compressed Coding in Sensory Systems. *PLoS Computational Biology*, *10*(8). https://doi.org/10.1371/journal.pcbi.1003793

Barrett, L. F., & Miller, E. K. (2025). *Categorization is "Baked" Into the Brain*. https://doi.org/10.31234/osf.io/bu4sn_v1

Barretto-García, M., de Hollander, G., Grueschow, M., Polanía, R., Woodford, M., & Ruff, C. C. (2023). Individual risk attitudes arise from noise in neurocognitive magnitude representations. *Nature Human Behaviour*, *7*(9). https://doi.org/10.1038/s41562-023-01643-4

Bernardi, S., Benna, M. K., Rigotti, M., Munuera, J., Fusi, S., & Salzman, C. D. (2020). The Geometry of Abstraction in the Hippocampus and Prefrontal Cortex. *Cell*, *183*(4). https://doi.org/10.1016/j.cell.2020.09.031

Billman, D., & Knutson, J. (1996). Unsupervised concept learning and value systematicitiy: A complex whole aids learning the parts. *Journal of Experimental Psychology: Learning, Memory, and Cognition*, *22*(2). https://doi.org/10.1037//0278-7393.22.2.458

Bowman, C. R., Iwashita, T., & Zeithamova, D. (2020). Tracking prototype and exemplar representations in the brain across learning. *ELife*, *9*, 1–47. https://doi.org/10.7554/eLife.59360

Bowman, C. R., & Zeithamova, D. (2018). Abstract memory representations in the ventromedial prefrontal cortex and hippocampus support concept generalization. *Journal of Neuroscience*, *38*(10), 2605–2614. https://doi.org/10.1523/JNEUROSCI.2811-17.2018

Bowman, C. R., & Zeithamova, D. (2020). Training Set Coherence and Set Size Effects on Concept Generalization and Recognition. *Journal of Experimental Psychology: Learning Memory and Cognition*. https://doi.org/10.1037/xlm0000824

Bowman, C. R., & Zeithamova, D. (2023). Coherent category training enhances generalization in prototype-based categories. *Journal of Experimental Psychology: Learning, Memory, and Cognition*. https://doi.org/10.1037/xlm0001243

Bowman, C. R., & Zeithamova, D. (2025a). High coherence among training exemplars promotes broad generalization of face families. *Journal of Experimental Psychology: Learning, Memory, and Cognition*. https://doi.org/10.1037/xlm0001478

Bowman, C. R., & Zeithamova, D. (2025b). *Ventromedial prefrontal cortex supports prototype representations in healthy older adults*. https://doi.org/10.1101/2025.09.24.678354

Brunec, I. K., Bellana, B., Ozubko, J. D., Man, V., Robin, J., Liu, Z. X., Grady, C., Rosenbaum, R. S., Winocur, G., Barense, M. D., & Moscovitch, M. (2018). Multiple Scales of Representation along the Hippocampal Anteroposterior Axis in Humans. *Current Biology*, *28*(13). https://doi.org/10.1016/j.cub.2018.05.016

Brunec, I. K., & Momennejad, I. (2022). Predictive Representations in Hippocampal and Prefrontal Hierarchies. *Journal of Neuroscience*, *42*(2). https://doi.org/10.1523/JNEUROSCI.1327-21.2021

Chen, Y., Hawkins, B., Puckett, H., Sharp, K., Lopez, A., Zeithamova, D., Xie, H., Verbalis, A., Van Meter, A. S., Gaillard, W. D., Kenworthy, L., & Vaidya, C. J. (2026). *From cognitive abstraction to adaptive behavior: neural bases of concept learning in autistic adolescents*. https://doi.org/10.64898/2026.01.17.700125

Chin-Parker, S., & Cantelon, J. (2017). Contrastive Constraints Guide Explanation-Based Category Learning. *Cognitive Science*, *41*(6). https://doi.org/10.1111/cogs.12405

Chin-Parker, S., & Ross, B. H. (2004). Diagnosticity and Prototypicality in Category Learning: A Comparison of Inference Learning and Classification Learning. *Journal of Experimental Psychology: Learning Memory and Cognition*, *30*(1). https://doi.org/10.1037/0278-7393.30.1.216

Clapper, J. P. (2006). When more is less: Negative exposure effects in unsupervised learning. *Memory and Cognition*, *34*(4). https://doi.org/10.3758/BF03193435

Clapper, J. P., & Bower, G. H. (1994). Category Invention in Unsupervised Learning. *Journal of Experimental Psychology: Learning, Memory, and Cognition*, *20*(2). https://doi.org/10.1037/0278-7393.20.2.443

Clapper, J. P., & Bower, G. H. (2002). Adaptive Categorization in Unsupervised Learning. *Journal of Experimental Psychology: Learning Memory and Cognition*, *28*(5). https://doi.org/10.1037/0278-7393.28.5.908

Cohen, A. L., Nosofsky, R. M., & Zaki, S. R. (2001). Category variability, exemplar similarity, and perceptual classification. *Memory and Cognition*, *29*(8). https://doi.org/10.3758/BF03206386

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

Daume, J., Kamiński, J., Salimpour, Y., Gómez Palacio Schjetnan, A., Anderson, W. S., Valiante, T. A., Mamelak, A. N., & Rutishauser, U. (2024). Persistent activity during working memory maintenance predicts long-term memory formation in the human hippocampus. *Neuron*, *112*(23). https://doi.org/10.1016/j.neuron.2024.09.013

Davis, T., Xue, G., Love, B. C., Preston, A. R., & Poldrack, R. A. (2014). Global neural pattern similarity as a common basis for categorization and recognition memory. *Journal of Neuroscience*, *34*(22). https://doi.org/10.1523/JNEUROSCI.3376-13.2014

DeBruine, L. M. (2002). Facial resemblance enhances trust. *Proceedings of the Royal Society B: Biological Sciences*, *269*(1498). https://doi.org/10.1098/rspb.2002.2034

DeBruine, L. M., Jones, B. C., Little, A. C., & Perrett, D. I. (2008). Social perception of facial resemblance in humans. *Archives of Sexual Behavior*, *37*(1). https://doi.org/10.1007/s10508-007-9266-0

Doya, K. (1999). What are the computations of the cerebellum, the basal ganglia and the cerebral cortex? *Neural Networks*, *12*(7–8). https://doi.org/10.1016/S0893-6080(99)00046-5

Dukes, W. F., & Bevan, W. (1967). STIMULUS VARIATION AND REPETITION IN THE ACQUISITION OF NAMING RESPONSES. *Journal of Experimental Psychology*, *74*(2 PART 1). https://doi.org/10.1037/h0024575

Ewing, L., Caulfield, F., Read, A., & Rhodes, G. (2015). Perceived trustworthiness of faces drives trust behaviour in children. *Developmental Science*, *18*(2). https://doi.org/10.1111/desc.12218

Ewing, L., Sutherland, C. A. M., & Willis, M. L. (2019). Children show adult-like facial appearance biases when trusting others. *Developmental Psychology*, *55*(8). https://doi.org/10.1037/dev0000747

Feldman, J. (2000). Minimization of Boolean complexity in human concept learning. *Nature*, *407*(6804). https://doi.org/10.1038/35036586

Feldman, J. (2003). The Simplicity Principle in Human Concept Learning. In *Current Directions in Psychological Science* (Vol. 12, Number 6). https://doi.org/10.1046/j.0963-7214.2003.01267.x

Feldman, J. (2006). An algebra of human concept learning. *Journal of Mathematical Psychology*, *50*(4). https://doi.org/10.1016/j.jmp.2006.03.002

Feldman, J. (2023). Probabilistic Origins of Compositional Mental Representations. *Psychological Review*, *131*(3). https://doi.org/10.1037/rev0000452

Findling, C., Skvortsova, V., Dromnelle, R., Palminteri, S., & Wyart, V. (2019). Computational noise in reward-guided learning drives behavioral variability in volatile environments. *Nature Neuroscience*, *22*(12). https://doi.org/10.1038/s41593-019-0518-9

Findling, C., & Wyart, V. (2021). Computation noise in human learning and decision-making: origin, impact, function. In *Current Opinion in Behavioral Sciences* (Vol. 38). https://doi.org/10.1016/j.cobeha.2021.02.018

Flade, F., Klar, Y., & Imhoff, R. (2019). Unite against: A common threat invokes spontaneous decategorization between social categories. *Journal of Experimental Social Psychology*, *85*. https://doi.org/10.1016/j.jesp.2019.103890

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). https://doi.org/10.1037/xhp0000073

Gabay, Y., & Holt, L. L. (2015). Incidental learning of sound categories is impaired in developmental dyslexia. *Cortex*, *73*. https://doi.org/10.1016/j.cortex.2015.08.008

Gabay, Y., Madlansacay, M., & Holt, L. L. (2023). Incidental auditory category learning and visuomotor sequence learning do not compete for cognitive resources. *Attention, Perception, and Psychophysics*, *85*(2). https://doi.org/10.3758/s13414-022-02616-x

Gabay, Y., Roark, C. L., & Holt, L. L. (2023). Impaired and Spared Auditory Category Learning in Developmental Dyslexia. *Clinical Psychological Science*, *34*(4). https://doi.org/10.1177/09567976231151581

Ganguli, S., & Sompolinsky, H. (2012). Compressed sensing, sparsity, and dimensionality in neuronal information processing and data analysis. In *Annual Review of Neuroscience* (Vol. 35). https://doi.org/10.1146/annurev-neuro-062111-150410

Giang, T., Bell, R., & Buchner, A. (2012). Does Facial Resemblance Enhance Cooperation? *PLoS ONE*, *7*(10). https://doi.org/10.1371/journal.pone.0047809

Hahn, U., Bailey, T. M., & Elvin, L. B. C. (2005). Effects of category diversity on learning, memory, and generalization. *Memory and Cognition*, *33*(2). https://doi.org/10.3758/BF03195318

Hewstone, M., Hantzi, A., & Johnston, L. (1991). Social categorization and person memory: The pervasiveness of race as an organizing principle. *European Journal of Social Psychology*, *21*(6). https://doi.org/10.1002/ejsp.2420210606

Hintzman, D. L. (1986). "Schema Abstraction" in a Multiple-Trace Memory Model. *Psychological Review*, *93*(4). https://doi.org/10.1037/0033-295X.93.4.411

Hu, M., & Nosofsky, R. M. (2024). High-variability training does not enhance generalization in the prototype-distortion paradigm. *Memory & Cognition*. https://doi.org/10.3758/s13421-023-01516-1

JD, W., KC, A., & EK, M. (2001). Single neurons in prefrontal cortex encode abstract rules. *Nature*, *411*(June).

Kahnt, T., Park, S. Q., Burke, C. J., & Tobler, P. N. (2012). How glitter relates to gold: Similarity-dependent reward prediction errors in the human striatum. *Journal of Neuroscience*, *32*(46). https://doi.org/10.1523/JNEUROSCI.2383-12.2012

Kerr, R., & Booth, B. (1978). Specific and varied practice of motor skill. *Perceptual and Motor Skills*, *46*(2). https://doi.org/10.1177/003151257804600201

Kim, S., & Rehder, B. (2011). How prior knowledge affects selective attention during category learning: An eyetracking study. *Memory and Cognition*, *39*(4). https://doi.org/10.3758/s13421-010-0050-3

Kruschke, J. K. (1992). ALCOVE: An exemplar-based connectionist model of category learning. *Psychological Review*, *99*(1). https://doi.org/10.1037/0033-295X.99.1.22

Kurtz, K. J. (2007). The divergent autoencoder (DIVA) model of category learning. *Psychonomic Bulletin and Review*, *14*(4). https://doi.org/10.3758/BF03196806

Kurtz, K. J. (2015). Human Category Learning: Toward a Broader Explanatory Account. *Psychology of Learning and Motivation - Advances in Research and Theory*, *63*. https://doi.org/10.1016/bs.plm.2015.03.001

Lee, J. K., Rouault, M., & Wyart, V. (2023). Adaptive tuning of human learning and choice variability to unexpected uncertainty. *Science Advances*, *9*(13). https://doi.org/10.1126/sciadv.add0501

Lee, P. C. (1990). How monkeys see the world: Inside the mind of another species. *Trends in Ecology & Evolution*, *5*(12). https://doi.org/10.1016/0169-5347(90)90037-e

Li, V., Herce Castañón, S., Solomon, J. A., Vandormael, H., & Summerfield, C. (2017). Robust averaging protects decisions from noise in neural computations. *PLoS Computational Biology*, *13*(8). https://doi.org/10.1371/journal.pcbi.1005723

Lim, S. J., Fiez, J. A., & Holt, L. L. (2019). Role of the striatum in incidental learning of sound categories. *Proceedings of the National Academy of Sciences of the United States of America*, *116*(10). https://doi.org/10.1073/pnas.1811992116

Love, B. C. (2002). Comparing supervised and unsupervised category learning. *Psychonomic Bulletin and Review*, *9*(4). https://doi.org/10.3758/BF03196342

Love, B. C. (2003). The multifaceted nature of unsupervised category learning. *Psychonomic Bulletin and Review*, *10*(1). https://doi.org/10.3758/BF03196484

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). https://doi.org/10.1016/j.cub.2013.08.035

Maner, J. K., Miller, S. L., Moss, J. H., Leo, J. L., & Ashby Plant, E. (2012). Motivated social categorization: Fundamental motives enhance people's sensitivity to basic social categories. *Journal of Personality and Social Psychology*, *103*(1). https://doi.org/10.1037/a0028172

Mansouri, F. A., Freedman, D. J., & Buckley, M. J. (2020). Emergence of abstract rules in the primate brain. In *Nature Reviews Neuroscience* (Vol. 21, Number 11). https://doi.org/10.1038/s41583-020-0364-5

Markman, A. B., & Ross, B. H. (2003). Category Use and Category Learning. In *Psychological Bulletin* (Vol. 129, Number 4). https://doi.org/10.1037/0033-2909.129.4.592

McKinley, S. C., & Nosofsky, R. M. (1995). Investigations of Exemplar and Decision Bound Models in Large, Ill-Defined Category Structures. *Journal of Experimental Psychology: Human Perception and Performance*, *21*(1). https://doi.org/10.1037/0096-1523.21.1.128

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

Minda, J. P., Desroches, A. S., & Church, B. A. (2008). Learning Rule-Described and Non-Rule-Described Categories: A Comparison of Children and Adults. *Journal of Experimental Psychology: Learning Memory and Cognition*, *34*(6). https://doi.org/10.1037/a0013355

Minda, J. P., Roark, C. L., Kalra, P., & Cruz, A. (2024). Single and multiple systems in categorization and category learning. In *Nature Reviews Psychology* (Vol. 3, Number 8). https://doi.org/10.1038/s44159-024-00336-7

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). https://doi.org/10.1037/0278-7393.27.3.775

Minda, J. P., & Smith, J. D. (2002). Comparing Prototype-Based and Exemplar-Based Accounts of Category Learning and Attentional Allocation. *Journal of Experimental Psychology: Learning Memory and Cognition*, *28*(2). https://doi.org/10.1037/0278-7393.28.2.275

Nosofsky, R. M. (1986). Attention, similarity, and the identification-categorization relationship. *Journal of Experimental Psychology: General*, *115*(1). 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). https://doi.org/10.1037/0278-7393.13.1.87

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). https://doi.org/10.1037/0096-1523.17.1.3

Nosofsky, R. M. (1992). Exemplar, prototypes, and similarity rules. *From Learning Theory to Connectionist Theory: Essays in Honor of William K. Estes, Vol. 1*.

Nosofsky, R. M., Gluck, M. A., Palmeri, T. J., Mckinley, S. C., & Glauthier, P. (1994). Comparing modes of rule-based classification learning: A replication and extension of Shepard, Hovland, and Jenkins (1961). *Memory & Cognition*, *22*(3). https://doi.org/10.3758/BF03200862

Nosofsky, R. M., & Kruschke, J. K. (1992). Investigations of an exemplar-based connectionist model of category learning. *Psychology of Learning and Motivation - Advances in Research and Theory*, *28*(C). https://doi.org/10.1016/S0079-7421(08)60491-0

Nosofsky, R. M., Palmeri, T. J., & McKinley, S. C. (1994). Rule-plus-exception model of classification learning. *Psychological Review*, *101*(1). https://doi.org/10.1037/0033-295X.101.1.53

Nosofsky, R. M., & Zaki, S. R. (1998). Dissociations Between Categorization and Recognition in Amnesic and Normal Individuals: An Exemplar-Based Interpretation. *Psychological Science*, *9*(4), 247–255. https://doi.org/10.1111/1467-9280.00051

Nosofsky, R. M., & Zaki, S. R. (2002). Exemplar and Prototype Models Revisited: Response Strategies, Selective Attention, and Stimulus Generalization. *Journal of Experimental Psychology: Learning Memory and Cognition*, *28*(5). https://doi.org/10.1037/0278-7393.28.5.924

O'Keefe, J., & Dostrovsky, J. (1971). The hippocampus as a spatial map. Preliminary evidence from unit activity in the freely-moving rat. *Brain Research*, *34*(1), 171–175. https://doi.org/10.1016/0006-8993(71)90358-1

Paniukov, D., & Davis, T. (2018). The evaluative role of rostrolateral prefrontal cortex in rule-based category learning. *NeuroImage*, *166*. https://doi.org/10.1016/j.neuroimage.2017.10.057

Peterson, M. J., Meagher, R. B., Chait, H., & Gillie, S. (1973). The abstraction and generalization of dot patterns. *Cognitive Psychology*, *4*(3). https://doi.org/10.1016/0010-0285(73)90019-4

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

Poppenk, J., & Moscovitch, M. (2011). A hippocampal marker of recollection memory ability among healthy young adults: Contributions of posterior and anterior segments. *Neuron*, *72*(6). https://doi.org/10.1016/j.neuron.2011.10.014

Posner, M. I., Goldsmith, R., & Welton, K. E. (1967). PERCEIVED DISTANCE AND THE CLASSIFICATION OF DISTORTED PATTERNS. *Journal of Experimental Psychology*, *73*(1). https://doi.org/10.1037/h0024135

Posner, M. I., & Keele, S. W. (1968). ON THE GENESIS OF ABSTRACT IDEAS. *Journal of Experimental Psychology*, *77*(3 PART 1). https://doi.org/10.1037/h0025953

Prat-Carrabin, A., & Woodford, M. (2022). Efficient coding of numbers explains decision bias and noise. *Nature Human Behaviour*, *6*(8). https://doi.org/10.1038/s41562-022-01352-4

Quilici, J. L., & Mayer, R. E. (1996). Role of Examples in How Students Learn to Categorize Statistics Word Problems. *Journal of Educational Psychology*, *88*(1). https://doi.org/10.1037/0022-0663.88.1.144

Quiroga, R. Q. (2012). Concept cells: the building blocks of declarative memory functions. In *Nature Reviews Neuroscience* (Vol. 13, Number 8, pp. 587–597). https://doi.org/10.1038/nrn3251

Quiroga, R. Q., Reddy, L., Kreiman, G., Koch, C., & Fried, I. (2005). Invariant visual representation by single neurons in the human brain. *Nature*, *435*(7045), 1102–1107. https://doi.org/10.1038/nature03687

Rabi, R., Miles, S. J., & Minda, J. P. (2015). Learning categories via rules and similarity: Comparing adults and children. *Journal of Experimental Child Psychology*, *131*. https://doi.org/10.1016/j.jecp.2014.10.007

Rabi, R., & Minda, J. P. (2014). Rule-based category learning in children: The role of age and executive functioning. *PLoS ONE*, *9*(1). https://doi.org/10.1371/journal.pone.0085316

Rabi, R., & Minda, J. P. (2016). Category learning in older adulthood: A study of the Shepard, Hovland, and Jenkins (1961) tasks. *Psychology and Aging*, *31*(2). https://doi.org/10.1037/pag0000071

Raviv, L., Lupyan, G., & Green, S. C. (2022). How variability shapes learning and generalization. In *Trends in Cognitive Sciences* (Vol. 26, Number 6). https://doi.org/10.1016/j.tics.2022.03.007

Rehder, B., & Hoffman, A. B. (2005a). Eyetracking and selective attention in category learning. *Cognitive Psychology*, *51*(1). https://doi.org/10.1016/j.cogpsych.2004.11.001

Rehder, B., & Hoffman, A. B. (2005b). Thirty-something categorization results explained: Selective attention, eyetracking, and models of category learning. In *Journal of Experimental Psychology: Learning Memory and Cognition* (Vol. 31, Number 5). https://doi.org/10.1037/0278-7393.31.5.811

Reinert, S., Hübener, M., Bonhoeffer, T., & Goltstein, P. M. (2021). Mouse prefrontal cortex represents learned rules for categorization. *Nature*, *593*(7859). https://doi.org/10.1038/s41586-021-03452-z

Rey, H. G., Gori, B., Chaure, F. J., Collavini, S., Blenkmann, A. O., Seoane, P., Seoane, E., Kochen, S., & Quian Quiroga, R. (2020). Single Neuron Coding of Identity in the Human Hippocampal Formation. *Current Biology*, *30*(6), 1152-1159.e3. https://doi.org/10.1016/j.cub.2020.01.035

Roark, C. L., Lehet, M. I., Dick, F., & Holt, L. L. (2022). The Representational Glue for Incidental Category Learning Is Alignment With Task-Relevant Behavior. *Journal of Experimental Psychology: Learning Memory and Cognition*, *48*(6). https://doi.org/10.1037/xlm0001078

Rosch, E., & Mervis, C. B. (1975). Family resemblances: Studies in the internal structure of categories. *Cognitive Psychology*, *7*(4). https://doi.org/10.1016/0010-0285(75)90024-9

Salvador, A., Arnal, L. H., Vinckier, F., Domenech, P., Gaillard, R., & Wyart, V. (2022). Premature commitment to uncertain decisions during human NMDA receptor hypofunction. *Nature Communications*, *13*(1). https://doi.org/10.1038/s41467-021-27876-3

Schapiro, A. C., McDevitt, E. A., Rogers, T. T., Mednick, S. C., & Norman, K. A. (2018). Human hippocampal replay during rest prioritizes weakly learned information and predicts memory performance. *Nature Communications*, *9*(1). https://doi.org/10.1038/s41467-018-06213-1

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*. https://doi.org/10.1038/ncomms9151

Shepard, R. N. (1987). Toward a universal law of generalization for psychological science. *Science*, *237*(4820). https://doi.org/10.1126/science.3629243

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

Singer, Y., Teramoto, Y., Willmore, B. D. B., King, A. J., Schnupp, J. W. H., & Harper, N. S. (2018). Sensory cortex is optimised for prediction of future input. *ELife*, *7*. https://doi.org/10.7554/eLife.31557

Smith, J. D. (2014). Prototypes, exemplars, and the natural history of categorization. *Psychonomic Bulletin and Review*, *21*(2). https://doi.org/10.3758/s13423-013-0506-0

Smith, J. D., & Minda, J. P. (1998). Prototypes in the mist: The early epochs of category learning. *Journal of Experimental Psychology: Learning Memory and Cognition*, *24*(6). https://doi.org/10.1037/0278-7393.24.6.1411

Smith, J. D., & Minda, J. P. (2000). Thirty Categorization Results in Search of a Model. *Journal of Experimental Psychology: Learning Memory and Cognition*, *26*(1). https://doi.org/10.1037/0278-7393.26.1.3

Smith, J. D., & Minda, J. P. (2002). Distinguishing Prototype-Based and Exemplar-Based Processes in Dot-Pattern Category Learning. *Journal of Experimental Psychology: Learning Memory and Cognition*, *28*(4). https://doi.org/10.1037/0278-7393.28.4.800

Smith, J. D., Minda, J. P., & Washbum, D. A. (2004). Category learning in rhesus monkeys: A study of the Shepard, Hovland, and Jenkins (1961) tasks. In *Journal of Experimental Psychology: General* (Vol. 133, Number 3). https://doi.org/10.1037/0096-3445.133.3.398

Sterling, P. (2012). Allostasis: A model of predictive regulation. *Physiology and Behavior*, *106*(1). https://doi.org/10.1016/j.physbeh.2011.06.004

Sterling, P. (2018). Predictive regulation and human design. *ELife*, *7*. https://doi.org/10.7554/eLife.36133

Sterling, P., & Laughlin, S. (2015). Principles of neural design. In *Principles of Neural Design*. https://doi.org/10.7551/mitpress/9395.003.0019

Strange, B. A., Witter, M. P., Lein, E. S., & Moser, E. I. (2014). Functional organization of the hippocampal longitudinal axis. In *Nature Reviews Neuroscience* (Vol. 15, Number 10). https://doi.org/10.1038/nrn3785

Tandoc, M. C., Dong, C. V., & Schapiro, A. C. (2024). Object Feature Memory Is Distorted by Category Structure. *Open Mind*, *8*, 1348–1368. https://doi.org/10.1162/opmi_a_00170

Tricomi, E., Delgado, M. R., McCandliss, B. D., McClelland, J. L., & Fiez, J. A. (2006). Performance feedback drives caudate activation in a phonological learning task. *Journal of Cognitive Neuroscience*, *18*(6). https://doi.org/10.1162/jocn.2006.18.6.1029

Unger, L., & Sloutsky, V. M. (2022). Ready to Learn: Incidental Exposure Fosters Category Learning. *Psychological Science*, *33*(6). https://doi.org/10.1177/09567976211061470

Vukatana, E., Graham, S. A., Curtin, S., & Zepeda, M. S. (2015). One is Not Enough: Multiple Exemplars Facilitate Infants' Generalizations of Novel Properties. *Infancy*, *20*(5). https://doi.org/10.1111/INFA.12092

Wittgenstein, L. (1953). Philosophical investigations. Philosophische Untersuchungen. In *Philosophical investigations. Philosophische Untersuchungen.*

Yamauchi, T., Love, B. C., & Markman, A. B. (2002). Learning Nonlinearly Separable Categories by Inference and Classification. *Journal of Experimental Psychology: Learning Memory and Cognition*, *28*(3). https://doi.org/10.1037/0278-7393.28.3.585

Yamauchi, T., & Markman, A. B. (1998). Category learning by inference and classification. *Journal of Memory and Language*, *39*(1). https://doi.org/10.1006/jmla.1998.2566

Yamauchi, T., & Markman, A. B. (2000). Inference Using Categories. *Journal of Experimental Psychology: Learning Memory and Cognition*, *26*(3). https://doi.org/10.1037/0278-7393.26.3.776

Zaki, S. R., & Nosofsky, R. M. (2001). A single-system interpretation of dissociations between recognition and categorization in a task involving object-like stimuli. *Cognitive, Affective and Behavioral Neuroscience*, *1*(4). https://doi.org/10.3758/CABN.1.4.344

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). https://doi.org/10.1016/j.neuron.2012.05.010

Zeithamova, D., Houser, T. M., & Bowman, C. R. (2026). Evaluating Cognitive Models With Permutation Testing: A Case Study of Prototype and Exemplar Categorization. *Advances in Methods and Practices in Psychological Science*, *9*(2). https://doi.org/10.1177/25152459261430729

Zeithamova, D., Mack, M. L., Braunlich, K., Davis, T., Seger, C. A., Kesteren, M. T. R. V., & Wutz, A. (2019). Brain mechanisms of concept learning. *Journal of Neuroscience*, *39*(42). https://doi.org/10.1523/JNEUROSCI.1166-19.2019

Chapter 2

Low exemplar coherence during category learning limits both exemplar memory and concept generalization

Category learning is a ubiquitous cognitive process involving connecting individual items to a shared category label based on rules or similarity. Concepts, or mental representations of categories, are highly adaptive as they enable one to transfer and generalize category knowledge to novel situations. Given the adaptive nature of generalization, identifying the factors that facilitate the formation of generalizable knowledge is of both theoretical and practical importance.

One such factor affecting generalization that has received particular attention is the variability of training exemplars, a topic of interest across cognitive psychology (Gorman & Goldstone, 2022; Posner & Keele, 1968), machine learning (Manenti et al., 2023; Tenenbaum & Griffiths, 2001), language (Perry et al., 2010), sports training (Hernández-Davo et al., 2014), and developmental psychology (Rost & McMurray, 2010). Most commonly, it has been assumed that learning a category using more varied exemplars initially hinders learning but enhances generalization (Raviv et al., 2022). Intuitively, having more variable instances may teach one the category boundaries, enabling better subsequent generalization (Nosofsky et al., 2019).

Early evidence seemingly aligned with these notions. In the human categorization literature, this topic was first explored using the dot pattern prototype distortion paradigm (Posner & Keele, 1968) and then replicated multiple times with other stimuli (Homa & Vosburgh, 1976; Peterson et al., 1973). These results continue shaping current theoretical frameworks for how to best educate individuals (Raviv et al., 2022). Although the question

seemed settled decades ago, the evidence ended up not as clear-cut. As one of the first voices of caution, Hintzman (1984) noted an important limitation of the original studies, namely, that participants were trained to a performance criterion prior to assessing generalization. When training duration is equated, Hintzman (1984) predicted, based on his computational model of category learning MINERVA, that training with more typical (more coherent, less distorted) exemplars should facilitate better generalization. In MINERVA, each training instance is stored as a separate memory trace. When a new stimulus is encountered, it activates stored traces based on similarity. The output of the network, referred to as an echo, is a weighted average of the activated traces. Typical traces closer to the central tendency of a category will overlap more in feature space, activating each other more strongly, leading to stronger and more coherent echoes that are more useful for generalization. In contrast, high variability traces are less similar to each other, and thus do not reinforce each other in memory, leading to a weaker and noisier echo. With limited training, typical examples give the most representative information in the shortest amount of time.

Bowman and Zeithamova (2020) tested Hintzman's prediction using cartoon animal stimuli with separable, binary-valued features. Participants learned to categorize exemplars into one of two categories via trial-and-error. The number of features shared with each training instance's respective prototype differed between groups: the high coherence group was trained with instances that shared more features than the low coherence group. Importantly, the number of training trials were equated between groups. Consistent with Hintzman's (1984) prediction, the high coherence group exhibited not just faster learning but also better generalization of category knowledge. The benefits of training with *less* variable exemplars have since been

replicated with not only binary dimension stimuli (Bowman & Zeithamova, 2023) but a dot pattern prototype distortion paradigm as well (Hu & Nosofsky, 2024).

Notably, in most studies that manipulate coherence (variability) of the training exemplars, higher variability is associated with more noise (e.g., dot pattern stimuli) or a greater proportion of features associated with the opposing category (e.g., binary dimension stimuli). This makes less coherent, less typical examples more difficult to categorize because there is less category-defining information available. To test whether benefits of high coherence, low variability training extend to situations when increased variability does not add noise, Bowman and Zeithamova (2025) used face blend stimuli that combined shared category features (shared parent faces) with stimulus-specific features (unique parent faces) that were *also* perfectly predictive of category membership. Despite both shared and unique features being perfectly diagnostic of their respective category, participants in the high coherence group trained on 50/50 blends (50% shared parent and 50% unique parent) learned category structure faster and generalized better than those in the low coherence group trained on 30/70 blends (30% shared parent and 70% unique parent). Thus, across different stimuli and category structures, recent work consistently shows that *low* variability (high coherence) rather than *high* variability (low coherence) promotes generalization when training length is equated (Bowman & Zeithamova, 2020, 2023, 2025a; Hu & Nosofsky, 2024).

Interestingly, the prior studies and the latest study by Bowman and Zeithamova (2025) included additional findings that were less expected and that led to additional questions about the types of representations participants form in the high and low coherence conditions. First, across studies, high coherence, low variability training always leads to superior generalization of items near the category center, with similar albeit weaker benefits for items further from the category

center (Bowman & Zeithamova, 2020, 2023; Hu & Nosofsky, 2024). This effect may extend to low coherence training items themselves. Bowman and Zeithamova (2025) directly compared categorization of low coherence faces that were training items in the low coherence group and generalization items in the high coherence group. The high coherence group was just as good (and numerically slightly better) as the low coherence group at categorizing test faces that were the low coherence group's training faces. This opens the question of whether it is possible that learning low variability items and then generalizing to high variability item can be superior compared to aiming to learn the same high variability items directly. The challenge to learn and remember category stimuli that do not have clear commonalities was also illustrated by another finding from Bowman and Zeithamova (2025). The study manipulated set size (3 vs. 6 training exemplars per condition) in addition to coherence. It was expected that coherence would play less of a role when set size is small, as participants should be able to simply memorize category labels for 6 training stimuli, 3 per category. However, set size had minimal effect, with the same poor learning and generalization in the low coherence condition irrespective of set size. This indicated that an exemplar memorization strategy was challenging to utilize.

Another observation from prior studies is that high coherence, low variability training may facilitate learning and generalization because it facilitates extraction of category shared features (Bowman & Zeithamova, 2020, 2023, 2025a). For example, in Bowman and Zeithamova (2025), the high coherence group appeared to form memory representations that were shifted towards the shared features and category-shared features dominated behavior even when stimulus-unique features were also category-relevant and more heavily weighted in test stimuli. First, when categorizing ambiguous face blends that combined shared features from one category with unique features from the opposing category, participants in the high coherence

group exhibited a reliable bias toward categorizing based on shared features, even for face blends that contained a greater proportion of unique than shared features. Second, during recognition testing, the high coherence group trained on 50/50 blends (50% shared face, 50% unique face) tended to false alarm to near-prototype parent blends (90% shared features) but be less likely than the low coherence group to confuse 30/70 blends with 50/50 blends. This pattern of results in the high coherence group suggests increased attention to shared features and/or the formation of memory representations shifted towards shared features.

The types of representations formed in the low coherence training group in Bowman and Zeithamova (2025) were more difficult to evaluate. Categorization success for unambiguous face blends was low and nearly uniform irrespective of the blend level (90% shared, 50% shared, 30% shared), showing no evidence for the abstraction of shared features or the central tendency of the category. Categorization of ambiguous face blends that combined shared features of one category with unique features from opposing category yielded a roughly equal number of responses consistent with shared features and those consistent with unique features. However, as there were only two response options, it was challenging to determine whether such responses reflect equal weighting of shared and unique features or whether they were simply random. Thus, it remains unclear to what degree low coherence training leads to a general learning failure, increased representation of unique features, and/or the extraction and bias towards shared features that is quantitatively but not qualitatively different from high coherence training.

In the current study, we adapted the design from Bowman and Zeithamova (2025) to address these open questions. First, the category structure consisted of three categories rather than two. We reasoned that adding a third category and lowering chance performance from 50% to 33% may further increase the sensitivity of this paradigm, allowing us to determine whether

high coherence training followed by generalization to low coherence items may be superior to direct training on low coherence items. Furthermore, the existence of a third response option allowed us to dissociate random responding to ambiguous items from responses that are equally driven by shared and unique features. Second, we used symmetrical blend levels to facilitate comparison across conditions and face blends and test whether asymmetry of false alarms (increased false alarms from low coherence faces to high coherence faces) replicates even then the stimulus structure itself is symmetric. Specifically, high coherence stimuli were created as 60/40 blends of shared and unique parent faces and the low coherence stimuli were created as 40/60 blends of shared and unique parent faces. Category set size was fixed at 3 faces per category, created by blending a single category-specific shared parent face with three different unique faces that were also category specific (Figure 2.1). Parent faces did not overlap across categories, making both shared and unique faces category diagnostic. Half of participants were trained on high coherence (60/40) stimuli and half on low coherence (40/60) stimuli to test how coherence affects category learning, generalization and recognition.

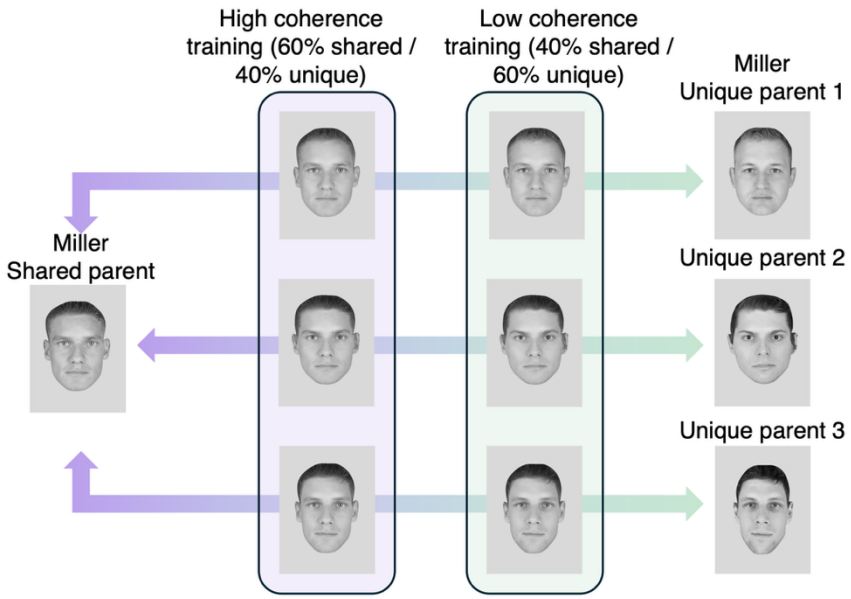

Figure 2.1. Training stimuli (faces in the middle boxes) were made by blending a never-shown shared parent face (left side) with a set of never-shown unique parent faces (right side).

High coherence training stimuli (purple box) were created as blends of 60% shared parent and 40% unique parent (60/40 blends), while low coherence training stimuli (green box) were created as blends of 40% shared parent and 60% unique parent (40/60 blends). During training, participants were trained only on high coherence training stimuli or only on low coherence training stimuli. All participants were exposed to both the 60/40 blends and 40/60 blends during subsequent tests.

Methods

Participants

There was a total of 173 participants (123 females) recruited from the University of Oregon, aged 18-30. The sample size was based on the study by Bowman and Zeithamova (2025), which showed significant effects for similar analyses with a smaller per-condition sample size. Participants were randomly assigned to either the high (n = 93) or low (n = 80) coherence condition. After excluding 6 participants that performed worse than chance (33%) during training, there were 91 participants in the high coherence group and 76 participants in the low coherence group. The mean age of the overall sample was 19.09 years (SD = 1.39). Participants received course credit for their participation and provided written informed consent. All experimental procedures were approved by the Institutional Review Board at the University of Oregon.

Materials

Stimuli were face images created by blending two "parent" faces using Abrosoft FantaMorph (https://www.fantamorph.com/) via a similar procedure used in previous studies (S. R. Ashby et al., 2020a; S. R. Ashby & Zeithamova, 2022; Bowman et al., 2021). Unlike our prior work that used a mixture of face stimuli from several databases, the parent faces for the current study were taken from an online repository used in (Morrison et al., 2017). Given that race and gender are pre-existing categories, all parent faces were from a single race and a single gender.

As in prior studies, we used male faces and the race was white given previous work showing that white Americans exhibit greater own-race effects in face recognition than non-white Americans (Herzmann et al., 2011; Meissner & Brigham, 2001) and because our sample was expected to be a majority white given the University of Oregon population demographics (64% white).

Training sets. The training set for both high and low coherence conditions comprised a total of 9 training faces. Each training face belonged to one of three categories (Miller, Wilson, or Smith families), with 3 training faces per category. Each training stimulus was created by blending one shared parent face and one unique parent face. For example, the three Miller training faces were created as blends of a Shared Miller parent face and Unique Miller 1, Unique Miller 2, and Unique Miller 3 (Figure 1). Unique parent faces did not overlap across categories, so the Wilson training faces would be created by blending a Shared Wilson face with Unique Wilson 1, Unique Wilson 2, and Unique Wilson 3. All the blended faces were created from a pool of 12 parent faces normed for perceived similarity, as tested in a separate pilot sample. The assignment of parent faces to categories and roles (shared vs. unique) was counterbalanced across participants to minimize item effects. Participants were randomly assigned to one of four pre-generated category sets.

The blend ratio of shared and unique parent face in the training set differed based on the assigned condition (high coherence training, low coherence training). In the high coherence training condition, all training faces were blended using 60% shared parent and 40% unique parent (referred to as 60/40 blend, or 60% shared parent face blends). In the low coherence condition, all training faces were blends of 40% of the shared parent and 60% a unique parent (referred to as 40/60 blend, or 40% shared parent face). Because unique parents were also associated with only one category, one can in principle learn the categories not only by learning

to associate the shared parent faces with their category labels (S. R. Ashby et al., 2020a; S. R. Ashby & Zeithamova, 2022; Bowman et al., 2021), but also by learning to associate the unique parent faces with their category labels (Bowman & Zeithamova, 2025).

		Category A training parent				Category B training parent				Category C training parent			
Test stimulus	Item type	S1	U1	U2	U3	S2	U4	U5	U6	S3	U7	U8	U9
A1	Old, unambiguous	60	40	0	0	0	0	0	0	0	0	0	0
A2		60	0	40	0	0	0	0	0	0	0	0	0
A3		60	0	0	40	0	0	0	0	0	0	0	0
B1		0	0	0	0	60	40	0	0	0	0	0	0
B2		0	0	0	0	60	0	40	0	0	0	0	0
B3		0	0	0	0	60	0	0	40	0	0	0	0
C1		0	0	0	0	0	0	0	0	60	40	0	0
C2		0	0	0	0	0	0	0	0	60	0	40	0
C3		0	0	0	0	0	0	0	0	60	0	0	40
T1	New, unambiguous	40	60	0	0	0	0	0	0	0	0	0	0
T2		40	0	60	0	0	0	0	0	0	0	0	0
T3		40	0	0	60	0	0	0	0	0	0	0	0
T4		0	0	0	0	40	60	0	0	0	0	0	0
T5		0	0	0	0	40	0	60	0	0	0	0	0
T6		0	0	0	0	40	0	0	60	0	0	0	0
T7		0	0	0	0	0	0	0	0	40	60	0	0
T8		0	0	0	0	0	0	0	0	40	0	60	0
T9		0	0	0	0	0	0	0	0	40	0	0	60
T10	New, ambiguous	60	0	0	0	0	40	0	0	0	0	0	0
T11		60	0	0	0	0	0	40	0	0	0	0	0
T12		60	0	0	0	0	0	0	40	0	0	0	0
T13		40	0	0	0	0	60	0	0	0	0	0	0
T14		40	0	0	0	0	0	60	0	0	0	0	0
T15		40	0	0	0	0	0	0	60	0	0	0	0
T16		0	0	0	0	60	0	0	0	0	40	0	0
T17		0	0	0	0	60	0	0	0	0	0	40	0
T18		0	0	0	0	60	0	0	0	0	0	0	40
T19		0	0	0	0	40	0	0	0	0	60	0	0
T20		0	0	0	0	40	0	0	0	0	0	60	0
T21		0	0	0	0	40	0	0	0	0	0	0	60
T22		0	40	0	0	0	0	0	0	60	0	0	0
T23		0	0	40	0	0	0	0	0	60	0	0	0
T24		0	0	0	40	0	0	0	0	60	0	0	0
T25		0	60	0	0	0	0	0	0	40	0	0	0
T26		0	0	60	0	0	0	0	0	40	0	0	0
T27		0	0	0	60	0	0	0	0	40	0	0	0
T28	90/10 blends	90	0	0	0	10	0	0	0	0	0	0	0
T29		10	0	0	0	90	0	0	0	0	0	0	0
T30		0	0	0	0	90	0	0	0	10	0	0	0
T31		90	0	0	0	0	0	0	0	10	0	0	0
T32		10	0	0	0	0	0	0	0	90	0	0	0
T33		0	0	0	0	10	0	0	0	90	0	0	0

Table 2.1. Example structure of test stimuli.

This example test stimulus set is based on high training coherence, reflected in the "old" label for the unambiguous 60/40 blends and "new" label for the unambiguous 40/60 blends. Participants in the low training coherence condition were tested using the same stimuli but "old, unambiguous" and "new, unambiguous" were swapped. S = shared parent; U = unique parent; A = Category A training example; B = Category B training example; C = Category C training example; T =transfer stimulus.

Recognition and categorization test sets. The structure of test stimuli is presented in Table 1. The same test stimuli were used in both conditions, and for both recognition and categorization. In addition to the training stimuli, new generalization stimuli were included whose category membership was either unambiguous (clearly belonging to one of the trained families) or ambiguous. Unambiguous stimuli included the 9 training faces as well as 9 blends that only differed from training faces by blend level. For example, for the high coherence condition, unambiguous test stimuli included the nine 60/40 blends used for training and nine 40/60 blends of the same parent faces, which served as training faces in a yoked low coherence condition (Figure 2.1). These stimuli were unambiguous because both parent faces (e.g., shared Miller face and Unique Miller face 1) were diagnostic of the same category.

To test the degree to which participants generalize to category center (prototype), we also included 90/10 blends (Fig. 2). The 90/10 blends were created from two shared parents (e.g., 90% Shared Wilson parent and 10% Shared Miller parent) but are considered unambiguous due to one shared parent being dominating in these blends (Figure 2.2A). We chose to use 90/10 blends instead of directly using the original shared parent face to control for any visual effects created by the blending process itself. All 6 possible 90/10 blends of the 3 shared parent faces were included in the test set.

Ambiguous test stimuli comprised of 18 novel faces that were a blend of the shared parent from one category and the unique parent from a *different* category (Figure 2.2B). For example, the shared Miller parent face might be blended with a unique face from the Wilson family, making this novel blend's category membership ambiguous. Specifically, the Miller shared parent face was blended with each Wilson unique face at both blend levels (60 Miller/40 Wilson and 40 Miller/60 Wilson), the Wilson shared parent face was blended with each Smith

unique face at both blend levels (60 Wilson/40 Smith and 40 Wilson/60 Smith), and the Smith shared parent face was blended with each Miller unique face at both blend levels (60 Smith/40 Miller, 40 Smith/60 Miller). The inclusion of ambiguous test stimuli allowed us to determine the extent to which people categorized faces relying on shared or unique parent features at both blend levels.

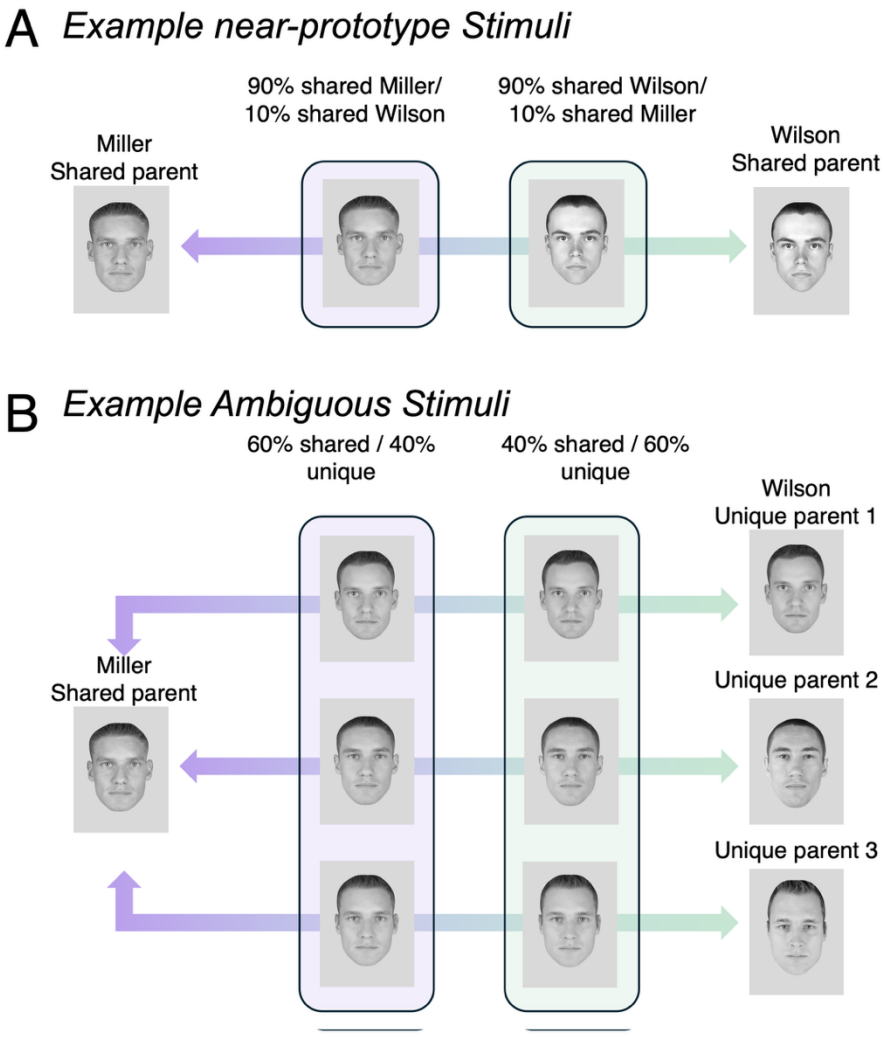

Figure 2.2. Example near-prototype and ambiguous stimuli.

Novel stimuli included in recognition and categorization tests. Non-ambiguous 90/10 (near-prototype) stimuli (A) and ambiguous stimuli (B) (faces in the rectangles) were made by blending a never-shown shared parent face (left side) with a set of never-shown unique parent faces (right side). (A) Non-ambiguous near-prototype stimuli were created as blends of 90% one shared parent and 10% another shared parent (90/10 blend). (B) High coherence

ambiguous stimuli were created as blends of 60% shared parent and 40% unique parent, while low coherence ambiguous stimuli were created as blends of 40% shared parent and 60% unique parent.

Procedure

Training. Participants first underwent feedback-based training (Figure 2.3, left), learning to categorize the 9 training faces into one of three family categories: the Millers, the Wilsons, or the Smiths. For each trial, a training stimulus was presented with three response options underneath corresponding to the three family categories. Timing was self-paced. A response was pressing either 7, 8, or 9 on the keyboard, with most participants using index, middle and ring finger of their right hand to make the responses. Following a categorization response was corrective feedback ("Correct!" or "Incorrect.") presented for 1.5 seconds followed by a 0.5 second fixation cross. There was a total of 72 training trials, distributed across 8 training blocks. In each block, participants were shown all 9 training faces 1 time each. Presentation order was randomized within each repetition.

Recognition test. After training, participants first completed a recognition test (Figure 2.3, middle). The recognition test featured a single block of 42 trials belonging to one of 5 trial types: 6 unambiguous 90/10 blend, 9 unambiguous 60/40 blend, 9 unambiguous 40/60 blend, 9 ambiguous 60/40 blend, and 9 ambiguous 40/60 blend. On each trial, participants were prompted to choose whether the face was old or new by pressing J or K on the keyboard, respectively. For high coherence group, the 60/40 unambiguous blends were 'old' and all other were 'new', while for the low coherence group, the 40/60 unambiguous blends were 'old' and all other were 'new'. We were primarily interested in the pattern of endorsements for the different stimulus types to determine the type of representations participants are relying on for their judgment. The timing

was self-paced and after a response was made, a 0.5 second fixation cross was shown prior to beginning the next trial.

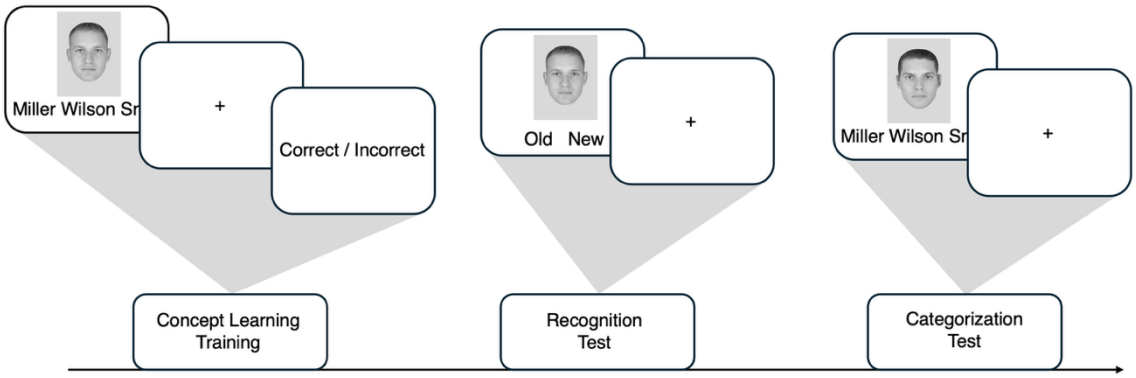

Figure 2.3. Example experimental design.

During training, participants underwent a concept learning paradigm (left side), where they viewed blended faces and categorized them into one of three families. Following a response, participants received corrective feedback. Participants in the high coherence training group only saw faces blended with 60% shared parent face and 40% unique parent face from the same family (60/40 blend) while those in the low coherence training group only saw faces blended with 40% shared parent face and 60% unique parent face from the same family (40/60 blend). Following training, participants were administered a recognition test (middle of the figure), where they were presented with old and new faces and prompted to choose whether they had seen the face during training ("old") or not ("new"). Following recognition, participants completed a categorization test, where they categorized training and new faces without feedback. Same stimuli were used in recognition and categorization. The category membership of test stimuli was either unambiguous or ambiguous. Unambiguous stimuli were the training stimuli used for both groups. Because participants only saw either 60/40 blends or 40/60 blends during training, half of the unambiguous stimuli were old stimuli from training and the other half were new but very similar stimuli (Figure 1). Ambiguous stimuli were blends of a shared parent from one category (e.g., Miller shared parent) and a unique parent from another category (e.g., Wilson unique parent). Also included in the test sets were 90/10 blends to test categorization and recognition of faces near category center.

Categorization test. After recognition, participants completed a categorization test (Figure 2.3, right), which presented participants in a random order the same 24 unambiguous and 18 ambiguous blend stimuli as seen in recognition. On each trial, participants were shown a face and prompted to choose which of 3 families the face belonged to. Timing was self-paced and there was no corrective feedback. Following a response, a 0.5 second fixation cross preceded beginning the next trial.

Statistical analyses

Training. For training, we conducted a 2 (high, low coherence) x 8 (1-8 blocks) repeated measures ANOVA to evaluate whether learning occurred overall and whether learning curves differed between groups.

Categorization test, unambiguous faces. Categorization of unambiguous faces at different blend levels allowed us to evaluate generalization gradients (i.e., increasing accuracy with increasing shared parent percentage), a signature of category learning of similarity-based categories. To test for differences in unambiguous face categorization following training on high versus low coherence stimuli, we ran a 3 (90/10, 60/40, 40/60 test blend) x 2 (high, low training coherence) ANOVA.

Categorization test, ambiguous faces. Ambiguous faces did not have defined "correct" category membership assignments as they were a blend of a shared parent from one category and a unique parent from a different category. Instead, we calculated for each trial type (ambiguous 60/40, ambiguous 40/60) the proportions of responses that were aligned with the membership of the shared parent, membership of the unique parent, or neither. We then used the proportion of categorization responses that do not align with either parent (e.g., responding Smith to Wilson/Miller blend) to estimate guessing-corrected proportion of responses aligned with the shared parent (e.g., responding Wilson to Shared Wilson/Unique Miller blend) and guessing-corrected proportion of responses aligned with the unique parent (e.g., responding Miller to Shared Wilson/Unique Miller blend), separately for each test blend level (60/40, 40/60). For example, the guessing-corrected rate of shared parent response for 60/40 test blends was calculated as the proportion of 60/40 trial responses aligned with the shared parent minus the proportion of 60/40 trial responses not aligned with either shared nor unique parent. To evaluate how training coherence and test blend level affected proportion of response aligned with shared

and unique features, we ran a 2 (high, low training coherence) x 2 (60/40, 40/60 ambiguous test blend) x 2 (shared, unique response type) ANOVA. Additionally, single sample t-tests were used to compare corrected responses proportions against 0 within each condition, to test whether participants were reliably using shared and unique features more than guessing.

Recognition test. Overall recognition was calculated as corrected hit rate, which is the proportion of 'old' responses to old stimuli minus the proportion of 'old' responses to new stimuli. Of main interest was to compare tendencies to respond old to stimuli at different blend levels as a function of training coherence condition. This was done by running two repeated measures ANOVAs, separately for unambiguous and ambiguous faces. The unambiguous model predicted proportions of old responses from shared parent blend level (90%, 60%, 40%) and training coherence. The ambiguous model predicted proportions of old responses from shared parent blend level (60%, 40%) and training coherence. We ran follow-up t-tests and repeated measures ANOVAs when indicated.

All code and materials required to recreate statistical analyses and visualizations from this paper can be found at https://osf.io/74mk9. This study was not preregistered.

Results

Training performance

Overall, participants in both groups performed well above chance (high coherence M = 77%, SD = 13%; low coherence M = 56%, SD = 11%; chance = 33%). As expected, the block x group ANOVA showed strong main effects of both block ($F(7,1155) = 111.03$, $\eta_p^2 = 0.40$, $p < 0.001$) and coherence condition ($F(1,165) = 128.00$, $\eta_p^2 = 0.44$, $p < 0.001$), indicating that participants learned over time and that the high coherence group performed better than the low coherence group (Figure 2.4). There was also a significant block x group interaction ($F(7,1155) = 7.35$, η_p^2

= 0.04, p < 0.001), denoting steeper learning curves in the high coherence group. Overall, high coherence training led to faster learning and higher asymptote than low coherence training.

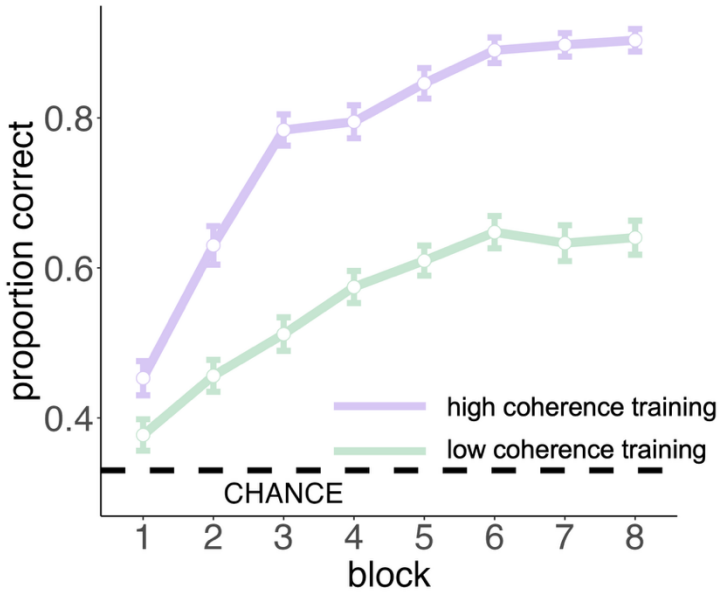

Figure 2.4. Categorization training accuracy.

Training accuracy for each block for high (purple) and low (green) coherence training groups. Each block contained three repetitions of training stimuli. White dots are average proportion correct per block across subjects. Error bars depict the standard error of the mean across subjects. The black dashed line denotes chance level performance (1 / 3 options = 0.33).

Categorization performance

Categorization accuracy for unambiguous trials. Ambiguous and unambiguous trials were analyzed separately as ambiguous faces do not have an objectively accurate response. Results of the unambiguous categorization are presented in Figure 2.5. A mixed design ANOVA with blend level (90%, 60%, 40% shared parent) x training coherence (high, low) as factors predicting average categorization accuracy for unambiguous faces showed a main effect of blend level ($F(2,330) = 16.59$, $\eta_p^2 = 0.09$, $p < 0.001$), a main effect of coherence ($F(1,165) = 28.2$, $\eta_p^2 = 0.15$, $p < 0.001$), as well as a significant blend x coherence interaction ($F(2,330) = 3.33$, $\eta_p^2 = 0.04$, $p = 0.002$). The main effect of coherence was driven by overall better performance by the

high coherence condition (M (high – low) = 0.15, t(165) = 5.31, $p < 0.001$). Follow-up pair-wise comparisons within each blend level indicate the high coherence group outperformed the low coherence group at every blend level, including 40/60 blend level, which were the training stimuli of the low coherence group (Figure 2.5). This means that it was easier for the high coherence group to learn the 60/40 blends and then generalize their knowledge to 40/60 blends than it was for the low coherence group to directly learn the category labels for the 40/60 blends.

The main effect of blend level was driven by overall greater accuracy for 90/10 (75%) and 60/40 blends (74%) compared to 40/60 blends (68%, both ts > 4.46, $p < .001$) but was qualified by an interaction. The blend x coherence interaction indicates that blend level effect was disproportionately stronger in the high than low coherence group (Figure 2.5). Specifically, the high coherence group performed marginally or significantly better with each increase in shared parent percentage ($F(2,180) = 24.6$, $\eta_p^2 = 0.22$, $p < 0.001$; M (90% - 60%) = 0.04, $t(165) = 2.39$, $p = 0.057$; M (60% - 40%) = 0.09, $t(165) = 4.56$, $p < 0.001$). In contrast, the effect of blend level was not significant in the low coherence group level ($F(2,150) = 1.79$, $\eta_p^2 = 0.02$, $p = 0.171$). Interestingly, post-hoc tests indicated that 60/40 blends were easier to categorize than 40/60 blends even in the low coherence group ($t(75) = 2.22$, $p_{uncorrected} = 0.029$, $p_{bonferroni} = 0.088$).

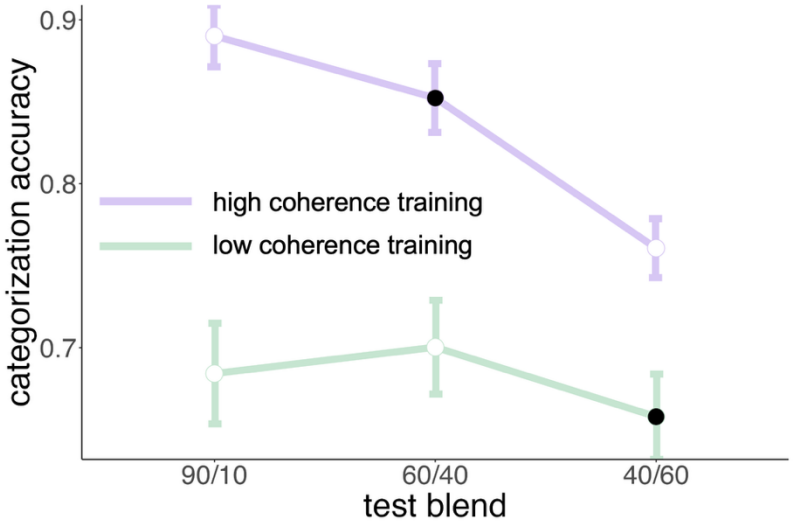

Figure 2.5. Unambiguous categorization performance.

Categorization accuracy for unambiguous test items, separated by training coherence group (purple = high coherence training, green = low coherence training). Dots denote the average proportion of correct responses across subjects. Error bars denote the standard error of the mean across subjects. Black dots indicate training faces for each respective training condition.

Categorization of ambiguous blends. Figure 2.6A visualizes the proportion of responses consistent with relying on shared or unique features, as well as the rate at which participants relied on neither, reflecting guessing. Prior work (Bowman & Zeithamova, 2025a) showed bias towards shared parent features after high coherence training. However, with only two categories, it was challenging to establish whether lack of bias towards shared vs. unique features in some conditions reflected random responses or participants actually favoring shared vs. unique faces equally. The existence of three categories allowed us to test both for shared bias as well as determine the degree to which participants are using unique parent features for classification on trials when they are not using the shared parent features.

Figure 2.6B illustrates the proportion of responses consistent with shared vs. unique parent membership after correcting for guessing (subtracting proportion of responses consistent

with the third, unrelated category within each blend level). First, we tested whether participants were making responses consistent with shared and unique features above chance, using one-sample t-tests comparing the corrected proportions to 0. The proportion of responses consistent with shared features was well above chance for both 60/40 and 40/60 blends in both high-coherence training and low-coherence training groups (all $t > 7$, $p < 0.001$). The proportion of responses consistent with unique features was also above chance in all conditions (all $t > 2.38$, all $p < 0.05$). Thus, in all conditions, there was evidence that participants' responses were sensitive to both shared and unique features to some degree.

To formally evaluate the degree of bias towards shared vs. unique features and how such bias depended on training coherence and blend level, we ran a 2x2x2 mixed ANOVA with training coherence (high, low) as a between-subject factor, and test stimulus blend level (60/40, 40/60) and response type proportion (corrected shared, corrected unique) as within-subject factors. The full results are presented in Table 2.2. We found a main effect of training coherence ($F(1,165) = 22.3$, $\eta_p^2 = 0.12$, $p < 0.001$; M (high – low) = 0.12, $t(165) = 4.73$, $p < 0.001$), indicating more non-guessing responses (shared and/or unique) after high-coherence than low-coherence training. We also found a main effect of blend level ($F(1,165) = 14.68$, $\eta_p^2 = 0.08$, $p < 0.001$), driven by a greater proportion of non-guessing responses (shared and/or unique) to 60/40 blends than 40/60 blends ($M_{difference} = 0.07$, $t(165) = 3.83$, $p < 0.001$; Figure 2.6B).

Of main interest were the effect of response type and any interactions with response type. The main effect of response type was significant and strong ($F(1,165) = 288.09$, $\eta_p^2 = 0.64$, $p < 0.001$), with responses much more likely aligning with the shared parent category than unique parent category ($M_{difference} = 0.35$, $t(165) = 17.0$, $p < 0.001$; Figure 2.6B). Response type also strongly and significantly interacted with blend level ($F(1,165) = 144.17$, $\eta_p^2 = 0.47$, $p < 0.001$),

driven by disproportionately greater shared parent reliance when test stimuli had greater proportion of the shared parent (corrected proportion of shared parent response for 60/40 blend M = 0.57, for 40/60 blend = 0.33). Interestingly, post-hoc tests indicated greater proportion of responses reflecting more reliance on shared parent than unique parent even when unique parent was more prominent in the test stimulus (60/40 blends: M (shared-unique) = 0.51, $t(165) = 18.9$, $p < 0.001$; 40/60 blends: M (shared-unique) = 0.18, $t(165) = 8.26$, $p < 0.001$).

Response type also significantly interacted with training coherence ($F(1,165) = 55.12$, $\eta_p^2 = 0.25$, $p < 0.001$), with greater bias towards shared than unique responses for high coherence group (M (shared – unique) = 0.50, $t(165) = 18.08$, $p < 0.001$) than low coherence group (M (shared – unique) = 0.20, $t(165) = 6.47$, $p < .001$). Interestingly, even low coherence group showed reliable bias towards shared over unique features. The three-way interaction did not reach significance. The marginal trend reflected that the bias towards shared features varied across training conditions and test stimulus types (Figure 2.6B). Pairwise comparison of shared vs. unique responses within each training and test condition showed a significant bias towards shared feature in all conditions (all $t > 8$, $p < 0.001$) except when low coherence training group categorized 40/60 ambiguous blends, where the numerical bias towards shared features did not reach significance ($t(75) = 1.74$, $p = .086$). Nevertheless, the direction of the effect in 40/60 ambiguous blends indicates that people relied more on shared features *even when both the training stimuli and the test stimulus were comprised of more unique features.*

Finally, a planned comparison tested whether the reliance on unique features increased following low coherence training. Contrary to our predictions, the corrected proportion of responses that aligned with unique features was not reliably different between low and high coherence training groups for either 60/40 or 40/60 test stimulus (both $t < 1.2$, $p > .26$). Thus,

low coherence training emphasizing unique features led to overall more random responding rather than a generalization strategy relying on unique features.

Predictor	F	*df*	*p*	η_p^2
Blend level	14.68	1,165	< 0.001	0.08
Response type	288.09	1,165	< 0.001	0.64
Training coherence	22.3	1,165	< 0.001	0.12
Blend level x Response type	144.17	1,165	< 0.001	0.47
Blend level x training coherence	1.82	1,165	0.179	0.01
Response type x training coherence	55.12	1,165	< 0.001	0.25
Blend level x response type x training coherence	3.01	1,165	0.084	0.02

Table 2.2. Mixed ANOVA predicting the proportion of responses consistent with shread vs. unique feature reliance.

2x2x2 mixed ANOVA results predicting shared vs. unique feature reliance when categorizing ambiguous face, corrected for guessing rate.

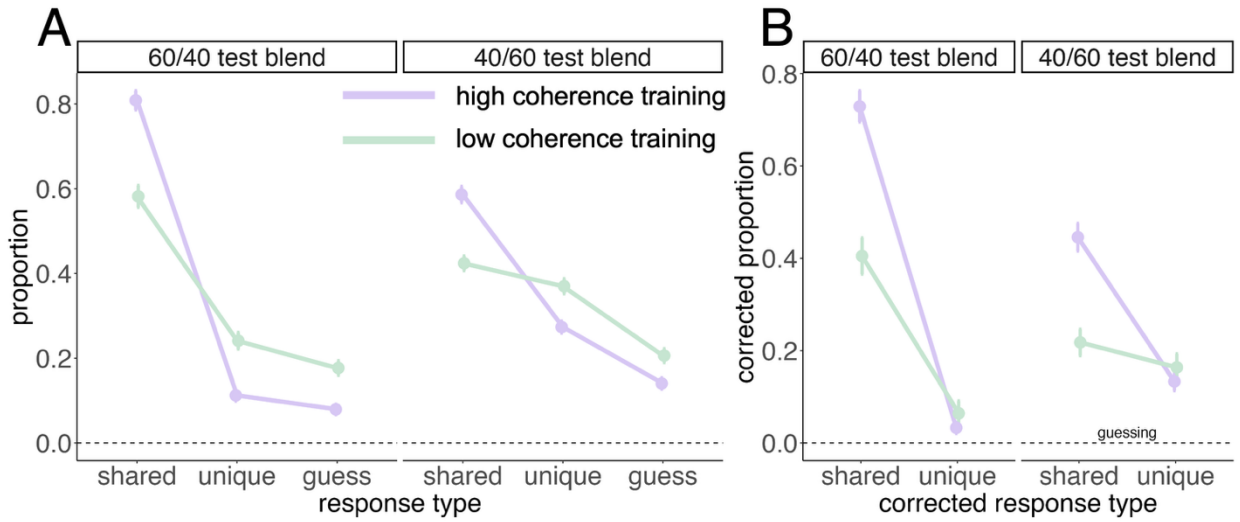

Figure 2.6. Categorization responses to ambiguous trials.

Categorization responses to ambiguous stimuli, separated by coherence training group and blend level of the test items. (A) Mean proportion of responses consistent with shared parent ("shared"), unique parent ("unique"), or neither ("guess"). For example, if the test stimulus was a blend of the Shared Miller parent with a Unique Wilson parent, response Miller would be coded as shared, response Wilson would be coded as unique, and response Smith would be coded as neither/guess. Left side of panel A shows responses to 60% shared/40% unique (60/40) ambiguous test blends. Right side of panel B shows responses to 40% shared/60% unique (40/60) ambiguous test blends. (B) Proportion of responses that were consistent with reliance on shared versus unique parent features after correcting for guessing rate. Correction was implemented by subtracting the proportion of guessing from shared and unique response proportions, separately for 60/40 and 40/60 blends. Error bars depict the standard error of the mean across subjects.

Recognition performance

Endorsement rates for all unambiguous and ambiguous trial types are presented in Figure 2.7. Overall, participants were able to successfully recognize their training stimuli (corrected hit rate M = 0.33 (SD = 0.19), one sample t-test against 0, $p < 0.001$). Both groups performed better than chance when analyzed separately (high coherence: M = 0.40, $p < 0.001$; low coherence: M = 0.25, $p < 0.001$), with the high coherence group achieving significantly better recognition scores than the low coherence group ($t(165) = 5.31$, $p < 0.001$).

Endorsement of unambiguous faces. Our prior study (Bowman & Zeithamova, 2025a) using a similar paradigm with 50/50 blends for high coherence training and 30/70 blends for low coherence training found two effects that we were interested to replicate and extend here. First, we found a high degree of false alarms to prototypes—a classic signature of category learning in many tasks--in the high coherence condition. Second, we found an asymmetric pattern of false alarms, where participants in the 50/50 training condition false alarmed to 30/70 blends much less often than participants in the 30/70 condition false alarmed to 50/50 blends. Here, we replicated both findings despite using symmetric blend levels (60/40 vs. 40/60) (Figure 2.7).

A repeated measures ANOVA with test stimulus blend level (90/10, 60/40, 40/60) as a within-subject factor and training coherence (high, low) as a between-subjects factor was used to assess recognition memory of unambiguous faces. Both main effects of blend level ($F(2,330) = 121.00$, $\eta_p^2 = 0.42$, $p < 0.001$) and training coherence ($F(1,165) = 5.68$, $\eta_p^2 = 0.03$, $p = 0.018$), as well as the blend level x coherence interaction ($F(2,330) = 183.00$, $\eta_p^2 = 0.53$, $p < 0.001$), were significant predictors of old endorsement rates. The main effect of blend level was driven by a greater tendency to endorse 60/40 blends as "old" than either 90/10 or 40/60 blends (M (90 - 60) = -0.27, $t(165) = -14.43$, $p < 0.001$; M (60 - 40) = 0.25, $t(165) = 14.99$, $p < 0.001$). The main

effect of training coherence was driven by an overall greater tendency of the high coherence group to endorse the unambiguous stimuli as "old" (M (high – low) = 0.05, $t(165) = 2.38$, $p = 0.018$). Importantly, these main effects were qualified by a significant blend level x coherence interaction that was driven by similar responding to 60/40 blends by both groups but different patterns of responding to both 90/10 blends and 40/60 blends. For 90/10 blends, the high coherence group showed much higher levels of false alarms than did the low coherence group (M(high – low) = 0.42, $t(165) = 11.30$, $p < 0.001$). The 60/40 blends were endorsed by both groups about equally (M(high – low) = 0.06, $t(165) = 2.30$, $p = 0.345$), even though they were actually old for the high coherence group but not for the low coherence group. For 40/60 blends, the low coherence group were much more likely to endorse them as "old" (M(high – low) = -0.33, $t(165) = -12.47$, $p < 0.001$), in line with these stimuli being actually old for the low coherence group.

Notably, despite the symmetrical blending pattern between high and low coherence stimuli in the current study, we replicated the asymmetric pattern of false alarms observed in our prior study. The high coherence group trained on 60/40 blends showed relatively low levels of false alarms to 40/60 blend; the low coherence group trained on 40/60 blends false alarmed to 60/40 blends much more often. Moreover, from Figure 2.7, we noted that the low coherence group endorsed 60/40 unambiguous faces more than 40/60 unambiguous faces which were their actual training faces (M (60 - 40) = 0.05, $t(75) = 2.10$, $p = 0.039$). This is particularly noteworthy because it means that the people trained with low coherence faces reported having seen the high coherence faces more than they reported seeing the actual faces they were trained with.

Endorsement of ambiguous faces. The endorsement of ambiguous faces is presented in Figure 2.7, right panel. A 2x2 mixed ANOVA with test stimulus blend level (60/40, 40/60) as a

within-subject factor and training coherence (high, low) as a between-subject factor, revealed a main effect of blend level ($F(1,165) = 195.00$, $\eta_p^2 = 0.54$, $p < 0.001$), no main effect of training coherence ($F(1,165) = 0.16$, $\eta_p^2 = 0.00$, $p = 0.686$), and a significant blend level x training coherence interaction ($F(1,165) = 136.00$, $\eta_p^2 = 0.45$, $p < 0.001$). The main effect of blend level was driven by greater endorsement of 60/40 ambiguous blends compared to 40/60 ambiguous blends. The blend level x coherence interaction was driven by the high coherence group's increased sensitivity to the percentage of shared parent face in ambiguous stimuli. Specifically, high coherence group endorsed 60/40 blends much more than 40/60 blends (M (60 – 40) = 41%) while the low coherence group was much less sensitive to the shared parent blend level (M (60 – 40) = 4%). Thus, the high coherence group showed increased false alarms to faces with greater proportion of shared features and lower false alarms to faces with greater proportion of unique features.

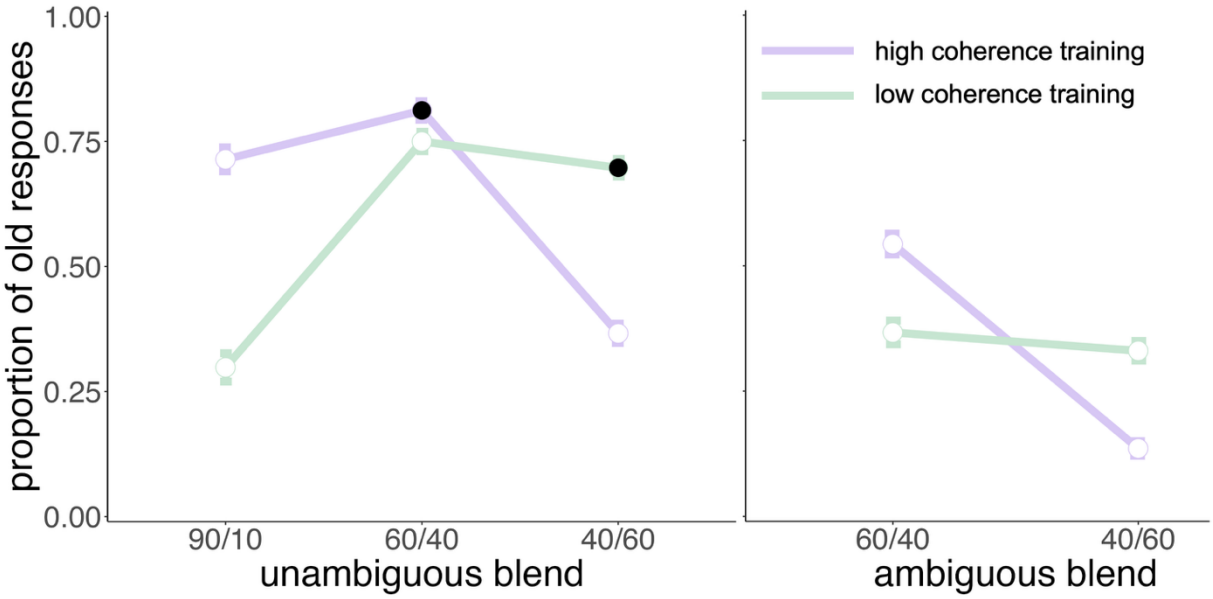

Figure 2.7. Recognition memory performance.

Proportion of "old" responses during the recognition test, separated by coherence training group, shared parent blend level, and ambiguity of the test item. Note that the actual old items for the high coherence training were unambiguous and 60% shared parent, while the actual old items for the low coherence training group were unambiguous and 40% shared parent. All other test items were new. Dots represent the average proportions across

subjects. Error bars denote the standard error of the mean across subjects. Black dots indicate training faces for each respective training condition.

Discussion

In the current study, we evaluated how the variability of training stimuli, operationalized as stimulus typicality or coherence, affects concept learning, recognition, and generalization. Extending prior work, we designed the category structure to further evaluate the nature of memory representations formed after training with low coherence exemplars. Participants learned to categorize a set of faces created as blends of a category shared face with an exemplar-unique face that was also perfectly predictive of category membership. High coherence training utilized faces with 60% shared and 40% unique face, while low coherence training utilized faces with 40% shared and 60% unique face blends. There was a clear advantage of high coherence training for learning category memberships of the training stimuli. Moreover, high coherence training facilitated more abstraction of commonalities across training stimuli, evidenced by their generalization gradient over novel stimuli at different distances from category prototypes and their high endorsement of category prototypes during recognition. Low coherence training led to worse learning, worse categorization test performance even for the training stimuli themselves, and fewer false alarms to prototypes. Nevertheless, we found novel evidence that shared features were also extracted by the low coherence group, even though they were less prominent in the training stimuli than unique features, and that memory representations appeared to shift towards the shared faces in both groups. Contrary to our expectation, we found little evidence that low coherence training would lead to better representation of unique features in memory; instead, low coherence training led to worse learning and higher guessing rates. These results replicate prior findings of the benefits of high coherence/low variability training and further elucidate the nature of memory representations formed when training exemplars are highly variable.

Early studies on categorization following concept learning indicated a benefit of learning from relatively variable category instances, presumably because they equip the learner with an understanding of the breadth of the category (Fried & Holyoak, 1984; Homa & Vosburgh, 1976; Posner & Keele, 1968). Importantly, however, these studies trained participants to a performance criterion prior to testing category knowledge. Because high coherence training facilitates much faster learning, the benefits observed for low coherence conditions in the early studies could be attributed to longer training instead of better abstraction of common features across stimuli (Hintzman, 1984).This suspicion has been tested and confirmed using binary dimension cartoon stimuli (Bowman & Zeithamova, 2020, 2023) and was subsequently replicated using the dot-pattern distortion paradigm (Hu & Nosofsky, 2024) and face-blend stimuli (Bowman & Zeithamova, 2025a).

Here, we used similar stimuli and category structure to Bowman and Zeithamova (2025) and replicated the finding that when training duration is equated, high coherence training facilitates not only faster learning but better generalization. Remarkably, people trained on high coherence stimuli were substantially better at categorizing low coherence training stimuli *than the low coherence group that was trained on them directly*. In other words, participants were more successful at categorizing less typical, low coherence stimuli when they were trained on high coherence stimuli and then asked to generalize to low coherence stimuli than when they studied the low coherence stimuli directly. This was observed even though the low coherence stimuli contained exemplar-unique features that were also perfectly diagnostic of a given category rather than simply adding noise. These findings run contrary to recent proposals (Raviv et al., 2022) and instead indicate that training on more coherent exemplars leads to a fast and efficient formation of category representations that support generalization.

Our study is a conceptual replication of Bowman and Zeithamova (2025) but extends it in several ways. First, incorporating a third category and moving chance performance from 50% to 33% allowed us to better evaluate how successful participants were in learning both shared and unique category features and further highlight the benefit of high coherence training during learning and test. Moreover, for ambiguous stimuli that were blends of a shared parent from one category and a unique parent from another category, it was challenging with just two categories to differentiate equal attention to shared and unique features from random responding. Here, including a third category as a baseline, we demonstrate that people use multiple types of stimulus features to categorize, as they responded by relying on shared and unique features reliably more often than responding randomly. While both groups used shared and unique features, they differed in the extent to which they relied on either. Specifically, the high coherence group used shared features more often than the low coherence group and the low coherence group guessed more often than the high coherence group. Both groups relied on unique features roughly an equal amount of time, indicating that the decreased reliance on shared features by the low coherence group is *not* driven by more unique feature reliance but rather a higher guessing rate. Ergo, the coherence, or typicality, of training stimuli is an important factor producing differences in categorization.

Interestingly, while the results showed that both groups can use unique features, we also found evidence that people trained with low coherence stimuli abstracted out commonalities across stimuli and emphasized them in memory, albeit to a lesser degree than those trained with high coherence stimuli. First, when categorizing unambiguous stimuli, people trained on low coherence stimuli were numerically *better* at categorizing 60/40 blends than their own (40/60) training stimuli. Second, for ambiguous test stimuli, the low coherence group remained biased

toward choosing the shared parent's category over unique parent's category. Finally, people trained on low coherence stimuli falsely recognized 60/40 blends more often than they correctly recognized their own 40/60 training stimuli. These results suggest that the low coherence training group abstracted out commonalities across training stimuli despite it being a more challenging process compared to the high coherence training group.

The current study also replicated recognition memory results from Bowman and Zeithamova (2025), in that the enhanced categorization performance in the high coherence group did not come at the cost of recognition memory. One interpretation may be in the framework of schema-based memory, where our results align with work indicating that schema-consistent items are better remembered than arbitrary information (Aghayan Golkashani et al., 2021; Amer et al., 2018; Kan et al., 2009; Van Kesteren et al., 2010). The category knowledge acquired better by high coherence training may serve as a scaffold for encoding, organizing, and retrieving detailed information. One line of evidence in support of this notion that abstract category structure scaffolds detailed memory comes from developmental research, where episodic memory in children has been shown to be facilitated by schemas (Bauer et al., 2016; Bauer & Dow, 1994; Cook et al., 2021; Goodman, 1980). Nevertheless, this interpretation would need to be tested in future studies that use recognition foils more directly tailored to this question.

Another way to interpret the recognition findings is that memory representations of both groups shifted towards the shared features. For example, the high coherence group was strongly endorsing both 90/10 and 60/40 stimuli, but not 40/60 stimuli despite their high physical similarity to 60/40 stimuli. The low coherence group was endorsing 60/40 stimuli numerically more often than 40/60 stimuli. This can occur through remembering the training blends to have a higher proportion of shared features than they had and/or attending to the shared features more

strongly than unique features, consistent with work demonstrating increased attention to category-relevant features following category learning (S. R. Ashby & Zeithamova, 2022; Nosofsky & Zaki, 2002; Rehder & Hoffman, 2005a, 2005b; Zaki et al., 2003). Here, we observed a shift towards shared features even though the unique features were also category diagnostic rather than irrelevant, and we observed the shift towards shared features even in the low coherence group, despite their general difficulty to extract them from the training blends.

One motivation for this study was to determine the extent to which low coherence training facilitates a systematic categorization strategy focused more strongly on unique features versus mere random responding. The results were nuanced in this question. First, we found that when the low coherence training group did not rely on common features across training stimuli, they instead relied on unique features to a degree greater than chance. However, the reliance on unique features was about the same as in the high coherence group once we corrected for guessing. Instead, the low coherence group guessed more often than the high coherence group, indicating that low coherence category training leads to a general learning deficit, where neither shared nor unique information diagnostic of category membership is learned. This finding highlights the human bias to seek commonalities among category members and the challenge of utilizing memorization of individual stimulus-label associations when commonalities are less prominent or absent (S. R. Ashby et al., 2023; Bowman & Zeithamova, 2023, 2025a; Shepard et al., 1961).

Conclusion

In the current study, we used naturalistic categories of face stimuli that could be learned via abstraction of common features across members of the same category and/or by learning the unique features of individual category members. We tested how the coherence of training

exemplars affects utilization of shared and unique features, as well as category learning and generalization success. We found that people extracted both shared and unique features to some degree but were largely biased to shared features unless both training and test stimuli emphasized unique features. High coherence training emphasizing shared features promoted better abstraction and utilization of shared features. Low coherence training emphasizing unique features led more likely to a general learning failure rather than increased utilization of unique features. Shared features seemed more emphasized in memory in both groups. This work contributes to our understanding of how the structure of experience shapes what information is extracted from our environment and subsequently used for adaptive decision-making.

References cited

Aghayan Golkashani, H., Leong, R. L. F., Wong, K. F., & Chee, M. W. L. (2021). Schema-driven memory benefits boost transitive inference in older adults. *Psychology and Aging*, *36*(4). https://doi.org/10.1037/pag0000586

Amer, T., Giovanello, K. S., Grady, C. L., & Hasher, L. (2018). Age differences in memory for meaningful and arbitrary associations: A memory retrieval account. *Psychology and Aging*, *33*(1). https://doi.org/10.1037/pag0000220

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 & Review*, *27*(4), 791–800. https://doi.org/10.3758/s13423-020-01754-3

Ashby, S. R., Chaloupka, B., & Zeithamova, D. (2023). Category bias in similarity ratings: the influence of perceptual and strategic biases in similarity judgments of faces. *Frontiers in Cognition*, *2*. https://doi.org/10.3389/fcogn.2023.1270519

Ashby, S. R., & Zeithamova, D. (2022). Category-Biased Neural Representations Form Spontaneously during Learning That Emphasizes Memory for Specific Instances. *Journal of Neuroscience*, *42*(5). https://doi.org/10.1523/JNEUROSCI.1396-21.2021

Bauer, P. J., & Dow, G. A. (1994). Episodic Memory in 16- and 20-Month-Old Children: Specifics Are Generalized but Not Forgotten. *Developmental Psychology*, *30*(3). https://doi.org/10.1037/0012-1649.30.3.403

Bauer, P. J., Stewart, R., White, E. A., & Larkina, M. (2016). A Place for Every Event and Every Event in Its Place: Memory for Locations and Activities by 4-Year-Old Children. *Journal of Cognition and Development*, *17*(2). https://doi.org/10.1080/15248372.2014.959521

Bowman, C. R., de Araujo Sanchez, M. A., Hou, W., Rubin, S., & Zeithamova, D. (2021). Generalization and False Memory in an Acquired Equivalence Paradigm: The Influence of Physical Resemblance Across Related Episodes. *Frontiers in Psychology*, *12*. https://doi.org/10.3389/fpsyg.2021.669481

Bowman, C. R., & Zeithamova, D. (2020). Training Set Coherence and Set Size Effects on Concept Generalization and Recognition. *Journal of Experimental Psychology: Learning Memory and Cognition*. https://doi.org/10.1037/xlm0000824

Bowman, C. R., & Zeithamova, D. (2023). Coherent category training enhances generalization in prototype-based categories. *Journal of Experimental Psychology: Learning, Memory, and Cognition*. https://doi.org/10.1037/xlm0001243

Bowman, C. R., & Zeithamova, D. (2025). High coherence among training exemplars promotes broad generalization of face families. *Journal of Experimental Psychology: Learning, Memory, and Cognition*. https://doi.org/10.1037/xlm0001478

Cook, K. M., You, X., Cherry, J. B., Merchant, J. S., Skapek, M., Powers, M. D., Pugliese, C. E., Kenworthy, L., & Vaidya, C. J. (2021). Neural correlates of schema-dependent episodic memory and association with behavioral flexibility in autism spectrum disorders and typical development. *Journal of Neurodevelopmental Disorders*, *13*(1). https://doi.org/10.1186/s11689-021-09388-9

Fried, L. S., & Holyoak, K. J. (1984). Induction of category distributions: A framework for classification learning. *Journal of Experimental Psychology: Learning, Memory, and Cognition*, *10*(2). https://doi.org/10.1037/0278-7393.10.2.234

Goodman, G. S. (1980). Picture memory: How the action schema affects retention. *Cognitive Psychology*, *12*(4). https://doi.org/10.1016/0010-0285(80)90017-1

Gorman, T. E., & Goldstone, R. L. (2022). An instance-based model account of the benefits of varied practice in visuomotor skill. *Cognitive Psychology*, *137*. https://doi.org/10.1016/j.cogpsych.2022.101491

Hernández-Davo, H., Urbán, T., Sarabia, J. M., Juan-Recio, C., & Javier Moreno, F. (2014). Variable training: effects on velocity and accuracy in the tennis serve. *Journal of Sports Sciences*, *32*(14). https://doi.org/10.1080/02640414.2014.891290

Herzmann, G., Willenbockel, V., Tanaka, J. W., & Curran, T. (2011). The neural correlates of memory encoding and recognition for own-race and other-race faces. *Neuropsychologia*, *49*(11). https://doi.org/10.1016/j.neuropsychologia.2011.07.019

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

Homa, D., & Vosburgh, R. (1976). Category breadth and the abstraction of prototypical information. *Journal of Experimental Psychology: Human Learning and Memory*, *2*(3). https://doi.org/10.1037/0278-7393.2.3.322

Hu, M., & Nosofsky, R. M. (2024). High-variability training does not enhance generalization in the prototype-distortion paradigm. *Memory & Cognition*. https://doi.org/10.3758/s13421-023-01516-1

Kan, I. P., Alexander, M. P., & Verfaellie, M. (2009). Contribution of prior semantic knowledge to new episodic learning in amnesia. *Journal of Cognitive Neuroscience*, *21*(5). https://doi.org/10.1162/jocn.2009.21066

Manenti, G. L., Dizaji, A. S., & Schwiedrzik, C. M. (2023). Variability in training unlocks generalization in visual perceptual learning through invariant representations. *Current Biology*, *33*(5). https://doi.org/10.1016/j.cub.2023.01.011

Meissner, C. A., & Brigham, J. C. (2001). Thirty years of investigating the own-race bias in memory for faces: A meta-analytic review. *Psychology, Public Policy, and Law*, *7*(1), 3–35. https://doi.org/10.1037/1076-8971.7.1.3

Morrison, D., Wang, H., Hahn, A. C., Jones, B. C., & DeBruine, L. M. (2017). Predicting the reward value of faces and bodies from social perception. *PLoS ONE*, *12*(9). https://doi.org/10.1371/journal.pone.0185093

Nosofsky, R. M., Sanders, C. A., Zhu, X., & McDaniel, M. A. (2019). Model-guided search for optimal natural-science-category training exemplars: A work in progress. *Psychonomic Bulletin and Review*, *26*(1). https://doi.org/10.3758/s13423-018-1508-8

Nosofsky, R. M., & Zaki, S. R. (2002). Exemplar and Prototype Models Revisited: Response Strategies, Selective Attention, and Stimulus Generalization. *Journal of Experimental Psychology: Learning Memory and Cognition*, *28*(5). https://doi.org/10.1037/0278-7393.28.5.924

Perry, L. K., Samuelson, L. K., Malloy, L. M., & Schiffer, R. N. (2010). Learn locally, think globally: Exemplar variability supports higher-order generalization and word learning. *Psychological Science*, *21*(12). https://doi.org/10.1177/0956797610389189

Peterson, M. J., Meagher, R. B., Chait, H., & Gillie, S. (1973). The abstraction and generalization of dot patterns. *Cognitive Psychology*, *4*(3). https://doi.org/10.1016/0010-0285(73)90019-4

Posner, M. I., & Keele, S. W. (1968). ON THE GENESIS OF ABSTRACT IDEAS. *Journal of Experimental Psychology*, *77*(3 PART 1). https://doi.org/10.1037/h0025953

Raviv, L., Lupyan, G., & Green, S. C. (2022). How variability shapes learning and generalization. In *Trends in Cognitive Sciences* (Vol. 26, Number 6). https://doi.org/10.1016/j.tics.2022.03.007

Rehder, B., & Hoffman, A. B. (2005a). Eyetracking and selective attention in category learning. *Cognitive Psychology*, *51*(1). https://doi.org/10.1016/j.cogpsych.2004.11.001

Rehder, B., & Hoffman, A. B. (2005b). Thirty-something categorization results explained: Selective attention, eyetracking, and models of category learning. In *Journal of Experimental Psychology: Learning Memory and Cognition* (Vol. 31, Number 5). https://doi.org/10.1037/0278-7393.31.5.811

Rost, G. C., & McMurray, B. (2010). Finding the signal by adding noise: The role of noncontrastive phonetic variability in early word learning. *Infancy*, *15*(6). https://doi.org/10.1111/j.1532-7078.2010.00033.x

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

Tenenbaum, J. B., & Griffiths, T. L. (2001). Generalization, similarity, and Bayesian inference. *Behavioral and Brain Sciences*, *24*(4). https://doi.org/10.1017/S0140525X01000061

Van Kesteren, M. T. R., Fernández, G., Norris, D. G., & Hermans, E. J. (2010). Persistent schema-dependent hippocampal-neocortical connectivity during memory encoding and postencoding rest in humans. *Proceedings of the National Academy of Sciences of the United States of America*, *107*(16). https://doi.org/10.1073/pnas.0914892107

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). https://doi.org/10.1037/0278-7393.29.6.1160

Chapter 3

Successful generalization of conceptual knowledge after training to remember specific events

From Houser, T.M., Resnick, A., & Zeithamova, D. (2024). Successful generalization of conceptual knowledge after training to remember specific events. *Frontiers in Cognition 3: 1324678.* Doi: 10.3389/fcogb.2024.1324678.

Categorization is important for many facets of cognition, including decision-making, object recognition, and language processing. The study of categorization has primarily utilized supervised category learning tasks that consist of showing participants a stimulus and teaching them its category membership through corrective feedback (Love, 2002; Nosofsky, Gluck, et al., 1994; Shepard et al., 1961). The ability to generalize category labels to new stimuli is then taken as evidence for the formation of category knowledge. This work has led to the development of several successful categorization models, including exemplar models that emphasize the role of memory for specific category exemplars (Medin & Schaffer, 1978; Nosofsky, 1986) and prototype models that emphasize the formation of abstract category representations (Posner & Keele, 1968; Smith & Minda, 1998). While we have learned much from this traditional approach, overt category decisions coupled with explicit feedback likely do not capture the range of ways in which we acquire category knowledge in the real world.

Importantly, category knowledge can also be acquired *incidentally,* without explicit instruction. For example, imagine trying an assortment of dishes from a foreign country. Perhaps, the dishes have similar spices, aromas, and flavors. Then, your friend tells you that the dishes are part of Thai cuisine. In the future, you will likely be able to differentiate Thai from other

cuisines, indicating that, while you may not have known what type of food you were eating, your brain was implicitly grouping similar sensory information together to form a unique category.

Incidental perceptual category learning studies in the auditory (Gabay et al., 2015; Gabay, Madlansacay, et al., 2023; Lim et al., 2019; Roark et al., 2022) and visual (Aizenstein et al., 2000; Bozoki et al., 2006; Kéri et al., 2001; Love, 2002; Reber et al., 2003; Wattenmaker, 1993) domains have demonstrated how category structure can be extracted in the absence of explicit instruction. In a typical incidental perceptual categorization task, participants are shown a series of stimuli and then informed that all of those stimuli form a category. They are then asked to indicate for a series of new stimuli whether or not each stimulus belongs to the same category (referred to as an A/not A task; e.g., Ashby & Maddox, 2011; Zeithamova et al., 2008). Less frequent has been work on unsupervised acquisition of category structures containing two or more categories (F. G. Ashby et al., 1999; Love, 2003). While learning in these paradigms is unsupervised, lacking feedback, the aim to discover the underlying category structure is typically explicitly instructed and its success is variable, influenced for example by the order in which the stimuli are presented (Clapper, 2006; Clapper & Bower, 1994, 2002; Zeithamova & Maddox, 2009).

More recently, a question arose whether it is possible to incidentally learn categories when explicitly focused on individuating each stimulus and when categories are not clearly dictated by an underlying similarity structure among stimuli. A recent study utilized the presence of a shared label (family name) as one way to provide an opportunity to spontaneously form category knowledge (S. R. Ashby et al., 2020a). Participants learned a set of paired associates, each consisting of a face and their full names (e.g., “Peter Miller”). To generate a similarity structure among the faces, face stimuli were created by morphing together two never-studied

"parent" faces. Some of the faces with a shared parent were assigned a shared family name while other faces with a shared parent were assigned different family names, enabling to tease apart the effects of physical similarity and category membership. After learning, participants demonstrated a category bias in perceptual similarity ratings such that faces with shared family names were rated as more similar than faces that were equated for physical similarity but had different family names. Participants were also highly successful in generalizing family names to never-studied faces, with the degree of generalization being predicted by the emerged perceptual biases. This finding was later replicated in a separate fMRI scanned sample, which provided evidence for category representations forming spontaneously in the brain while participants were still learning the unique full names for each face (S. R. Ashby & Zeithamova, 2022). The nature of those category representations, such as whether they tend to be exemplar-based or generalized, is however not straightforward to adjudicate for such stimuli.

Here, we build upon and extend this work in two directions. First, we tested whether people spontaneously acquire category knowledge even when no shared label is provided, through a mere association of each stimulus with other information that itself has category structure. Second, we utilized binary-dimension stimuli, a type of stimuli used extensively in prior categorization studies and suitable for the use of formal categorization models that can estimate underlying category representations from the patterns of participants' responses. Participants were asked to remember a set of paired associates, consisting of a cartoon animal and its preferred habitat (a background scene unique to each animal). Though participants needed to remember the specific scene for each cartoon animal, the scenes fell into two categories, mountains and forests, providing the opportunity for the participants to spontaneously organize the cartoon animals in line with their associated scene type. After paired-associates training, we

measured category knowledge by testing people's ability to generalize category membership (forest or mountain habitat) to novel animals. Further, the binary dimension stimuli in a form of cartoon animals allowed us to utilize two well-established categorization models to determine the type of representations people rely on after they acquired category information incidentally. The *exemplar model* posits that people store individual instances in memory (Fig. 1A) and categorize new instances based on their similarity to all stored exemplars (Medin & Schaffer, 1978; Nosofsky, 1986). By contrast, the *prototype model* posits that people extract the central tendency across category exemplars (Fig. 1A) which subsequently guides generalization behavior (Posner & Keele, 1968; Smith & Minda, 1998). We hypothesized that people trained to remember specific associations will successfully form and generalize category knowledge, but may increasingly rely on memory for specific exemplars rather than abstracted central tendencies (prototypes). To test this hypothesis, we utilized a category structure that leads to predominantly prototype representations when learned through traditional feedback-based classification training (Bowman & Zeithamova, 2018), and assessed (1) people's ability to incidentally acquire and subsequently generalize category knowledge, and (2) the impact that incidental learning has on generalization strategy. Finally, we explored whether the type of category representations that people formed (prototype or exemplar) related to how well they remember specific animal-scene associations requiring them to individuate category members, given the theoretical proposals suggesting a trade-off between memory specificity and generalization (McClelland & Goddard, 1996; Shohamy & Wagner, 2008; Varga et al., 2019). Overall, this work moves us closer to understanding the nature of category learning under more naturalistic demands.

Experiment 1

Methods

Participants. Healthy participants (n = 71; 50 females) were undergraduates recruited from the University of Oregon via the university's SONA research system. Ages ranged from 18-30 (M = 19) years. Participants received course credit for their participation. All participants provided written informed consent and experimental procedures were approved by Research Compliance Services at the University of Oregon.

Materials.

Stimuli. The complete stimulus set is freely available through the Open Science Framework (http://osf.io/8bph2). Stimuli consisted of cartoon animals that varied along 8 binary feature dimensions: color (yellow/gray), shape of feet (clawed/webbed), shape of body (squared/circular), shape of tail (devil tail/feather tail), dot orientation on body (vertical/horizontal), pattern on neck (stripes/horns), head shape (with beak/with horn), and head orientation (forward/up). Two stimuli with the maximally different number of features were the prototypes for Category A and Category B for all participants (Fig. 1A). Physical similarity between stimuli was defined as the number of shared features (Fig. 1B). The inverse, physical distance, was defined as the number of differing features. Category A stimuli shared more features with prototype A than prototype B and vice versa. Stimuli that were equidistant from both prototypes were not used.

Training set. Four training sets were created and assigned to participants randomly. All training sets included four stimuli per category (8 training stimuli in total), each differing from their respective category prototype by two features. An example training set with the corresponding prototypes is presented in Figure 1A. Prototypes themselves were never presented during training. Because the majority of the training exemplars differed from each other by four

features, whether they belonged to the same or opposing category, the similarity structure among stimuli did not fully dictate category membership. Instead, the membership was determined via the associated scenes.

Training scenes. Each training animal was paired with an animal-unique scene (8 scenes total) for the paired-associate learning task (Fig. 1C). The scenes were distinctive from each other, but fell into two categories, mountains and forests. Training animals that belonged to one category were all paired with scenes from one scene type, training animals that belong to the other category were all paired with scenes from the other scene type. In total, there were 8 unique animal-scene paired-associates used during training.

Testing set. Four testing sets were created and assigned to participants based on which training set they were administered. All testing sets included the 8 training stimuli (each presented twice) and 50 novel stimuli. Novel stimuli consisted of 8 randomly selected animals at each distance from the category prototypes (e.g., 8 new animals one feature different from prototype A, 8 new animals 2 features different from prototype A, and so on), excluding equidistant stimuli. The test sets also included the two previously unseen prototypes, each shown twice. Repeating the training stimuli and prototypes helps ensure dissociable prototype and exemplar model predictions (Bowman & Zeithamova, 2018; Kéri et al., 2001; Smith et al., 2008).

Experimental design.

Training. Participants first completed a training session where they were instructed to remember which background scene is the habitat for each unique training cartoon animal. Importantly, while each of the 8 background scenes was unique, all scenes fell into two categories: mountains or forests, with the habitat type aligning with category structure. In other

words, all four category A exemplars were paired with one of the four unique forest scenes while all four category B exemplars were paired with one of the four unique mountain scenes. Participants were not informed about the existence of the scene or animal categories and instead the task resembled a paired-associate learning task as used in episodic memory studies.

On each trial, participants were first shown one of eight background scenes that filled the screen. The associated cartoon animal then took two seconds to materialize by gradually increasing the image's opacity. Participants were then prompted to report how confident they are that they would remember the scene habitat for the particular cartoon animal using a button press corresponding to "1 = Definitely Forget", "2 = Maybe Forget", "3 = Maybe Remember", or "4 = Definitely Remember". Participants had 5 seconds to make a decision but the trial advanced if a response was registered sooner. An example training trial is shown in Figure 1C. Participants completed five blocks of training, each containing five repetitions of all training items with self-paced breaks between the blocks. The total number of training trials was two hundred (8 animal-scene pairs X 5 repetitions per block X 5 blocks = 200 trials). Stimulus order was randomized within each block.

Associative memory test. Following training, participants were administered an associative memory test, where they were shown each training stimulus along with four scene options and instructed to respond which scene they think the stimulus was paired with during training (Fig. 1D). Scene options always consisted of two mountain and two forest scenes: the correct scene, a lure of the same type, and two foils from the other habitat type. That way, the choices provided did not offer any cues with respect to which scene type (forest or mountain) is correct. Participants were also given the option of responding "Mountains" or "Forests" if they could not recall the specific scene.

Category generalization test. As the final portion of the session, participants were administered a surprise generalization test. On each trial, they were shown a training or a new (generalization) cartoon animal and instructed to respond either "Mountains" or "Forest" via button press (Fig. 1E). The categorization test consisted of one block of 68 trials with 5 second stimulus presentation and 7 second fixation. Trials advanced earlier if responses were registered before 5 seconds. There was no feedback during the generalization test.

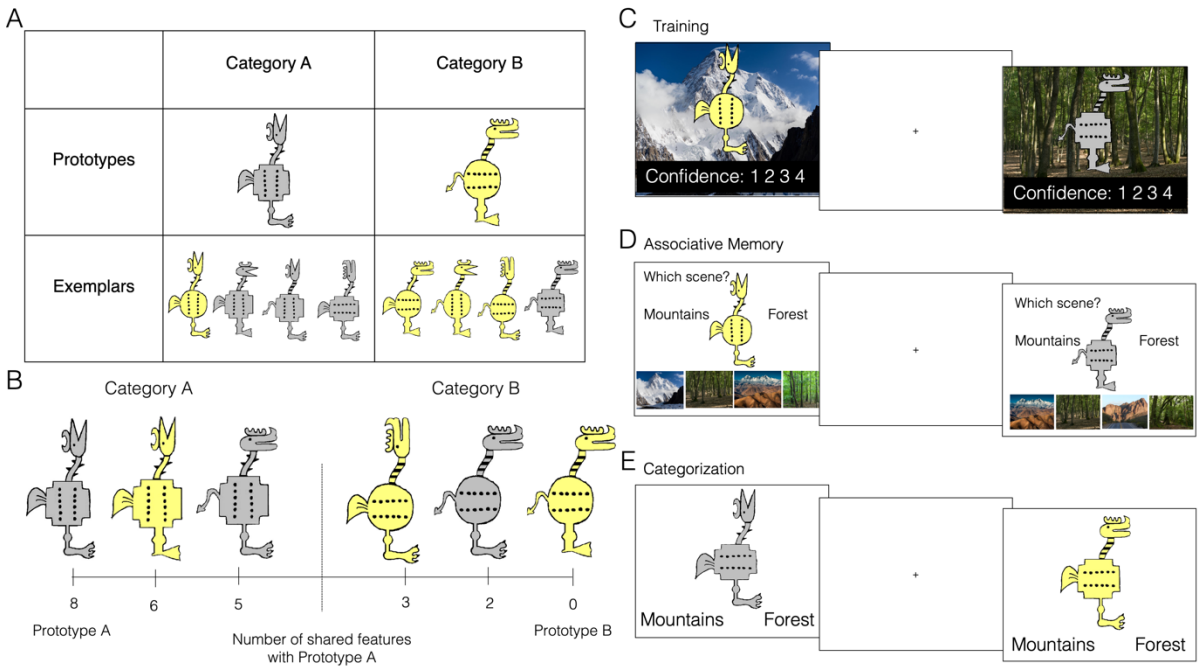

Figure 3.1. Experiment design.

A. Example stimulus set and the corresponding prototypes. Top row: category prototypes. Bottom row: example training set, divided into their correct categories. B. Stimuli were cartoon animals varying along 8 binary dimensions (features). Two maximally distinct stimuli (different on all 8 dimensions) served as category prototypes. Each stimulus can be characterized by the number of features by which it differs from each prototype (referred to as physical distance). C. Example training trial. Background and response options appeared first. The stimulus took 2 s to gradually appear, at which point subject can report how confident they are they will remember the animal-scene association. Fixation screen immediately follows response. D. Example associative memory test trial. Each training stimulus was shown with 6 possible response options: Correct scene, lure scene from the same category, two foil scenes from the other category, and two verbal category labels ("Mountains", "Forests") if a participant remembers the scene category but not the exact scene. E. Example category generalization test trial. Participants were asked to guess the correct scene category for a set of training and new (generalization) animals.

Statistical analysis. Raw data and the code necessary to reproduce the plots for Experiment 1 can be found at osf.io/dhmks.

Prospective memory judgments during training. Prospective memory judgements at training were mainly collected to ensure continued attention during observational learning and were not of primary interest. Nevertheless, we conducted a repeated-measures ANOVA to see if participants' confidence in their memory for the scene-animal associations increased over time. For each participant and each block, we computed the mean confidence rating. We expected to see an effect of training block. Only 49 participants were included in the prospective judgments analysis: due to technical error, data from the final training block was not saved for 21 participants and 1 participant made no responses for the first training block.

Associative memory. To evaluate how well participants remembered the specific animal-scene associations, we calculated the proportion of correct scene responses during the associative memory test for each participant. For completeness, we also computed and report the proportion of responses to all other unique options (i.e., to lure scene, specific foil scenes, and correct and incorrect general category labels).

Categorization accuracy during generalization test. To estimate how well participants generalized incidentally learned category information, we calculated the average accuracy of responses for each participant and compared these averages to chance level (50%) using a single sample t-test. Next, we tested to what degree generalization responses are modulated by the similarity to prototypes and whether the data exhibits a similarity gradient as commonly reported in category learning studies. For each participant, we first calculated the mean categorization accuracy across all generalization (new) trials at each distance from its prototype. These scores were submitted to a repeated measures ANOVA to test for an effect of distance on generalization

performance. Additionally, we compared categorization accuracy for old (training) animals and new (generalization) animals at the same distance from the prototypes (2 features difference, as all training animals varied from their respective prototypes by 2 features) using a paired t-test. An old-advantage, or greater accuracy for training (old) than generalization (new) stimuli at the same distance, indicates a contribution of the memory for specific instances to categorization and is often observed even when behavior is otherwise dominated by a prototype strategy (Bowman & Zeithamova, 2018). Given that we instructed participants to remember specific events, we predicted that old animals would be categorized more accurately than new animals.

Categorization strategies during generalization test. To identify the generalization strategies that people used, we fit prototype and exemplar models to the trial-by-trial responses from the generalization phase for each participant. Prototype models assume that category structures are represented by their prototype, and thus, they compute the similarity of each generalization stimulus to each prototype. Perceptual similarity is modeled as an exponential decay function of physical similarity (Shepard, 1958, 1987), while taking into account differences in attention to specific features. Formally, the similarity S of a stimulus x to category A is:

$$S_A(x) = \exp\left[-c\sum(w|x - proto_A|^r)^{\frac{1}{r}}\right],$$

where c is a sensitivity parameter that controls the rate of decay (estimated from the data), w is a vector of attention weights with length equal to the number of stimulus features (estimated from the data and constrained to sum to one), and r is the distance metric (set to 1 to reflect city-block distance metric typically used for binary features). In total, there are 9 participant-specific

parameters estimated from their pattern of responses: one c parameter and eight attention weights, one for each stimulus dimension. Thus, the models can account for intra-individual and inter-individual variability in feature saliency.

Exemplar models assume that categories are represented by their exemplars, or individual instances. These models calculate similarity between each stimulus and a category by computing the summed similarity of the stimulus and all the training stimuli from the given category (Nosofsky, 1987):

$$S_A(x) = \sum_{y \in A} \exp\left[-c \sum (w|x - y|^r)^{\frac{1}{r}}\right].$$

Here, y represents a training stimulus from category A and all other parameters are the same as in the prototype model.

To transform similarity scores to probabilities, we used a Luce choice rule (Luce, 1963):

$$P(A|x) = \frac{S_A(x)}{S_A(x) + S_B(x)}.$$

Using these equations, we obtained best fitting parameters from both models for each participant via maximum likelihood techniques. Specifically, we minimized the negative log likelihood by adjusting parameter estimates with gradient descent. To obtain best fitting parameter estimates, we used the *Rsolnp::solnp* function (Ghalanos & Theussl, 2015) as implemented in R, which optimizes model fit with general nonlinear optimization augmented with Lagrange multipliers.

We used the method of Lagrange multipliers to impose an equality constraint on the attention weight parameters such that they had to sum to 1. We used noninformative priors, set all lower bounds to 0, and the upper bound for the sensitivity parameter was set to 20.

Model selection. After optimization, prototype and exemplar model fits were compared at the individual level. A number of studies have demonstrated how group-level trends in generalization can be obtained by superimposing subgroup trends (F. G. Ashby et al., 1994; J. C. Lee et al., 2018; J. C. Lee & Livesey, 2018; Lovibond et al., 2019), highlighting the need for individualized assessment of generalization behavior (Zaman et al., 2023). Model fits were compared to each other and to chance to estimate the best fitting categorization strategy for each individual participant. We used a Monte Carlo simulation approach for model selection, as described previously (Bowman & Zeithamova, 2018; Kroese & Rubinstein, 2012). Each participant was assigned one of four strategy labels: exemplar, prototype, comparable fit, or random. First, we shuffled empirical responses 10000 times and obtained model fit values for each permutation to generate subject-specific null distributions for both computational models. We then compared empirical model fits against the subject-specific distribution of these null model fits. If the empirically observed fits were better than 95% of the simulated fits, we conclude that the model describes the data better than assuming that the participant responded randomly ($p < 0.05$, one-tailed). Participants who had neither prototype nor exemplar fit reliably better than chance were given "random" as their strategy label. For the remaining participants, we then directly compared the prototype and exemplar model fits to each other. To test for differences in model fits for each participant, we compared relative model fit differences, $[(fit_{prototype} - fit_{exemplar}) / (fit_{prototype} + fit_{exemplar})]$ to the relative model fit differences of simulated fits. One model was deemed a winner (i.e., subject was deemed a "prototypist" or "exemplarist")

if the difference in empirical fits appeared by chance with a frequency of less than 25% ($p < 0.25$, two-tailed). This approach has been shown to be more suitable for model comparison than simply comparing raw fits for two reasons: (a) the exemplar model tends to fit even random data slightly better than the prototype model, suggesting that equal number of parameters does not guarantee equal model flexibility, and (b) utilizing a confidence threshold before one model is deemed a winner helps differentiate negligible model fit differences from more meaningful ones (see Bowman et al., 2020; Bowman & Zeithamova, 2018). We preregistered using the raw fit differences, but verified that strategy assignment remained the same whether we used raw fit differences or relative fit differences.

Results

Prospective memory judgments during training. We found a main effect of block ($F(4,192) = 48.2$, $\eta^2 = 0.21$, $p < 0.001$), driven by participants increasing confidence in their ability to remember the animal-scene associations over time (Figure 3.2A). This effect indicates that participants paid attention to the task.

Associative memory. We assessed how often participants correctly recalled each training exemplar's paired background scene. Overall, people chose the correct scene most often (M = 64%, SD = 27%) indicating that participants successfully bound training exemplars to their unique background scenes. The full distribution of responses to all options is visualized in Figure 3.2B and suggests that when participants did not remember the specific correct scene association, they did not seem to remember the general category either. For example, among the three incorrect scenes, they would pick the single lure scene (9.5% of trials) just as often as one of the two foil scenes from the wrong category (18.1% for the two foil scenes = approximately 9% for each foil; Figure 3.2B). Moreover, the general habitat labels were not used frequently but also

did not show any evidence that participants would favor the correct habitat over wrong habitat ($t(70) = -0.50$, 95% $CI_{difference} = [-0.03, 0.02]$, $d = 0.06$, $p = 0.616$). Thus, participants seemed to most often encode the specific scene-animal associations and did not seem to be encoding the scenes at the generic category level.

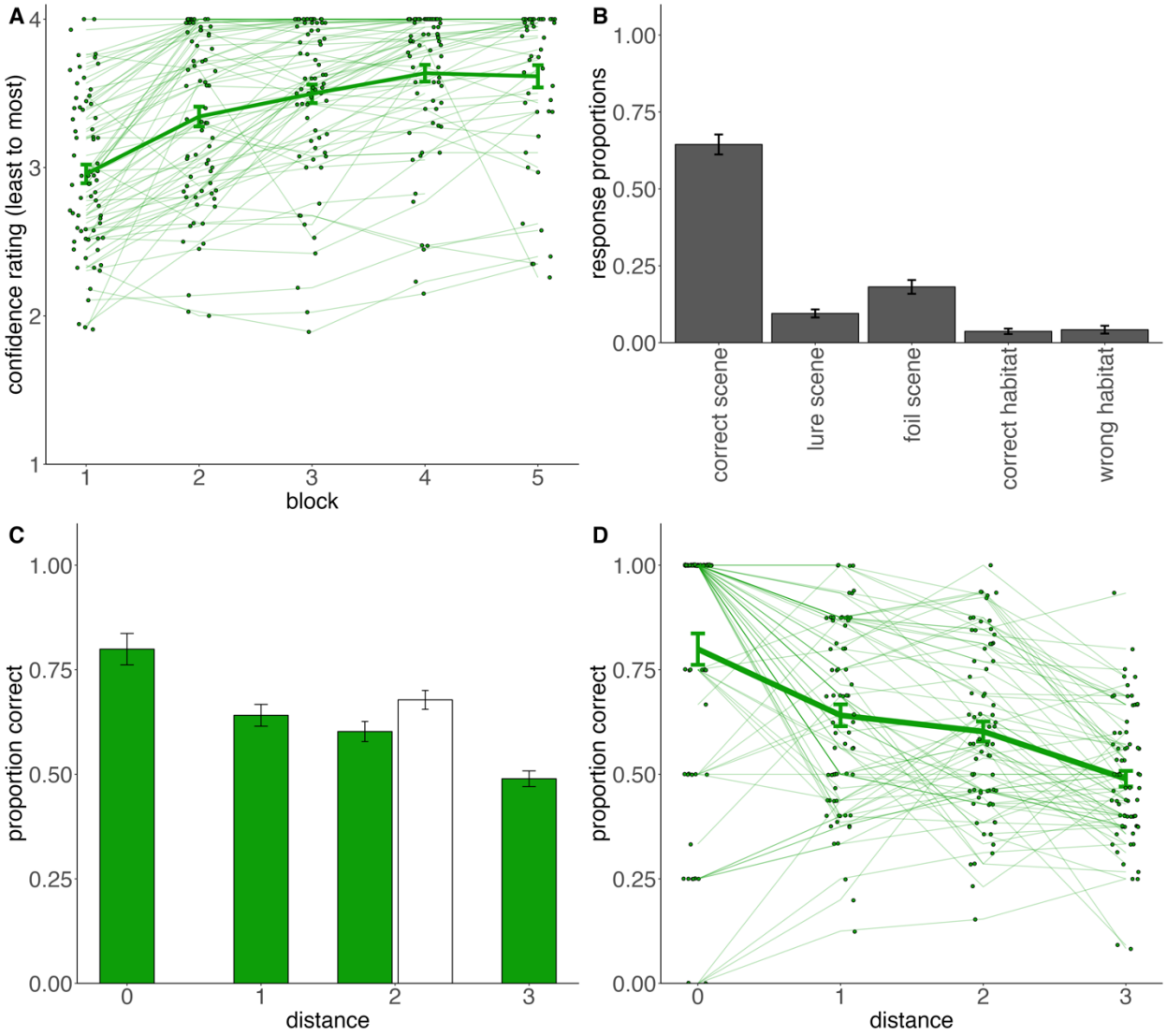

Figure 3.2. Experiment 1 training and generalization performance.

A. Mean confidence ratings for each training block. Thick line indicates the group average ratings, illustrating the increased confidence with training time. Thin lines depict average ratings for each participant. B. Associative memory performance. Bar heights indicate proportion of associative test responses. C. Generalization phase accuracy per each distance from the prototypes. Color indicates stimulus type (white = old; green = new). All

training (old) items were at the distance 2 from their respective prototypes. D. Generalization phase accuracy per each distance from the prototypes, including visualization of individual participants. The thick green line indicates the group average (same as panel C). Dots and connecting thin lines represent individual subjects. In all panels, error bars represent standard errors of the mean.

Categorization accuracy. The overall average categorization accuracy was 62% (SD = 16%), which was well above chance ($t(70) = 6.19$, 95% CI = [0.58, 0.66], $d = 0.73$, $p < 0.001$). Participants not only remembered the habitat type for the training animals (M = 68%, SD = 19%), but also tended to correctly guess habitats for new animals (M = 60%, SD = 17%). Categorization accuracy split by the distance to the prototype is presented in Figure 2C,D. To test for generalization gradients among the new test exemplars, we used a repeated measures ANOVA with distance from the category prototype as a predictor and average categorization accuracy as an outcome. We found a main effect of distance (0-3) ($F(3,210) = 39.6$, $\eta^2 = 0.189$, $p < 0.001$), with accuracy decreasing with the distance from prototypes (Figure 3.2D).

As all old (training) exemplars were at distance 2 from the prototypes, we also compared categorization for old vs. new distance 2 exemplars. Consistent with prior work, we found higher categorization accuracy for old than new exemplars at distance 2 from prototypes ($t(70) = 3.02$, 95% $CI_{difference}$ = [0.03, 0.13], $d = 0.39$, $p = 0.004$; Figure 3.2B). Nevertheless, participants were still more successful categorizing the never-seen prototypes than the directly studied training exemplars ($t(70) = 4.52$, 95% CI = [0.07, 0.17], $d = 0.46$, $p < 0.001$). Thus, the paired-associate learning task produced a pattern of categorization performance similar to a traditional supervised training task, including the categorization advantage for prototypes over training items themselves.

Categorization strategies. While a number of sophisticated categorization models exists, prototype and exemplar models represent two points on a continuum of conceptual knowledge

from generality to specificity. Thus, we used them here to augment our analysis to determine what kinds of representations people form and utilize for categorization decisions.

The conceptual illustration of the prototype and exemplar models is presented in Figure 3.3A. The proportions of participants that received each label (prototype, exemplar, comparable fit, random) is depicted in Figure 3B. Approximately 58% (41 out of 71) of participants had responses best fit by the prototype model, 20% (14 out of 71) were best fit by the exemplar model, and 1% (1 out of 71) of participants had comparable prototype and exemplar model fits (Figure 3.3B). The remaining 21% (15 out of 71) of participants had neither prototype nor exemplar model fit reliably above chance and were classified as using a random strategy. Thus, contrary to our prediction, the majority of participants were still best fit by the prototype model, suggesting that they abstracted the central tendency of the category rather than primarily relying on their memory for specific animal-scene associations.

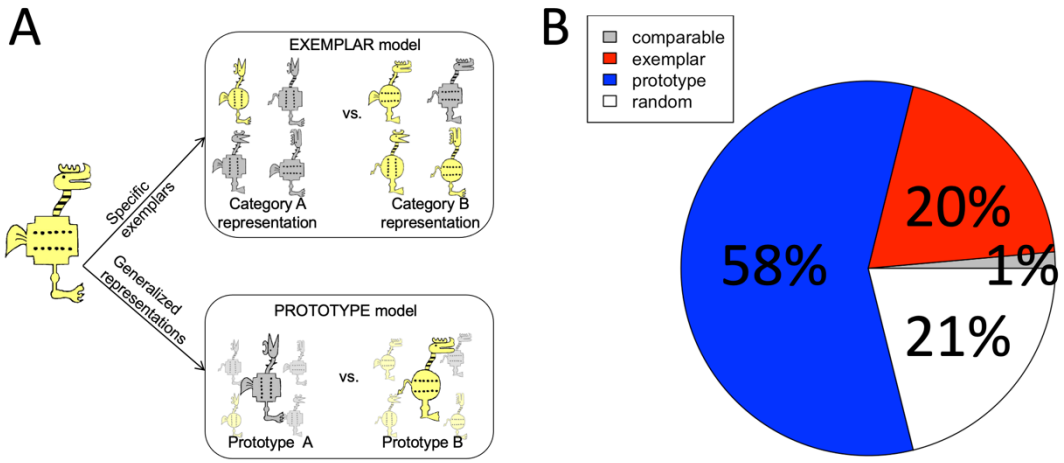

Figure 3.3. Experiment 1 computational model fitting results.

A. Example of the two accounts of generalization that we tested for. Upon seeing a novel stimulus (animal on the left side), one can either compare it to all training exemplars and use the summed similarities within categories to

make a response (exemplar model) or compare it to abstracted prototypes to make a response (prototype model). B. The proportion of subjects classified as generalizing according to the prototype strategy, exemplar strategy, having a comparable prototype and exemplar model fit, or having fits not exceeding chance (random).

Discussion

Overall, the results from Experiment 1 provide strong evidence for people's ability to form category knowledge incidentally through association with information that itself has category structure. Notably, participants acquired the category structure even when the task at hand required them to differentiate the specific associated scenes (e.g., all the forests from each other) rather than treating them as equivalent category members. Furthermore, participants seemed to base their categorization judgments on abstract category representations based on the category center (prototypes) rather than the specific associations they acquired for the training stimuli. First, participants showed a typical categorization gradient, and categorized novel prototypes better than the actual training exemplars that they were instructed to remember. Intuitively, generalization gradients are thought to emerge from prototype-based categorization representations but exemplar representations can also produce such a gradient (Jäkel et al., 2007, 2008; Nosofsky & Kruschke, 1992). Because only a formal comparison can determine whether the model predictions provide a good quantitative rather than just qualitative fit (Smith, 2002), we employed formal model fitting to directly estimate each participant's generalization strategy from their pattern of responses. Contrary to our prediction, we found that a majority of participants' responses were fit best by the prototype model. The dominance of prototype strategies in this task is consistent with a previous study (Bowman & Zeithamova, 2018) that used identical stimuli and category structures but employed traditional feedback-based classification training. It was however contrary to our prediction that generalization after the paired-associate learning would be based primarily on exemplar representations. While both the

current study and the prior study with the same stimuli found prototype dominance, the specific proportions were numerically shifted towards exemplar representations in our data; we found 58% participants best fit by the prototype strategy compared to 73% in Bowman & Zeithamova (2018) and 20% participants best fit by exemplar model, compared to just 10% in Bowman & Zeithamova (2018). Unsure of whether these are meaningful differences driven by the paired-associate learning task in comparison to the traditional feedback-based classification training, we preregistered and designed Experiment 2. Our goal was to replicate Experiment 1 findings as well as directly compare paired associate learning with feedback-based classification training using a between-subjects design.

Experiment 2

Method

Participants. Our pre-registered sample size was N=110 (55 for each group; https://osf.io/dzr6v). This was based on a power analysis for a Chi squared test of independence, to provide 80% power to detect medium effects (Effect size w = 0.3) for the difference in strategies between the two groups, as determined in G-Power (Faul et al., 2007, 2009), with a cushion for potential exclusion. The final sample size was larger than our pre-registered sample size as we were posting timeslots in batches every two weeks, as planned in our pre-registration.

Healthy young undergraduate students (n = 135; 95 females) were recruited from the University of Oregon via the university's SONA research system and received a course credit for their participation. Participants ages ranged from 18-25 (M = 19) years. All participants provided

written informed consent and experimental procedures were approved by Research Compliance Services at the University of Oregon.

Materials.

Stimuli. The logical structure of the stimuli was identical to Experiment 1 but the appearance and specific features were different to further test replicability of Experiment 1 with new stimuli. The cartoon animals in Experiment 2 varied along the following 8 binary feature dimensions: color (purple/red), neck (short/long), tail (straight/curled), foot shape (claws/round), snout (rounded/pig), head (ears/antennae), body shape (angled/round), and design on the body (polka dots/stripes). Two example prototypes are presented in Figure 3.4A. The complete stimulus set is freely available through the Open Science Framework (http://osf.io/8bph2). The structure of the training set and testing set were identical to Experiment 1.

Experimental design. Experiment 2 utilized a between-subjects design, where participants were randomly assigned into a paired-associate learning group (N = 67) or a feedback-based classification training group (N = 68).

Participants in the paired-associate group completed the exact same tasks (paired-associate learning, associative memory test, and category generalization test) as participants in Experiment 1. As in Experiment 1, the paired-associates learning task required participants to memorize 8 unique animal-scene association over 5 blocks of training, with 5 repetitions of each association per block (200 learning trials in total). Training animals from one category were associated with four different mountain scenes, training animals from the other category were associated with four different forest scenes. To ensure attention and equate response requirement with the feedback-based classification group, participants were asked to indicate on each trial how likely they are to remember each animal-scene association on a subsequent test. The paired-

associate learning task was followed by an associative memory task for the 8 animal-scene associations, identical to Experiment 1.

Participants in the feedback-based classification training group were learning about the two categories directly, using traditional supervised category learning task with the same stimuli (no scenes). Prior to training, participants were told they would learn the habitats in which different animals lived. During training, participants viewed individual cartoon animals on a computer screen, with a prompt "Does this guy live in the forest or the mountains?" displayed above the animal and two response options on the left ("Mountains") and the right ("Forest") side of the screen. Participants choose one of the two categories by pressing either 1 or 0 on the keyboard. Participants had 5 seconds to make a decision but the trial advanced if a response was registered sooner. After making a response, participants were given feedback for 3 s as to whether they were correct or incorrect (e.g., "Correct!" or "Incorrect. This one was from the mountains."). As in the paired-associate learning group, there were 5 training blocks with 5 repetitions of each of the 8 training exemplars in each block (200 training trials in total).

The category generalization test was the same as in Experiment 1 and identical for both groups. On each trial, participants saw a cartoon animal and had to guess the habitat ("Mountains" or "Forest") of that animal. The test was self-paced and no feedback was given.

Statistical analysis. Raw data and the code necessary to reproduce the plots for Experiment 2 can be found at osf.io/dhmks.

Training. To test whether participants in the paired-associate learning group maintained attention and participants in the feedback-based classification training group learned over time, we conducted two separate repeated measures ANOVAs using mean confidence ratings (paired-

associate learning group) and mean accuracy (feedback-based classification training group) for each participant and training block.

Associative memory. We computed associative memory measures for participants in the paired-associate group the same way that we did in Experiment 1.

Categorization test accuracy and strategy analysis. For the final category generalization test, we reported the same analyses as in Experiment 1 within each group separately and then compared the groups on each measure. Categorization accuracy was compared using an independent samples t-test (overall) and two mixed design ANOVAs (generalization gradient and old/new comparisons). To compare strategies, we used a Chi square test of independence to determine whether the proportions of people classified as using one strategy versus another differed between groups.

Results

Training data. There was a main effect of block in both the paired-associate learning group ($F(4,248) = 73.9$, $\eta^2 = 0.193$, $p < 0.001$) and feedback-based classification training group ($F(4,264) = 41.2$, $\eta^2 = 0.173$, $p < 0.001$). Thus, participants trained to remember specific associations became more confident in their ability to remember these associations over time while participants trained to categorize exemplars became more accurate over time (Figure 3.4B,C). Ergo, both groups were engaged in their respective tasks.

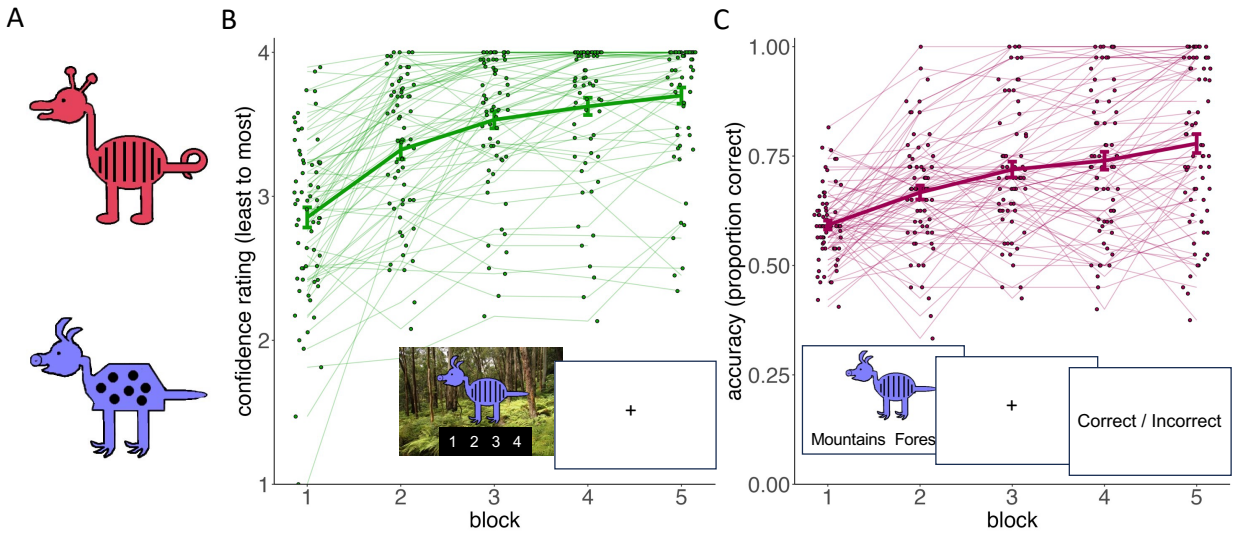

Figure 3.4. Experiment 2 stimuli and training performance.

A. Two example prototypes used in Experiment 2, illustrating the 8 binary features on which the cartoon animals can differ. B. Training responses for paired-associates learning group (confidence ratings). Inset shows an example training trial in the paired-associates learning condition, consisting on observing animal-scene association and rating how likely the association will be remembered on a future test. C. Training performance for feedback-based classification training group (proportion correct). Inset shows an example supervised classification training trial, consisting of categorizing an animal into one of two categories and receiving feedback. In both B and C, thick lines represent group averages and the thin lines and dots represent individual subjects (dots jittered to avoid overlap). Error bars are for standard errors of the mean.

Associative memory (paired-associate group only). As in Experiment 1, we assessed how often participants correctly recalled each training exemplar's paired background scene. Overall, people chose the correct scene most often (M = 71%, SD = 24%) indicating that participants successfully bound training exemplars to their unique background scenes. The full distribution of responses to all options is visualized in Figure 3.5A and, as in Exp. 1, suggests that when participants did not remember the specific correct scene association, they did not seem to remember the general category either. Thus, we replicated the finding from Exp. 1 that participants seemed to most often encode the specific scene-animal associations.

Categorization accuracy. During the final category generalization test, participants performed well-above chance in both the paired-associate learning group (M = 72%, SD = 13%, t(66) = 14.272, 95% CI = [0.69, 0.76], d = 1.74, $p < 0.001$) and feedback-based classification training group (M = 75%, SD = 13%, t(67) = 15.786, 95% CI = [0.72, 0.79], d = 1.91, $p < 0.001$). The overall accuracy did not significantly differ between the groups (t(133) = 1.38, 95% CI = [-0.01, 0.08], d = 0.24, p = 0.171). Thus, categorization during the generalization test was nearly as successful after the paired-associate learning task as after traditional feedback-based classification training. This was quite surprising given that participants in the paired-associate learning group focused on learning specific scene associates of specific animals while participants in the feedback-based classification training group were directly learning to categorize the animals into two categories.

Categorization accuracy for all novel items at each distance from the prototypes is presented in Figure 3.5B, and with accuracy for old items included in Figures 3.5C,D. To test for the gradient effect among novel test exemplars and any distance-specific differences between groups, we employed a mixed design ANOVA with distance from the prototype as a within-subject factor and training group (paired-associate learning vs. feedback-based classification training) as a between-subject factor (Figure 3.5B). We found a main effect of distance ($F(3,399) = 96.05$, $\eta^2 = 0.226$, $p < 0.001$), indicating that performance decreased with distance from the prototype, no main effect of group ($F(1,133) = 2.52$, $\eta^2 = 0.008$, $p = 0.115$), and a significant group-by-distance interaction ($F(1,399) = 2.98$, $\eta^2 = 0.007$, $p = 0.031$). To follow up on this interaction, we ran post-hoc independent samples t-tests, comparing group performance at each distance (0-3) from the prototypes. The feedback-based classification training group showed better categorization of distance two items (t(133) = 2.54, 95 % CI = [0.01, 0.12], d = 0.44, p =

0.012), with a similar trend for prototypes (t(133) = 1.88, 95% CI = [0.01, 0.16], d = 0.32, p = 0.062), while distance 1 and distance 3 items were comparable across groups (both $|t| < 0.53$, $d < 0.09$, $p > 0.59$). Thus, although we found no reliable differences in the overall accuracy, there seemed to be a small generalization advantage after feedback-based classification training for some of the generalization stimuli.

We next looked at old/new differences among distance 2 items and how they compare between groups. We found a main effect of stimulus type ($F(1,133) = 20.418$, $\eta^2 = 0.028$, $p < 0.001$), such that participants categorized training (old) stimuli better than novel ones. We also found a main effect of group ($F(1,133) = 4.73$, $\eta^2 = 0.027$, $p = 0.031$), with greater accuracy after feedback-based classification training than paired-associates learning. There was no group-by-old/new interaction ($F(1,133) = 0.94$, $\eta^2 = 0.001$, $p = 0.333$), suggesting that the old-advantage was comparable between groups.

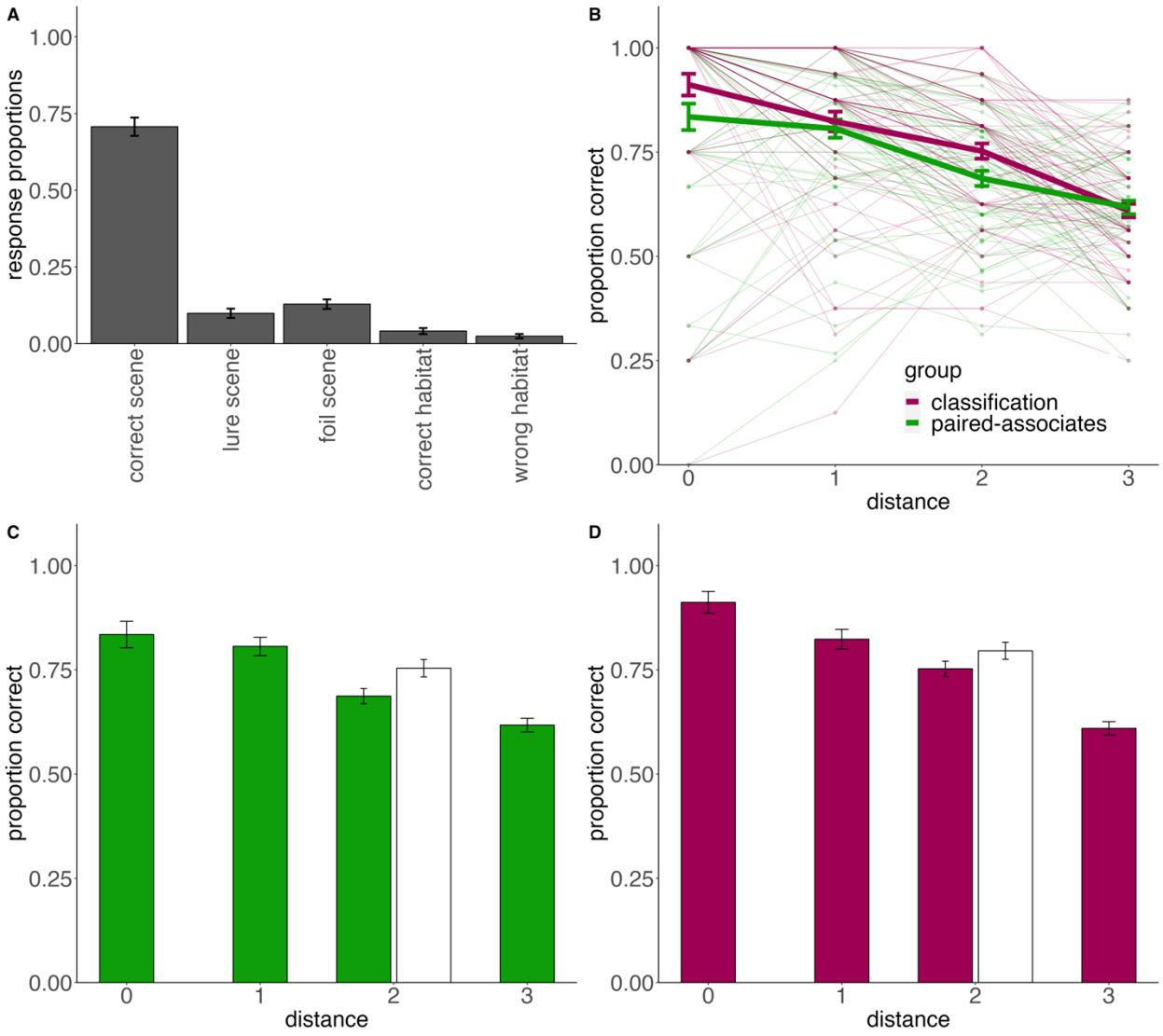

Figure 3.5. Experiment 2 test performance.

A. Associative memory responses for paired-associate group. Bar height indicates the proportion of each type of response. B. Categorization performance for both groups for novel stimuli at each distance to the prototypes. Thick lines represent group averages and thin lines and dots represent individual subject data. C. Categorization performance for just the paired-associates learning group. D. Categorization performance for just the feedback-based classification training group. In both C and D, bar height denotes accuracy at the given distance, averaged across subjects. White bars are old (training) exemplars. In all plots, error bars denote the standard error of the mean.

Categorization strategies. The percentage of participants labeled with each strategy label is presented in Figure 3.6. To compare the proportions of exemplarists and prototypists across groups, we used a Chi-squared test of independence. Only participants that were

confidently assigned one or the other strategy were included in the analysis, participants with comparable fits or that used a random strategy were excluded from this analysis. The results showed a trending relationship between generalization strategy and group ($\mathcal{X}^2$ = 5.73, p = 0.057). Thus, while generalization strategies following incidental category acquisition through paired-associates learning were similar to generalization strategies following traditional feedback-based category learning, there was a numerical shift toward exemplar-based strategies.

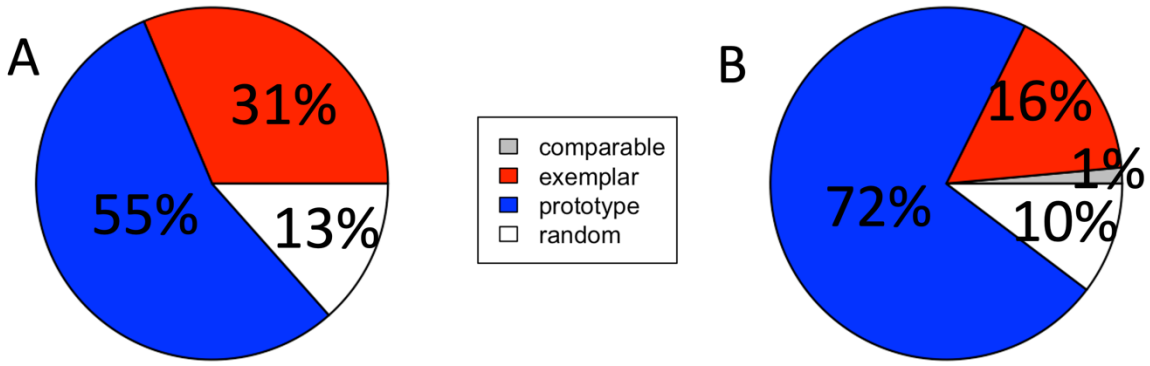

Figure 3.6. Experiment 2 categorization strategies.

The percentage of participants in the (A) paired-associate learning group and (B) feedback-based classification training group that were assigned each of four possible categorization strategy labels.

Relationship between specific and general knowledge (Exp 1 and 2). We next conducted a (preregistered) exploratory analysis that attempted to better understand the relationship between specific and general knowledge. Many theories suggests that generalization and specificity may trade off, such that formation of generalized memory representations comes at the expense of memory for specific details (McClelland & Goddard, 1996; Shohamy & Wagner, 2008; Varga et al., 2019). Because the paired-associates training group was tested not

only on concept generalization but also on their memory for specific animal-scene associations requiring them to distinguish among similar animals and among similar scenes, we utilized the opportunity to explore the potential trade-off at the level of individual differences.

To look for a potential trade off in our data, we first tested whether participants who successfully formed generalized knowledge were less able to remember specific details. We reasoned that if prototypists formed general knowledge while exemplarists formed specific memories, then exemplarists should demonstrate superior associative memory performance.

Using a two sample t-test comparing associative memory between participants classified as "exemplarists" vs. "prototypists", we found significantly greater associative memory for exemplarists than prototypists in Exp. 1 ($t(53) = 2.68$, $d = 0.83$, $p = 0.010$; Figure 3.7A) and numerically greater associative memory for exemplarists than prototypists in Exp. 2 ($t(56) = 1.17$, $d = 0.32$, $p = 0.248$; Figure 3.7B). The difference between exemplarists and prototypists was also significant when data were collapsed across experiments ($t(111) = 2.65$, $d = 0.54$, $p = 0.009$; Fig. 7C). Thus, we found some evidence that the formation of generalized category representation, operationalized here as the use of a prototype strategy, came at the expense for memory for specific animal-scene associations.

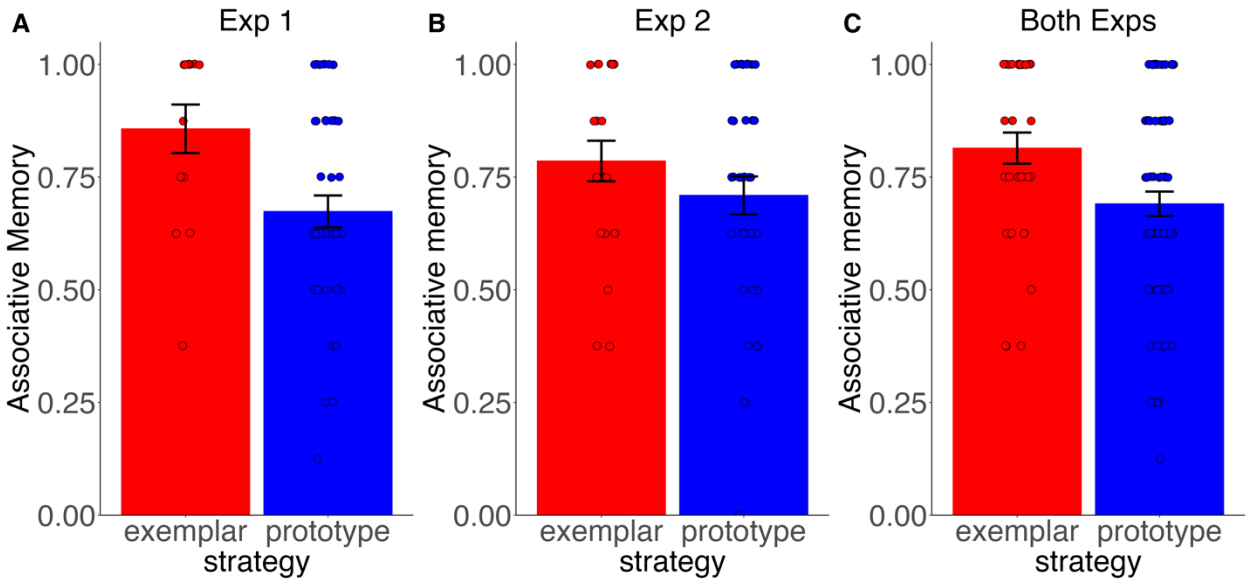

Figure 3.7. Associative memory and subsequent strategy use.

Associative memory performance for prototype and exemplar strategy users in Exp. 1 (A), Exp. 2 (B), and when collapsed across both experiments (C). Error bars are standard errors of the mean. Dots denote individual participants. Colors denote generalization strategy (red = exemplarists; blue = prototypists).

Next, we were interested whether memory for specific details and generalization success are related in our sample, and whether this relationship differs depending on categorization strategy. While generalization decisions can be based both on specific memories and generalized knowledge, making predictions for the specificity-generalization relationship uncertain, the formal models are helpful in estimating the underlying memory representations underlying categorization behaviors. For example, specificity and generalization may go hand-in-hand instead of trading off when they are both dependent on the same memory representations (Richter et al., 2016). We hypothesized that, if people used an exemplar-based strategy to categorize stimuli, their generalization and associative memory performance should be positively related. Such a relationship may not exist among prototypists as they presumably use different representations for the associative memory and generalization tasks. To test these ideas, we

computed across-subjects correlations between associative memory scores and generalization accuracy during the final categorization test, separately among prototypists and exemplarists. The results were inconsistent across experiments. In Exp 1, we found no relationship between associative memory and generalization for either exemplarists (r = -0.08, 95% CI = [-0.59, 0.47], p = 0.782) or prototypists (r = 0.07, 95% CI = [-0.24, 0.37], p = 0.670) (Figure 3.8A). In Exp. 2, we found a strong positive correlation between associative memory and generalization for exemplarists (r = 0.66, 95% CI = [0.32, 0.85], p = 0.001) but not for prototypists (r = 0.22, 95% CI = [-0.11, 0.51], p = 0.181) (Figure 3.8B). Regression analysis directly comparing the prototype and exemplar strategy users confirmed that the relationship between associative memory and generalization was reliably stronger for exemplarists than prototypists (strategy * associative memory interaction, β = 0.26, p = 0.013). Thus, we found partial evidence that the relationship between specific and general knowledge does differ depending on strategy, though the results were inconsistent across experiments.

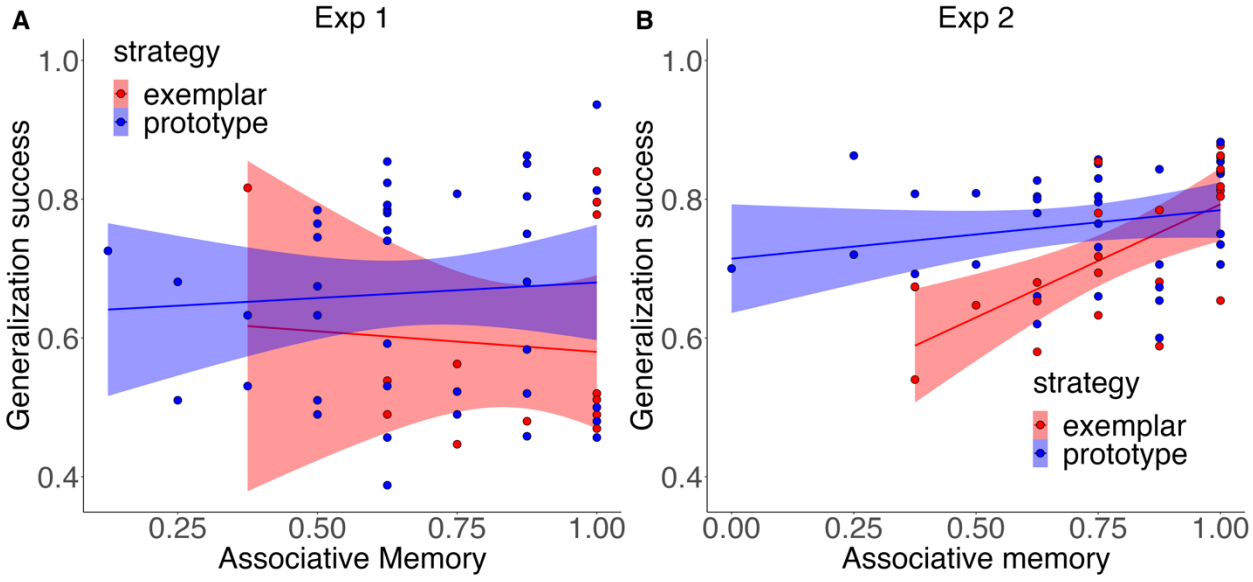

Figure 3.8. Generalization-associative memory relationship.

Lines represent correlations between correct scene responses (y-axis) and generalization accuracy (x-axis) and ribbons denote 95% confidence intervals for Exp. 1 (A), Exp. 2 (B). Dots denote individual participants. Colors denote generalization strategy (red = exemplarists; blue = prototypists).

General Discussion

The complexity of the world allows us to acquire category knowledge in a number of ways. Most work evaluating category knowledge acquisition trains people to learn categories using category response options coupled with explicit feedback, such that people are trained to associate stimuli with category labels through trial-and-error. In the current paper, we were interested in whether the same category knowledge could be acquired incidentally, without explicit instruction, by learning to associate individual category members with other specific stimuli that can be organized into categories based on preexisting knowledge. We found that people did indeed acquire category knowledge despite training to remember specific paired associates, reaching nearly identical generalization performance as those explicitly trained to categorize. Moreover, similarly to those who underwent explicit feedback-based categorization training, paired associate learners showed a prototype strategy dominance during categorization test. Nevertheless, a marginal trend suggested a tendency towards more exemplar-based strategies in people trained to remember specific information. These findings inform our understanding of how memories for experiences are spontaneously organized to serve multiple forms of cognition.

Results from the current study underscore important similarities and differences between incidental and more traditional category learning. Incidental category learning may provide a more naturalistic framework to characterize categorization processes (Broadbent et al., 2018; Gabay et al., 2015; Gabay, Madlansacay, et al., 2023; Roark et al., 2022; Unger & Sloutsky, 2022). For example, a child that hears the word “ball” might already be focusing on its red and

spherical properties, and thus, will incidentally learn to associate the category *ball* with red and sphere features. Indeed, a number of studies have implicated incidental category structure in learning relational information in children (Gentner & Namy, 2006; Vasilyeva et al., 2018). In the current study, we administered a novel category learning paradigm that resembles incidental learning, in that participants never received either explicit instruction to categorize stimuli or corrective feedback during training. However, unlike the majority of prior incidental categorization studies, the categories were not uniquely dictated by a similarity structure among training exemplars. Despite that, we found evidence for robust incidental concept learning after learning to remember specific associations. While the categorization performance was slightly better in the feedback-based classification training group for some of the stimuli, including the prototypes, the overall accuracy was comparable. Participants trained to remember specific associations were able to generalize incidentally-aligned category structure to novel stimuli, and showed similar generalization gradients as typically found in traditional category training studies, such as higher categorization success for prototypes than the stimuli they were trained on. As such, this study contributes to our understanding of concept learning under more naturalistic demands and indicates that many aspects of category learning are robust with respect to how the categories are learned.

Prior work on incidental learning has provided evidence that the learning systems and cognitive mechanisms engaged by traditional, feedback-based learning may be distinct from those employed during incidental category learning (Lim et al., 2019; Tricomi et al., 2006). Furthermore, research on episodic memory has shown how task goals and other conditions may bias participants to encode related information into separated memories vs. integration representations (Chanales et al., 2019, 2021; Schlichting et al., 2014, 2015; Schlichting &

Preston, 2015; Zeithamova & Preston, 2017). Here, we hypothesized there would be a difference in generalization strategy between feedback-based classification training group and paired-associate learning group. Specifically, we used prototype and exemplar models of categorization, both of which can provide good fits to behavior (Bowman et al., 2022; Bowman & Zeithamova, 2020, 2023; Heit, 1992; Hintzman, 1984; Nosofsky, 1986; Nosofsky, Gluck, et al., 1994; Posner & Keele, 1968) and neural activity (Bowman et al., 2020; Bowman & Zeithamova, 2018; Mack et al., 2013). We predicted an increase in reliance on exemplars after paired-associate learning emphasizing memory for individual stimuli, but the results were mixed in that regard. Although we found a marginally significant shift towards exemplar strategies after paired-associate learning, more than half of the participants were still fit best by the prototype model. As a second approach to comparing the role of specific memories in generalization success, we evaluated the degree of "old-advantage", or better categorization performance for old items than new items at the same distance from the prototypes, which can indicate a role for exemplar memorization in the overall categorization success. Old-advantage in both experiments was reliably above zero, but comparable between paired-associate learning group and feedback-based classification training group. Thus, the role of exemplar memorization, as indicated by the old-advantage analysis, did not seem to increase after the paired-associate learning. One possible reason for a lack of increased exemplar memorization following paired-associate learning is that the common scenes facilitated more contextual-based memory than memorization of associations (Hayes et al., 2007). That is, perhaps visual context (i.e., mountains or forest) facilitated retrieval of central tendencies within each context.

Exploratory analyses focused on the relationship between specificity and generalization. First, we hypothesized that given their reliance on specific memories, exemplarists may have

better memory for animal-scene associations than those relying on prototypes. The data were relatively well aligned with this prediction. This may indicate that there is a trade-off between specificity and generalization, such that the formation of a prototype comes at the expense of memories for specific associations. Alternatively, those with superior associative memory may be more likely to adopt an exemplar strategy. Second, we hypothesized that individual differences in associative memory may track generalization success in those relying on an exemplar strategy, given that both tasks are presumably supported by the same representations. In contrast, such a relationship may be less pronounced in those using a prototype strategy, as they presumably use distinct representations for the two tasks. The results of Experiment 2 aligned with this prediction: we observed a strong positive correlation between associative memory and generalization success in exemplarists, which is predicted by single-system theories assuming that the same type of memory representations supports both specific and generalized judgments (Zeithamova & Bowman, 2020). However, this correlation was only observed for Exp. 2 and Exp. 1, meaning the results are inconclusive. The relationship between associative memory and generalization success in those relying on prototypes was not significant in either experiment, which may reflect the use of distinct representations across the two tasks. However, each experiment was underpowered with respect to individual differences analyses, preventing strong conclusions from these null findings. Future studies may revisit this question in a larger sample, as within-subject correlations can contribute to our understanding of the shared and unique processes across different cognitive measures.

The grouping of participants into exemplar- and prototype-based strategy users assumes that people form representations of either prototypes or exemplars. Decades of debate have reinforced this dichotomy, yet, more recently, it has been theorized that prototypes and exemplars

form in parallel in the brain (Bowman et al., 2020; Zeithamova & Bowman, 2020). Forming representations at multiple levels of specificity can promote flexibility in future decision-making, as it is not always clear how current experience will become relevant. Intriguingly, the hippocampal axis has been posited to track multiple scales of information content (Brunec et al., 2018; Maurer & Nadel, 2021), consistent with a gradient of place cell receptive field sizes (Strange et al., 2014). Indeed, neuroimaging studies have found prototype representations specifically in the anterior portion of the hippocampus (Bowman et al., 2020; Bowman & Zeithamova, 2018). Functional connectivity along the hippocampal axis also aligns with the specificity-generalization continuum (Frank et al., 2019) and Guo & Yang (2023) show that the ventromedial prefrontal cortex is more strongly connected with the posterior hippocampus when retrieving specific associations but more strongly connected with the anterior hippocampus when retrieving schemas. While category representations can be found elsewhere in the brain (Mansouri et al., 2020), it is apparent that, across both tasks (Blank & Bayer, 2022; Mack et al., 2013) and methodologies (Sherman et al., 2023), specific and general representations of experience may form in parallel.

The current findings are broadly consistent with the notion that generalized representations may form in parallel with specific ones during learning to be flexibly utilized during subsequent decision-making depending on task goals. While we hypothesized that people trained to remember specific information would increasingly use an exemplar-based strategy, we instead found reliance on prototypes in the majority of participants. Importantly, because the associative test required differentiation among similar animals and background scenes from the same category, it could not be based on just general knowledge about scene categories. Thus, individuals appeared to both encode specific information and construct general representations,

and were able to use either type of memory in a flexible, task-dependent manner. Prototype strategy may dominate categorization task in most participants because it is cognitively more efficient. Nevertheless, the increased attention to differentiating details of individual exemplars elicited by the paired-associate learning task may result in somewhat greater exemplar strategy utilization than observed after traditional supervised training.

Conclusion

In the current study, we have demonstrated that people trained to remember specific associations can incidentally acquire knowledge of category structures revealed by association with other material that itself is structured. Furthermore, generalization of such knowledge to novel stimuli relied on similar representations as when acquired during traditional feedback-based classification training. When considered in tandem with previous literature, these results suggest that people may encode experiences at multiple levels of specificity, maintaining specific details while spontaneously organizing related memories to form more generalized knowledge. This may allow them to flexibly switch among representations, focusing on differentiating details of specific memories vs. generalized knowledge, in response to task demands.

References cited

Aizenstein, H. J., MacDonald, A. W., Stenger, V. A., Nebes, R. D., Larson, J. K., Ursu, S., & Carter, C. S. (2000). Complementary category learning systems identified using event-related functional MRI. *Journal of Cognitive Neuroscience*, *12*(6). https://doi.org/10.1162/08989290051137512

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

Ashby, F. G., Maddox, W. T., & Lee, W. W. (1994). On the dangers of averaging across subjects when using multidimensional scaling or the similarity-choice model. *Psychological Science*, *5*(3). https://doi.org/10.1111/j.1467-9280.1994.tb00651.x

Ashby, F. G., Queller, S., & Berretty, P. M. (1999). On the dominance of unidimensional rules in unsupervised categorization. *Perception and Psychophysics*, *61*(6). https://doi.org/10.3758/BF03207622

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 & Review*, *27*(4), 791–800. https://doi.org/10.3758/s13423-020-01754-3

Ashby, S. R., & Zeithamova, D. (2022). Category-Biased Neural Representations Form Spontaneously during Learning That Emphasizes Memory for Specific Instances. *Journal of Neuroscience*, *42*(5). https://doi.org/10.1523/JNEUROSCI.1396-21.2021

Blank, H., & Bayer, J. (2022). Functional imaging analyses reveal prototype and exemplar representations in a perceptual single-category task. *Communications Biology*, *5*(1). https://doi.org/10.1038/s42003-022-03858-z

Bowman, C. R., Iwashita, T., & Zeithamova, D. (2020). Tracking prototype and exemplar representations in the brain across learning. *ELife*, *9*, 1–47. https://doi.org/10.7554/eLife.59360

Bowman, C. R., Iwashita, T., & Zeithamova, D. (2022). The Effects of Age on Category Learning and Prototype and Exemplar-Based Generalization. *Psychology and Aging*, *37*(7). https://doi.org/10.1037/pag0000714

Bowman, C. R., & Zeithamova, D. (2018). Abstract memory representations in the ventromedial prefrontal cortex and hippocampus support concept generalization. *Journal of Neuroscience*, *38*(10), 2605–2614. https://doi.org/10.1523/JNEUROSCI.2811-17.2018

Bowman, C. R., & Zeithamova, D. (2020). Training Set Coherence and Set Size Effects on Concept Generalization and Recognition. *Journal of Experimental Psychology: Learning Memory and Cognition*. https://doi.org/10.1037/xlm0000824

Bowman, C. R., & Zeithamova, D. (2023). Coherent category training enhances generalization in prototype-based categories. *Journal of Experimental Psychology: Learning, Memory, and Cognition*. https://doi.org/10.1037/xlm0001243

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

Broadbent, H. J., White, H., Mareschal, D., & Kirkham, N. Z. (2018). Incidental learning in a multisensory environment across childhood. *Developmental Science*, *21*(2). https://doi.org/10.1111/desc.12554

Brunec, I. K., Bellana, B., Ozubko, J. D., Man, V., Robin, J., Liu, Z. X., Grady, C., Rosenbaum, R. S., Winocur, G., Barense, M. D., & Moscovitch, M. (2018). Multiple Scales of Representation along the Hippocampal Anteroposterior Axis in Humans. *Current Biology*, *28*(13). https://doi.org/10.1016/j.cub.2018.05.016

Chanales, A. J. H., Dudukovic, N. M., Richter, F. R., & Kuhl, B. A. (2019). Interference between overlapping memories is predicted by neural states during learning. *Nature Communications*, *10*(1). https://doi.org/10.1038/s41467-019-13377-x

Chanales, A. J. H., Tremblay-McGaw, A. G., Drascher, M. L., & Kuhl, B. A. (2021). Adaptive Repulsion of Long-Term Memory Representations Is Triggered by Event Similarity. *Psychological Science*, *32*(5). https://doi.org/10.1177/0956797620972490

Clapper, J. P. (2006). When more is less: Negative exposure effects in unsupervised learning. *Memory and Cognition*, *34*(4). https://doi.org/10.3758/BF03193435

Clapper, J. P., & Bower, G. H. (1994). Category Invention in Unsupervised Learning. *Journal of Experimental Psychology: Learning, Memory, and Cognition*, *20*(2). https://doi.org/10.1037/0278-7393.20.2.443

Clapper, J. P., & Bower, G. H. (2002). Adaptive Categorization in Unsupervised Learning. *Journal of Experimental Psychology: Learning Memory and Cognition*, *28*(5). https://doi.org/10.1037/0278-7393.28.5.908

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 3: 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

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). https://doi.org/10.1162/jocn_a_01457

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). https://doi.org/10.1037/xhp0000073

Gabay, Y., Madlansacay, M., & Holt, L. L. (2023). Incidental auditory category learning and visuomotor sequence learning do not compete for cognitive resources. *Attention, Perception, and Psychophysics*, *85*(2). https://doi.org/10.3758/s13414-022-02616-x

Gentner, D., & Namy, L. L. (2006). Analogical processes in language learning. In *Current Directions in Psychological Science* (Vol. 15, Number 6). https://doi.org/10.1111/j.1467-8721.2006.00456.x

Ghalanos, A., & Theussl, S. (2015). *Rsolnp: General Non-linear Optimization Using Augmented Lagrange Multiplier Method* (1.16). R.

Guo, D., & Yang, J. (2023). Reactivation of schema representation in lateral occipital cortex supports successful memory encoding. *Cerebral Cortex (New York, N.Y. : 1991)*, *33*(10). https://doi.org/10.1093/cercor/bhac475

Hayes, S. M., Nadel, L., & Ryan, L. (2007). The effect of scene context on episodic object recognition: Parahippocampal cortex mediates memory encoding and retrieval success. *Hippocampus*, *17*(9). https://doi.org/10.1002/hipo.20319

Heit, E. (1992). Categorization using chains of examples. *Cognitive Psychology*, *24*(3). https://doi.org/10.1016/0010-0285(92)90011-P

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

Jäkel, F., Schölkopf, B., & Wichmann, F. A. (2007). A tutorial on kernel methods for categorization. In *Journal of Mathematical Psychology* (Vol. 51, Number 6). https://doi.org/10.1016/j.jmp.2007.06.002

Jäkel, F., Schölkopf, B., & Wichmann, F. A. (2008). Generalization and similarity in exemplar models of categorization: Insights from machine learning. In *Psychonomic Bulletin and Review* (Vol. 15, Number 2). https://doi.org/10.3758/PBR.15.2.256

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

Kroese, D. P., & Rubinstein, R. Y. (2012). Monte Carlo methods. *Wiley Interdisciplinary Reviews: Computational Statistics*, *4*(1). https://doi.org/10.1002/wics.194

Lee, J. C., Hayes, B. K., & Lovibond, P. F. (2018). Peak shift and rules in human generalization. *Journal of Experimental Psychology: Learning Memory and Cognition*, *44*(12). https://doi.org/10.1037/xlm0000558

Lee, J. C., & Livesey, E. J. (2018). Rule-based generalization and peak shift in the presence of simple relational rules. *PLoS ONE*, *13*(9). https://doi.org/10.1371/journal.pone.0203805

Lim, S. J., Fiez, J. A., & Holt, L. L. (2019). Role of the striatum in incidental learning of sound categories. *Proceedings of the National Academy of Sciences of the United States of America*, *116*(10). https://doi.org/10.1073/pnas.1811992116

Love, B. C. (2002). Comparing supervised and unsupervised category learning. *Psychonomic Bulletin and Review*, *9*(4). https://doi.org/10.3758/BF03196342

Love, B. C. (2003). The multifaceted nature of unsupervised category learning. *Psychonomic Bulletin and Review*, *10*(1). https://doi.org/10.3758/BF03196484

Lovibond, P. F., Lee, J. C., & Hayes, B. K. (2019). Stimulus Discriminability and Induction as Independent Components of Generalization. *Journal of Experimental Psychology: Learning Memory and Cognition*. https://doi.org/10.1037/xlm0000779

Luce, R. D. (1963). A Threshold Theory for Simple Detection Experiments. *Psychological Review*, *70*(1). https://doi.org/10.1037/h0039723

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). https://doi.org/10.1016/j.cub.2013.08.035

Mansouri, F. A., Freedman, D. J., & Buckley, M. J. (2020). Emergence of abstract rules in the primate brain. In *Nature Reviews Neuroscience* (Vol. 21, Number 11). https://doi.org/10.1038/s41583-020-0364-5

Maurer, A. P., & Nadel, L. (2021). The Continuity of Context: A Role for the Hippocampus. In *Trends in Cognitive Sciences* (Vol. 25, Number 3). https://doi.org/10.1016/j.tics.2020.12.007

McClelland, J. L., & Goddard, N. H. (1996). Considerations arising from a complementary learning systems perspective on hippocampus and neocortex. *Hippocampus*, *6*(6). https://doi.org/10.1002/(SICI)1098-1063(1996)6:6<654::AID-HIPO8>3.0.CO;2-G

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

Nosofsky, R. M. (1986). Attention, similarity, and the identification-categorization relationship. *Journal of Experimental Psychology: General*, *115*(1). 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). https://doi.org/10.1037/0278-7393.13.1.87

Nosofsky, R. M., Gluck, M. A., Palmeri, T. J., Mckinley, S. C., & Glauthier, P. (1994). Comparing modes of rule-based classification learning: A replication and extension of Shepard, Hovland, and Jenkins (1961). *Memory & Cognition*, *22*(3). https://doi.org/10.3758/BF03200862

Nosofsky, R. M., & Kruschke, J. K. (1992). Investigations of an exemplar-based connectionist model of category learning. *Psychology of Learning and Motivation - Advances in Research and Theory*, *28*(C). https://doi.org/10.1016/S0079-7421(08)60491-0

Posner, M. I., & Keele, S. W. (1968). ON THE GENESIS OF ABSTRACT IDEAS. *Journal of Experimental Psychology*, *77*(3 PART 1). https://doi.org/10.1037/h0025953

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

Richter, F. R., Chanales, A. J. H., & Kuhl, B. A. (2016). Predicting the integration of overlapping memories by decoding mnemonic processing states during learning. *NeuroImage*, *124*. https://doi.org/10.1016/j.neuroimage.2015.08.051

Roark, C. L., Lehet, M. I., Dick, F., & Holt, L. L. (2022). The Representational Glue for Incidental Category Learning Is Alignment With Task-Relevant Behavior. *Journal of Experimental Psychology: Learning Memory and Cognition*, *48*(6). https://doi.org/10.1037/xlm0001078

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*. https://doi.org/10.1038/ncomms9151

Schlichting, M. L., & Preston, A. R. (2015). Memory integration: Neural mechanisms and implications for behavior. In *Current Opinion in Behavioral Sciences* (Vol. 1). 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). https://doi.org/10.1002/hipo.22310

Shepard, R. N. (1958). Stimulus and response generalization: Tests of a model relating generalization to distance in psychological space. *Journal of Experimental Psychology*, *55*(6). https://doi.org/10.1037/h0042354

Shepard, R. N. (1987). Toward a universal law of generalization for psychological science. *Science*, *237*(4820). https://doi.org/10.1126/science.3629243

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

Sherman, B. E., Aljishi, A., Graves, K. N., Quraishi, I. H., Sivaraju, A., Damisah, E. C., & Turk-Browne, N. B. (2023). Intracranial Entrainment Reveals Statistical Learning across Levels of Abstraction. *Journal of Cognitive Neuroscience*, *35*(8), 1312–1328. https://doi.org/10.1162/jocn_a_02012

Shohamy, D., & Wagner, A. D. (2008). Integrating Memories in the Human Brain: Hippocampal-Midbrain Encoding of Overlapping Events. *Neuron*, *60*(2). https://doi.org/10.1016/j.neuron.2008.09.023

Smith, J. D. (2002). Exemplar theory's predicted typicality gradient can be tested and disconfirmed. *Psychological Science*, *13*(5). https://doi.org/10.1111/1467-9280.00477

Smith, J. D., & Minda, J. P. (1998). Prototypes in the mist: The early epochs of category learning. *Journal of Experimental Psychology: Learning Memory and Cognition*, *24*(6). https://doi.org/10.1037/0278-7393.24.6.1411

Smith, J. D., Redford, J. S., & Haas, S. M. (2008). Prototype Abstraction by Monkeys (Macaca mulatta). *Journal of Experimental Psychology: General*, *137*(2). https://doi.org/10.1037/0096-3445.137.2.390

Strange, B. A., Witter, M. P., Lein, E. S., & Moser, E. I. (2014). Functional organization of the hippocampal longitudinal axis. In *Nature Reviews Neuroscience* (Vol. 15, Number 10). https://doi.org/10.1038/nrn3785

The Role of Comparison in Children's Early Word Learning. (2018). In *Weaving A Lexicon*. https://doi.org/10.7551/mitpress/7185.003.0021

Tricomi, E., Delgado, M. R., McCandliss, B. D., McClelland, J. L., & Fiez, J. A. (2006). Performance feedback drives caudate activation in a phonological learning task. *Journal of Cognitive Neuroscience*, *18*(6). https://doi.org/10.1162/jocn.2006.18.6.1029

Unger, L., & Sloutsky, V. M. (2022). Ready to Learn: Incidental Exposure Fosters Category Learning. *Psychological Science*, *33*(6). https://doi.org/10.1177/09567976211061470

Varga, N., Gaugler, T., & Talarico, J. (2019). Are mnemonic failures and benefits two sides of the same coin?: Investigating the real-world consequences of individual differences in memory integration. *Memory and Cognition*, *47*(3). https://doi.org/10.3758/s13421-018-0887-4

Vasilyeva, N., Gopnik, A., & Lombrozo, T. (2018). The development of structural thinking about social categories. *Developmental Psychology*, *54*(9). https://doi.org/10.1037/dev0000555

Wattenmaker, W. D. (1993). Incidental Concept Learning, Feature Frequency, and Correlated Properties. *Journal of Experimental Psychology: Learning, Memory, and Cognition*, *19*(1). https://doi.org/10.1037/0278-7393.19.1.203

Zaman, J., Yu, K., & Verheyen, S. (2023). The idiosyncratic nature of how individuals perceive, represent, and remember their surroundings and its impact on learning-based generalization. *Journal of Experimental Psychology: General*. https://doi.org/10.1037/xge0001403

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

Zeithamova, D., & Maddox, W. T. (2009). Learning Mode and Exemplar Sequencing in Unsupervised Category Learning. *Journal of Experimental Psychology: Learning Memory and Cognition*, *35*(3). https://doi.org/10.1037/a0015005

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

Zeithamova, D., & Preston, A. R. (2017). Temporal proximity promotes integration of overlapping events. *Journal of Cognitive Neuroscience*, *29*(8). https://doi.org/10.1162/jocn_a_01116

Chapter 4

Incidental category learning leads to exemplar representations in the ventromedial prefrontal cortex

A core goal of cognitive science is to understand how category knowledge is acquired and subsequently generalized to novel scenarios. The overwhelming majority of research investigating this provides people with explicit instruction and has them make overt category judgments during learning which is followed by corrective feedback. However, this is not the only way that people learn categories (Markman & Ross, 2003). Arguably, incidental learning of categories occurs far more often in the real world than learning via supervised feedback (Clapper, 2006; Clapper & Bower, 1994, 2002; Love, 2002, 2003; Unger & Sloutsky, 2022). Does incidental learning lead to similar category representations in behavior and the brain?

Incidental category learning studies using auditory (Gabay et al., 2015; Gabay, Madlansacay, et al., 2023; Gabay & Holt, 2015; Lim et al., 2019; Roark et al., 2022) and visual (Aizenstein et al., 2000; Bozoki et al., 2006; Kéri et al., 2001; Love, 2002; Reber et al., 2003; Wattenmaker, 1993) stimuli have revealed that category structure information can be extracted despite not receiving explicit instruction. Extending this work, Ashby et al. (2020) tested the extent to which categories can be learned incidentally (via the presence of a shared label across exemplars) despite being explicitly trained to focus on individuating details of category exemplars. Participants were able to extract shared features across exemplars within a category and use them to generalize to novel stimuli (S. R. Ashby et al., 2020b). This finding was subsequently replicated in a separate fMRI sample (S. R. Ashby & Zeithamova, 2022).

Despite being trained to memorize individuating details, the participants from Ashby et al. likely extracted shared feature information by associating within-category stimuli to the

shared category label. Houser et al. (2024) followed this line of research by not only evaluating whether people can incidentally learn categories despite being trained to focus on individuating details but with the additional absence of a shared category label. Specifically, people were instructed to memorize eight unique exemplar-background scene associations. Participants were not informed that each exemplar belonged to one of two categories; however, scene type (mountains or forests) aligned with category assignment, meaning one could implicitly group exemplars with shared scene type associations. Indeed, participants could successfully acquire category knowledge, as indexed by standard generalization gradients, where categorization accuracy decreases with dissimilarity to category prototypes (Houser et al., 2024). Moreover, Ashby et al. (2020) used naturalistic face stimuli, which made it difficult to adjudicate between possible category representations that people may have formed during training, whereas Houser et al. used easily quantifiable stimuli to compare model fits to generalization behavior.

Houser et al. (2024) fit prototype and exemplar models to people's generalization data to compare two competing approaches to characterizing the category representations that people use during generalization. The exemplar model suggests that people store individual exemplars as memory traces and categorize new exemplars based on their similarity to all exemplars stored in memory (Medin & Schaffer, 1978; Nosofsky, 1986). On the other hand, the prototype model posits that people extract out a central tendency across encountered exemplars, which then guides generalization behavior (Posner & Keele, 1968; Smith & Minda, 1998). Whereas previous work utilizing the same stimuli in a feedback-based paradigm found that roughly 73% of people relied on prototypes to generalize (Bowman & Zeithamova, 2018), Houser et al. observed only 58 and 55% of people relying on prototypes across two experiments and a corresponding increase in the number of people relying on exemplars (20 and 31% in Houser et al. compared to

10% in Bowman & Zeithamova). Together, the incidental associative learning paradigm led to successful acquisition of category knowledge but through encoding of more exemplars.

Here, we tested the degree to which incidentally acquired category representations in the brain share overlapping neural substrates with feedback-based ones as identified in prior studies. Given that previous work in the lab has found robust evidence for prototype representations in the brain in a sample predominantly characterized best by the prototype model (Bowman & Zeithamova, 2018), but that incidental encoding shifts behavioral strategies towards more exemplar-reliance, it is possible that exemplar representations in the brain can be decoded following incidental category learning. These questions were addressed in the current paper using a combination of a novel category learning paradigm, computational modeling, and fMRI.

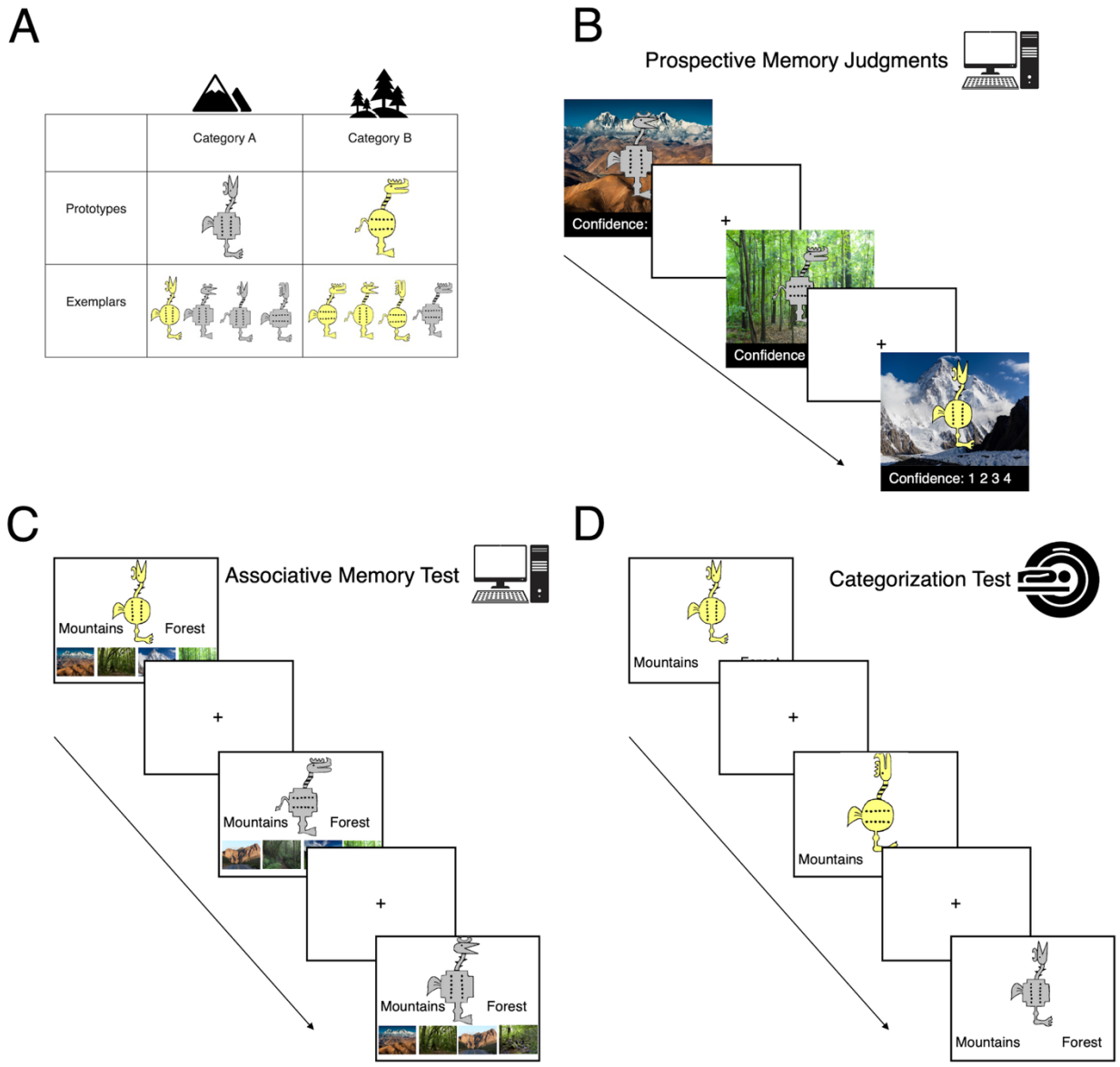

Figure 4.1. Experimental design.

A. Example stimulus set and corresponding prototypes. Top row: category prototypes. Bottom row: example training set, divided into their correct categories. Stimuli were cartoon animals varying along 8 binary features. Two maximally distinct stimuli (different on all 8 dimensions) served as the prototypes. Each stimulus, from training and testing, can be characterized by the number of features different from their prototype. B. Example training trial. Background scene and response options appeared immediately and the training exemplar took 2 s to gradually appear, at which point the participant could indicate with button press how confident they are that they will remember the stimulus-scene association. A fixation screen immediately followed response. C. Example associative memory test trial. Each training exemplar was shown with six possible response options: correct scene, lure scene from the same habitat category as the correct scene, two foil scenes from the other category, and two verbal category labels (“Mountains”, “Forests”) for if the participant remembers the scene category but not the exact scene. D.

Example category test trial. Participants were asked to guess the correct scene category for the training exemplars intermixed randomly with new (generalization) exemplars.

Methods

Healthy young adult participants (n = 45) were recruited from the University of Oregon and the surrounding community of Eugene, Oregon. Due to a coding error, demographic information for 10 participants was not saved. Of the 35 participants for which demographic information was saved (21 females), ages ranged from 18 to 27 ($M = 20.3 \pm 2.65$) years. Participants were compensated with $25. All participants provided written informed consent and experimental procedures were approved by the Institutional Review Board at the University of Oregon. We excluded 3 participants from all analyses due to guessing (accuracy less than or equal to 50%) during the categorization test (n = 2) and collinear GLM regressors (n = 1), defined as Pearson correlation coefficients greater than 0.8 (see below).

Materials

Stimuli. The stimulus set used in the current study is freely available through the Open Science Framework (https://osf.io/va5p8). Stimuli were cartoon animals that varied along eight binary feature dimensions: color (yellow/gray), shape of feet (clawed/webbed), shape of body (squared/circular), shape of tail (devil tail/feather tail), dot orientation on body (vertical/horizontal), pattern on neck (stripes/horns), head shape (with beak/with horn), and head orientation (forward/up). Two of these stimuli with the maximally different number of features served as the prototypes for Category A and Category B for all participants. The number of distinct features between any two given stimuli was defined as physical dissimilarity (i.e., distance), making the inverse equivalent to physical, or perceptual, similarity. Category A stimuli

shared more features with Category A's prototype than Category B's prototype and vice versa. Equidistant stimuli were not used.

Training set. Four training sets were created a priori and randomly assigned to participants. Each training set included four stimuli per category (eight total). Notably, prototypes themselves were never presented during training. Moreover, the majority of training exemplars differed from each other by four features regardless of whether they belonged to the same or opposing category, meaning that the similarity structure did not fully dictate category membership. Category membership could, however, be learned incidentally via associated scenes (see below). Each training exemplar differed from its prototype by two features and these two features were different for each exemplar. This ensures that three out of four exemplars within a category share any given prototype feature (i.e., during training, each feature is 75% diagnostic of category membership). This encourages attention to all features and biases participants away from using unidimensional rule-based strategies for categorizing. Training sets following the exact same structure have been used previously (Bowman & Zeithamova, 2018; Houser et al., 2024).

Training scenes. Each training exemplar was paired with an exemplar-unique scene (eight unique scenes in total). The scenes were distinctive from each other but drawn from two categories: mountains or forests. Training exemplars that belonged to one category were all paired with scenes from one scene type, or habitat, while training exemplars that belonged to the other category were all paired with scenes from the other habitat. There were eight total exemplar-scene associations used during training.

Testing set. Four testing sets were generated a priori and assigned based on which training set a participant was administered. Each testing set included the eight training exemplars

(each presented twice) and 50 novel exemplars. Novel exemplars consisted of 8 randomly selected stimuli at each distance from the category prototypes (i.e., eight new exemplars one feature different from prototype A, eight new exemplars two features different from prototype A, and so on), excluding equidistant stimuli. Test sets also included the two previously unseen prototypes, each shown twice. Repeating the training exemplars and prototypes enables better ability to dissociate between prototype and exemplar model predictions (Bowman & Zeithamova, 2018; Kéri et al., 2001; Smith et al., 2008).

Procedure

Training. Participants first completed a training session, where they were instructed to remember the background scene associated with each training exemplar. Notably, while each training exemplar and each scene was unique, all scenes fell into two categories: mountains or forests, with the habitat type aligning with category structure. That is, all four Category A exemplars were paired with one of the four unique forest scenes and all four Category B exemplars were paired with one of the four mountain scenes. Participants were not informed about the category alignment with habitat types. Instead, the task resembled a paired-associate learning task, typically used in studies of episodic memory.

On each trial, participants were shown one of eight background scenes that filled the screen. The associated training exemplar then took 2 seconds to materialize by gradually increasing the image's opacity. Participants were then prompted to report how confident they were that they could remember the scene habitat for the presented training exemplar using a button press ("1 = Definitely Forget", "2 = Maybe Forget", "3 = Maybe Remember", or "4 = Definitely Remember"). Participants were given 5 seconds to make a confidence judgment but the trial advanced if a response was registered sooner. Participants completed a total of five

training blocks, where each block consisted of five repetitions of all training exemplars (8 exemplars x 5 repetitions x 5 blocks = 200 total training trials). Participants were allowed to take self-paced breaks between blocks. The order of stimulus presentation was randomized within each block and repetition.

Associative memory test. Following training, participants were administered an associative memory test, where they were shown each training stimulus along with four scene options and instructed to choose which scene was paired with the presented stimulus during training. Scene options always consisted of two mountain and two forest scenes: the correct scene, a lure of the same habitat type, and two foils from the other habitat type. Thus, choices provided did not offer any cues regarding which habitat type is correct. Participants were also given the option to respond "Mountains" or "Forests" if they could not recall the specific scene.

Category generalization test. Lastly, participants entered the scanner where they were administered a surprise generalization test. They completed a total of four runs while undergoing fMRI. Each run comprised 17 trials, where a stimulus was presented for 5 seconds followed by a 7 second intertrial interval. On each trial, participants were instructed to categorize the stimulus/exemplar presented into one of two categories: mountains or forests using a button press made on an MRI compatible button box.

FMRI data acquisition. Neuroimaging was done using a 3T Siemens MAGNETOM Skyra at the University of Oregon Lewis Center for Neuroimaging using a 32-channel head coil. Head motion was minimized via foam padding and images were projected onto a screen at the head of the bore using a mirror attached to the head coil. The scanning session began with a localizer scan, followed by a T1-weighted MPRAGE anatomical image (TR = 2500 ms; TE = 3.43 ms; TI = 1100 ms; flip angle = 7°; matrix size = 256x256; 176 contiguous slices; FOV =

256 mm; slice thickness = 1mm; voxel size = 1x1x1 mm; GRAPPA factor = 2) and four functional runs using a multiband gradient echo pulse sequence (TR = 2000 ms; TE = 26 ms; flip angle = 90°; matrix size = 100x100; 72 contiguous slices oriented 15° off the anterior commissure-posterior commissure line to reduce prefrontal signal dropout; interleaved acquisition; FOV = 200 mm; voxel size = 2x2x2 mm; GRAPPA factor = 2).

Statistical analysis.

Prospective memory judgments during training. Prospective memory judgments, or confidence ratings, during training were collected to ensure continued attention during the incidental learning and were not of primary interest; however, we conducted a repeated measures ANOVA to see if participants' confidence in their memory for the exemplar-scene associations increased over time. For each participant and block, mean confidence was calculated. We expected to see an effect of training block. Notably, 5 participants had missing confidence ratings in one or more blocks due to technical error. To ensure this did not affect the ANOVA results, we also ran a linear mixed effects model which can handle missing data.

Associative memory. To assess how accurate people could recall specific exemplar-scene associations, the proportion of correct scene responses during the associative memory test was calculated for each participant. For completeness, we also computed and report the full distribution of responses (i.e., proportion of lure scene, foil scene, and correct and incorrect general category label decisions).

Categorization accuracy during generalization test. To evaluate how well participants could generalize incidentally learned category information, we calculated average correct responses for each participant and compared these averages to chance level (50%) using a single sample t-test. To then test the degree to which generalization follows a gradient over distance

from prototypes, average accuracy was calculated across all novel stimuli at each distance from its prototype. These scores were then submitted to a repeated measures ANOVA to test for an effect of distance on generalization performance. Additionally, an "old advantage" was tested by subtracting from the categorization accuracy for old, training exemplars accuracy for new exemplars at the same distance from the prototypes (two feature difference, as all training exemplars varied from their respective prototype by two features). The significance of old advantage scores was tested with a single sample t-test against 0. An old advantage, or greater accuracy for training exemplars than novel ones at the same distance, indicates a contribution of the memory for specific instances to categorization and is often observed even when behavior is otherwise dominated by prototype-based categorization strategies (Bowman & Zeithamova, 2018; Houser et al., 2024). Given that participants were instructed to remember specific events, it was predicted that participants would demonstrate a significant old advantage.

Categorization strategies during generalization test. To identify the generalization strategy each participant used to categorize exemplars during the test phase, prototype and exemplar models were fit to the trial-by-trial response data. Prototype models assume that category structures are represented by their prototype. Therefore, the prototype model computes the similarity between each generalization test stimulus and prototypes. Perceptual similarity is modeled with an exponential decay function of physical similarity (Shepard, 1958, 1987), while taking into account differences in attention to specific features. Specifically, the similarity S of a test stimulus x to Category A is:

$$S_A(x) = \exp\left[-c \sum_{m=1}^{M} \left(w_m \left| x_m - proto_{A,m} \right|\right)\right], \quad \text{Equation 1}$$

where c is a sensitivity parameter that determines the rate of similarity decay with physical distance. W is a vector of attention weights with length equal to the number of stimulus features and each feature is indexed by m.

Exemplar models assume that categories are represented by their exemplars and thus calculate similarity between each stimulus and a category by computing the summed similarity to all training exemplars j from a given category (Nosofsky, 1986, 1987):

$$S_A(x) = \sum_{y \in A} \exp\left[-c \sum_{m=1}^{M} \left(w_m \left| x_m - j_{y,m} \right|\right)\right]. \quad \text{Equation 2}$$

The probability that x is categorized in Category A is then obtained by transforming similarity with a Luce choice rule (Luce, 1963):

$$P(A|x) = \frac{S_A(x)}{S_A(x) + S_B(x)}. \quad \text{Equation 3}$$

In total, there are nine participant-specific free parameters for both models estimated from their pattern of responses: one c parameter and eight attention weights, one for each stimulus feature.

Using these equations, best fitting parameters were obtained from both models per participant via minimizing the negative log likelihood with gradient descent. This was done using the *Rsolnp::solnp* function (Ghalanos & Theussl, 2015) in R, which optimizes model fit with general nonlinear optimization augmented with Lagrange multipliers. An equality constraint was imposed on attention weights, such that they had to sum to 1. Parameter lower bounds were all set to 0, with an upper bound of 1 for the attention weights and 20 for the sensitivity parameter.

Model selection. Following the optimization procedure, model fits were compared at the individual level. Several studies have indicated that group-level trends in generalization can be obtained by superimposing subgroup trends (F. G. Ashby et al., 1994; J. C. Lee et al., 2018; J. C. Lee & Livesey, 2018; Lovibond et al., 2019), highlighting the need for individualized assessment of generalization behavior (Zaman et al., 2023). As such, model fits were compared to each other and to chance to determine the model best characterizing a given participant. A Monte Carlo simulation approach was employed, as described previously (Bowman & Zeithamova, 2018; Houser et al., 2024; see Zeithamova et al., 2026for a tutorial).

Participants could be assigned one of four possible strategy labels: exemplar, prototype, similar fit, or random. First, empirical responses during the generalization test were shuffled 5000 times and model fits were recalculated after each permutation (Zeithamova et al., 2025). This yields participant-specific null distributions for both model fits. If the empirically observed fits were better than (i.e., less than) 95% of the simulated fits, we concluded that the model describes the data better than chance ($p < 0.05$, one-tailed). If neither model fit reliably better than chance, the participant was labeled "random". For the remaining participants, prototype and exemplar models were compared directly by comparing their relative fit difference

$$\frac{fit_{exemplar} - fit_{prototype}}{fit_{exemplar} + fit_{prototype}} \qquad \text{Equation 4}$$

to the relative fit differences of simulated fits. "Fit" in Equation 4 corresponds to negative log likelihood. One model was deemed a winner if the difference in empirical fits appeared by chance with a frequency of <25% ($p < 0.25$, two-tailed). This approach is more suitable for model comparison than simply comparing raw fits because the exemplar model is inherently more flexible and it facilitates differentiation between negligible model fit differences and more

meaningful ones (Bowman et al., 2020; Bowman & Zeithamova, 2018). It is also preferred to more standard parameter penalization approaches, such as AIC and BIC, as it better approximates the additional flexibility each freely estimated parameter adds to the models (Zeithamova et al., 2025).

fMRI preprocessing. fMRI images were preprocessed using *fMRIPrep* 23.1.4 (Esteban et al., 2019). T1-weighted (T1w) images were corrected for intensity non-uniformity and skull stripped using *antsBrainExtraction.sh.* Brain-extracted T1w images were segmented into cerebrospinal fluid, white matter, and brain matter using *fast* (Zhang et al., 2001). Cortical surfaces were reconstructed via *recon-all* from FreeSurfer 7.3.2 (Dale et al., 1999) and then normalized via volume-based spatial normalization to MNI space.

For each of the four BOLD runs per participant, a reference volume was generated and used to estimate head-motion parameters via *mcflirt* (Jenkinson et al., 2002). Fieldmaps were aligned to the target EPI reference run and used to perform slice-time correction via *3dTshift* from AFNI (Cox & Hyde, 1997). The BOLD reference was co-registered to the T1w reference via *bbregister* (Greve & Fischl, 2009). Framewise displacement was calculated per run and used to exclude individual runs with too much head motion. Images were then noise corrected using a principal component-based method (Avants et al., 2009; Behzadi et al., 2007), which generates probabilistic brain masks that are resampled into BOLD space and binarized by thresholding at 0.99. Finally, the BOLD timeseries was then resampled into participant-native space using the *antsApplyTransforms* function (Avants et al., 2009). BOLD timeseries images were entered into the FEAT (fMRI Expert Analysis Tool) in FSL for high-pass temporal filtering (100 s) and spatial smoothing using a 5 mm FWHM kernel. For group level analyses, functional data were

registered to standard stereotaxic space following coregistration with the T1 image using FLIRT (FMRIB's Linear Image Registration Tool) in FSL.

Regions of interest.

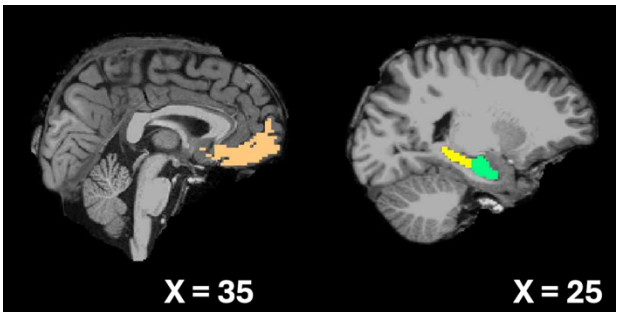

Figure 4.2. Regions of interest from a representative participant.

Regions were defined in native space using the automated segmentation tool in Freesurfer.

Regions of interest (ROIs) were defined anatomically in each participant's native space using the cortical (vmPFC) and subcortical (hippocampus) segmentation from Freesurfer 6. ROIs were collapsed across hemispheres (Figure 4.2). vmPFC was expected to track prototype representations based on its role in episodic generalization (Zeithamova et al., 2012) and previous work finding precisely that (Bowman et al., 2020; Bowman & Zeithamova, 2018; Seger et al., 2025). Previous research has additionally indicated that generalization in the hippocampus may be specific to its anterior half, while the posterior hippocampus tracks more fine-grained information (Collin et al., 2015; Poppenk et al., 2013; Poppenk & Moscovitch, 2011; Schlichting et al., 2015; Strange et al., 2014). Thus, anterior/posterior differences were tested by dividing each participant's hippocampal ROI into anterior and posterior halves, splitting at the middle slice. If a participant had an odd number of slices, the middle slice was assigned to the posterior portion. Following previous work (Bowman et al., 2020; Bowman & Zeithamova, 2018), the anterior hippocampus was predicted to track prototypes while the posterior hippocampus was predicted to track exemplars.

Model-based fMRI analysis. Voxel-wise univariate activation analyses were employed to evaluate brain activity induced by category representations with the general linear model (GLM).

A GLM with a design matrix including a total of 9 regressors was constructed for each run per participant. The 9 regressors included one item regressor, two parametric regressors derived from the exemplar and prototype models, and six motion regressors. The model-based parametric predictors corresponded to the (within-run demeaned) summed similarity to both categories. Prototype- and exemplar-based summed similarity to both categories corresponds to the denominator in Equation 3 for each respective model. Summed similarity indexes how similar a given stimulus during the generalization test is to both category representations combined, and thus, beta coefficients for each term measure how much evidence there is for one strategy versus the other, regardless of which category the stimulus actually belongs to (Figure 4.3A). Similar regressors have been used in several previous fMRI studies to dissociate generalization strategies successfully (Bowman et al., 2020; Bowman & Zeithamova, 2018; Davis et al., 2017; Mack et al., 2013). Including both exemplar and prototype regressors in the same GLM accounts for shared variance between both predictions. Additionally, exploratory analyses modeling both predictors in separate GLMs yielded qualitatively the same results. Moreover, participants (n = 1) were excluded if any pair of their regressors had correlation coefficients greater than 0.8, averaged across runs (Figure 4.3B shows the distribution of regressor correlations). This resulted in the exclusion of 1 participant. Events were modeled with a 5 second duration regardless of response times. Onsets were convolved with the canonical hemodynamic response function as implemented in FSL. This is the same GLM used in a prior intentional category learning study with identical stimulus structure (Bowman & Zeithamova, 2018) allowing for easy comparison.

ROI analyses were conducted in native space. Mean parameter estimates from each region were extracted per run per participant and divided by the corresponding standard deviation of parameter estimate to derive the effect size of how much the BOLD signal tracked

each model predictor. 12 individual runs were excluded due to incompleteness (n = 3) and framewise displacement being greater than 2mm (n = 9). Normalized βs were then averaged across runs for each participant and submitted to group level analyses. Given our a priori hypotheses, we ran three single sample t-tests comparing each ROI's βs to implicit baseline (baseline estimate = 0). P-values were Bonferroni corrected for multiple comparisons ($\alpha = 0.05 / 3 = 0.016$).

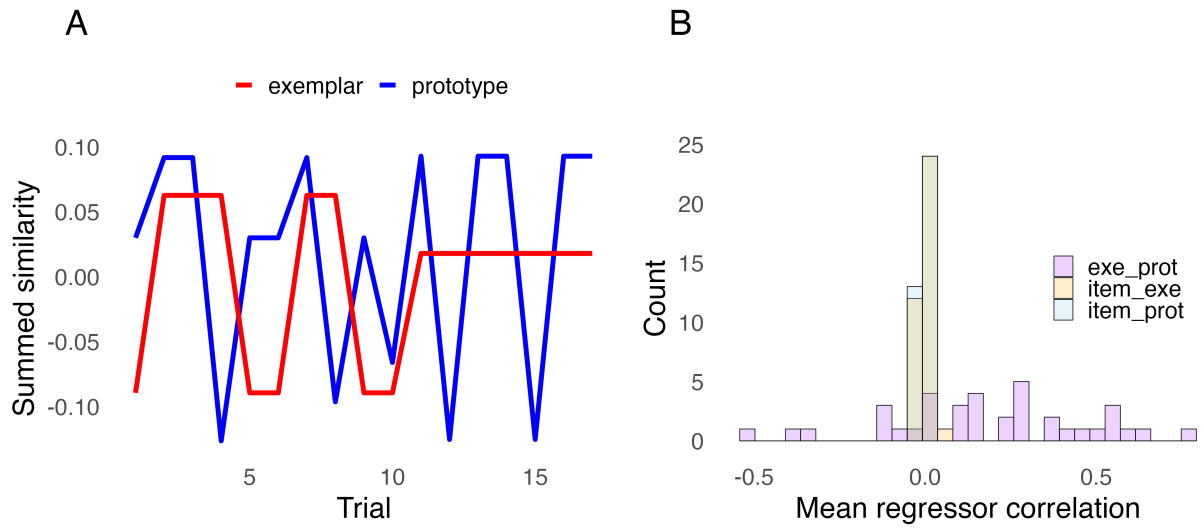

Figure 4.3. Summed similarity estimates and their correlations.

A. Summed similarity estimates (mean-centered) for a representative participant during a single block of the categorization test (17 trials). These estimates were entered as regressors into first level GLMs to model parametric modulation of the BOLD signal. B. Pairwise Pearson correlation coefficient for all combinations of fMRI GLM regressors (after filtering out participants with correlations higher than 0.8), averaged across categorization test blocks.

Exploratory whole brain analysis. Whole brain maps were created by averaging parameter estimates across runs (with less than 2mm of framewise displacement) within each participant. These participant level estimates were then used to estimate group-level contrasts using mixed effects FLAME 1 in FSL, comparing both prototype and exemplar effects to baseline to reveal clusters of voxels tracking their predictions. Whole brain maps were computed with a voxelwise threshold of $z > 2.3$ and cluster corrected at $p < 0.05$.

Exploratory brain-behavior analysis. Given that our previous work indicated that our novel incidental category learning paradigm shifted people's reliance to be more exemplar-based when generalizing, we tested the degree to which exemplar representations in the brain correlated with associative memory. Specifically, we ran two Spearman correlation tests evaluating the relationship between neural correlates of exemplar representations, derived from the ROI analysis, and associative memory for both people that relied on exemplars and those that relied on prototypes when generalizing.

Results

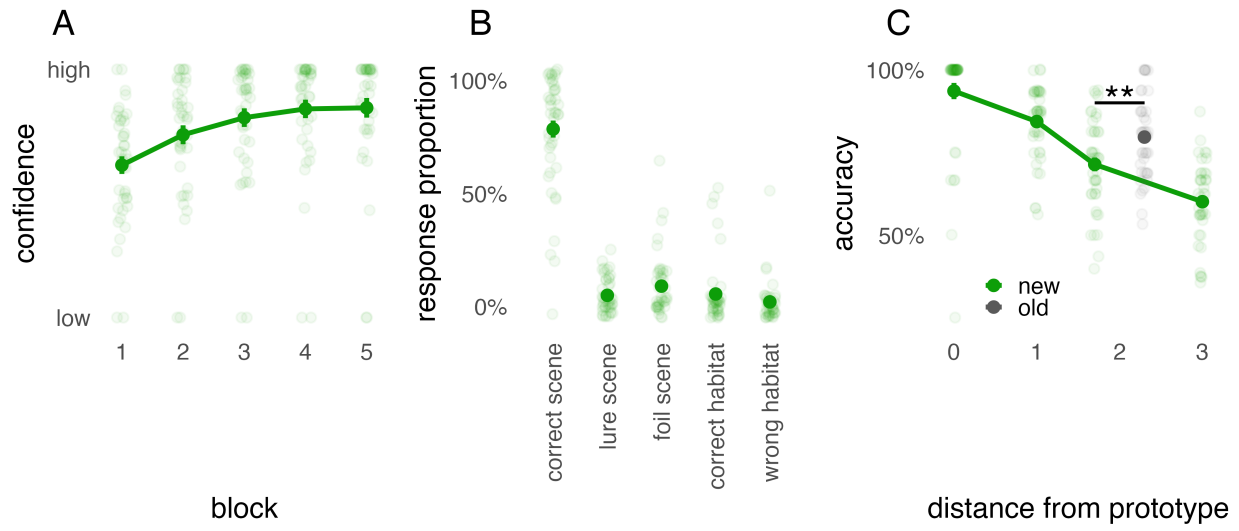

Figure 4.4. Behavioral performance.

A. Mean confidence ratings for each training block. B. Response distribution for associative memory. C. Categorization accuracy per each distance from the prototypes. Thick lines and larger dots indicate group averages. Smaller dots denote individual participants. Error bars represent standard errors of the mean. **p < 0.01.

Prospective memory judgments during training

There was a main effect of block ($F(4, 144) = 44.7$, $\eta_p^2 = 0.55$, $p < 0.001$), indicating participants trained to remember specific associations became more confident in their ability to

remember these associations over time (Figure 4.4A). The linear mixed effects model supported this conclusion (main effect of block $\beta = 0.35$, $p < 0.001$).

Associative memory

As in our previous work, people chose the correct scene during the associative memory test most often (M = 78% SD = 25%). This means that participants successfully paired training exemplars with their unique scene in memory. The full distribution of responses is visualized in Figure 4.4B.

Categorization accuracy during generalization test

During the generalization test, overall accuracy was well above chance (M = 75%, SD = 8%; one sample t-test against 50%: $t(42) = 19.5$, 95% CI = [0.73, 0.78], $p < 0.001$). Generalization of category knowledge to novel exemplars followed a typicality gradient, such that accuracy was best for the never-before-seen prototypes and worst for exemplars furthest from their prototypes (Figure 4.4C). This was clarified with a strong main effect of distance from the prototype ($F(3, 126) = 84.50$, $\eta_p^2 = 0.67$, $p < 0.001$).

Participants also demonstrated an old advantage (M = 0.08, SD = 0.17), indicating that they could categorize training exemplars significantly better than novel exemplars equated by distance from prototypes ($t(42) = 3.18$, 95% CI = [0.03, 0.13], $p = 0.003$). This advantage for categorizing training stimuli, however, did not extend to novel exemplars more similar to the prototype (mean difference in accuracy between distance of 1 exemplars and training exemplars = 5%; paired samples t-test: $t(42) = 2.22$, $p = 0.032$) or the prototypes themselves (mean difference in accuracy between prototypes and training exemplars = 14%; paired samples t-test: $t(42) = 4.96$, $p < 0.001$).

Categorization strategies

The distribution of participant labels, as assigned via our model selection procedure described above, is presented in Figure 4.5A. Overall, nobody showed signs of prototype and exemplar model fits being comparable. There was a single participant for which neither model fit better than chance. However, this participant was kept in subsequent analyses because they exhibited some learning (associative memory = 87.5%, categorization of training exemplars = 69%; overall test accuracy = 53%). 23 of the 43 participants (~53%) relied on prototypes, while 19 (~44%) relied on exemplars. This pattern of results is consistent with those found in Houser et al. (2024), where more people than in other previous studies were found to rely on exemplars for generalization. Notably, this is qualitatively distinct from the 10% of participants that relied on exemplars when trained with feedback-based learning (Bowman & Zeithamova, 2018).

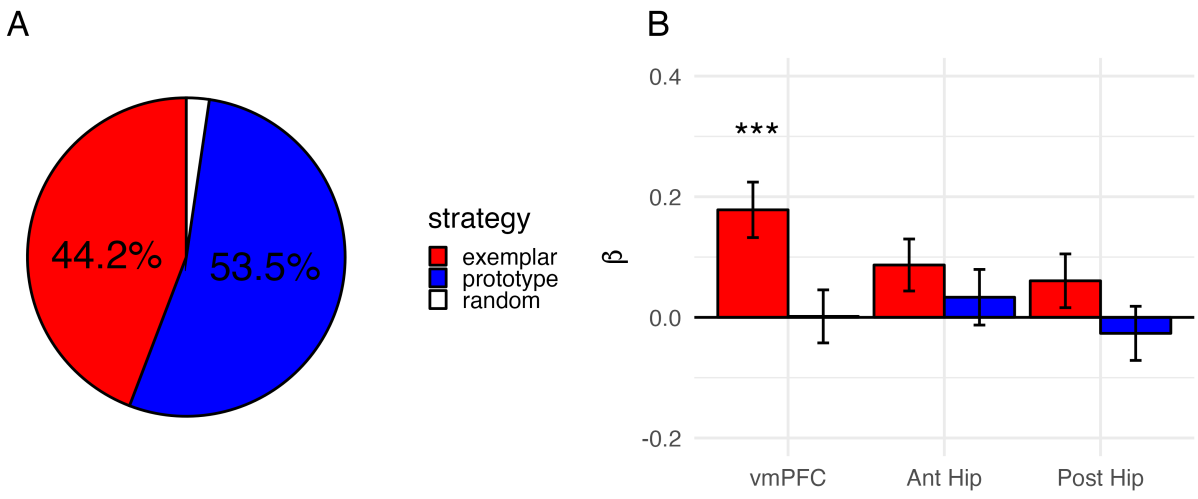

Figure 4.5. Generalization strategies in behavior and brain.

A. Distribution of winning models. Most participants were deemed "prototypists", though a substantial proportion were deemed "exemplarists". One participant was not fit well by either model (random). B. Normalized beta estimates for both models per each a priori selected brain region of interest. Error bars are standard errors of the mean. ***$p < 0.001$.

ROI analysis

Prototype and exemplar parameter estimates (normalized β values) for each a priori selected ROI are visualized in Figure 4.5B. Single sample t-tests revealed above-chance exemplar correlates in the vmPFC (M = 0.18, t(42) = 3.88, 95% CI = [0.09, 0.27], p < 0.001). Exemplar correlates in the vmPFC were also significantly greater than prototype correlates in the vmPFC (paired samples t-test: t(42) = 2.43, 95% CI = [0.03, 0.32], p = 0.020). There was no robust evidence that the anterior (M = 0.15, t(42) = 2.02, 95% CI = [-0.0001, 0.17], $p_{corrected}$ = 0.151) or posterior (M = 0.06, t(42) = 1.36, 95% CI = [-0.03, 0.15], p = 0.543) hippocampus tracked exemplar predictions. No ROI showed evidence of tracking prototype predictions (all uncorrected *p*s > 0.47).

Whole brain analysis

At the group level, there was one cluster in the right lateral parietal cortex (z = 3.81, peak MNI_{xyz}: 64, -20, 30; 458 voxels) that significantly tracked prototype model predictions (Figure 6). We additionally observed two clusters that significantly tracked exemplar model predictions (Figure 4.6). One cluster was in the left lateral parietal cortex (z = 3.78, peak MNI_{xyz}: -56, -44, 46; 128 voxels) and the other was in the vmPFC, specifically in the right frontal polar region (z = 3.33, peak MNI_{xyz}: 22, 56, -12; 102 voxels).

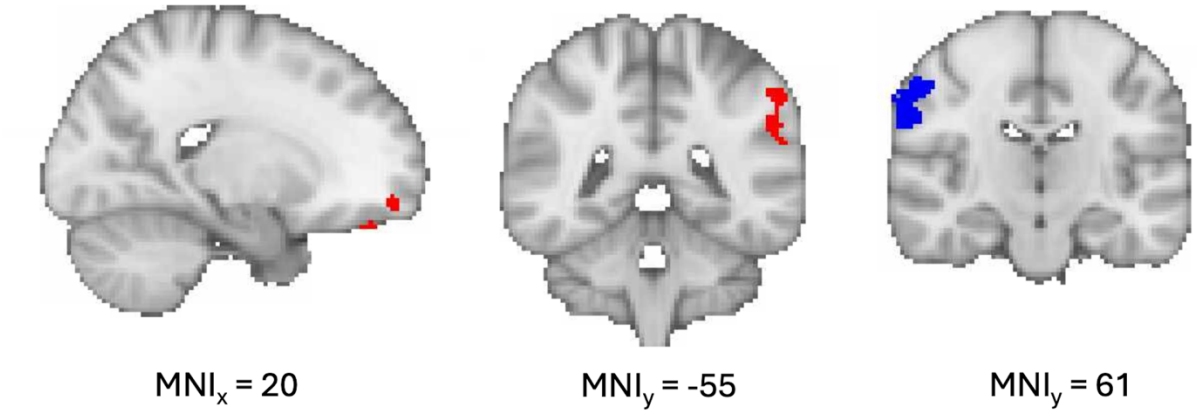

Figure 4.6. Clusters exhibiting significant activation in response to model predictions.

The left and middle images reveal voxels representing exemplars during the categorization test. The cluster in the left image is in the vmPFC, and particularly the right frontal pole. The cluster in the middle image is in the left lateral parietal cortex. The cluster on the right is in the right lateral parietal cortex and denotes prototype representations. Red = exemplars, blue = prototypes.

There were no clusters that significantly tracked negative exemplar predictions; however, there were seven clusters tracking negative prototype predictions (Table 4.1). Clusters spanned a distributed network across frontal, parietal, and visual regions (Figure 4.7).

Cluster	Anatomic region included in cluster	Voxels	Peak *Z*	MNI(*x*, *y*, *z*)	*P*
1	Anterior Cingulate Cortex	1860	5.08	0, 22, 36	<0.0001
2	Precuneus	908	4.23	-2, -74, 36	<0.0001
3	Frontal Pole / vmPFC	807	4.64	8, 70, -6	<0.0001
4	Lateral Occipital Cortex	502	3.56	40, -68, 36	0.001
5	Orbitofrontal Cortex	392	5.21	-30, 22, -8	0.008
6	Posterior Cingulate Cortex	336	4.10	0, -32, 34	0.022
7	Occipital Pole	296	3.80	-18, -104, -2	0.044

Table 4.1. Clusters significantly tracking negative prototype model predictions.

Results were whole-brain cluster corrected at a cluster-forming threshold of $z>2.3$ and $p<0.05$.

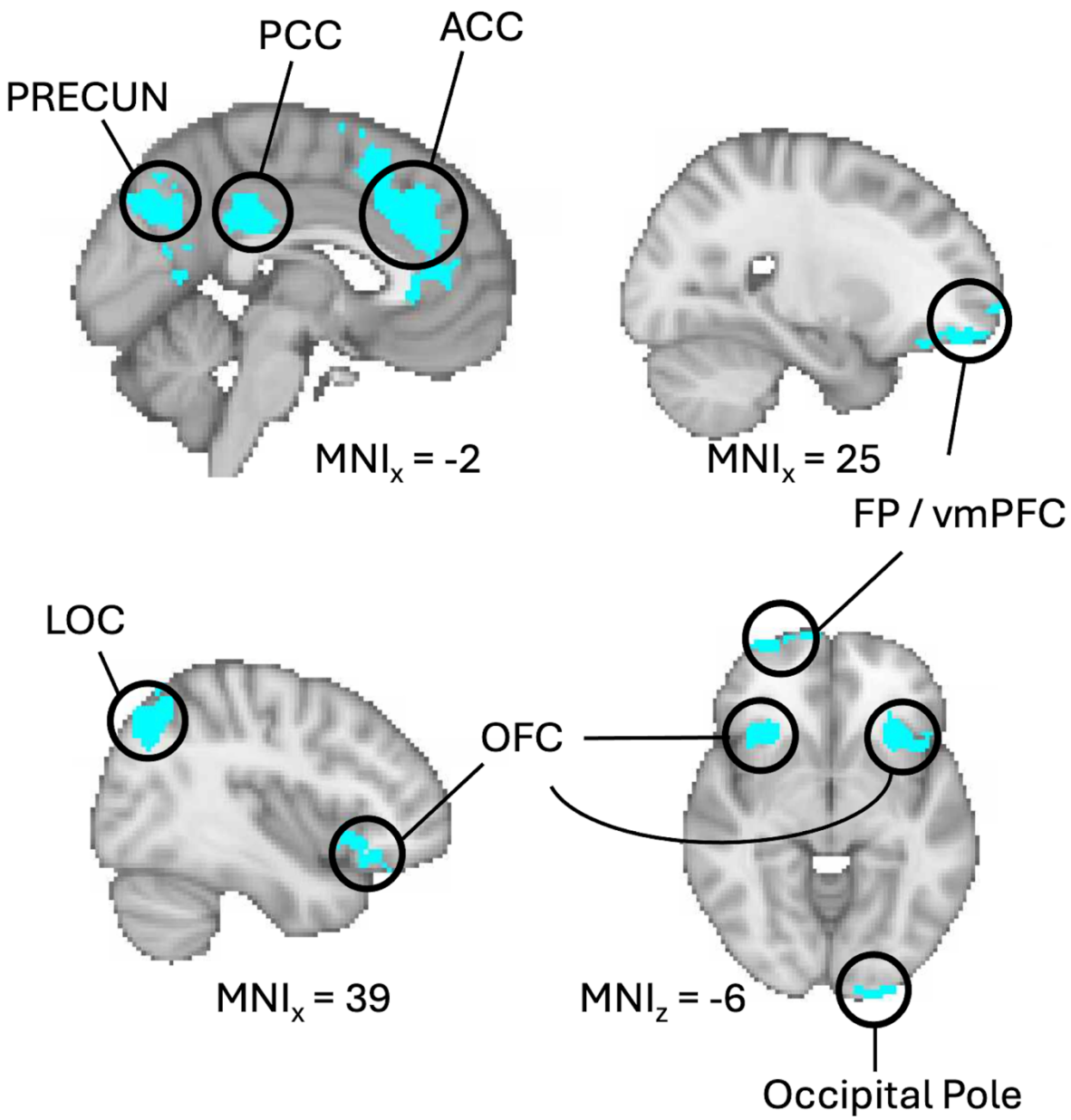

Figure 4.7. Visualization of clusters significantly tracking dissimilarity to the prototype.

PRECUN=precuneus; PCC=posterior cingulate cortex; ACC=anterior cingulate cortex; FP=frontal pole; vmPFC=ventromedial prefrontal cortex; LOC=lateral occipital cortex; OFC=orbitofrontal cortex.

Brain-behavior relationship

We tested the degree to which associative memory related to exemplar representations in the vmPFC (Figure 4.8). There was a strong positive relationship between these measures in

people who relied on exemplars when generalizing (rho = 0.58, p = 0.018, bootstrapped 95% CI = [0.17, 0.83]) and no evidence of a relationship in people who relied on prototypes when generalizing (rho = -0.07, p = 0.758) (Figure 4.8A). Notably, the significance of the relationship in exemplarists survives Bonferroni correction. However, because this analysis included only 19 participants, we sought to further assess the robustness of this result by rerunning Spearman correlation tests using a leave-one-subject-out analysis (Figure 4.8B). All 19 tests were significant (all *p*s < 0.024) and effect sizes were highly similar to the full sample effect size (M_{rho} = 0.58, SD = 0.04, range = [0.53, 0.70]). This confirms that no single exemplarist was driving the significant relationship between associative memory and vmPFC exemplar representations.

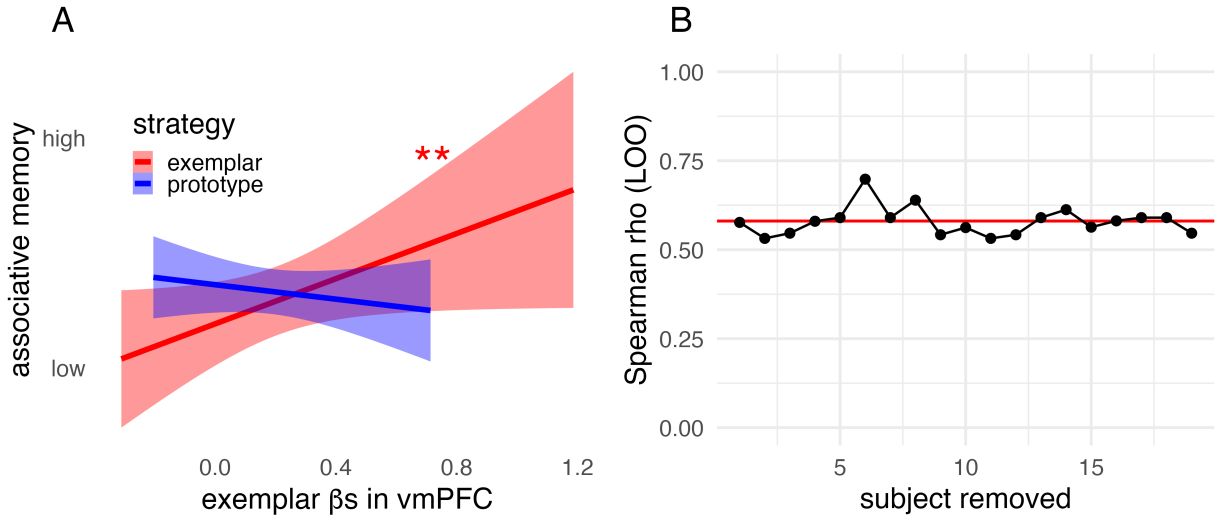

Figure 4.8. Brain-behavior relationship.

A. Correlation between exemplar representations in the vmPFC, indexed by normalized beta estimates, and associative memory, indexed by the proportion of correct scene responses during the associative memory test, separated by generalization strategy (exemplar, prototype). Ribbons denote 95% confidence intervals. **p < 0.01. B. Spearman correlation coefficients between exemplar representations in the vmPFC and associative memory for exemplarists after removing each participant one-by-one. Red line denotes the empirically observed correlation (as visualized in A).

Discussion

Here, we employed a novel category learning paradigm that resembles incidental learning, given that participants never received instruction to assign training exemplars to categories or corrective feedback. Unlike prior work on incidental learning, however, category membership was not uniquely determined by similarity. Additionally, emphasis was placed on memorizing paired associates and people indeed learned these associations, evidenced by their highly accurate associative memory. Despite this, people categorized old and new exemplars during a surprise categorization test accurately. People showed generalization gradients that are similar to those found in traditional category learning studies. Together, these results suggest that several aspects of feedback-based category learning are the same under more naturalistic demands. We further examined neural correlates of exemplars and prototypes while people generalized category knowledge and found exemplar representations in the vmPFC.

While an overwhelming majority of category learning research evaluates the acquisition of category knowledge by training people to learn categories via explicit feedback following responses, incidental category learning is common in the real world (Clapper, 2006; Clapper & Bower, 1994, 2002; Love, 2002, 2003; Unger & Sloutsky, 2022). That is, people do not typically receive instructions to search for features of experience that are diagnostic of category membership or experimenter-provided feedback. Beyond ecological validity, the investigation of incidental category learning is important because there is growing evidence that it draws on computational and neural underpinnings that are distinct from the more traditional feedback-based ones (Doya, 1999; Lim et al., 2019; Tricomi et al., 2006). Most of this prior work on incidental category learning, however, utilized auditory stimuli and evaluated category representations during acquisition of category knowledge (Tricomi et a., 2006; Lim et al., 2019).

In the current study, we assessed signatures of feedback-based category learning (e.g., generalization gradients) following explicit training to memorize exemplar-scene associations using visual stimuli. We found that people still abstracted out category structure across the associations and generalized this information to novel exemplars.

Consistent with our prior work (Houser et al., 2024), we found that more people relied on training exemplars when generalizing than expected if feedback was provided during category learning (Bowman & Zeithamova, 2018). Houser et al. (2024) directly compared proportions of exemplar versus prototype strategies across incidental and feedback-based category learning and found marginally significant evidence for a shift from prototype reliance in the feedback-based group to exemplar reliance in the incidental group. This was with 55% of the sample relying on prototypes in the incidental group and 31% relying on exemplars. Here, we found 54% of participants relied on prototypes and 44% relied on exemplars. Thus, the emphasis on paired associates, while keeping category knowledge intact, does seem to shift people to a more exemplar-based generalization strategy, highlighting the notion that people's category representations are highly flexible and can adapt to task demands (Markman & Ross, 2003).

Previous work from the lab using the same stimuli and category structure but with feedback and without background scenes found prototype representations in the vmPFC and anterior hippocampus (Bowman & Zeithamova, 2018), suggesting these brain regions support abstract category information (also see Bowman et al., 2020; Bowman & Zeithamova, 2025; Seger et al., 2025). Given the similar behavioral markers of generalization following incidental learning and the maintained majority of prototype users, we expected similar neural substrates in the current study. Unexpectedly, we found exemplar representations in the vmPFC. Intriguingly, recent studies have yielded converging results.

Mack et al. (2020) observed that the vmPFC category representations change with category structure, such that the low neural compression of information led to more distributed attention and high compression led to selective attention to a single feature. This study indicates that the vmPFC can flexibly represent categories across a continuum of representational granularity, perhaps indicating that it can represent prototypes and exemplars.

Blank and Bayer (2022) trained participants to learn a category with feedback, employing the dot pattern prototype distortion paradigm, which creates a prototype by placing an array of dots on a grid and connecting the dots to form a polygon (Posner & Keele, 1968). The polygon's vertices are then jittered randomly to create category exemplars. These exemplars are used to teach participants the extent of the category and then a generalization test has people judge whether old and new exemplars are part of the category or not. Interestingly, using this paradigm, Blank and Bayer (2022) observed more people that relied on training exemplars than prototypes when generalizing. This study utilized a multivariate approach to characterizing exemplar and prototype representations in the brain, which involved first obtaining neural representations of each training stimulus. Then, all training stimulus representations were correlated, creating an empirical representational similarity matrix, which was correlated with prototype and exemplar model matrices. The prototype model matrix assumes that all exemplars within the category are represented as equivalent in the brain while the exemplar model matrix assumes each exemplar is represented distinctly and thus exemplar-to-exemplar similarities should be decodable. The empirical representational similarity matrix obtained from vmPFC activity was significantly correlated with the exemplar model matrix (Blank & Bayer, 2022), indicating converging results when using a multivariate approach. Notably, however, this study used an A/not A paradigm, which could have also differentially shaped neural correlates.

Another study trained participants to learn rule-based and unstructured categories, where unstructured categories required participants to memorize individual exemplar-category label associations (Schlegelmilch et al., 2026). This study observed a positive correlation between outcome processing and vmPFC activity when learning the unstructured category, which converges with the finding that hippocampal-vmPFC functional connectivity supports dynamic updating of exemplar-based category representations (Mack et al., 2016). While these studies scanned participants during learning, they nevertheless entail that, like the current work, specific category representations can be supported by vmPFC. Consistently, Bowman and Zeithamova (2026) found both exemplar and prototype representations in the vmPFC while people generalized category knowledge. Additionally, the vmPFC has been shown to contain a mixture of integrated and separated representations during associative inference (Schlichting et al., 2015). Together, this study suggests that the role of vmPFC in category learning and generalization is highly flexible, being able to represent prototypes, as past work has shown, and exemplars following incidental category learning, as indicated here.

Exploratory group-level analyses further revealed both exemplar and prototype representations in the lateral parietal cortex, a region that has been implicated in stimulus-specific and categorical memory representations (Freedman & Assad, 2006; Kuhl & Chun, 2014; Liu et al., 2025; Sestieri et al., 2017; Yago & Ishai, 2006; Zhong et al., 2019). The representational differences in the current study mapped onto hemispheric differences, such that the left lateral parietal cortex tracked exemplars while the right lateral parietal cortex tracked prototypes. This hemispheric distinction is unexpected and will need to be tested more explicitly in future studies. Some accounts do advocate for visual representational differences between hemispheres rather than confining these differences to post-visual processing (Marsolek, 1995,

1999). One explanation for prototype representations in the right parietal cortex is that this signal reflects familiarity, as previous lesion studies have shown right hemisphere lesioned patients are impaired at processing familiar stimuli (Langguth et al., 2009; Marzi & Berlucchi, 1977; but see Gauthier et al., 1997 for conflicting results). However, other work detected exemplar representations in the right hemisphere's visual processing stream (Jüttner et al., 2021). Together, the evidence for hemispheric differences in occipitoparietal regions between prototype and exemplar representations remains mixed.

Lastly, several brain regions tracked negative prototype model predictions, meaning these regions were more active for atypical exemplars, or exemplars closer to the category boundary. The significantly modulated regions spanned frontal (frontal pole, vmPFC, anterior cingulate, and orbitofrontal cortex), parietal (posterior cingulate and precuneus), and visual (lateral occipital cortex and occipital pole) regions. This effect may reflect increased attention allocated to less diagnostic features. That is, while more typical exemplars can be encoded with less precision because their category assignment will not change even with some noise, atypical exemplars need sharpened representations, as any imprecision could lead to them being categorized inaccurately given their closeness to the category boundary. One study found that category learning facilitated more precise representations of exemplars near the category boundary than exemplars near the category center (O'Bryan et al., 2024; also see Folstein et al., 2013). Other studies have additionally implicated occipitoparietal (Bowman & Zeithamova, 2025b; Seger et al., 2015) and occipitotemporal (Braunlich et al., 2017) activation in tracking negative prototype model predictions. Several prefrontal regions were also responsible for tracking negative prototype model predictions in the current study. Negative prototype model predictions can also be framed as closeness to the category boundary, which has consistently

been shown to be represented in the prefrontal cortex using rule-based category structures (Esposito et al., 2025; Liu et al., 2023; Mansouri et al., 2020; Reinert et al., 2021). Thus, near-boundary exemplars may have recruited increased frontoparietal regions for more control over the decision boundaries. Notably, the current study observed a more widespread network of regions tracking negative prototype model predictions than in Bowman and Zeithamova (2026) where participants trained with a feedback-based category learning paradigm, which converges with the notion that incidental category learning emphasizes perceptual features and exemplar-unique perceptual feature combinations.

Conclusion

The current study administered an incidental category learning paradigm and tested participants on their ability to generalize category structure to novel exemplars while undergoing fMRI. We found that people's generalization performance closely mirrored typical generalization gradients observed following feedback-based category learning, but that people's generalization strategy emphasized more reliance on memorized exemplars than abstracted prototypes. We additionally found exemplar representations in the vmPFC, where prototypes have been observed previously. This stark contrast from previous work could suggest that the incidental category learning paradigm used here biases category representations in the vmPFC toward more fine-grained details. The lack of prototype correlates could have been due to interference from scene representations in the category signal, necessitating future studies to obtain more detailed neural signatures of the representations formed following our paired associates category learning task.

References cited

Aizenstein, H. J., MacDonald, A. W., Stenger, V. A., Nebes, R. D., Larson, J. K., Ursu, S., & Carter, C. S. (2000). Complementary category learning systems identified using event-related functional MRI. *Journal of Cognitive Neuroscience*, *12*(6). https://doi.org/10.1162/08989290051137512

Ashby, F. G., Maddox, W. T., & Lee, W. W. (1994). On the dangers of averaging across subjects when using multidimensional scaling or the similarity-choice model. *Psychological Science*, *5*(3). https://doi.org/10.1111/j.1467-9280.1994.tb00651.x

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 & Review*, *27*(4), 791–800. https://doi.org/10.3758/s13423-020-01754-3

Ashby, S. R., & Zeithamova, D. (2022). Category-Biased Neural Representations Form Spontaneously during Learning That Emphasizes Memory for Specific Instances. *Journal of Neuroscience*, *42*(5). https://doi.org/10.1523/JNEUROSCI.1396-21.2021

Avants, B., Tustison, N., & Johnson, H. (2009). Advanced Normalization Tools (ANTS). *Insight Journal*.

Behzadi, Y., Restom, K., Liau, J., & Liu, T. T. (2007). A component based noise correction method (CompCor) for BOLD and perfusion based fMRI. *NeuroImage*, *37*(1). https://doi.org/10.1016/j.neuroimage.2007.04.042

Blank, H., & Bayer, J. (2022). Functional imaging analyses reveal prototype and exemplar representations in a perceptual single-category task. *Communications Biology*, *5*(1). https://doi.org/10.1038/s42003-022-03858-z

Bowman, C. R., Iwashita, T., & Zeithamova, D. (2020). Tracking prototype and exemplar representations in the brain across learning. *ELife*, *9*, 1–47. https://doi.org/10.7554/eLife.59360

Bowman, C. R., & Zeithamova, D. (2018). Abstract memory representations in the ventromedial prefrontal cortex and hippocampus support concept generalization. *Journal of Neuroscience*, *38*(10), 2605–2614. https://doi.org/10.1523/JNEUROSCI.2811-17.2018

Bowman, C. R., & Zeithamova, D. (2025). *Ventromedial prefrontal cortex supports prototype representations in healthy older adults*. https://doi.org/10.1101/2025.09.24.678354

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

Braunlich, K., Liu, Z., & Seger, C. A. (2017). Occipitotemporal category representations are sensitive to abstract category boundaries defined by generalization demands. *Journal of Neuroscience*, *37*(32). https://doi.org/10.1523/JNEUROSCI.3825-16.2017

Clapper, J. P. (2006). When more is less: Negative exposure effects in unsupervised learning. *Memory and Cognition*, *34*(4). https://doi.org/10.3758/BF03193435

Clapper, J. P., & Bower, G. H. (1994). Category Invention in Unsupervised Learning. *Journal of Experimental Psychology: Learning, Memory, and Cognition*, *20*(2). https://doi.org/10.1037/0278-7393.20.2.443

Clapper, J. P., & Bower, G. H. (2002). Adaptive Categorization in Unsupervised Learning. *Journal of Experimental Psychology: Learning Memory and Cognition*, *28*(5). https://doi.org/10.1037/0278-7393.28.5.908

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

Cox, R. W., & Hyde, J. S. (1997). Software tools for analysis and visualization of fMRI data. *NMR in Biomedicine*, *10*(4–5). https://doi.org/10.1002/(SICI)1099-1492(199706/08)10:4/5<171::AID-NBM453>3.0.CO;2-L

Dale, A. M., Fischl, B., & Sereno, M. I. (1999). Cortical surface-based analysis: I. Segmentation and surface reconstruction. *NeuroImage*, *9*(2). https://doi.org/10.1006/nimg.1998.0395

Davis, T., Goldwater, M., & Giron, J. (2017). From concrete examples to abstract relations: The rostrolateral prefrontal cortex integrates novel examples into relational categories. *Cerebral Cortex*, *27*(4). https://doi.org/10.1093/cercor/bhw099

Doya, K. (1999). What are the computations of the cerebellum, the basal ganglia and the cerebral cortex? *Neural Networks*, *12*(7–8). https://doi.org/10.1016/S0893-6080(99)00046-5

Esposito, M., Abdul, L. S., Ghouse, A., Rodríguez Aramendía, M., & Kaplan, R. (2025). Flexible hippocampal representation of abstract boundaries supports memory-guided choice. *Nature Communications*, *16*(1), 2377. https://doi.org/10.1038/s41467-025-57644-6

Esteban, O., Markiewicz, C. J., Blair, R. W., Moodie, C. A., Isik, A. I., Erramuzpe, A., Kent, J. D., Goncalves, M., DuPre, E., Snyder, M., Oya, H., Ghosh, S. S., Wright, J., Durnez, J., Poldrack, R. A., & Gorgolewski, K. J. (2019). fMRIPrep: a robust preprocessing pipeline for functional MRI. *Nature Methods*, *16*(1). https://doi.org/10.1038/s41592-018-0235-4

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

Freedman, D. J., & Assad, J. A. (2006). Experience-dependent representation of visual categories in parietal cortex. *Nature*, *443*(7107). https://doi.org/10.1038/nature05078

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). https://doi.org/10.1037/xhp0000073

Gabay, Y., & Holt, L. L. (2015). Incidental learning of sound categories is impaired in developmental dyslexia. *Cortex*, *73*. https://doi.org/10.1016/j.cortex.2015.08.008

Gabay, Y., Madlansacay, M., & Holt, L. L. (2023). Incidental auditory category learning and visuomotor sequence learning do not compete for cognitive resources. *Attention, Perception, and Psychophysics*, *85*(2). https://doi.org/10.3758/s13414-022-02616-x

Gauthier, I., Anderson, A. W., Tarr, M. J., Skudlarski, P., & Gore, J. C. (1997). Levels of categorization in visual recognition studied using functional magnetic resonance imaging. *Current Biology*, *7*(9). https://doi.org/10.1016/S0960-9822(06)00291-0

Ghalanos, A., & Theussl, S. (2015). *Rsolnp: General Non-linear Optimization Using Augmented Lagrange Multiplier Method* (1.16). R.

Greve, D. N., & Fischl, B. (2009). Accurate and robust brain image alignment using boundary-based registration. *NeuroImage*, *48*(1). https://doi.org/10.1016/j.neuroimage.2009.06.060

Houser, T. M., Resnick, A., & Zeithamova, D. (2024). Successful generalization of conceptual knowledge after training to remember specific events. *Frontiers in Cognition*, *3*. https://doi.org/10.3389/fcogn.2024.1324678

Jenkinson, M., Bannister, P., Brady, M., & Smith, S. (2002). Improved optimization for the robust and accurate linear registration and motion correction of brain images. *NeuroImage*, *17*(2). https://doi.org/10.1016/S1053-8119(02)91132-8

Jüttner, M., Strasburger, H., Treutwein, B., Landis, T., & Rentschler, I. (2021). Complementary deficits in perceptual classification in pure alexia and acquired prosopagnosia – New insights from two classic cases. *Neuropsychologia*, *155*. https://doi.org/10.1016/j.neuropsychologia.2021.107820

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

Kuhl, B. A., & Chun, M. M. (2014). Successful remembering elicits event-specific activity patterns in lateral parietal cortex. *Journal of Neuroscience*, *34*(23). https://doi.org/10.1523/JNEUROSCI.4328-13.2014

Langguth, B., Jüttner, M., Landis, T., Regard, M., & Rentschler, I. (2009). Differential impact of posterior lesions in the left and right hemisphere on visual category learning and generalization to contrast reversal. *Neuropsychologia*, *47*(13). https://doi.org/10.1016/j.neuropsychologia.2009.06.023

Lee, J. C., Hayes, B. K., & Lovibond, P. F. (2018). Peak shift and rules in human generalization. *Journal of Experimental Psychology: Learning Memory and Cognition*, *44*(12). https://doi.org/10.1037/xlm0000558

Lee, J. C., & Livesey, E. J. (2018). Rule-based generalization and peak shift in the presence of simple relational rules. *PLoS ONE*, *13*(9). https://doi.org/10.1371/journal.pone.0203805

Lim, S. J., Fiez, J. A., & Holt, L. L. (2019). Role of the striatum in incidental learning of sound categories. *Proceedings of the National Academy of Sciences of the United States of America*, *116*(10). https://doi.org/10.1073/pnas.1811992116

Liu, Z., Liao, S., & Seger, C. A. (2023). Rule and Exemplar-based Transfer in Category Learning. *Journal of Cognitive Neuroscience*, *35*(4). https://doi.org/10.1162/jocn_a_01963

Liu, Z., Zhang, Y., Wen, C., Yuan, J., Zhang, J., & Seger, C. A. (2025). Emergence of Categorical Representations in Parietal and Ventromedial Prefrontal Cortex across Extended Training. *Journal of Neuroscience*, *45*(9). https://doi.org/10.1523/JNEUROSCI.1315-24.2024

Love, B. C. (2002). Comparing supervised and unsupervised category learning. *Psychonomic Bulletin and Review*, *9*(4). https://doi.org/10.3758/BF03196342

Love, B. C. (2003). The multifaceted nature of unsupervised category learning. *Psychonomic Bulletin and Review*, *10*(1). https://doi.org/10.3758/BF03196484

Lovibond, P. F., Lee, J. C., & Hayes, B. K. (2019). Stimulus Discriminability and Induction as Independent Components of Generalization. *Journal of Experimental Psychology: Learning Memory and Cognition*. https://doi.org/10.1037/xlm0000779

Luce, R. D. (1963). A Threshold Theory for Simple Detection Experiments. *Psychological Review*, *70*(1). https://doi.org/10.1037/h0039723

Mack, M. L., Love, B. C., & Preston, A. R. (2016). Dynamic updating of hippocampal object representations reflects new conceptual knowledge. *Proceedings of the National Academy of Sciences of the United States of America*, *113*(46). https://doi.org/10.1073/pnas.1614048113

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). https://doi.org/10.1016/j.cub.2013.08.035

Mack, M. L., Preston, A. R., & Love, B. C. (2020). Ventromedial prefrontal cortex compression during concept learning. *Nature Communications*, *11*(1). https://doi.org/10.1038/s41467-019-13930-8

Mansouri, F. A., Freedman, D. J., & Buckley, M. J. (2020). Emergence of abstract rules in the primate brain. In *Nature Reviews Neuroscience* (Vol. 21, Number 11). https://doi.org/10.1038/s41583-020-0364-5

Markman, A. B., & Ross, B. H. (2003). Category Use and Category Learning. In *Psychological Bulletin* (Vol. 129, Number 4). https://doi.org/10.1037/0033-2909.129.4.592

Marsolek, C. J. (1995). Abstract Visual-Form Representations in the Left Cerebral Hemisphere. *Journal of Experimental Psychology: Human Perception and Performance*, *21*(2). https://doi.org/10.1037/0096-1523.21.2.375

Marsolek, C. J. (1999). Dissociable neural subsystems underlie abstract and specific object recognition. *Psychological Science*, *10*(2). https://doi.org/10.1111/1467-9280.00117

Marzi, C. A., & Berlucchi, G. (1977). Right visual field superiority for accuracy of recognition of famous faces in normals. *Neuropsychologia*, *15*(6). https://doi.org/10.1016/0028-3932(77)90005-7

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

Nosofsky, R. M. (1986). Attention, similarity, and the identification-categorization relationship. *Journal of Experimental Psychology: General*, *115*(1). 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). https://doi.org/10.1037/0278-7393.13.1.87

O'Bryan, S. R., Jung, S., Mohan, A. J., & Scolari, M. (2024). Category Learning Selectively Enhances Representations of Boundary-Adjacent Exemplars in Early Visual Cortex. *Journal of Neuroscience*, *44*(13). https://doi.org/10.1523/JNEUROSCI.1039-23.2023

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

Poppenk, J., & Moscovitch, M. (2011). A hippocampal marker of recollection memory ability among healthy young adults: Contributions of posterior and anterior segments. *Neuron*, *72*(6). https://doi.org/10.1016/j.neuron.2011.10.014

Posner, M. I., & Keele, S. W. (1968). ON THE GENESIS OF ABSTRACT IDEAS. *Journal of Experimental Psychology*, *77*(3 PART 1). https://doi.org/10.1037/h0025953

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

Reinert, S., Hübener, M., Bonhoeffer, T., & Goltstein, P. M. (2021). Mouse prefrontal cortex represents learned rules for categorization. *Nature*, *593*(7859). https://doi.org/10.1038/s41586-021-03452-z

Roark, C. L., Lehet, M. I., Dick, F., & Holt, L. L. (2022). The Representational Glue for Incidental Category Learning Is Alignment With Task-Relevant Behavior. *Journal of Experimental Psychology: Learning Memory and Cognition*, *48*(6). https://doi.org/10.1037/xlm0001078

Schlegelmilch, R., Dinu, A., Joue, G., Gläscher, J., & Sommer, T. (2026). Neurocomputational modeling of rule abstraction and memorization during probabilistic stimulus-reward learning. *IScience*, *29*(1). https://doi.org/10.1016/j.isci.2025.113950

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*. https://doi.org/10.1038/ncomms9151

Seger, C. A., Braunlich, K., Wehe, H. S., & Liu, Z. (2015). Generalization in category learning: The roles of representational and decisional uncertainty. *Journal of Neuroscience*, *35*(23). https://doi.org/10.1523/JNEUROSCI.0654-15.2015

Seger, C. A., Xie, W., Li, P., Han, L., & Liu, Z. (2025). Complex Rule Transfer Recruits Rostral Prefrontal and Ventromedial Prefrontal Neural Networks. *Human Brain Mapping*, *46*(12). https://doi.org/10.1002/hbm.70327

Sestieri, C., Shulman, G. L., & Corbetta, M. (2017). The contribution of the human posterior parietal cortex to episodic memory. In *Nature Reviews Neuroscience* (Vol. 18, Number 3). https://doi.org/10.1038/nrn.2017.6

Shepard, R. N. (1958). Stimulus and response generalization: Tests of a model relating generalization to distance in psychological space. *Journal of Experimental Psychology*, *55*(6). https://doi.org/10.1037/h0042354

Shepard, R. N. (1987). Toward a universal law of generalization for psychological science. *Science*, *237*(4820). https://doi.org/10.1126/science.3629243

Smith, J. D., & Minda, J. P. (1998). Prototypes in the mist: The early epochs of category learning. *Journal of Experimental Psychology: Learning Memory and Cognition*, *24*(6). https://doi.org/10.1037/0278-7393.24.6.1411

Smith, J. D., Redford, J. S., & Haas, S. M. (2008). Prototype Abstraction by Monkeys (Macaca mulatta). *Journal of Experimental Psychology: General*, *137*(2). https://doi.org/10.1037/0096-3445.137.2.390

Strange, B. A., Witter, M. P., Lein, E. S., & Moser, E. I. (2014). Functional organization of the hippocampal longitudinal axis. In *Nature Reviews Neuroscience* (Vol. 15, Number 10). https://doi.org/10.1038/nrn3785

Tricomi, E., Delgado, M. R., McCandliss, B. D., McClelland, J. L., & Fiez, J. A. (2006). Performance feedback drives caudate activation in a phonological learning task. *Journal of Cognitive Neuroscience*, *18*(6). https://doi.org/10.1162/jocn.2006.18.6.1029

Unger, L., & Sloutsky, V. M. (2022). Ready to Learn: Incidental Exposure Fosters Category Learning. *Psychological Science*, *33*(6). https://doi.org/10.1177/09567976211061470

Wattenmaker, W. D. (1993). Incidental Concept Learning, Feature Frequency, and Correlated Properties. *Journal of Experimental Psychology: Learning, Memory, and Cognition*, *19*(1). https://doi.org/10.1037/0278-7393.19.1.203

Yago, E., & Ishai, A. (2006). Recognition memory is modulated by visual similarity. *NeuroImage*, *31*(2). https://doi.org/10.1016/j.neuroimage.2005.12.014

Zaman, J., Yu, K., & Verheyen, S. (2023). The idiosyncratic nature of how individuals perceive, represent, and remember their surroundings and its impact on learning-based generalization. *Journal of Experimental Psychology: General*. https://doi.org/10.1037/xge0001403

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). https://doi.org/10.1016/j.neuron.2012.05.010

Zeithamova, D., Houser, T., & Bowman, C. (2025). *Monte Carlo approaches to model selection: Application to the prototype and exemplar models of categorization*. https://doi.org/10.31234/osf.io/vhm7r_v2

Zeithamova, D., Houser, T. M., & Bowman, C. R. (2026). Evaluating Cognitive Models With Permutation Testing: A Case Study of Prototype and Exemplar Categorization. *Advances in Methods and Practices in Psychological Science*, *9*(2). https://doi.org/10.1177/25152459261430729

Zhang, Y., Brady, M., & Smith, S. (2001). Segmentation of brain MR images through a hidden Markov random field model and the expectation-maximization algorithm. *IEEE Transactions on Medical Imaging*, *20*(1). https://doi.org/10.1109/42.906424

Zhong, L., Zhang, Y., Duan, C. A., Deng, J., Pan, J., & Xu, N. long. (2019). Causal contributions of parietal cortex to perceptual decision-making during stimulus categorization. *Nature Neuroscience*, *22*(6). https://doi.org/10.1038/s41593-019-0383-6

Chapter 5

General discussion

Category learning has long been a central focus of cognitive science because of its ubiquity in everyday life. Its voluminous literature has benefitted from some of the most prominent thinkers of history. In particular, we can demarcate major shifts in the study of category learning with Plato and Aristotle, who both sought out the objective nature of categories apart from any mental influence, and Ludwig Wittgenstein, whose work led to considering how categories in the world are represented in the mind. One view focuses on objective categories in the real world while the other focuses on subjective categories in the mind of the beholder. However, it is perhaps not appreciated at first glance how even the second approach assumes an objective nature to categories. That is, the second approach assumes that there is a "correct" way to *represent* categories. The current dissertation makes no assumption of objective categories in the world or correct categories in the mind. Instead, here we set out to determine the degree to which category representations are flexible. Specifically, I tested how the world, or experience, carves out categories in the mind and how this process differs depending on category structure and the manner in which one learns categories. The present dissertation aimed 1) to understand how variability within categories shapes whether people rely on shared versus stimulus-unique features to generalize to novel information, 2) to evaluate the extent to which incidental category learning biases people to use exemplars to generalize in behavior and (3) the brain.

How experience plus goals shape category learning

Variability in experience is inevitable and a fundamental aspect of life. Every tree that we come across is a little bit different. Even the same tree will change from day to day. What does

this variance do to our category representations? It has been hypothesized to facilitate broader generalization to novel stimuli (Raviv et al., 2022). Thus, seeing several highly distinguishable trees is expected to facilitate better generalization to a novel, atypical tree. However, one explanation for evidence in favor of this notion has suggested that it is due to an imbalance in training trials (Hintzman, 1986). That is, because typical category learning studies investigating effects of variability train people until they reach a performance criterion, highly variable category training often requires many more training trials than less variable training. Nascent work shows that, when training duration is equated, less variable training instances facilitates broader generalization (Bowman & Zeithamova, 2020, 2023, 2025a; Hu & Nosofsky, 2024).

What do these results mean for how variable experiences shape representations? It suggests that, under time constraints, variable instances do not provide enough scaffolding for supporting category knowledge. Even when variability is category diagnostic, as in Chapter 2, rather than mere noise, people trained with less variability generalize more successfully. One explanation is that people trained with more variable instances memorize each instance but cannot link them into a coherent category structure because of the reduced number of shared features. This is unlikely to be the case because recognition memory of variable instances is poor compared to that for less variable instances. Specifically, people trained with more variable instances falsely remembered seeing more typical instances more often than they correctly recognized their own training instances. Further, people trained with more variable instances relied on shared features much more frequently to categorize ambiguous faces than they relied on exemplar-unique features.

These findings point toward a different conclusion from what I hypothesized. Specifically, it was predicted that highly variable training would lead to people relying on unique

features for generalization. Due to a more challenging abstraction process, it was assumed that these people would have to depend on memorizing exemplar-unique information. Instead, they relied on shared features, just like the people trained with less variable instances, albeit to a lesser degree. In other words, training variability led to very similar *kinds* of category representations, though the representations following less variability were of much better quality, facilitating better performance across training, generalization, and recognition.

These findings therefore cannot be explained solely as a function of category structure in the world. The way in which people engaged with the information seemingly contributed much. That is, it seems likely that people trained to group information together default to a strategy of attempting to abstract out common features from across multiple exemplars. This means that, rather than use the unique features to their advantage, both groups ignored the unique features because of the goal of learning categories. Given the built-in difference of shared features between groups, this goal naturally led to stark performance differences.

Another interpretation of these results is that abstraction is automatic (Murphy, 2016) and cannot be overridden by the world or a combination of experience and goals. Indeed, many recent studies on category learning indicate that abstract summaries dominate people's representations of categories. Building off incidental category learning research, in Chapter 3 we tested how generalization strategy might change following memorization of paired associates. Unbeknownst to participants, the associates had a similarity structure, enabling incidental acquisition of category knowledge. Notably, these people did not have the goal of abstracting out shared features to form a category. Using computational modeling, I revealed a trend of people's generalization strategy shifting from prototype-based (i.e., abstract) following feedback-based learning to exemplar-based following incidental learning.

The work from Chapter 3 therefore underscores the flexibility of category representations, rather than an automatic abstraction process, highlighting how a more supervised category learning paradigm, where goals of assigning exemplars to correct categories are clear, leads to prototype reliance and how incidental category learning, where goals are believed to be memorization of paired associates, leads to exemplar reliance. An intriguing future direction would be to evaluate the degree to which feedback-based versus incidental category learning facilitate category representations at the time of encoding versus retrieval. That is, when goals of categorization are clear, abstract summary representations can form rapidly as people learn (Bowman et al., 2020; Schapiro et al., 2018; Sučević & Schapiro, 2023). However, it is less clear whether this should be the case during incidental learning: discrete associations could have been learned during training after which within-category exemplars are flexibly retrieved to meet categorization demands or discrete associations could have been learned while organizing exemplars based on similarity in a latent psychological space.

Does the specific tradeoff with the abstract?

Substantial work in category learning and episodic memory suggests that abstract and specific representations tradeoff with one another, such that better abstraction is tied to worse specificity and vice versa (Boyle et al., 2024; Hindy et al., 2026; Lutz et al., 2017). One crucial thread of research in support of this notion draws on systems consolidation theory, where specific memories encoded in the hippocampus get transferred to the neocortex during sleep, and in the process, become more gist-like and abstract (McClelland et al., 1995). Other frameworks claim that the hippocampus can learn both the abstract and the specific (Schapiro et al., 2017; Sučević & Schapiro, 2023), though evidence for this in a single study is lacking. In fact, this was

a core motivation for Chapter 2, where it was predicted that more variability would lead to more reliance on shared features than unique features and that less variability would lead to more reliance on unique features than shared features. Having instead found that both groups rely on shared features, one conclusion would be that a tradeoff is still possible but that everyone was simply biased toward using shared features and thus it could not be observed. However, the experimental design of Chapter 2 enables evaluation of this notion from a different angle: across tasks rather than groups. That is, if a tradeoff between the abstract and the specific exists, people who relied on shared features should have poor recognition memory. While the case can indeed be made that both groups did false alarm quite a bit, they both also correctly recognized their training exemplars. Moreover, we would expect that people who relied on shared features *more*, false alarmed more, but instead people who relied most often on shared features (the less variable training group) had better recognition.

While Chapter 2 therefore provided no clear indication of a tradeoff, Chapter 3 highlights a couple of positive signs. First, people that used an exemplar-based generalization strategy had better associative memory. This is evidence of a tradeoff because it shows that people that formed abstract category representations (i.e., built prototypes) had worse memory specificity. It is therefore possible that the tradeoff exists at the item level but not the feature level.

Incidental category learning leads to exemplar representations in the ventromedial prefrontal cortex

Previous work in the lab has elucidated prototype representations in the ventromedial prefrontal cortex (vmPFC) (Bowman et al., 2020; Bowman & Zeithamova, 2018), which is consistent with this brain region's prominent role in schema representations (Gilboa & Marlatte,

2017). Specifically, a wealth of research indicates that vmPFC is involved in abstract representations (Cortese et al., 2021; De Martino & Cortese, 2023; Desai et al., 2018; Ghosh & Gilboa, 2014; Gilboa & Marlatte, 2017; Moneta et al., 2024; Vaidya & Badre, 2022), which encompasses many functions including memory integration (Spalding et al., 2018), relational memory (Wing et al., 2021), conjunctive representations (Binder, 2016), coarse grained temporal representations (Brunec & Momennejad, 2022), and abstract cognitive maps (D. Chen et al., 2021; Liang et al., 2023; Veselic et al., 2025). However, emerging evidence underscores vmPFC's role beyond abstraction and conceptualization to include more fine-grained cognitive representations, such as in vivid, scene-based imagination (Achmed Ali et al., 2026; Ciaramelli et al., 2019; Lieberman et al., 2019; McCormick et al., 2018; St-Laurent et al., 2015) or specific category representations (Bowman & Zeithamova, 2025; Mack et al., 2020; Schlegelmilch et al., 2026). Together, this emerging literature combined with the fMRI study presented in Chapter 4 suggest that vmPFC function flexibly shifts depending on task demands.

This result also elucidates behavioral and neural differences in category representations. In Chapter 2, the representations that people formed following differently variable training instances did not differ in kind, only quality, and in Chapter 3, while more people used an exemplar-based generalization strategy following incidental than feedback-based learning, there was still a large number of participants using the prototype-based strategy. Despite these subtle differences in behavior, Chapter 4 elucidates much more prominent neural differences in category representations, suggesting that representational changes may be more pronounced in the brain than is visible in behavior.

Furthermore, the study in Chapter 4 points to a network of brain regions that respond more strongly to atypical than typical exemplars, perhaps reflecting greater allocation of

cognitive resources to enhance the signal-to-noise ratio. As we would expect under this framework, the network was comprised of frontoparietal and visual processing regions, which may indicate more attention and visual precision.

Broader implications

Quixotic categories

The current dissertation furthers our understanding of how generalization of category knowledge depends on the structure of our experience and provides evidence for how computational processes and neural activity can be used to predict category representations that people use to generalize. Deepening our understanding of how people generalize is important, for generalization is a ubiquitous phenomenon. Not only is generalization ubiquitous, but it is also critical for our survival. One turn down a creepy, dark alley will likely prevent us from frequenting such alleys, where danger may lurk. When visiting a new restaurant, we can expect a menu, a meal, and paying for the meal in that order because we can generalize knowledge acquired from previous restaurants. Generalization seems to guide our behavior from moment to moment. Despite its never-ending influence, an immense, scientific literature, and long, historical tradition of philosophical inquiry, generalization and categorization remain mysteries. This is no doubt in large part due to the fact that researchers must use indirect methods to measure them. Additionally, conceptual knowledge operates over an unusually large number of distinct cognitive domains, ranging from perception to language and thought to social and affective processes, making the pinning down of its nature an almost insurmountable feat.

Given this looming challenge, Barsalou (2016) has suggested that what he calls "Quixotic dead ends" are especially tempting to conclude. Quixotic dead ends are approaches to conceptual knowledge that are overly idealistic, while being impractical and unrealistic (Barsalou, 2016; hence *Quixotic*). For example, anti-representation accounts of concepts suggest that all conceptual knowledge can be reduced to sensorimotor contingencies (O'Regan & Noë, 2001), which is closely related to theories of affordances (Gibson, 2014), and grounded and embodied cognition (Varela et al., 2016; Weber & Varela, 2002). Such accounts lead to the conclusion that the brain cannot represent abstract concepts not grounded in the external world (Barsalou, 2008), which contradicts our knowledge of such things as geometry, which is not inherently tied to the external world. At the other end of this argument is the idea that concepts are amodal, operating in a space divorced from sensorimotor contingencies (Leshinskaya & Caramazza, 2016; Mahon, 2015), though empirical work overwhelmingly finds multimodal rather than amodal representations in the brain (Martin, 2016). Moreover, amodal concepts are essentially a black box, for it is not clear exactly what such a representation is (Barsalou, 2016).

Though not mentioned in Barsalou (2016), I would argue that the work from this dissertation underscores additional Quixotic dead ends wrought from shared vs. unique feature reliance and prototype vs. exemplar theory. Both areas of focus can be distilled to a single dissociation between category representations that are abstract and ones that are specific (which ironically may be mapped onto amodal versus sensorimotor, though this is beyond the scope of this discussion). Reliance on shared versus unique features speaks more to the dissociation between abstract and specific information during attention while prototype versus exemplar theory dissociates abstract and specific information in memory. This dissertation shows how

neither shared nor unique features are the sole focus of attention and that neither prototypes nor exemplars are the sole representations in memory.

Chapter 2 may suggest a subtle discrepancy between its results and the attention optimization account that dominates category research. The attention optimization account says that people allocate attention to features to optimize categorization accuracy (Nosofsky, 1986, 1987; Rehder & Hoffman, 2005a, 2005b). For example, if a set of stimuli have two features (color and shape), where color is the same for all members of a category but shape is random, then people should devote all their attention to color and ignore shape. Interestingly, the face stimuli used in Chapter 2 were comprised of category-shared and exemplar-unique features, but both were perfectly diagnostic of which category a given face belonged in. The attention optimization account predicts that people rely on shared and unique features roughly equally because they are all diagnostic; however, I still found that the overwhelming majority of attention was allocated to shared features. This points to the possibility that attention optimization is not a unidimensional optimization process of categorization accuracy but rather optimizes both accuracy and resource expenditure simultaneously. Specifically, while both shared and unique features can optimize accuracy, there are more unique features to attend to, thus requiring more effortful attention. Attending to shared features instead offers the same level of categorization accuracy while minimizing the amount of effortful attention. Nascent evidence in behavior, eyetracking, and neuroimaging supports this dual optimization approach to attention in category learning (Galdo et al., 2022; Gao et al., 2024; Weichart et al., 2024; Weichart, Evans, et al., 2022; Weichart, Galdo, et al., 2022).

This dual optimization approach to attention in category learning illustrates why relying on either shared or unique features to categorize (and why optimizing attention solely for

categorization accuracy) are Quixotic dead ends. There is no a priori answer to which features people will use to categorize and generalize. I found that people rely on shared features, perhaps because of their low cost, meaning if we were to impose additional costs on attending to shared features, people would be expected to shift attention to unique features. More work will be needed to confirm this empirical prediction.

Situating conceptual knowledge

Categorization and generalization are often studied in isolation (Barsalou et al., 2018). A word for a concept might be presented on a screen and information that becomes active in response is evaluated with behavioral and neural measures. As such, theories of categorization and generalization tend to focus on representational components underlying concepts cued in isolation. This ignores the role that concepts play as psychological elements built for our adaptive functioning. Assuming concepts serve a more (environmentally and biologically) situated role than organizing information for information's sake, studying concepts in isolation is likely to limit and distort what is known about them. In the extreme, studying concepts completely divorced from the real world will likely lead to the Aristotelian premise that all-or-none conditions are required for defining categories, in which case the only conceptual knowledge that is real are geometric forms. These forms, initially posited by Plato, do not serve a real world, biologically situated function. In and of themselves, they are irrelevant to our survival.

Most research in the cognitive sciences must sacrifice some natural validity, though this does not mean we cannot approximate it better than tradition. According to Barsalou et al. (2018), situational elements are integrated via temporal co-occurrence. As such, concepts

emerging via temporal co-occurrence may be one route to studying them in a more situated, biologically-relevant manner. It is not obvious why temporal co-occurrence is automatically biologically-relevant to category learning. This notion relies on a predictive processing account of categorization (Michel, 2023; also see Clark, 2013b, 2013a; Corcoran et al., 2020; K. Friston, 2010; K. J. Friston, 2019) that argues that information is categorized in the brain insofar as the category predicts goal-relevant information. If we combine the situational and predictive aspects, we can see that reliable temporal co-occurrence will be predictive in a given situation, and thus, lumping a co-occurrence into a category is a useful heuristic to serve adaptive and efficient behavior. For example, a birthday party is comprised of numerous sensory details, like the smell of the cake, the sounds of children screaming, the visuals of balloons and streamers, though their temporal co-occurrence allows the brain to group all these distinct details into a single category. The utility of this is that, in the future, the brain does not have to note every minute detail to understand the situation; it can instead recognize the category "birthday party" and proceed accordingly.

This bottom-up form of (statistical) learning comes with a caveat: The brain does not—at least initially—know what is important and what is not. Chapter 3 of this dissertation showed that learning categories through temporally co-occurring exemplar-scene information facilitated more exemplar-based generalization strategies, consistent with the notion that people's attention was distributed widely to memorize exemplar-unique details. Chapter 3 focused more on exemplar memory and exemplar-based generalization strategies, as opposed to their prototype counterpart, and thus more targeted studies on how temporal co-occurrence shapes attention will be useful for deepening our knowledge of incidental category learning.

Not knowing which temporal co-occurrences are important during encoding leads naturally to the question of how subsequent consolidation processes shape the incidentally-acquired category knowledge. For example, plenty of research indicates that sleep enhances gist-like abstraction across multiple experiences (Cowan et al., 2020; Durrant et al., 2015; Gilboa & Marlatte, 2017; Klinzing et al., 2019; Lewis & Durrant, 2011; Rasch & Born, 2013), meaning people might rely on prototypes for generalization if tested after a night of sleep. Other work indicates that weakly encoded features are prioritized for consolidation (Schapiro et al., 2018), which could further lead even more people to use exemplars when generalizing. A third option is that, because the brain does not know what is important, both shared and unique features are consolidated during sleep, leading to more people with comparable strategies. One recent study testing veridical and gist representations after incidental encoding and consolidation supports this possibility (Wernette & Fenn, 2023).

Back to Don Quixote

In the spirit of Quixotic dead-ends, let us now push this idea that category representations are flexible to its extreme. Rips (2009) asked people to judge whether a circular object of three inches was more similar to a quarter or a pizza. Most people said quarter. Then he asked them if it was more likely to be a quarter or a pizza and they said pizza! It has been rationalized that these diverging answers is the result of comparing the object based on similarity in the first question and based on theory in the second question (because pizzas are more likely to vary in size than quarters) (Michel, 2023). This is an example of just how dynamic our category representations can be. If the conceptual content represented for a category varies infinitely

across context, people, time, and space, how do we ever find common ground? How do multiple people share a category representation? How do we communicate?

One solution is realizing that, despite minute, often irrelevant discrepancies, experiences are quite similar in a given situation. People in the same neighborhood will see the same visual information, talk to the same people, be subject to the same social norms. The variance of category representations from person to person thus likely is a function of (social, temporal, and/or spatial) similarity. To illustrate, all else being equal, the same situation will likely generate more similar concepts between neighbors in Eugene, Oregon than between one person that lives in Eugene and one person that likes in Tokyo.

Goal-relevance is likely another factor that contributes to finding common ground. Detected discrepancies can be ignored in the service of a broader goal. For example, we understand spoken language despite regional accents because accents do not subtract from the communicability of the words being spoken. More relatedly, concepts like 'death' might have many different associations for two different cultures (American and Mexican) but saying somebody died can be effectively communicated between the two because the goal of communicative success does not depend on the discrepancies. Ergo, as long as concepts and constructs are similar *enough*, common ground can be found.

Technical implications

Research has consistently demonstrated the utility of mathematical modeling of cognitive processes and behavior (Farrell & Lewandowsky, 2010, 2018; M. D. Lee & Wagenmakers, 2013). Such models can predict future decision-making, unobservable psychological representations, and clinical disorders (Lewandowsky & Oberauer, 2018; Pitt et al., 2002). Here,

I used computational modeling to elucidate differences in generalization strategy between people who nevertheless behave qualitatively the same. Further, I elucidated how such generalization strategies differ under distinct category learning contexts. Across three empirical studies, compelling evidence is offered for how to bias individuals toward using exemplars to generalize category knowledge.

Moreover, functional magnetic resonance imaging (fMRI) is another technique used in the current dissertation that has great potential to better understand how the brain works and to improve diagnosis and treatment of neurobehavioral disorders (Detre & Wang, 2002). Using computational model-informed fMRI, I highlighted exemplar representations in the brain that also predicted measures of associative memory. This finding challenges the notion that vmPFC explicitly codes for abstract information and weakens proposals that specific and general memories are localized to distinct brain regions (F. G. Ashby et al., 1998; McClelland et al., 1995).

Limitations and Future Directions

The dissertation explored how category knowledge and generalization is shaped by task demands. While the current dissertation examined differences between exemplar and prototype representations, there are other models that may better capture people's behavior, such as SUSTAIN (Love et al., 2004), which is a cluster-based category learning model. SUSTAIN groups exemplars into clusters and represent these clusters as centroids, like the prototype model; however, when prediction errors are high, SUSTAIN recruits additional clusters, meaning, in the limit, it can reduce to the exemplar model, where a "cluster" is formed for each individual exemplar. Thus, SUSTAIN captures both ends of the spectrum of representational granularity in

addition to everything in between. This high degree of resolution may allow SUSTAIN to better capture idiosyncratic generalization patterns, which could be useful in characterizing behavior following incidental category learning.

The present dissertation has shown that vmPFC function may extend to specific memories; however, other fMRI approaches can be used to strengthen this finding. For example, comparing schema related exemplars to their schema is often underpinned by functional connectivity between vmPFC and the anterior hippocampus (Audrain & McAndrews, 2022; Guo et al., 2023; Sommer et al., 2022), and thus, assuming that the findings from Chapter 4 can be interpreted under this framework, functional connectivity analyses can be used for verification. Additionally, a multivariate approach to characterizing category representations in the brain could be used (Blank & Bayer, 2022), which is a more sensitive and more informative method for detecting functional organization than univariate analyses (Haxby, 2012).

General conclusion

Category learning is fundamental to human cognition. It has been a topic of central scientific focus for thousands of years. Despite such an extensive history and widespread interest, the study of category learning has overwhelmingly relied on a single approach that consists of motivating people to attend to category diagnostic features by making category judgments, for which they receive corrective feedback. The present dissertation diverged from this standard approach to category learning by evaluating categorization following (1) training emphasizing either category-shared or exemplar-unique features (Chapter 2) and (2) incidental learning of category structure in behavior and (3) the brain. Overall, the results underscore both similarities and differences in categorization following these lesser-used approaches relative to

the more standard approach. Differences in generalization strategy and neural correlates following incidental category learning in particular highlight the need for further research into the flexibility of both generalization and category representations in the brain. The current dissertation provides substantial support for the flexibility of category learning, category representations, and generalization processes, all of which appeared to depend on a mixture of goal-relevance, the structure of the world, and biological constraints.

References cited

Achmed Ali, S., Leelaarporn, P., Stirnberger, R., Bilzer, M., Abdel Kafi, N., Taube, J., Sagik, Y., & McCormick, C. (2026). Seeing more than schemas: the vmPFC represents imagery-rich mental scenarios. *Neuropsychologia*, *222*. https://doi.org/10.1016/j.neuropsychologia.2026.109370

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). https://doi.org/10.1037/0033-295X.105.3.442

Audrain, S., & McAndrews, M. P. (2022). Schemas provide a scaffold for neocortical integration of new memories over time. *Nature Communications*, *13*(1). https://doi.org/10.1038/s41467-022-33517-0

Barsalou, L. W. (2016). On Staying Grounded and Avoiding Quixotic Dead Ends. In *Psychonomic Bulletin and Review* (Vol. 23, Number 4). https://doi.org/10.3758/s13423-016-1028-3

Barsalou, L. W., Dutriaux, L., & Scheepers, C. (2018). Moving beyond the distinction between concrete and abstract concepts. In *Philosophical Transactions of the Royal Society B: Biological Sciences* (Vol. 373, Number 1752). https://doi.org/10.1098/rstb.2017.0144

Binder, J. R. (2016). In defense of abstract conceptual representations. *Psychonomic Bulletin and Review*, *23*(4). https://doi.org/10.3758/s13423-015-0909-1

Blank, H., & Bayer, J. (2022). Functional imaging analyses reveal prototype and exemplar representations in a perceptual single-category task. *Communications Biology*, *5*(1). https://doi.org/10.1038/s42003-022-03858-z

Bowman, C. R., Iwashita, T., & Zeithamova, D. (2020). Tracking prototype and exemplar representations in the brain across learning. *ELife*, *9*, 1–47. https://doi.org/10.7554/eLife.59360

Bowman, C. R., & Zeithamova, D. (2018). Abstract memory representations in the ventromedial prefrontal cortex and hippocampus support concept generalization. *Journal of Neuroscience*, *38*(10), 2605–2614. https://doi.org/10.1523/JNEUROSCI.2811-17.2018

Bowman, C. R., & Zeithamova, D. (2020). Training Set Coherence and Set Size Effects on Concept Generalization and Recognition. *Journal of Experimental Psychology: Learning Memory and Cognition*. https://doi.org/10.1037/xlm0000824

Bowman, C. R., & Zeithamova, D. (2023). Coherent category training enhances generalization in prototype-based categories. *Journal of Experimental Psychology: Learning, Memory, and Cognition*. https://doi.org/10.1037/xlm0001243

Bowman, C. R., & Zeithamova, D. (2025a). High coherence among training exemplars promotes broad generalization of face families. *Journal of Experimental Psychology: Learning, Memory, and Cognition*. https://doi.org/10.1037/xlm0001478

Bowman, C. R., & Zeithamova, D. (2025b). *Ventromedial prefrontal cortex supports prototype representations in healthy older adults*. https://doi.org/10.1101/2025.09.24.678354

Boyle, L. M., Posani, L., Irfan, S., Siegelbaum, S. A., & Fusi, S. (2024). Tuned geometries of hippocampal representations meet the computational demands of social memory. *Neuron*, *112*(8). https://doi.org/10.1016/j.neuron.2024.01.021

Brunec, I. K., & Momennejad, I. (2022). Predictive Representations in Hippocampal and Prefrontal Hierarchies. *Journal of Neuroscience*, *42*(2). https://doi.org/10.1523/JNEUROSCI.1327-21.2021

Chen, D., Kunz, L., Lv, P., Zhang, H., Zhou, W., Liang, S., Axmacher, N., & Wang, L. (2021). Theta oscillations coordinate grid-like representations between ventromedial prefrontal and entorhinal cortex. *Science Advances*, *7*(44). https://doi.org/10.1126/sciadv.abj0200

Ciaramelli, E., De Luca, F., Monk, A. M., McCormick, C., & Maguire, E. A. (2019). What "wins" in VMPFC: Scenes, situations, or schema? In *Neuroscience and Biobehavioral Reviews* (Vol. 100). https://doi.org/10.1016/j.neubiorev.2019.03.001

Clark, A. (2013a). Expecting the world: Perception, prediction, and the origins of human knowledge. *Journal of Philosophy*, *110*(9). https://doi.org/10.5840/jphil2013110913

Clark, A. (2013b). Whatever next? Predictive brains, situated agents, and the future of cognitive science. *Behavioral and Brain Sciences*, *36*(3). https://doi.org/10.1017/S0140525X12000477

Corcoran, A. W., Pezzulo, G., & Hohwy, J. (2020). From allostatic agents to counterfactual cognisers: active inference, biological regulation, and the origins of cognition. *Biology and Philosophy*, *35*(3). https://doi.org/10.1007/s10539-020-09746-2

Cortese, A., Yamamoto, A., Hashemzadeh, M., Sepulveda, P., Kawato, M., & De Martino, B. (2021). Value signals guide abstraction during learning. *ELife*, *10*. https://doi.org/10.7554/eLife.68943

Cowan, E., Liu, A., Henin, S., Kothare, S., Devinsky, O., & Davachi, X. L. (2020). Sleep spindles promote the restructuring of memory representations in ventromedial prefrontal cortex through enhanced hippocampal–cortical functional connectivity. *Journal of Neuroscience*, *40*(9). https://doi.org/10.1523/JNEUROSCI.1946-19.2020

De Martino, B., & Cortese, A. (2023). Goals, usefulness and abstraction in value-based choice. In *Trends in Cognitive Sciences* (Vol. 27, Number 1). https://doi.org/10.1016/j.tics.2022.11.001

Desai, R. H., Reilly, M., & Van Dam, W. (2018). The multifaceted abstract brain. In *Philosophical Transactions of the Royal Society B: Biological Sciences* (Vol. 373, Number 1752). https://doi.org/10.1098/rstb.2017.0122

Detre, J. A., & Wang, J. (2002). Technical aspects and utility of fMRI using BOLD and ASL. In *Clinical Neurophysiology* (Vol. 113, Number 5). https://doi.org/10.1016/S1388-2457(02)00038-X

Durrant, S. J., Cairney, S. A., McDermott, C., & Lewis, P. A. (2015). Schema-conformant memories are preferentially consolidated during REM sleep. *Neurobiology of Learning and Memory*, *122*. https://doi.org/10.1016/j.nlm.2015.02.011

Farrell, S., & Lewandowsky, S. (2010). Computational models as aids to better reasoning in psychology. *Current Directions in Psychological Science*, *19*(5). https://doi.org/10.1177/0963721410386677

Farrell, S., & Lewandowsky, S. (2018). Computational modeling of cognition and behavior. In *Computational Modeling of Cognition and Behavior*. https://doi.org/10.1017/CBO9781316272503

Friston, K. (2010). The free-energy principle: A unified brain theory? In *Nature Reviews Neuroscience* (Vol. 11, Number 2). https://doi.org/10.1038/nrn2787

Friston, K. J. (2019). Waves of prediction. *PLoS Biology*, *17*(10). https://doi.org/10.1371/journal.pbio.3000426

Galdo, M., Weichart, E. R., Sloutsky, V. M., & Turner, B. M. (2022). The quest for simplicity in human learning: Identifying the constraints on attention. *Cognitive Psychology*, *138*. https://doi.org/10.1016/j.cogpsych.2022.101508

Gao, M., Turner, B. M., & Sloutsky, V. M. (2024). The Role of Attention in Category Representation. *Cognitive Science*, *48*(4). https://doi.org/10.1111/cogs.13438

Ghosh, V. E., & Gilboa, A. (2014). What is a memory schema? A historical perspective on current neuroscience literature. In *Neuropsychologia* (Vol. 53, Number 1). https://doi.org/10.1016/j.neuropsychologia.2013.11.010

Gibson, J. J. (2014). The theory of affordances (1979). In *The People, Place, and Space Reader*. https://doi.org/10.4324/9781315816852

Gilboa, A., & Marlatte, H. (2017). Neurobiology of Schemas and Schema-Mediated Memory. In *Trends in Cognitive Sciences* (Vol. 21, Number 8). https://doi.org/10.1016/j.tics.2017.04.013

Guo, D., Chen, H., Wang, L., & Yang, J. (2023). Effects of prior knowledge on brain activation and functional connectivity during memory retrieval. *Scientific Reports*, *13*(1). https://doi.org/10.1038/s41598-023-40966-0

Haxby, J. V. (2012). Multivariate pattern analysis of fMRI: The early beginnings. In *NeuroImage* (Vol. 62, Number 2). https://doi.org/10.1016/j.neuroimage.2012.03.016

Hindy, N. C., Coleman, R., & Selvam, S. R. M. (2026). Hippocampal Reactivation Trades Episodic Detail for Semantic Gist in Human Memory. *Journal of Cognitive Neuroscience*, 1–13. https://doi.org/10.1162/JOCN.a.2592

Hintzman, D. L. (1986). “Schema Abstraction” in a Multiple-Trace Memory Model. *Psychological Review*, *93*(4). https://doi.org/10.1037/0033-295X.93.4.411

Hu, M., & Nosofsky, R. M. (2024). High-variability training does not enhance generalization in the prototype-distortion paradigm. *Memory & Cognition*. https://doi.org/10.3758/s13421-023-01516-1

Klinzing, J. G., Niethard, N., & Born, J. (2019). Mechanisms of systems memory consolidation during sleep. In *Nature Neuroscience* (Vol. 22, Number 10). https://doi.org/10.1038/s41593-019-0467-3

Lee, M. D., & Wagenmakers, E. J. (2013). Bayesian cognitive modeling: A practical course. In *Bayesian Cognitive Modeling: A Practical Course*. https://doi.org/10.1017/CBO9781139087759

Leshinskaya, A., & Caramazza, A. (2016). For a cognitive neuroscience of concepts: Moving beyond the grounding issue. In *Psychonomic Bulletin and Review* (Vol. 23, Number 4). https://doi.org/10.3758/s13423-015-0870-z

Lewandowsky, S., & Oberauer, K. (2018). Computational Modeling in Cognition and Cognitive Neuroscience. In *Stevens' Handbook of Experimental Psychology and Cognitive Neuroscience*. https://doi.org/10.1002/9781119170174.epcn501

Lewis, P. A., & Durrant, S. J. (2011). Overlapping memory replay during sleep builds cognitive schemata. In *Trends in Cognitive Sciences* (Vol. 15, Number 8). https://doi.org/10.1016/j.tics.2011.06.004

Liang, Z., Wu, S., Wu, J., Wang, W., Qin, S., & Liu, C. (2023). Social navigation: distance and grid-like codes support navigation of abstract social space in human brain. *BioRxiv*.

Lieberman, M. D., Straccia, M. A., Meyer, M. L., Du, M., & Tan, K. M. (2019). Social, self, (situational), and affective processes in medial prefrontal cortex (MPFC): Causal, multivariate, and reverse inference evidence. In *Neuroscience and Biobehavioral Reviews* (Vol. 99). https://doi.org/10.1016/j.neubiorev.2018.12.021

Love, B. C., Medin, D. L., & Gureckis, T. M. (2004). SUSTAIN: A Network Model of Category Learning. *Psychological Review*, *111*(2). https://doi.org/10.1037/0033-295X.111.2.309

Lutz, N. D., Diekelmann, S., Hinse-Stern, P., Born, J., & Rauss, K. (2017). Sleep Supports the Slow Abstraction of GIST from Visual Perceptual Memories. *Scientific Reports*, *7*. https://doi.org/10.1038/srep42950

Mack, M. L., Preston, A. R., & Love, B. C. (2020). Ventromedial prefrontal cortex compression during concept learning. *Nature Communications*, *11*(1). https://doi.org/10.1038/s41467-019-13930-8

Mahon, B. Z. (2015). What is embodied about cognition? *Language, Cognition and Neuroscience*, *30*(4). https://doi.org/10.1080/23273798.2014.987791

Martin, A. (2016). GRAPES—Grounding representations in action, perception, and emotion systems: How object properties and categories are represented in the human brain. *Psychonomic Bulletin and Review*, *23*(4). https://doi.org/10.3758/s13423-015-0842-3

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). https://doi.org/10.1037/0033-295X.102.3.419

McCormick, C., Ciaramelli, E., De Luca, F., & Maguire, E. A. (2018). Comparing and Contrasting the Cognitive Effects of Hippocampal and Ventromedial Prefrontal Cortex Damage: A Review of Human Lesion Studies. In *Neuroscience* (Vol. 374). https://doi.org/10.1016/j.neuroscience.2017.07.066

Michel, C. (2023). A Hybrid Account of Concepts Within the Predictive Processing Paradigm. *Review of Philosophy and Psychology*, *14*(4). https://doi.org/10.1007/s13164-022-00648-8

Moneta, N., Grossman, S., & Schuck, N. W. (2024). Representational spaces in orbitofrontal and ventromedial prefrontal cortex: task states, values, and beyond. In *Trends in Neurosciences* (Vol. 47, Number 12). https://doi.org/10.1016/j.tins.2024.10.005

Murphy, G. L. (2016). Is there an exemplar theory of concepts? *Psychonomic Bulletin and Review*, *23*(4). https://doi.org/10.3758/s13423-015-0834-3

Nosofsky, R. M. (1986). Attention, similarity, and the identification-categorization relationship. *Journal of Experimental Psychology: General*, *115*(1). 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). https://doi.org/10.1037/0278-7393.13.1.87

O'Regan, J. K., & Noë, A. (2001). What it is like to see: A sensorimotor theory of perceptual experience. *Synthese*, *129*(1). https://doi.org/10.1023/A:1012699224677

Pitt, M. A., Myung, I. J., & Zhang, S. (2002). Toward a method of selecting among computational models of cognition. *Psychological Review*, *109*(3). https://doi.org/10.1037/0033-295X.109.3.472

Rasch, B., & Born, J. (2013). About sleep's role in memory. *Physiological Reviews*, *93*(2). https://doi.org/10.1152/physrev.00032.2012

Raviv, L., Lupyan, G., & Green, S. C. (2022). How variability shapes learning and generalization. In *Trends in Cognitive Sciences* (Vol. 26, Number 6). https://doi.org/10.1016/j.tics.2022.03.007

Rehder, B., & Hoffman, A. B. (2005a). Eyetracking and selective attention in category learning. *Cognitive Psychology*, *51*(1). https://doi.org/10.1016/j.cogpsych.2004.11.001

Rehder, B., & Hoffman, A. B. (2005b). Thirty-something categorization results explained: Selective attention, eyetracking, and models of category learning. In *Journal of Experimental Psychology: Learning Memory and Cognition* (Vol. 31, Number 5). https://doi.org/10.1037/0278-7393.31.5.811

Rips, L. J. . (2009). Similarity, typicality, and categorization. In *Similarity and Analogical Reasoning*. https://doi.org/10.1017/cbo9780511529863.004

Schapiro, A. C., McDevitt, E. A., Rogers, T. T., Mednick, S. C., & Norman, K. A. (2018). Human hippocampal replay during rest prioritizes weakly learned information and predicts memory performance. *Nature Communications*, *9*(1). https://doi.org/10.1038/s41467-018-06213-1

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: Biological Sciences*, *372*(1711). https://doi.org/10.1098/rstb.2016.0049

Schlegelmilch, R., Dinu, A., Joue, G., Gläscher, J., & Sommer, T. (2026). Neurocomputational modeling of rule abstraction and memorization during probabilistic stimulus-reward learning. *IScience*, *29*(1). https://doi.org/10.1016/j.isci.2025.113950

Sommer, T., Hennies, N., Lewis, P. A., & Alink, A. (2022). The assimilation of novel information into schemata and its efficient consolidation. *Journal of Neuroscience*, *42*(30). https://doi.org/10.1523/JNEUROSCI.2373-21.2022

Spalding, K. N., Schlichting, M. L., Zeithamova, D., Preston, A. R., Tranel, D., Duff, M. C., & Warren, D. E. (2018). Ventromedial prefrontal cortex is necessary for normal associative inference and memory integration. *Journal of Neuroscience*, *38*(15). https://doi.org/10.1523/JNEUROSCI.2501-17.2018

St-Laurent, M., Abdi, H., & Buchsbaum, B. R. (2015). Distributed patterns of reactivation predict vividness of recollectio. *Journal of Cognitive Neuroscience*, *27*(10). https://doi.org/10.1162/jocn_a_00839

Sučević, J., & Schapiro, A. C. (2023). A neural network model of hippocampal contributions to category learning. *ELife*, *12*. https://doi.org/10.7554/eLife.77185

Vaidya, A. R., & Badre, D. (2022). Abstract task representations for inference and control. In *Trends in Cognitive Sciences* (Vol. 26, Number 6). https://doi.org/10.1016/j.tics.2022.03.009

Varela, F. J., Thompson, E., Rosch, E., & Kabat-Zinn, J. (2016). The embodied mind: Cognitive science and human experience. In *The Embodied Mind: Cognitive Science and Human Experience*. https://doi.org/10.29173/cmplct8718

Veselic, S., Muller, T. H., Gutierrez, E., Behrens, T. E. J., Hunt, L. T., Butler, J. L., & Kennerley, S. W. (2025). A cognitive map for value-guided choice in the ventromedial prefrontal cortex. *Cell*, *188*(12). https://doi.org/10.1016/j.cell.2025.03.038

Weber, A., & Varela, F. (2002). Life after Kant: Natural purposes and the autopoietic foundations of biological individuality. *Phenomenology and the Cognitive Sciences*, *1*(2). https://doi.org/10.1023/A:1020368120174

Weichart, E. R., Evans, D. G., Galdo, M., Bahg, G., & Turner, B. M. (2022). Distributed Neural Systems Support Flexible Attention Updating during Category Learning. *Journal of Cognitive Neuroscience*, *34*(10). https://doi.org/10.1162/jocn_a_01882

Weichart, E. R., Galdo, M., Sloutsky, V. M., & Turner, B. M. (2022). As Within, so Without, as Above, so Below: Common Mechanisms Can Support Between and Within-Trial Category Learning Dynamics. *Psychological Review*, *129*(5). https://doi.org/10.1037/rev0000381

Weichart, E. R., Unger, L., King, N., Sloutsky, V. M., & Turner, B. M. (2024). "The Eyes Are the Window to the Representation": Linking Gaze to Memory Precision and Decision Weights in Object Discrimination Tasks. *Psychological Review*, *131*(4). https://doi.org/10.1037/rev0000475

Wernette, E. M. D., & Fenn, K. M. (2023). Consolidation without intention: Sleep strengthens veridical and gist representations of information after incidental encoding. *Psychonomic Bulletin and Review*, *30*(4). https://doi.org/10.3758/s13423-023-02247-9

Wing, E. A., D'angelo, M. C., Gilboa, A., & Ryan, J. D. (2021). The role of the ventromedial prefrontal cortex and basal forebrain in relational memory and inference. *Journal of Cognitive Neuroscience*, *33*(9). https://doi.org/10.1162/jocn_a_01722