Journal of Experimental Psychology:
Learning, Memory, and Cognition

ISSN: 0278-7393

2024, Vol. 50, No. 3, 458–483
https://doi.org/10.1037/xlm0001257

Structure Shapes the Representation of a Novel Category

Sarah H. Solomon and Anna C. Schapiro
Department of Psychology, University of Pennsylvania

Concepts contain rich structures that support flexible semantic cognition. These structures can be characterized by patterns of feature covariation: Certain features tend to cluster in the same items (e.g., *feathers*, *wings*, *can fly*). Existing computational models demonstrate how this kind of structure can be leveraged to slowly learn the distinctions between categories, on developmental timescales. However, it is not clear whether and how we leverage feature structure to quickly learn a novel category. We thus investigated how the internal structure of a new category is first extracted from experience, with the prediction that feature-based structure would have a rapid and broad influence on the learned category representation. Across three experiments, novel categories were designed with patterns of feature associations determined by carefully constructed graph structures, with Modular graphs—exhibiting strong clusters of feature covariation—compared against Random and Lattice graphs. In Experiment 1, a feature inference task using verbal stimuli revealed that Modular structure broadly facilitated category learning. Experiment 2 replicated this effect in visual categories. In Experiment 3, a statistical learning paradigm revealed that this Modular benefit relates to high-level structure rather than pairwise feature associations and persists even when category structure is incidental to the task. A neural network model was readily able to account for these effects, suggesting that correlational feature structure may be encoded within rapidly learned, distributed category representations. These findings constrain theories of category representation and link theories of category learning with structure learning more broadly.

Keywords: concepts, categories, structure, learning, neural network models

Supplemental materials: https://doi.org/10.1037/xlm0001257.supp

Mental representations are only useful insofar as they capture the content and structure of the environment. Structure—the organization of relations among units in a system—can reflect different kinds of relations in different domains. For example, structure within the visual domain may reflect co-occurrences of visual features across scenes, spatial structure may reflect the physical relationships between various landmarks, and semantic structure may reflect the pattern of semantic feature associations on multiple conceptual scales. Recently, in part inspired by the idea of "cognitive maps" (Tolman, 1948), researchers have sought to understand the formation of structural representations and their influence on human cognition. Here we focus on the semantic domain, and the feature-based structure that underlies concepts and categories. Our approach is to synthesize ideas from structure learning and category learning to investigate the nature of structured category representations.

We use the term *concept* to reflect the internal representation of a *category*, which is a set of ideas or items in the external world (Komatsu, 1992; Rips et al., 2012; K. O. Solomon et al., 1999). Research on concepts typically focuses on the rich, mature representations humans already carry with them around the world (e.g., FIREFLY, ROSE, TRUTH). On the other hand, research on category learning typically focuses on newly constructed categories defined by abstract perceptual stimuli (e.g., visual shapes, auditory tones) or simplified figures (e.g., cartoon animals, line drawings), lacking the richness in structure that eventually supports our mature concept representations. Here we attempt to bridge these two literatures to target fundamental questions of concept representation—when a new category is learned, how do the initially formed representations lead to the rich, structured concepts that compose semantic memory?

Structure Within and Between Concepts

A feature-based view of concepts enables us to link theories of semantic knowledge with those of novel category learning. While this empiricist tradition is widely adopted, there are also dissenting views. For example, proponents of the "theory theory" of concepts argue that concepts are not primarily collections of features, but rather abstract knowledge structures that contain explanations, counterfactuals, causal relations, predictions, and rules (e.g., Ahn & Kim, 2000; Carey, 1985; Gopnik, 1988; Keil, 1992). Nevertheless, feature-based approaches to conceptual knowledge give us traction in understanding the human semantic system. Our present goal is to examine whether and how structure in the semantic environment becomes embedded within our semantic representations.

Structure within the semantic domain can explain various aspects of semantic cognition, including the developmental trajectory of semantic learning (Rogers & McClelland, 2004; Unger et al., 2020) and human performance on conceptual tasks such as property verification and similarity judgments (Cree et al., 1999; McRae et al., 1997, 1999; Rosch, 1975; Tyler et al., 2000). Classic studies suggest

This article was published Online First June 15, 2023.

Stimulus materials and data will be made publicly available on OSF (https://osf.io/te96s/?view_only=66bf6f9aab544938b80903b3f7b4a287).

Correspondence concerning this article should be addressed to Sarah H. Solomon, Department of Psychology, University of Pennsylvania, 425 South University Avenue, Philadelphia, PA 19104, United States. Email: sarahsol@sas.upenn.edu

that concepts are not represented as collections of independent features, but as the "web of relationships in which these properties participate" (Medin et al., 1987). For example, one study found that participants considered material and size dimensions to be correlated within the SPOON concept: Wooden spoons are typically large, and metal spoons are typically small (Medin & Shoben, 1988). This was interpreted to suggest that semantic features are not independent but are structured via associative relations. Classic empirical findings on the interpretation of combined concepts (e.g., METAL ROSE) are often used to support the claim that concepts contain structured sets of features and/or relations (Sloman et al., 1998; Wisniewski & Gentner, 1991). The structure of individual concepts also appears to have consequences for how those concepts are flexibly used in language. For example, in recent work we found evidence that the feature-based structure of a concept predicts its semantic variability as measured within language corpora (S. H. Solomon et al., 2019).

Correlational feature structure plays an important role in computational models of concept learning and representation. Feature associations can explain why people consider certain kinds of features (i.e., functional vs. perceptual) as more important for some concepts versus others (i.e., artifacts vs. living things; Tyler et al., 2000; Tyler & Moss, 2001) and can predict performance on property verification, similarity judgment, and typicality judgment tasks (Cree et al., 1999; McRae et al., 1997, 1999). These models of real-world concepts, among others, were inspired by the seminal McClelland and Rumelhart's (1985) neural network models of semantic learning. Neural network models capitalize on semantic feature correlations to learn and represent category information. For example, Rogers and McClelland (2004) demonstrate that the structure of the semantic environment—that is, patterns of semantic feature co-occurrences across items—can be used by neural network models to learn hierarchical semantic relationships. The models build representations that differentiate superordinate categories (e.g., ANIMALS vs. PLANTS), basic-level categories (e.g., FLOWER vs. TREE), and individual items (e.g., ROSE vs. PINE TREE) in a manner that accounts for patterns of semantic cognition in humans: for example, the order in which concepts are learnt, the special status of basic level concepts, and the acquisition and deterioration of semantic knowledge observed in infants and dementia patients (Rogers & McClelland, 2004).

An important principle underlying the success of the models is that they leverage "coherent covariation" among clusters of features to build internal representations that capture the structure of the semantic environment (Cree et al., 1999; McClelland & Rogers, 2003; Rogers & McClelland, 2004; Saxe et al., 2019). Coherent covariation refers to the tendency of features to co-occur across different semantic items. For example, the presence of feathers and wings and the capability of flight are features that tend to co-occur, facilitating the formation of a BIRD concept. Neural network models represent these feature associations via the fine-tuning of weights during learning. By the end of learning, the model has integrated information across items such that the model's weights store patterns of feature associations rather than storing traces of the individual items that were observed. In other words, semantic structure is encoded directly in the neural network model's learned representation.

Internal Structure of Novel Categories

Categories (and concepts) do much more than categorize. Once we have classified an item as belonging to some category or another, we use relevant category information to make inferences about its other features, predict its behaviors or uses, and communicate with others. Feature inferences rely on an understanding of a category's structure, that is, on feature frequencies and correlations. Imagine that, on a hike through a rainforest, you find two creatures and classify one as a frog and the other as a bird. Based on what you know about frogs, you infer that it would feel slimy to the touch and because it is brightly colored you also infer that it is potentially poisonous. Based on what you know about birds, you infer that this one is lightweight and because it is covered in a grey fuzz you also infer that it is young and therefore that its mother is likely nearby.

Categories can be learned in either classification or inference tasks, and evidence suggests that these different learning conditions result in different category representations. In a classification task, participants learn to associate exemplars with the correct category label. In an inference task, participants are shown exemplars with one feature missing and learn to select the identity of the missing feature. Whereas classification tasks encourage people to learn and represent the diagnostic features for each category, inference tasks encourage people to focus on within-category information and consequently to learn a category's internal structure (Anderson et al., 2002; Chin-Parker & Ross, 2002; Lassaline & Murphy, 1996; Markman & Ross, 2003; Rehder & Ross, 2001). Thus, different learning tasks can result in different learning patterns and learned representations; traditional models of category learning that are trained on categorization tasks perform poorly on subsequent inference tasks (Markman & Ross, 2003; Yamauchi & Markman, 1998). In contrast, neural network models of semantic learning (Rogers & McClelland, 2004; Rumelhart, 1990) were developed to explain feature-based inference, rendering them especially useful in the study of internal category structure.

In studies that contrast classification with inference learning, multiple categories are employed (there must at least be two categories to test classification decisions). However, feature inference is distinct from classification—after an item has been categorized as belonging to a certain category, the internal structure of that category can then guide feature inference. For this reason, studies investigating feature inference and feature correlations often do not include classification tasks and sometimes participants are only exposed to one category during learning (Franks & Bransford, 1971; Medin et al., 1982; Neumann, 1974; Wattenmaker, 1991, 1993).

Humans are sensitive to feature correlations in novel category learning tasks. Medin et al. (1982) exposed participants to a novel disease category in which patients presented with correlated symptoms. In a following transfer test, the study participants were more likely to diagnose a new patient with the disease if correlated symptoms were present, even if those symptoms were less characteristic of the disease overall. In another series of experiments, Wattenmaker (1991, 1993) created categories of hypothetical people in which several pairs of social characteristics were perfectly correlated. Across a range of different learning conditions, behavior at test revealed knowledge of category-specific feature correlations. Sensitivity was observed for pairs as well as triplets of correlated features, when feature correlations were pitted against feature frequencies, when only one category was presented, and in both intentional and incidental learning conditions. Hayes et al. (1996) further suggest that prior knowledge boosts sensitivity to feature correlations. These results provide strong evidence that humans are sensitive to both feature frequencies and feature correlations during category learning.

Embedding Structure in Models of Category Learning

Feature-based structure thus appears to influence category learning and a variety of conceptual judgments. This has interesting implications for theories of semantic representation—are feature correlations directly embedded within category representations, or are they inferred during retrieval processes? To explore this question, we can consider different theories of category representation. Broadly speaking, there are two primary classes of category models: exemplar models and abstraction models (Barsalou, 1990). Both classes of models are powerful and can, in principle, contain the same kinds of semantic information. The crucial difference between these models is not *what* information is stored, but *how* the information is stored.

Exemplar models store traces of observed exemplars as distinct items in memory. These exemplar traces exhibit information duplication (the same feature appears multiple times across stored exemplars) and, typically, no information revision (exemplar representations are not updated as new information becomes available). The context theory of classification learning (Medin & Schaffer, 1978) and the associated generalized context model (GCM; Nosofsky, 1986, 2011) are classic examples of this view. Exemplar models have had significant success in predicting behavioral patterns of category learning (e.g., Hintzman, 1984; Medin & Schaffer, 1978; Nosofsky, 1984, 1986; Nosofsky et al., 1992, 2018) as well as neural responses during category learning (Mack et al., 2013).

Abstraction models, on the other hand, integrate information across exemplars into a centralized category representation. Information revision is intrinsic to these models, and there is no information duplication. Prototype models are classic examples of abstraction models, with centralized category representations that contain information about distinct category features (Rosch, 1973; Rosch & Mervis, 1975). There are many variations of "prototype" style models—for example, modal prototype models represent the category's most frequent features, whereas average distance models represent the average feature values across exemplars. In all cases, exemplar information is integrated to form a centralized representation of feature frequencies or probabilities, and the individual exemplar traces are discarded. Prototypes can explain the fuzziness of category boundaries (Hampton, 1979) and can predict a wide range of category judgments and categorization phenomena (Rosch, 1973; Rosch & Mervis, 1975). Models instantiating prototype theory have also had success in predicting human categorization performance (Bowman et al., 2020; Devraj et al., 2021; Smith & Minda, 1998, 2002). It has been argued that prototype- as well as exemplar-based representations emerge in the brain during category learning (Bowman et al., 2020).

How do exemplar and abstraction models stack up in their ability to explain human sensitivity to feature correlations? If the class of abstraction models is reduced to the well-known prototype models, exemplar-based theories are unequivocally supported (Medin et al., 1982; Wattenmaker, 1991). Neither exemplar nor prototype models directly represent feature correlations, but while feature co-occurrences can be extracted from exemplar representations at retrieval, this information is lost in prototype models. However, the class of abstraction models extends beyond classic "prototypes" and some models do represent feature co-occurrences in addition to, or instead of, feature frequencies (Gluck & Bower, 1988; Hayes-Roth & Hayes-Roth, 1977; Neumann, 1974; Reitman & Bower, 1973). For example, the "configural cue" model is a simple one-layer network in which the powerset of a category's features is represented as localist input nodes (Gluck & Bower, 1988). We will refer to this class of models, in which information is abstracted away from exemplars and feature associations are encoded, as "relational abstraction" models (Barsalou, 1990).

Relational abstraction theories that posit localist representations of feature associations are criticized as implausible, given that a category's powerset exponentially increases in size as each new feature is added. However, neural network models of semantic learning can also be characterized as relational abstraction models and use distributed, instead of localist, representations (McClelland & Rogers, 2003; McClelland & Rumelhart, 1985; Rogers & McClelland, 2004). Neural network models can represent pairwise feature associations as well as larger feature clusters and thus are promising models in the investigation of category structure representations. Various network models have been used to account for category learning in humans that make different representational and processing assumptions ("adaptive network model," Gluck & Bower, 1988; "ALCOVE," Kruschke, 1992; "SUSTAIN," Love et al., 2004; "DIVA," Kurtz, 2007). Here, our arguments are based on neural network models such as McClelland and Rumelhart (1985) and Kurtz (2007), which are multilayer networks with fixed architectures that receive semantic feature information as inputs. In the context of feature correlation learning, these network characteristics are crucial—a learning process fine-tunes the weights between the input/output and hidden layers, such that feature associations are stored.

The range of exemplar and abstraction representation models is vast, and because they can be paired with a wide range of processing assumptions, it is hard to empirically support one theory of representation over the other (Barsalou, 1990). It is clear that humans are sensitive to feature frequencies and can develop category representations that include feature co-occurrence information. However, sensitivity to feature correlations could be explained by processes that occur at either encoding or retrieval. It could be that abstraction occurs during encoding, such that feature correlations are stored within the category representation, but it is also possible that only exemplar traces are stored during encoding and postencoding processes applied to these traces at retrieval result in the extraction of feature co-occurrence information. Some researchers have attempted to make claims regarding category representation by manipulating the learning task to focus on category-level or item-level information (e.g., intentional vs. incidental encoding; Hayes et al., 1996; Wattenmaker, 1991, 1993). For example, based on Wattenmaker's (1991) finding that incidental learners were more sensitive to feature correlations than intentional learners, he concluded that feature correlations are extracted from item-level representations at retrieval rather than encoded as abstracted rules. Other researchers have argued for exemplar-style category representations based on findings that exemplar models can explain human category learning behavior better than prototype models (e.g., Medin & Schaffer, 1978). However, the dismissal of prototype models should not result in the dismissal of all abstraction models. Indeed, these researchers often concede that abstraction models that store feature correlations or relations would be able to explain the findings ("In fact, most of the qualitative predictions of the context model examined in the present experiments are shared by relational frequency models"; Medin & Schaffer, 1978).

We propose a relational abstraction model of category representation in which feature correlations are stored within a distributed system. This model can be instantiated in a multilayer neural network model—when features are represented on the input layer, a learning process fine-tunes the weights between input/output and hidden layers such that useful feature associations are stored. This is an abstraction model because exemplar information is integrated during learning, and is relational because feature correlations are encoded. Our experiments aim to test how well a neural network, relational abstraction model of category learning can explain patterns of human category learning behavior.

Learning Categories Within a Distributed System

Multilayer neural network models rely on *distributed representations*, in which units are responsive to many related features of the environment, to learn and represent information. We hypothesize that learning categories within a distributed representational system can lead to structured category representations containing meaningful feature associations. This kind of representation crucially contributes to the successful semantic development observed in neural network models, though (a) the structures learned by existing models are within and across larger semantic domains (e.g., ANIMALS, PLANTS) as opposed to within items (e.g., ROSE; c.f., McClelland & Rumelhart, 1985) and (b) the semantic learning simulated in the models is presumed to occur over longer time scales (e.g., years of development in infants). Here we aim to extend the principles of these prior models to examine whether *internal* category structure influences how categories are quickly learned and later used.

We will use neural network model simulations to explore how distributed representations can support the rapid initial learning of novel category structure. Though our neural network model is a learning model, our goal is not to propose a theoretical model of particular learning mechanisms and our hypotheses are not related to learning dynamics. Rather, we use the model as a tool for demonstrating how internal category structure emerges rapidly in a distributed representational system. When we refer to distributed representations, we mean distributed representations that have emerged to represent the structure and content of the input environment, as opposed to distributed representations of the individual inputs themselves.

According to complementary learning systems (CLS) theory, experiences are first encoded as sparse, nonoverlapping representations in the hippocampus, and only during subsequent consolidation are they integrated into distributed representations in cortex (McClelland et al., 1995). However, our recent computational model of the hippocampus argues that distributed representations quickly emerge in subfield CA1, separate from the sparse representations of Dentate Gyrus and CA3 (Schapiro et al., 2017). We have also shown that the hippocampus is sensitive to overlapping associations and higher-level structure in the environment (Schapiro et al., 2016). This prior work leads us to predict that distributed representations may rapidly develop during initial encounters with a rich category structure.

Our approach will also enable us to test whether the same principles that govern learning of broad semantic domains also govern learning of individual categories. If leveraging coherent covariation of semantic features facilitates the formation of representations of semantic categories, it seems likely that coherent covariation of features *within* a category might similarly benefit learning. In a category with clusters of reliably covarying features, it might be relatively easier to learn the features that are most important for that category, or there may be additional benefits such as improved category generalization (Bowman & Zeithamova, 2021). We already know that humans are sensitive to the statistical structure found within many cognitive domains, as discussed below. An open question is whether the feature-based structure of a category influences learning and category use more broadly, and further whether specific structures benefit or interfere with category learning and the building of useful category representations.

Structure Learning Across Cognitive Domains

Structure can manifest in many ways. It can be operationalized in terms of statistical co-occurrences, transitional probabilities, and physical proximities, among others. Structure learning in the context of novel categories was classically construed in terms of hypothesis testing and logical rule formation (Martin & Caramazza, 1980; Nosofsky et al., 1994; Ward & Scott, 1987). In terms of internal category structure, participants were presumed to form hypotheses regarding which category features co-occurred (Wattenmaker, 1991, 1993). However, decades of research on statistical learning (SL) has revealed that structure can be learned implicitly without the use of explicit hypotheses or rules. Some of the earliest empirical evidence that humans are sensitive to statistical structure was provided in the domain of language development. Saffran et al. (1996) reported that after brief exposure to a stream of nonsense syllables, infants were able to learn the pseudoword boundaries based solely on the temporal contingencies between syllables. This SL phenomenon has been reported across many domains (Frost et al., 2019). We will explore the idea that internal category structure might be learned implicitly through similar mechanisms.

Humans are sensitive to higher-level structure that goes beyond pairwise statistics. Researchers often describe high-level temporal structure in terms of graphs, or networks (Karuza et al., 2016; Lynn & Bassett, 2020). Graph nodes correspond to different stimuli, and the edges between nodes specify possible transitions between stimuli. Within an experiment, the topography of the graph is carefully designed in order to target elements of the learning process. A network topography that is often used in these experiments contains clusters of nodes, or "communities," such that the stimuli within a community tend to occur within close temporal proximity (Kahn et al., 2018; Kakaei et al., 2021; Karuza et al., 2017, 2019; Lynn et al., 2020; Mark et al., 2020; Pudhiyidath et al., 2020; Schapiro et al., 2013, 2016). The extent to which networks or graphs contain these communities is quantified in terms of "modularity"—networks with denser clusters of nodes exhibit higher modularity (Rubinov & Sporns, 2011). Behavioral performance on SL tasks reveals that humans are sensitive to community structure underlying visual (Kakaei et al., 2021; Schapiro et al., 2013), motor (Kahn et al., 2018), and navigational (Mark et al., 2020) tasks. It has been argued that modular structure leads to more accurate representations of a network (Lynn & Bassett, 2020).

While structure learning has been witnessed across many cognitive domains—including language, visual perception, motor actions, and navigation—structure learning paradigms have not yet been applied to novel category learning. When a category is modeled as a graph containing features and feature associations

(S. H. Solomon et al., 2019), modular structure describes coherent covariation across features. That is, a modular "community" would, in this case, reflect a set of features that tend to co-occur with high probability across exemplars. Within a community, features have a tendency to co-occur but may not be perfectly correlated. This kind of structure is more naturalistic than the perfectly correlated feature pairs used in previous category learning designs (e.g., Hayes et al., 1996; Medin et al., 1982; Wattenmaker, 1991). Importantly, this model of category representation differs considerably from prototype and family resemblance models, neither of which capture feature co-occurrences. A prototype contains one set of category-diagnostic features, and its primary goal is to differentiate one category from another (Rosch & Mervis, 1975); family resemblance models similarly represent a category in terms of nonnecessary features that may be shared in any combination across category members (Wittgenstein, 2010). Conversely, a graph-based model can contain multiple feature communities and its primary goal is to describe within-category variance. It is not the presence of individual features that is diagnostic, but rather their patterns of co-occurrence. This approach will reveal aspects of category and concept learning specifically, but also will inform theories and models of structure learning more broadly.

Overview of Experiments

We aimed to synthesize structure learning and category learning paradigms to examine how structured categories may give rise to structured representations. Evidence that structured category representations are rapidly built during learning would forge a link between novel category representations and the rich representations known to underlie established concepts. In order to address these questions, we developed a paradigm distinct from traditional category learning paradigms. In a series of behavioral experiments, human participants learned novel animal categories whose feature association statistics were determined by carefully manipulated underlying graph structures. We specifically compared a Modular graph structure like the ones described above—containing clusters of covarying features—with two non-Modular structures (i.e., Random, Lattice). Within each novel category, the specific features assigned to the graph nodes were randomized for each participant in order to isolate the effects of the category structure and eliminate any bias due to prior knowledge of feature correlations (e.g., Hayes et al., 1996). We also ran simulations of a neural network model to test whether a relational abstraction model of category learning—in which feature associations are encoded into the learned representation—can explain patterns of human category learning.

An important aspect of our design is that all categories contained an identically structured set of three high-frequency "core" features. That is, while the "peripheral" structures differed—resulting in Modular, Random, and Lattice categories—the core structure across all categories was identical. We examined participants' ability to learn each category's core features in order to assess the general influence of category structure on category learning. Experiments 1 and 2 employed a classic missing feature inference task, whereas Experiment 3 was inspired by more recent SL designs. In all cases, we compared core feature learning across Modular and non-Modular categories.

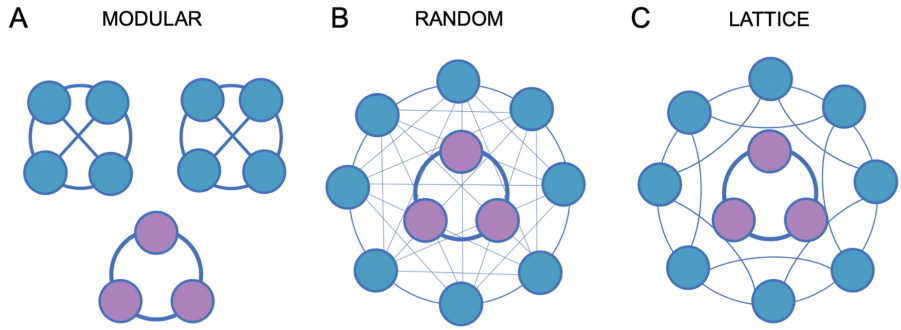
We predicted that humans would be able to learn the feature-based category structures, based on previous research in the category learning domain suggesting that humans can learn pairs or triplets of correlated features (Hayes et al., 1996; Medin et al., 1982; Wattenmaker, 1991, 1993). More importantly, we predicted that the specific structure of a category (i.e., Modular vs. non-Modular) would influence the ease with which that category is learned. Previous research in the structure learning domain suggests that modularity within a structure, or clusters of reliably co-occurring information, benefits learning (Lynn & Bassett, 2020; Rogers & McClelland, 2004). If the mechanisms involved in learning an individual category parallel those involved in structure learning more broadly, we would expect increased performance on Modular categories relative to Random and Lattice categories. That is, we predicted that increased feature clustering among the Modular category's peripheral features would improve participants' ability to detect the high-frequency core features. Further, we predicted that a neural network model—instantiating a relational abstraction model of category learning—would also reveal this influence of internal category structure since it is designed to encode feature correlations and is especially attuned to coherent covariation among features (Rogers & McClelland, 2004).

We tested these predictions across three experiments, examining human behavior and model simulations. In Experiment 1, we designed two novel animal categories defined by lists of verbal features whose co-occurrence statistics were dictated by a Modular or Random graph structure. Since our empirical questions relate to internal category structure as opposed to category distinctions, we exposed participants to the two categories in a missing feature task, an inferential task that promotes within category comparisons (Anderson et al., 2002; Chin-Parker & Ross, 2002, 2004; Erickson et al., 2005; Markman & Ross, 2003; Yamauchi & Markman, 1998). To preview our results, we found support for the prediction that core feature learning was improved in the Modular relative to the Random category and that our neural network model could account for this effect. In Experiment 2, we replicated the results of Experiment 1 using novel visual insect stimuli to show that this effect replicates across stimulus modality. In Experiment 3, we compared Modular and Lattice category learning using our novel insects in a SL paradigm to test whether high-level category structure effects emerge even when structure is incidental to all aspects of the learning task. We again found evidence that a Modular category structure improved core feature learning and that the neural network model could account for the behavioral results. In sum, these experiments reveal that coherent feature clusters facilitate category learning in intentional and incidental learning tasks. The model simulations suggest that the influence of internal category structure on category learning can be explained by the abstraction of feature associations during encoding.

Experiment 1

In Experiment 1, human participants were exposed to two novel animal categories whose features were presented in a verbal format. Participants learned the features of these categories in an inferential missing features task, in which a set of features (i.e., an "exemplar") was presented and participants selected the feature they believed was missing out of three possible options. Unbeknownst to the participants, the internal structure (i.e., pattern of feature correlations) differed between the two categories: the Random category only contained one cluster of "core" features with no structure among its peripheral features (Figure 1B), whereas the Modular category

Figure 1
Category Structures Used Across the Experiments



Note. Nodes indicate features and edges indicate their co-occurrence statistics. All graphs contained the same core structure, corresponding to three high-frequency features (purple nodes) sharing strong co-occurrence statistics (indicated by the thick edges). All core nodes share an edge with all peripheral nodes (not visualized for simplicity). (A) The Modular structure contained two clusters of peripheral features such that features in one module could never co-occur with features in the other module. (B) The Random structure contained peripheral features (blue nodes) with no clustering; any peripheral feature could co-occur with any other peripheral feature. (C) The Lattice structure, like the Random structure, contained no clustering in its periphery, but each peripheral feature could only co-occur with a subset of all peripheral features. See the online article for the color version of the figure.

consisted of an identical core structure in addition to two clusters of correlated peripheral features (Figure 1A). There are two important contrasts in the design of these category structures. First, core features are more frequent than peripheral features; this frequency differential is identical across the Modular and Random structures. Second, the peripheral features in the Modular structure contain reliable feature correlations, whereas the peripheral features in the Random structure do not. This experiment allows us to evaluate whether peripheral feature correlations influence participants' overall category representations.

An advantage for the Modular category would suggest that people are sensitive to the coherent covariation of features created by the modules, in line with previous computational work (Rogers & McClelland, 2004). We therefore expected a neural network model, which instantiates a relational abstraction model of category learning, to reveal a similar pattern of behavior.

Experiment 1: Behavioral

Method

Participants. Forty participants recruited from Amazon Mechanical Turk contributed data to Experiment 1 (Age: $M = 37.5$, $SD = 11.3$; 64% female) and were compensated $4.50 for their time. An additional five participants did not pass attentional checks (see catch trials below) and were excluded from analyses. Consent was obtained for all participants in accordance with the University of Pennsylvania Institutional Review Board (IRB).

Categories and Features. Two "species" of novel animals were created, each defined by 11 features that were presented in verbal form. While classic category learning experiments use a smaller set of features, increasing the number of feature dimensions brings us closer to understanding the representation of natural categories. In fact, prior work has revealed that categories with more feature dimensions results in humans learning more about those categories (Hoffman & Murphy, 2006). One of our species contained the features *large*, *two legs*, *solitary*, *blue eyes*, *bushy tail*, *sleeps in caves*, *has horns*, *growls*, *brown fur*, *drinks water*, and *striped*. The other species contained the features *small*, *four legs*, *social*, *grey eyes*, *hairless tail*, *sleeps in trees*, *has claws*, *roars*, *black fur*, *drinks milk*, and *spotted*. The dimensions specified by the 11 features in each species were approximately matched (e.g., size, eye color, markings). In order to add additional variability to the category exemplars, for each species we generated an additional 40 features of the form "*eats—*" (e.g., *eats lemons*, *eats lilacs*, *eats corn*). We also chose an additional six features per species to use as catch features during the behavioral task (e.g., *has fangs*, *white feet*). No features overlapped between species. All features used in Experiment 1 are shown in Table S1 in the online supplemental materials. Assignment of species to category label (i.e., *Timbo*, *Sudex*) was randomized for each participant.

Category Structures. Experiments 1 and 2 compared Random and Modular structures (Figure 1). Both graphs contained the same core structure: three high-frequency features that are found in all exemplars. This identical core structure—shared by all structure conditions—means that different approaches to the task (e.g., focusing on only a subset of features) are unlikely to bias core feature behavior differently across categories. However, the peripheral structures of the graphs differed: In the Modular structure, the peripheral features were divided into two modules, or clusters, such that features from one cluster never co-occurred with features from the other cluster within the same exemplar. In the Random structure, every peripheral feature could co-occur with any other peripheral feature within an exemplar. Note that our Random structure is not named to reflect a "random network" topology (Watts & Strogatz, 1998), but rather the fact that exemplars can have any random combination of peripheral features. Importantly, the Modular peripheral structure contained reliable feature correlations whereas the Random peripheral structure did not. Assignment of species to structure was randomized across participants. In Experiment 1 the features assigned to the core

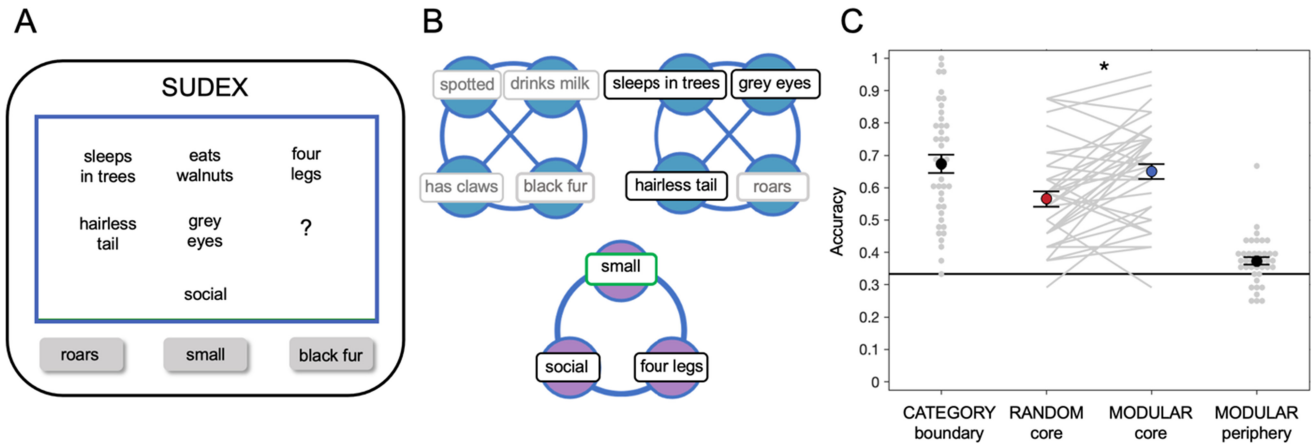
nodes were fixed within species (i.e., *large*, *two legs*, and *solitary* were always the core features for one category and *small*, *four legs*, and *social* were always the core features for the other category). See Figure 2A for an example of a Modular category in Experiment 1. The assignment of species (i.e., feature set) to structure was counterbalanced, so any prior knowledge related to core-feature assignment did not influence comparisons between Modular and Random categories. The assignment of features to the peripheral nodes was randomized such that the features corresponding to each of the clusters in the Modular category differed across participants. Finally, category labels ("sudex," "timbo") were randomly assigned to each category.

Category Exemplars. Each exemplar was originally defined by six features, the combination of which was determined by the underlying graph structure (i.e., three core, three peripheral). Within these parameters, the Modular graph generates eight unique exemplars (${}_4C_3 \times 2$ Modules) and the Random graph generates 56 unique exemplars (${}_8C_3$). To reach the 72 total exemplars per category needed for the behavioral task, we used nine sets of the eight unique exemplars from the Modular category; for the Random category, we used the 56 unique exemplars and subsampled an additional 16 from the same set to reach a total of 72 exemplars. Each exemplar was also assigned an idiosyncratic *eats* feature; each *eats* feature was seen in only one or two exemplars throughout the experiment. Thus, each exemplar contained a total of seven features: three core features, three peripheral features, and one idiosyncratic feature.

Task Design. Each of the 72 exemplars from each category corresponded to a trial in the behavioral task. On each trial, one of the exemplar's features was removed and three possible features were presented as response options. The identity of the missing features and the identity of the three feature options were determined by one of four task conditions: category boundary, core structure (Random and Modular), and peripheral structure (Modular). In category boundary trials, a Random category exemplar was presented with a peripheral feature missing; the three options included the correct peripheral feature from the Random category and two Modular category peripheral features. These trials tested participants' ability to learn which features defined each category. To test core structure in both Random and Modular categories, an exemplar was presented with a core feature missing; the three options included the correct core feature and two peripheral features from that category. These trials tested participants' ability to learn that the three core features were necessary for each exemplar. (Note that in the Modular category, because each module contained only four features, one of the two incorrect features was from an exemplar-inconsistent module.) To test peripheral structure in the Modular category, an exemplar was presented with a peripheral feature missing (e.g., from module1); the three options included the correct module1 feature and two incorrect features from module2. These trials tested participants' ability to learn that features from separate modules could not co-occur. The 144 experimental trials were evenly divided between the boundary, core, and peripheral conditions: 48 category boundary trials (Random), 48 core structure trials (24 Random, 24 Modular), and 48 peripheral structure trials (Modular). An additional six catch trials (three random, three modular) were designed in which a peripheral feature was removed and the options included the correct feature in addition to two features never before seen by the participant. Five participants who responded incorrectly to three or more of the six catch trials were excluded from analyses. Trial order was pseudorandomized such that exemplars from a single category were seen approximately five times in a row before switching to the other category, and all catch trials occurred in the second half of the experiment.

The frequencies with which features appeared on exemplars were balanced within and across structure conditions: Each core and peripheral feature in the Modular category appeared in 64 and 21

Figure 2
Experiment 1: Design and Results



Note. (A) On each trial of the missing feature task in Experiment 1, the relevant category label was displayed (e.g., "sudex," "timbo") above a box containing an exemplar's features. Five structure-consistent features were displayed in addition to an idiosyncratic "eats" feature. A missing feature was indicated by a question mark and participants decided which of three possible features belonged to the exemplar. The outlined nodes in (B) indicate the exemplar's shown features (black) and the correct feature (green). (C) Human participants successfully learned the category boundary, Random core structure, Modular core structure, and Modular periphery structure. Accuracy was significantly higher for Modular versus Random core structure, $t(39) = 3.8$, $p = .0004$, even though core structure was identical across categories. Error bars reflect standard error of the mean. See the online article for the color version of the figure.

exemplars, respectively; these frequencies were identical in the Random category. Core and peripheral features appeared as the correct response option on eight and six trials, respectively, in both structure conditions. However, the category boundary condition made it impossible to balance the extent to which features appeared as an incorrect response: Each Random peripheral feature appeared as an incorrect response on six trials, whereas each Modular peripheral feature appeared as an incorrect response between 27 and 33 times. Core features were never presented as incorrect response options.

Verbal Missing Feature Task. Participants completed 150 trials of the missing feature task. Before the task began, the participants were told: "We will train you on the animal categories by showing you category members with one feature missing. On each trial, you will be given three features and you will choose the feature that you think is the missing one." On each trial, the category label was presented above a box in the center of the screen, within which an exemplar's features were presented. Three feature options were presented beneath the box (Figure 2B). The order of the exemplar's features inside the box and the order of the three response options were randomized on each trial. Participants could take as long as they needed to select a feature by clicking the corresponding button. When a participant made a correct response, their selection would be outlined in green and they would proceed to the next trial. When a participant made an incorrect response, their selection would be outlined in red and the trial was repeated until the correct feature was chosen. The positions of exemplar features and response options were randomized each time a trial was repeated. Feedback was presented immediately after a feature option was clicked and remained on screen for 1,000 ms; the subsequent trial began 1,500 ms after the response was made.

Statistical Analysis. Each trial was coded as accurate (1) or inaccurate (0) based on the first response given (i.e., a trial was similarly marked as incorrect whether it took two or more than two times to choose the correct feature). Mean accuracies for category boundary, Random core structure, Modular core structure, and Modular peripheral structure trials were calculated for each participant. Accuracy across participants was compared to chance (0.33), and core structure knowledge in Random and Modular categories was compared using a paired t-test. Relationships between different kinds of category knowledge were assessed using Pearson correlations. Data from Experiment 1 are available on OSF.

Results

Accuracy for Verbal Missing Features. Mean accuracies across participants for the four structure conditions are shown in Figure 2C. Above chance accuracy was observed in the category boundary, $M = 67.4\%$, $SD = 18\%$, $t(39) = 11.9$, $p < .0001$; random core, $M = 56.6\%$, $SD = 15.2\%$, $t(39) = 9.7$, $p < .0001$; modular core, $M = 65.0\%$, $SD = 14.8\%$, $t(39) = 13.5$, $p < .0001$; and Modular periphery, $M = 37.3\%$, $SD = 7.4\%$, $t(39) = 3.4$, $p = .002$, conditions. A one-way analysis of variance (ANOVA) revealed a significant difference between conditions, $F(3) = 35.9$, $p < .0001$. Pairwise dependent t-tests revealed that accuracy for Modular periphery structure was significantly lower than all other kinds of structure knowledge: category boundary, $t(39) = 10.3$, $p < .0001$; Random core, $t(39) = 7.9$, $p < .0001$; and Modular core, $t(39) = 11.5$, $p < .0001$. Category boundary accuracy exceeded Random core accuracy, $t(39) = 3.3$, $p = .002$, but was not significantly higher than Modular core accuracy ($p > .4$). Most interestingly, Modular core accuracy was significantly higher than Random core accuracy, $t(39) = 3.8$, $p = .0004$, 95% confidence interval (CI): [0.04–0.129], Cohen's $d = 0.56$, despite the fact that core structure was identical across categories.

Relationships Between Kinds of Structure Knowledge. Accuracy for the Random core and Modular core conditions was significantly correlated across participants, $r(39) = 0.57$, $p = .0001$, suggesting that some participants are generally more sensitive to feature frequencies than others. No relationship between Modular core and periphery accuracy was observed ($p > .2$). Category boundary accuracy did not reliably predict Random, $r(39) = 0.25$, $p = .13$, or Modular, $r(39) = 0.30$, $p = .06$, core accuracy; it notably also did not predict the increase in accuracy for Modular core versus Random core structure, $r(39) = 0.05$, $p = .74$. This suggests that the observed Modular core benefit does not emerge merely because participants learned to avoid Modular peripheral features on category boundary trials. Similarly, Modular periphery accuracy did not predict the Modular core structure benefit ($p > .9$), further suggesting that higher accuracy on Modular core trials is unlikely to be due to participants' ability to exclude peripheral features from the inconsistent module on Modular core trials. Thus, the results suggest that participants were more sensitive to core features in the Modular category, which contained increased feature correlations in the periphery.

Experiment 1: Neural Network Model

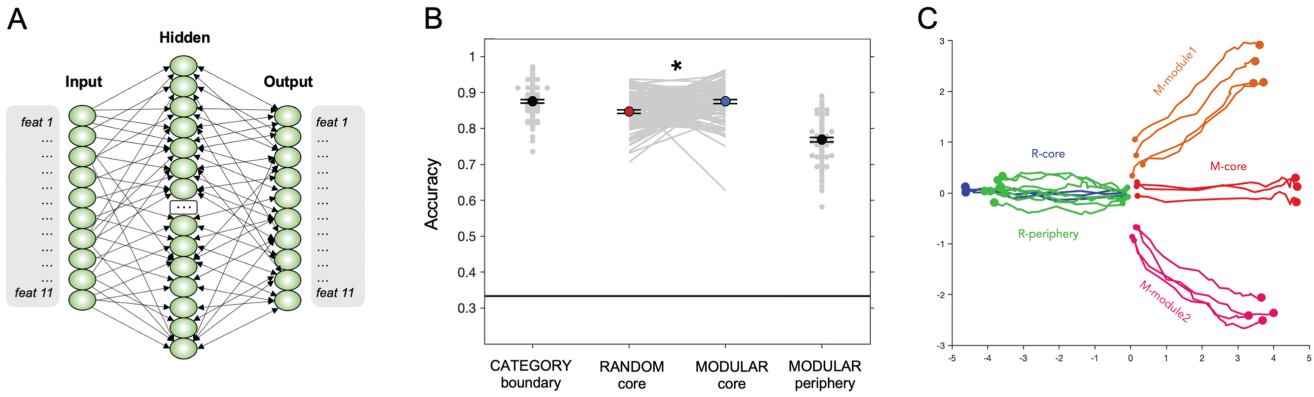
Method

Model Architecture and Parameters. We constructed our model in the Emergent simulation environment (O'Reilly et al., 2020). A schematic of our model architecture is shown in Figure 3A. The model comprised an input layer (22 units), one hidden layer (400 units), and an output layer (22 units). The units in the input and output layers corresponded to the 22 total features across the two categories. The input layer had full feedforward connectivity to the hidden layer, and there was full bidirectional connectivity between the hidden layer and output layer. We used Emergent's default parameters and training regime, which implements the eXtended Contrastive Attractor Learning rule. We expect, however, that any standard learning rule (e.g., backpropagation) operating on distributed representations would behave similarly (O'Reilly, 1996). The neural network model we report here is a simple demonstration of how we expect all models in this class to behave. We did not optimize any hyperparameters and used the default Emergent settings including a learning rate of $\lambda = 0.04$.

Training Protocol. The model was trained as an autoencoder on full exemplars that corresponded with the 144 experimental trials used in the behavioral experiment (i.e., the five shown features plus the one correct feature). That is, training consisted of 144 trials in which the model was presented with a six-feature exemplar on the input layer and learned to recreate the six-feature exemplar on the output layer. As our neural network model is not impacted by the kinds of attentional effects that likely drive the blocking advantage in humans, and because blocking in neural network models tends to produce interference, there was no reason to use the same pseudorandomized trial order used in human behavior. Rather, the order of training trials was randomized.

Figure 3

Neural Network Model Simulation of the Missing Feature Task



Note. (A) The model architecture contained 22 feature nodes on the input and output layers and a 400-unit hidden layer. (B) Across 100 simulations, the model replicated patterns of human performance such that higher accuracy was observed in the Modular core versus Random core conditions, $t(99) = 3.5$, $p = .0006$. Error bars reflect standard error of the mean. (C) MDS solution revealing how feature representations change from the beginning (center small dots) to the end (large dots) of training for Random (cool colors) and Modular (warm colors) categories. Modular core (red) and peripheral (orange, pink) features are more strongly differentiated by the end of learning relative to the Random core (blue) and peripheral (green) features. MDS = multidimensional scaling. * indicates $p < .05$. See the online article for the color version of the figure.

Testing Protocol. Test trials were analogous to the experimental trials in the missing feature task. A total of 144 trials tested category boundary knowledge (48), Random core structure knowledge (24), Modular core structure knowledge (24), and Modular peripheral structure knowledge (48). On each trial, five features were presented on the input layer (i.e., an exemplar with a missing feature) and we analyzed activity in the output layer to determine whether the model was able to successfully activate the correct feature. We restricted this analysis to the same three features that were presented to human participants; a test trial was coded as accurate (1) if the model activated the correct feature more strongly than the two incorrect features, and incorrect (0) if one of the incorrect features was the most strongly activated unit. The full set of test trials (144) was run after every eight training trials for a total of 18 test epochs per simulation.

Model Assessment. We ran the model 100 times, with a new initialization of weights and randomized trial order for each simulation. Within each simulation, the mean accuracy for each of the four conditions was calculated within each test epoch. We calculated mean accuracy across test epochs for category boundary, Random core, Modular core, and Modular peripheral conditions for each simulation. Accuracy was compared to chance (0.33), and paired *t*-tests were used to compare Random and Modular core structure knowledge.

Analyzing Feature Representations. In order to assess how feature representations in the model transformed during learning, we recorded the pattern of activity evoked across the hidden layer units on test trials where each feature was presented by itself. Thus, after every eight training trials we obtained patterns of activity representing the 22 category features. For each simulation, we simultaneously ran multidimensional scaling (MDS) on the 18 Test Epochs × 22 Feature Representations using Euclidean distance, enabling us to visualize how the features became more or less similar to each other across learning (Figure 3C). Owing to the simultaneous calculation of the distances across all time points, the initial starting points of the 22 features appear to be clustered by category but are not in fact meaningfully differentiated.

Results

Mean model accuracy across simulations is shown in Figure 3B. As with human participants, the model was able to learn the category boundary ($M = 87.5\%$, $SD = 4.7\%$), Random core ($M = 84.7\%$, $SD = 5.2\%$), Modular core ($M = 87.5\%$, $SD = 5.3\%$), and Modular periphery ($M = 76.9\%$, $SD = 6.1\%$) at above chance levels ($ps < .0001$). Most interestingly, the model replicated the Modular core benefit observed in human participants: Accuracy was significantly greater for Modular core structure than for Random core structure, $t(99) = 3.5$, $p = .0006$, 95% CI = [0.012–0.043], Cohen's $d = 0.53$, meaning that the neural network model was more sensitive to the core features in the Modular versus the Random structure. Visualization of the training course MDS solution reveals that the model learns to more strongly differentiate the core from peripheral features in the Modular categories relative to the Random categories (Figure 3C). The model also learns to differentiate the Modular categories' two peripheral modules from each other. The reduction of interference resulting from this differentiation of the feature clusters in the Modular categories likely underlies the superior performance when filling in a missing core feature in this category.

Experiment 1: Discussion

In Experiment 1, we found that the rich internal structure of a novel category influences category learning more broadly, suggesting that the structure of a category can warp the emergent category representation. Most notably, participants found it easier to learn the high frequency core features of the Modular category relative to the Random category, even though core structure and core feature frequencies were identical across categories. In line with our predictions, these results support the hypothesis that increased coherence among semantic features benefits learning of individual concepts and categories.

What kind of category representation underlies this effect? One possibility is that feature associations are encoded into the category

representation during learning. We therefore predicted that a neural network model—which instantiates this relational abstraction theory—would mirror patterns of human behavior and similarly reveal a Modular core benefit. Indeed, we found that our simple three-layer network found it easier to learn the high frequency core features of the Modular category than the non-Modular category. The alignment of behavioral and neural network behavior is consistent with a theory of category representation in which feature associations are encoded into the representation itself, and supports a role for distributed representations in novel category learning. These simulations in Experiment 1 do not eliminate other theories of category representation, but rather show how an abstraction theory of representation can succeed in explaining how humans learn and represent feature associations in novel categories. Indeed, we found that an exemplar-based model is capable of explaining the Modular core benefit in Experiment 1 (online supplemental materials).

Analyses of the learned internal representations in the neural network model revealed that the peripheral features were more differentiated from core features in the Modular category and that the two clusters of peripheral features were highly differentiated from one another. This provides the hypothesis that a similar kind of representational differentiation may be underlying the Modular core benefit observed in humans. This prediction could be tested through measurement of neural representations or by collecting postlearning feature similarity ratings from humans and relating it to their task performance.

The simulation results are consistent with prior demonstrations that neural network models with distributed representations are highly attuned to coherent covariation among features at larger semantic scales (Rogers & McClelland, 2004). Learning the structure of large semantic domains (e.g., ANIMALS) and learning the internal structure of individual concepts (e.g., FOX) might involve similar mechanisms of extracting the correlational structure of features across items or exemplars (McClelland & Rumelhart, 1985). Just as the clustering of reliably co-occurring features provides traction on learning the boundaries *between* concepts, as has been previously shown, clusters of reliably co-occurring features *within* a concept aid the formation of a learned, structured category representation that can be used to generalize across category members and support feature inference. Both humans and the neural network model learned this new structure *rapidly*, orders of magnitude more rapidly than in the learning of large semantic domains (e.g., Mikolov et al., 2013; Rogers & McClelland, 2008; Rumelhart, 1990). This suggests that neither humans nor neural network models require large amounts of training to build rich distributed representations.

The model's representation of the task was abstract and not tied to the particular verbal nature of the stimuli, suggesting that the Modular core benefit observed in Experiment 1 should manifest in other domains as well. We thus predicted that we would observe a similar behavioral effect when categories are presented not as lists of features but as visual objects.

Experiment 2

The results of Experiment 1 suggest that it is easier for humans to learn the core structure of a Modular category relative to a Random category, indicating a sensitivity to coherent covariation among the clustered peripheral features. Experiment 2 aimed to replicate this effect using visually rather than verbally presented categories, based on our expectation that category structure learning effects emerge from a domain-general learning mechanism. Experiment 2 was thus designed to test whether the Modular core benefit is robust across feature types.

Experiment 2: Behavior

Method

Participants. Forty participants recruited from Amazon Mechanical Turk contributed data to Experiment 2 (age: $M = 40.6$, $SD = 12.0$; 53% female) and were compensated $4.50 for their time. An additional eight participants did not pass the attentional checks and were excluded from analyses. Consent was obtained for all participants in accordance with the University of Pennsylvania IRB.

Categories and Features. Two species of novel insects were created with 11 features each, now displayed in visual form ("beetle" and "butterfly"; Figure 4). Each species consisted of an insect base on which 11 various animal features could be attached. The general kind and location of features were matched across the species: horns/antennae (three), sets of wings (two), sets of arms (two), colored markings (two), tail (one), and toe features (one). These stimuli were designed such that any subset of the 11 features could be added to the base to form a potential category exemplar. An additional six features were generated for each species to use in the catch trials. The category base and features were created in Adobe Illustrator by cropping and editing open-source images of real animals found on the internet. Category stimuli are publicly available for use (https://osf.io/te96s/?view_only=66bf6f9aab544938b80903b3f7b4a287). No category labels were used in Experiment 2.

Category Structures and Exemplars. The same Random and Modular structures used in Experiment 1 were used in Experiment 2. Assignment of species to structure was randomized across participants, and assignment of features to graph nodes was fully randomized for each participant (i.e., any feature could be a core feature). The set of visual exemplars in Experiment 2 were generated using an identical procedure to the one used in Experiment 1, resulting in 72 six-feature exemplars within both the Random and Modular categories. Instead of assigning an idiosyncratic feature to the visual exemplars, the color of the insect base was randomly adjusted slightly on each trial to add additional variability across exemplars.

Figure 4
Visual Stimuli in Experiments 2 and 3

Note. The "beetle" (left) and "butterfly" (right) species each consisted of an insect base upon which 11 features could be presented. See the online article for the color version of the figure.

Task Design. The design of Experiment 2 was identical to Experiment 1, resulting in a total of 144 experimental trials and six catch trials. Participants who did not respond correctly to at least four of the six catch trials ($N = 8$) were excluded from subsequent analyses.

Visual Missing Feature Task. Participants completed 150 trials of the missing feature task, and were given the same instructions used in Experiment 1. On each trial, the visual exemplar with five features was presented in the center of the screen, and the three feature options were displayed below (Figure 5A and 5B). The feature options were presented on a desaturated insect base above a numbered button (on category boundary trials, the insect base consistent with the displayed exemplar was used even though it displayed a feature from the incorrect category). Participants could take as long as they needed to click on the button corresponding to their feature choice. Feedback was presented in an identical way to Experiment 1 and trials were repeated until the correct feature choice was made.

Statistical Analysis. Mean accuracy for each participant and each condition was calculated as in Experiment 1. Accuracy for each condition was compared to chance (0.33), and conditions were compared with paired *t*-tests and Pearson correlations. Data from Experiment 2 are available on OSF.

Results

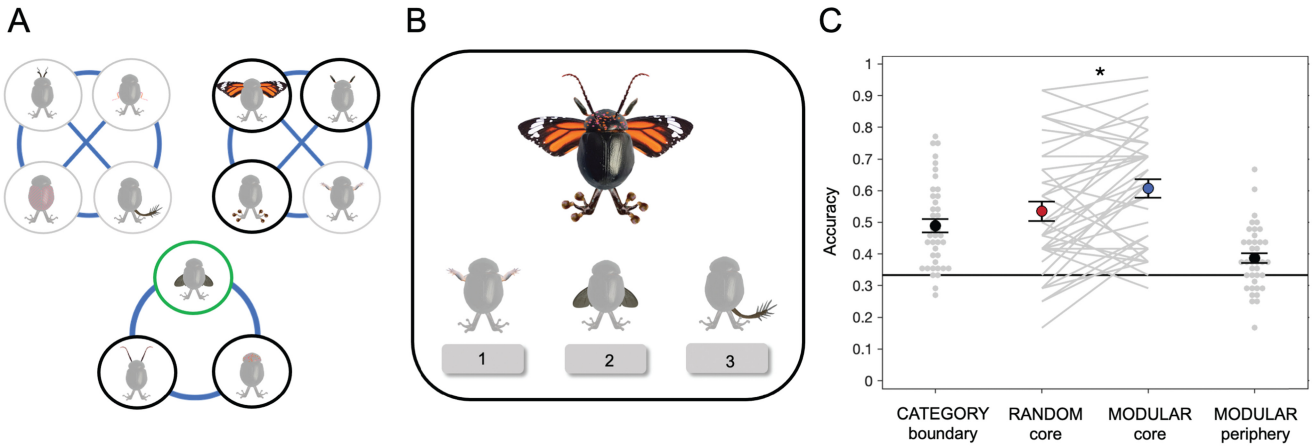
Accuracy for Visual Missing Features. Mean accuracies across participants for the four structure conditions are shown in Figure 5D. As in Experiment 1, above chance accuracy was observed in the category boundary, $M = 48.9\%$, $SD = 13.5\%$, $t(39) = 7.3$, $p < .0001$; Random core, $M = 53.5\%$, $SD = 19.4\%$, $t(39) = 6.6$, $p < .0001$; Modular core, $M = 60.7\%$, $SD = 18.1\%$, $t(39) = 9.6$, $p < .0001$; and Modular periphery, $M = 38.7\%$, $SD = 9.9\%$, $t(39) = 3.4$, $p = .002$, conditions. A one-way ANOVA revealed a significant difference between conditions, $F(3) = 13.9$, $p < .0001$. Pairwise dependent *t*-tests revealed that accuracy for Modular periphery structure was significantly lower than all other kinds of structure knowledge: category boundary, $t(39) = 4.2$, $p < .001$; Random core, $t(39) = 5.0$, $p < .001$; and Modular core, $t(39) = 7.7$, $p < .001$. Category boundary accuracy did not significantly differ from Random core accuracy, $t(39) = 1.5$, $p = .15$, but was significantly lower than Modular core accuracy, $t(39) = 3.7$, $p < .001$. Most importantly, we replicated the Modular core benefit using visual categories: Modular core accuracy was significantly higher than Random core accuracy, $t(39) = 2.8$, $p = .007$, 95% CI = [0.02–0.123], Cohen's $d = 0.38$, despite the fact that core structure and core feature frequencies were identical across categories.

Relationships Between Kinds of Structure Knowledge. Accuracy for Random core and Modular core structures was significantly correlated, $r(39) = 0.63$, $p < .0001$. No significant relationship between Modular core and periphery accuracy was observed, $r(39) = 0.26$, $p = .10$. Category boundary accuracy did not reliably predict Random, $r(39) = 0.31$, $p = .052$, or Modular, $r(39) = 0.19$, $p = .23$, accuracy; it notably also did not predict the increase in accuracy for Modular core versus Random core structure, $r(39) = -0.16$, $p > .3$. Similarly, Modular periphery accuracy did not predict the Modular core structure benefit, $r(39) = -0.10$, $p > .5$.

Experiment 2: Discussion

Using visual category stimuli, Experiment 2 replicated the behavioral Modular core benefit observed in Experiment 1. Participants were exposed to two novel insect categories defined by sets of discrete, nonoverlapping features in a missing feature task. Even though the structure and frequencies of core features were identical across the Modular and Random categories, participants found it easier to learn the three high frequency core features of the Modular category. These results further support the claim that humans leverage feature correlations across category members to build structured

Figure 5
Visual Missing Feature Task in Experiment 2



Note. (A) For each participant, assignment of structure (Modular vs. Random) to species (Beetle vs. Butterfly) was randomized, as were the feature-node assignments. (B) On each trial, an exemplar with five features was presented and participants selected which of three possible features they thought belonged. The outlined nodes in (A) indicate the shown exemplar's features on this trial (black) and the correct feature response (green). (C) The Modular core benefit was replicated in Experiment 2 such that participants' accuracy was significantly higher for the Modular core versus Random core trials, $t(39) = 2.8$, $p = .007$. * indicates $p < .05$. See the online article for the color version of the figure.

category representations and that these representations support successful feature inference. The replication across stimulus types suggests that a domain-general learning mechanism is at play.

While results from Experiments 1 and 2 converge on the finding that feature correlations benefit the learning of a category more broadly, we are left with three questions. The first relates to the distribution of features in the missing feature task. In the displayed exemplars, feature frequencies were exactly balanced—core features appeared with equal frequencies in the Random and Modular categories, as did peripheral features. The frequency with which core and peripheral features appeared as the correct feature response was similarly balanced across category structures. However, the feature frequencies in the set of incorrect feature responses were not balanced—Modular peripheral features appeared more frequently as incorrect options than did Random peripheral features. It is possible that participants learned to reject Modular peripheral features in the missing feature task, thereby increasing the likelihood that they would respond correctly on the Modular core trials. This cannot explain the Modular benefit observed in the model simulations, and our correlational analyses on human behavior do not lend support to this explanation. However, we aimed to design another experiment that would eliminate this possibility entirely.

The second question relates to pairwise correlations versus high-order structure. The Modular and Random category structures differed on both counts—the Random structure contained no reliable pairwise feature correlations in its periphery and therefore no higher-order structure. The Modular structure contained higher-order feature clusters, but also had increased pairs of correlated features as a result. Consequently, the results from Experiments 1 and 2 suggest that, at minimum, pairwise feature correlations influence learning, but we cannot make any claims about the encoding of higher-level structure. We thus aimed to design a new non-Modular graph structure that still contained no higher-order peripheral clustering but was better matched with the Modular graph in terms of pairwise feature correlations. The presence of peripheral structure in this new graph would additionally enable a more balanced task design and would help remove the constraints that contributed to the feature frequency imbalance in Experiments 1 and 2.

The third question relates to the differentiation of encoding and retrieval processes. In Experiments 1 and 2, encoding and retrieval were not cleanly differentiated in time, since categories were learned via feedback on inferential questions that tapped into category structure knowledge. A stronger claim could be made if encoding and retrieval processes were perfectly disentangled. This is another goal of Experiment 3.

Experiment 3

In Experiments 1 and 2, categories were learned in an inferential missing features task, which has previously been shown to induce learning of internal category structure (Anderson et al., 2002; Chin-Parker & Ross, 2002, 2004; Markman & Ross, 2003; Yamauchi & Markman, 1998). This task can be classified as an intentional learning condition, since the task encouraged participants to focus on the structure of each category. That is, successful performance required participants to integrate information across exemplars to learn about which features are most frequent and which features tend to co-occur. It could be that feature associations are only encoded under intentional conditions when they are relevant to the learning task. In Experiment 3, participants were exposed to novel categories in a task that did not encourage category-related processing in any way, resulting in a purely incidental learning condition. We know that incidental learning can result in sensitivity to feature correlations at retrieval, but it is unclear whether a relational abstraction model of category representation can explain this effect (Wattenmaker, 1991, 1993). We also know that humans can learn statistical structure in other domains even when it is irrelevant to the explicit task. In Experiment 3, we embed category feature correlations in a SL paradigm to ask whether feature associations are abstracted into category representations even under purely incidental learning conditions.

Whereas the Modular structure was previously contrasted with a Random structure with no peripheral pairwise feature correlations, in Experiment 3 we contrasted the Modular structure with a Lattice structure in which pairs of peripheral features are correlated but do not reveal any higher-order clustering (Figure 1C). This enabled us to test both core and peripheral structure knowledge in all categories, and to evaluate the influence of higher-level feature structure on the representations that emerge during category learning.

Participants were exposed to a single category in a SL paradigm in which a temporal stream of two-feature exemplars were presented. The only task given to participants was an orthogonal one-back repeat detection task; no focus on frequent or correlated features was encouraged or required to succeed at this task. We then tested participants' category structure knowledge in a two-alternative forced choice (2AFC) task and in a feature selection task, in which they were asked to select the three features most important for the category. Whereas the 2AFC task indirectly taps into feature correlation and frequency information, the feature selection task more explicitly taps into feature frequency sensitivity. We again ran simulations of the neural network model to test whether an abstraction model of category representation could replicate patterns of human learning.

Thus, this experiment provides a more conservative test of our prediction that feature correlations are abstracted during encoding. It is more conservative in three ways. First, the category structures differ in higher-order clustering rather than simple pairwise correlations: a Modular core benefit would imply that the feature-based structures encoded into category representations are more complex than simple pairwise associations. Second, category structure is completely incidental to the learning task, and therefore an observed Modular core benefit would imply that structure is automatically encoded into category representations. Third, encoding and retrieval stages are cleanly separated in time, enabling a clear evaluation of what is abstracted during category encoding.

Experiment 3: Behavioral

Method

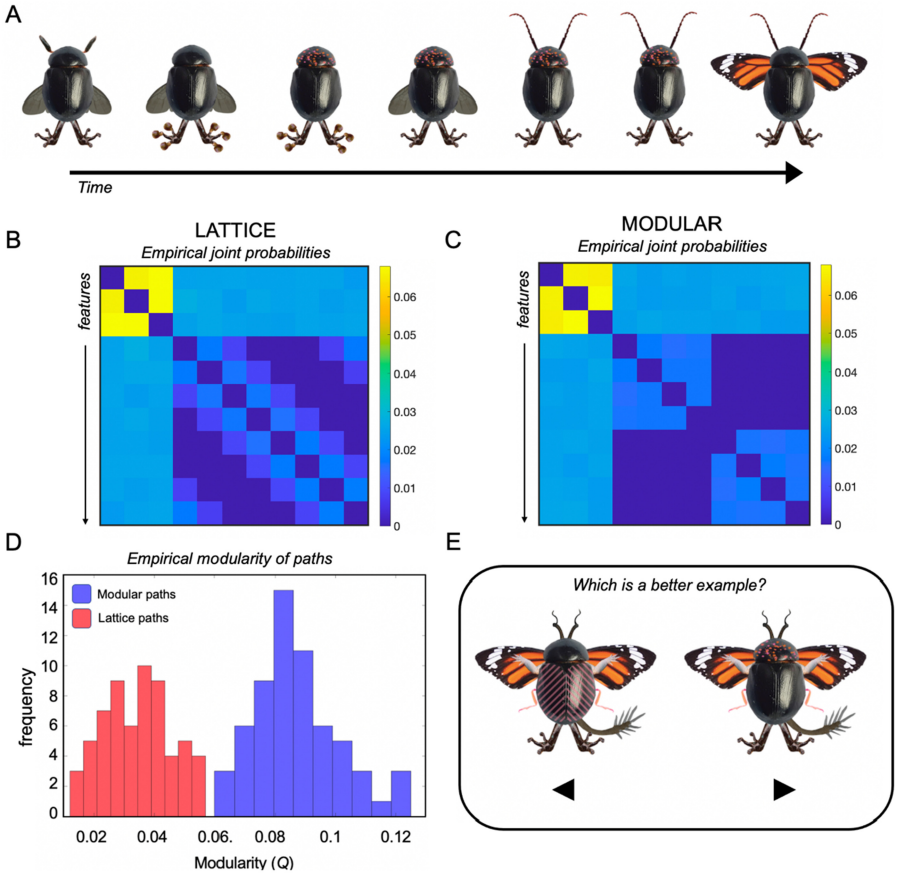
Participants. One hundred participants contributed data to Experiment 3 and were recruited from Amazon Mechanical Turk (Mean age: $M = 39.2$, $SD = 10.1$; 45% female). An additional 16 participants were excluded from analysis based on poor performance on the orthogonal task. Participants received a base pay of $2 in addition to a bonus of up to $6 based on performance on the orthogonal task (M=$4.79 bonus for 100 participants). Consent was obtained for all participants in accordance with the University of Pennsylvania IRB.

Categories and Features. The beetle and butterfly species in Experiment 2 were used in Experiment 3. No catch features or category labels were used.

Category Structures. Experiment 3 examined the same Modular structure from Experiments 1 and 2, but replaced the Random structure with a Lattice structure (Figure 1C). The Lattice graph contains the same core structure as the other two graphs, but the peripheral structure differs. Like the Random graph, the Lattice graph has no clustering in its periphery. However, unlike the Random graph, the peripheral nodes are not fully connected, meaning that each peripheral node is only connected to a subset of all peripheral nodes. There were two motivations for using a Lattice instead of a Random graph: (a) The structure of the Lattice graph enables peripheral structure knowledge to be trained and tested, and (b) the degree of pairwise feature correlations is more balanced with the Modular graph. That is, in our generated streams of categorical stimuli described below, each peripheral node can only be followed by three other peripheral nodes in both the Modular and Lattice structures. Comparing Modular and Lattice structures thus enables a more direct assessment of the specific influence of higher-level feature clustering on core structure learning. Assignment of species to structure and subsequent assignment of features to graph nodes was fully randomized for each participant. That is, any of the 11 species-specific features could become a core feature or a peripheral feature within a category.

Generating Stimuli for the SL Task. We designed an SL task that exposed participants to a category's feature co-occurrence structure by embedding co-occurrence statistics in a temporal stream of stimuli (Figure 6A). Each SL task trial consisted of an exemplar with two features (x and y) that were directly connected on the graph. The subsequent trial would contain one of the same features (y) and a new feature (z), directly connected to y. The order of the

Figure 6
Experiment 3: Tasks and Design



Note. Participants were exposed to a single category with either a Lattice or Modular structure. (A) In the SL task, participants viewed a stream of 550 two-feature exemplars. On each trial one feature was swapped for another, and the order of features was determined by a random walk over the category structure graph. Category structure was incidental to the orthogonal one-back repeat detection task. The actual joint probabilities between features successfully reflected the (B) Lattice and (C) Modular structures. (D) The modularity distinction between structure conditions was successfully maintained in the stimulus streams. (E) After the SL task participants completed a 2AFC task, in which they were asked to select which of two exemplars was a better example of the learned category. SL = statistical learning; 2AFC = two-alternative forced choice. See the online article for the color version of the figure.

features was thus determined by a random walk across the assigned graph, with two features on the walk displayed per trial. These random walks contained 550 steps (i.e., stimuli) and were generated based on transitional probability matrices that maintained the underlying feature-based Modular and Lattice structures. In both Modular and Lattice categories, each peripheral feature could be followed by three other peripheral features (in addition to the three core features). In order to ensure that other various properties of the random walks were matched across Modular and Lattice structures, we first generated a set of 100 walks per graph which all contained balanced frequencies of core and peripheral features. All paths contained a total of 273–277 presentations of core features—Lattice: $M = 274.8$; Modular: $M = 274.9$; $t(122) = 0.36$, $p = .72$. In the set of Modular walks, we balanced the frequency of module1 and module2 features such that features from each module were seen 135–140 times. From this set of possible paths, we subsampled 62 paths for each graph structure such that Lattice and Modular structures were exactly matched on the number of one-back repeat trials ($M = 79.2$, $SD = 6.7$). To ensure that the feature transition statistics in these walks mirrored the intended Modular and Lattice structures, we used the feature transition statistics within each of the final walks to derive the actual joint feature probabilities for each structure (Figure 6B and 6C). We calculated the modularity (Q) of these derived graphs to ensure that the structures underlying the Modular walks were indeed more modular than the structures underlying the Lattice walks (Figure 6D). Modularity was calculated using the Brain Connectivity Toolbox (brain-connectivity-toolbox.net; Rubinov & Sporns, 2010).

Generating Stimuli for the 2AFC Task. For each category structure, we generated 24 correct, structure-consistent exemplars with six features each. Each of these correct exemplars was paired with an incorrect, structure-inconsistent exemplar, in which one of the correct exemplar's features was swapped for another feature that violated the assigned category structure (Figure 6E).

Using six-feature exemplars, the Lattice category structure allows for 24 unique correct exemplars and the Modular category structure allows for eight unique correct exemplars; to match the number of exemplars across categories, three sets of the eight unique Modular exemplars were used. Within each category, each of the 24 correct exemplars was assigned to test either core (eight) or peripheral (16) structure knowledge. For each correct exemplar, the exact feature that was replaced to create the paired incorrect exemplar depended on the kind of structure (i.e., core, peripheral) the pair was designed to test. We will refer to these different correct/incorrect exemplar pairs as "Core pairs" and "Peripheral pairs."

In Core pairs, the incorrect exemplar was created by replacing one of the three core features in the correct exemplar with a peripheral feature that did not violate the peripheral structure. For example, in a Modular category exemplar containing features from module1, a core feature would be replaced with another module1 feature. Thus, the features of the incorrect exemplar are consistent with the category's peripheral structure, while the core structure is violated. A similar strategy was used to generate Core pairs for the Lattice category.

In Peripheral pairs, the incorrect exemplar was created by replacing one of the three peripheral features in the correct exemplar with another peripheral feature that did violate the peripheral structure. For example, a module1 feature might be replaced with a module2 feature; since module1 and module2 features never co-occur, the peripheral structure is violated while the core structure remains intact. A similar strategy was used to generate Peripheral pairs for the Lattice category, in which all incorrect exemplars contained a feature pair with 0% joint probability, never appearing together in the SL task.

The exemplar pairs were designed such that feature frequencies were balanced within Modular and Lattice categories. Across all exemplar pairs within each category structure, each peripheral feature appeared nine times in correct exemplars and 10 times in incorrect exemplars; each core feature appeared 24 times in correct exemplars and 21–22 times in incorrect exemplars. The specific consistent features that were swapped with inconsistent features were also balanced: In both category structures, each peripheral feature was the consistent feature twice (i.e., the feature removed in the incorrect exemplar) and appeared three times as the inconsistent feature in incorrect exemplars. Each core feature was the consistent feature two to three times; core features were never used as inconsistent features.

Task Protocols. Each participant was exposed to a single category whose species (i.e., Beetle or Butterfly) and structure (i.e., Modular or Lattice) was randomly assigned. The assignment of specific features to graph nodes was also randomized for each participant. Participants were told that they were part of a scientific research team that had discovered a new insect species and that "the overall goal of the project is to better understand the characteristics and features that define this new species." However, participants were also informed that "the collected specimens are old and sometimes they are missing body parts (e.g., wings, legs) or their color markings have faded." As an initial exposure to the category, the participants were shown four full category exemplars, each with six features that were consistent with the assigned category structure. The participants then completed three tasks: (a) an SL task, (b) a 2AFC task, and (c) an explicit feature selection task.

SL Task. Participants viewed a stream of "partial" category exemplars with two features each (Figure 6A). The insect base remained identical throughout the experiment, but one of the two features was replaced by another feature on each subsequent trial based on one of the random walks described above; the specific walk shown to each participant was chosen randomly out of the 62 options. The SL task comprised 550 trials: In each trial the exemplar remained on screen for 1,800ms, followed by a 200 ms interstimulus interval with a blank screen. The participants performed an orthogonal one-back task in which they pressed a key on each trial to indicate whether the exemplar on that trial was identical to the exemplar on the previous trial; participants were instructed to press the right arrow key for repeats and the left arrow key otherwise. Repeat exemplars were presented on ~14% of the trials. Feedback was given for hits, misses, and false alarms: a green star appeared beneath the exemplar on a correct hit and a red cross appeared for misses and false alarms. A d-prime sensitivity measure was calculated and updated on each trial, transformed into a percentage score (0%–100%), and displayed beneath the insect stimuli for the entirely of the SL task to motivate participants to pay attention to the stimuli and successfully complete the one-back task. The participants were told ahead of time that their final score on the last SL trial determined their bonus payment. A set of practice trials using abstract shapes was given to participants before the main SL task to acclimate them to the task and clarify instructions.

2AFC Task. In the 2AFC task, participants were instructed to use their new knowledge of the species to decide what full specimens should look like. Specifically, they were asked to decide on each trial

which of the two specimens is a better example of the newly discovered insect (Figure 6E). Each of the 24 trials consisted of one of the Core or Peripheral exemplar pairs described above, in which one exemplar was consistent with the category's feature-based structure and the other was inconsistent. The two exemplars were shown side by side in the center of a white screen; left/right placement of correct/incorrect exemplars was randomized on each trial. Stimuli remained on the screen until participants made a response by pressing the left or right button on their keyboard. The category stimuli disappeared from the screen 500 ms after a response, and 1,000 ms later the next trial began.

Explicit Structure Task. In the final task, participants were asked to "click on the three features that are most important for this species." The 11 category features were presented individually, each one attached to the greyscale insect base. The 11 features appeared in a grid on the screen, and participants were asked to click on three features to make their response. When a feature was selected, a blue outline appeared around the feature so the participants could keep track of their responses. Once three features were selected, the task immediately ended.

Statistical Analysis. In the SL task data, we analyzed participants' repeat detection sensitivity in order to exclude participants who did not pay sufficient attention to the stimuli. We used a d-prime sensitivity measure, in which sensitivity was defined in terms of the hit rate and false alarm rate. We excluded 15 participants whose sensitivity score was below 50% by the final trial and one additional participant who self-reported taking notes during the experiment, resulting in 16 participants (separate from the final $N = 100$) that were excluded from subsequent analyses.

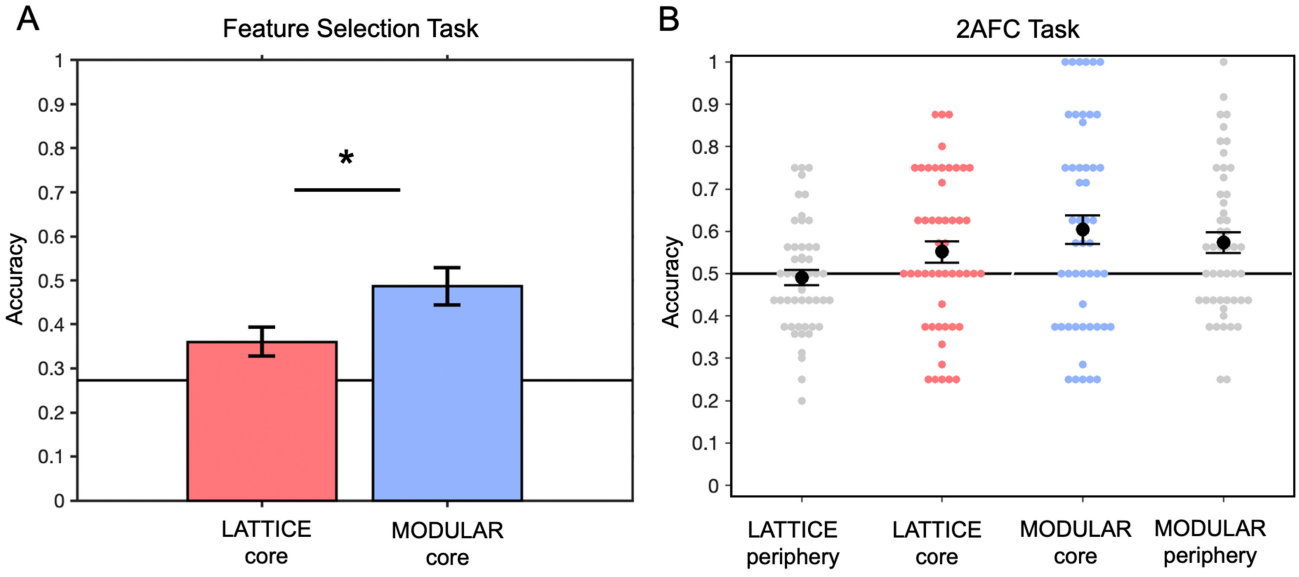
We analyzed reaction time (RT) and accuracy on the 2AFC task. We used RT to exclude trials in which participants' responses were unlikely to be meaningful (RT $<$ 250 ms or RT $>$ 10,000 ms). In the remaining trials, we determined whether participants chose the correct (1) over the incorrect exemplar (0). We averaged these accuracy scores separately for Core and Peripheral structure trials, resulting in a mean Core accuracy and mean Peripheral accuracy for each participant. Independent *t*-tests were used to compare mean Modular versus Lattice core structure accuracy, and Pearson correlations were used to assess relationships between conditions.

In the Explicit structure task, we calculated accuracy based on the number of correct core features each participant selected (0–3) and compared it against chance-level performance (0.27). Independent *t*-tests were used to compare accuracy between Modular and Lattice structures. Data from Experiment 3 is available on OSF.

Results

Accuracy on 2AFC Task. We determined whether participants successfully learned core and peripheral structures of Lattice and Modular categories using data from the 2AFC task (Figure 7B). While participants' ability to learn Lattice core structure was inconclusive, $M = 55\%$, $SD = 18.1\%$, $t(49) = 2.0$, $p = .05$, accuracy for Modular core structure was reliably above chance, $M = 60.4\%$, $SD = 24.1\%$, $t(49) = 3.1$, $p = .004$; however, there was no reliable difference between Lattice and Modular core accuracy, $t(98) = 1.2$, $p > .2$. Our data also suggests that participants were not able to learn the Lattice periphery structure, $M = 49.1\%$, $SD = 12.9\%$, $t(49) = 0.5$, $p > .6$, but were successful at learning Modular periphery structure, $M = 57.4\%$, $SD = 17.6\%$, $t(49) = 2.9$, $p = .005$; accuracy was significantly higher for Modular versus Lattice periphery structure, $t(49) = 2.7$, $p = .01$. No reliable relationship was found between core and periphery accuracy in either the Lattice, $r(49) = 0.11$, $p > .4$, or Modular, $r(49) = 0.14$, $p > .3$, category.

Figure 7
Behavioral Results From Experiment 3



Note. (A) In the explicit feature selection task, participants found it easier to select the three core features from the Modular versus the Lattice category, $t(98) = 2.3$, $p = .02$, replicating the preference observed in Experiments 1 and 2. (B) The 2AFC task did not reveal a significant difference in accuracy between Modular core and Lattice core trials. 2AFC = two-alternative forced choice. * indicates $p < .05$. See the online article for the color version of the figure.

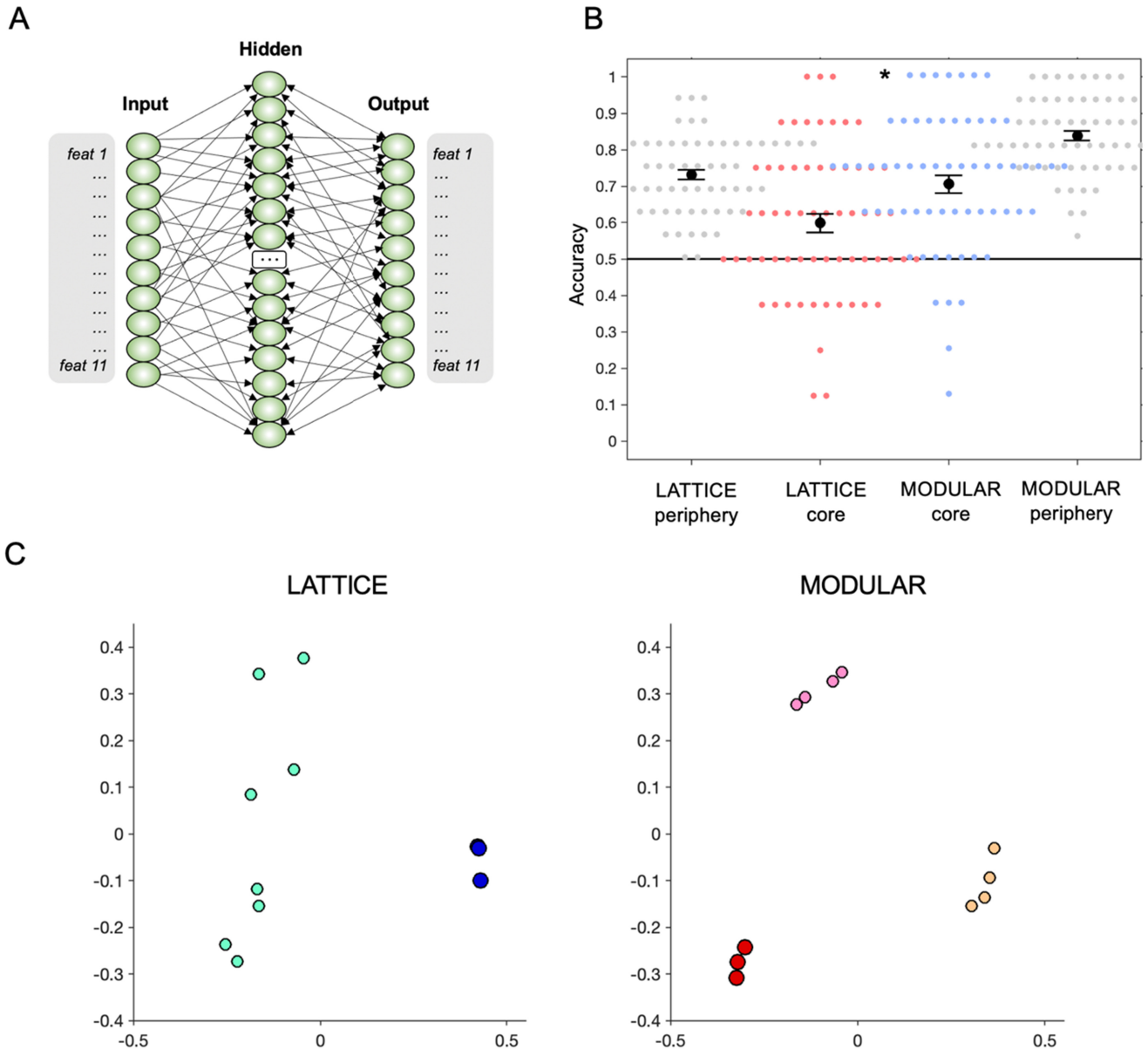
Accuracy on Explicit Feature Selection Task. We also examined core structure learning in the explicit feature selection task, in which participants selected the three features that they thought were most important for the category (Figure 7A). Accuracy was reliably above chance for both Lattice, $t(49) = 2.7$, $p = .01$, and Modular, $t(49) = 5.0$, $p < .0001$, categories. Consistent with results from the previous behavioral experiments, accuracy was significantly higher for Modular relative to Lattice core structure, $t(98) = 2.3$, $p = .02$, 95% CI = [0.02–0.234], Cohen's $d = 0.47$.

Experiment 3: Neural Network Model

Method

Model Architecture and Parameters. The model used in Experiment 1 was modified slightly for Experiment 3. The input and output layers were restricted to 11 features each since the model was trained to learn only one category at a time (Figure 8A). We also removed the inhibition on the output layer so that the model was able to activate all features simultaneously

Figure 8
Neural Network Model Simulations in Experiment 3



Note. (A) The model architecture contained 11 feature units on the input and output layers and a 400-unit hidden layer. (B) After training on the SL stimuli, the model performed more accurately on Modular core vs. Lattice core 2AFC trials, $t(122) = 3.0$, $p = .003$. (C) MDS scaling solutions for the final feature representations in the Lattice and Modular categories. Lattice core features are blue; Modular core features are red. SL = statistical learning; 2AFC = two-alternative forced choice; MDS = multidimensional scaling. * indicates $p < .05$. See the online article for the color version of the figure.

despite learning to only activate two features during training (see below).

Training Protocol. The model was trained on the 62 Lattice and 62 Modular paths generated for the SL task, resulting in 62 Lattice and 62 Modular simulations. Training consisted of 550 trials in which the model was presented with two features on the input layer and learned to replicate the same two features on the output layer. The training trials were presented in the same order shown to human participants.

Testing Protocol. Testing trials were analogous to the 2AFC trials shown to human participants. On each of the 24 trials, the model was presented with the five features shared by the correct and incorrect exemplars. We observed the resulting activity on the output layer to determine whether the correct exemplar's extra feature (structure-consistent) or the incorrect exemplar's extra feature (structure-inconsistent) was more strongly activated. Binary accuracy was assigned on each trial based on whether the consistent (1) or inconsistent (0) feature was more strongly activated.

Model Assessment. We used the 62 Modular and 62 Lattice paths to train and test the models, resulting in 62 simulations per structure. For each structure, and within each simulation, we calculated the mean accuracy separately for core structure and peripheral structure trials. Accuracy was compared against chance (0.5) and independent *t*-tests were used to assess differences in core and peripheral structure learning across structure conditions.

Analyzing Feature Representations. After the full training protocol, the model was probed with each of the 11 category features by setting each individual feature unit's activation to 1 on the input layer. Settled activation in the 20×20 hidden layer was extracted, resulting in 11 final feature representations. We assessed feature similarity using cosine distance and extracted network measures from each run's learned feature structure using the Brain Connectivity Toolbox (Rubinov & Sporns, 2010). For each network, we calculated its modularity, and also the local strength of each node (i.e., weighted connections of each feature to all other features). Local feature strengths were averaged to provide a network-level measure of overall feature connectivity. MDS was used to visualize the space of learned feature representations.

Results

Mean model accuracy across runs for Lattice and Modular categories is shown in Figure 8B. While human participants were able to learn the Modular but not the Lattice periphery structure, the model was able to learn the peripheral structure of both the Modular, $M = 83.9\%$, $SD = 10.5$, $t(61) = 25.5$, $p < .0001$, and Lattice categories, $M = 73.2\%$, $SD = 10.7$, $t(61) = 17.1$, $p < .0001$. The model also learned the core structure of the Modular, $M = 70.6\%$, $SD = 19.0$, $t(61) = 8.5$, $p < .0001$, and Lattice, $M = 59.9\%$, $SD = 20.3$, $t(61) = 3.8$, $p = .0003$, categories at above chance levels. Most notably, model accuracy for Modular core structure was greater than accuracy for Lattice core structure, $t(122) = 3.0$, $p = .003$, 95% CI = [0.037–0.177], Cohen's $d = 0.41$.

MDS solutions to the final learned feature representations, averaged across runs, are visualized in Figure 8C. On visual inspection, in both categories, core features are separated from peripheral features and the unique peripheral structures are discernable, demonstrating that the learned distributed representations were impacted by the patterns of feature associations. As in Experiment 1, it is possible that less interference among feature groups in the Modular categories can account for the observed Modular core benefit. To compare the shape of the learned representational spaces across category structures, we quantified differences between the Modular and Lattice runs using network strength to assess the degree of pairwise connectedness between the features and modularity to characterize each network's higher-level structure. We observed no difference in network strength, $t(122) = 1.01$, $p > .3$, indicating that pairwise feature association strengths in the Modular and Lattice conditions were equated overall. However, the modularity of the learned feature networks in Modular ($M = 0.22$) and Lattice ($M = 0.18$) conditions differed significantly, $t(122) = 2.82$, $p = .006$. This suggests that the Modular advantage is driven, as in Experiment 1, by the differentiated representations of the feature clusters in that category.

Experiment 3: Discussion

In Experiment 3, we replicated the Modular core benefit observed in Experiments 1 and 2 in a different task paradigm and with a new graph structure. Participants were each exposed to either a Modular or Lattice category in a temporal SL paradigm in which category structure was entirely incidental to the orthogonal task. Participants subsequently completed a 2AFC task which tested their ability to detect structure-consistent exemplars, and a feature selection task which tested sensitivity to the three high-frequency core category features.

The 2AFC results demonstrated that participants successfully learned Modular core structure, while ability to learn Lattice core structure was inconclusive. However, Modular core accuracy was not significantly greater than Lattice core accuracy in this task. Given that Experiment 3 instantiated a very conservative test of the previously observed Modular core benefit, it is not that surprising that a smaller difference was observed. This finding can be interpreted in a couple different ways. First, the unreliable difference between Lattice and Modular core structure knowledge in the 2AFC task can be interpreted to reflect the fact that both structures contained pairwise correlations between peripheral features. These pairwise feature covariations likely facilitated Lattice category learning overall, relative to the Random structure used in the first two experiments. These results suggest that pairwise feature correlations are abstracted during category encoding even in the absence of higher-level feature clustering. Note, however, that participants exposed to a Lattice category did not reliably learn the core feature structure, whereas participants exposed to a Modular category were reliably above chance on tests of core structure knowledge. While the difference between Modular and Lattice core knowledge on the 2AFC task was not reliable, the pattern suggests that higher-level structure—beyond pairwise associations—may be relevant during novel category encoding. The second interpretation is that the purely incidental nature of the task did not encourage focus on internal category structure and therefore feature associations were not encoded. However, participants in the Modular condition did exhibit sensitivity to peripheral structure in the 2AFC test, and the results of the feature selection task show that the categories' peripheral structure did indeed influence category learning more broadly.

In the feature selection task, participants were asked to select the three most important category features. Correctly selecting the three

core features reflects participants' sensitivity to the category's core structure. Replicating the Modular core benefit from Experiments 1 and 2, participants were more sensitive to core features in the Modular versus Lattice condition. Even though Experiment 3 was a conservative test of our hypotheses, these findings suggest that feature associations are encoded during category exposure and that increased feature clustering benefits learning. Increased clustering in novel categories thus appears to benefit category learning across stimulus modalities and experimental tasks.

We also found that a neural network model was able to successfully replicate the Modular core benefit observed in human behavior. The model was exposed to the same pairs of features presented to human participants in the SL task, and a subsequent test analogous to the human 2AFC task revealed that the model was more likely to generate correct relative to incorrect exemplars. In human behavior, the Modular core benefit emerged in the Feature Selection task rather than the 2AFC task. However, an overall facilitation for Modular categories was observed in both humans and neural network model simulations in Experiment 3. Furthermore, a graph-based analysis of the model's learned feature representations suggests that this difference is driven by the high-level modular structure, rather than pairwise feature associations. While these simulations alone cannot rule out alternative theories of category representation, we provide additional simulations in the online supplemental materials suggesting that the neural network model is better able to explain these findings than standard exemplar or prototype models, which did not exhibit the Modular core benefit. Taken together, these findings suggest that the Modular core benefits observed in Experiments 1, 2, and 3 emerge from a shared mechanism that encodes feature co-occurrence statistics across episodes, regardless of stimulus modality or task design.

General Discussion

Across three experiments, we aimed to determine whether certain structures facilitate category learning and whether internal category structure is encoded into category representations. We hypothesized that, if feature correlations are encoded into category representations during learning, this should have consequences for category learning and representation more broadly. We predicted that different patterns of feature associations (peripheral structure) would shape the learned category representation and influence learning of a separate set of features (core structure). These predictions are consistent with a relational abstraction model of category representation, in which feature associations are encoded but traces of the original exemplars are lost. To test these predictions, we designed novel categories characterized by different graph structures such that some categories contained multiple communities of reliably co-occurring features (Modular) whereas other categories did not (Random, Lattice). All categories had a separate set of high-frequency core features, and testing these features let us examine any potential influence of feature correlations on the emergent category representations. Human behavior in the three experiments revealed that reliable feature covariation facilitated category learning. A neural network model—which instantiates the relational abstraction theory—revealed the same learning patterns observed in humans. These behavioral and modeling results are consistent with an abstraction theory of category learning, in which feature correlations are embedded within category representations themselves.

Coherent Covariation and Structure Learning

Our observed benefit for Modular categories adds to our understanding of how humans extract structure from their environment more generally. Outside of the semantic domain, recent work has highlighted humans' sensitivity to structure, characterized by underlying graph structures like the ones used here (Kakaei et al., 2021; Karuza et al., 2019; Lynn & Bassett, 2020; Lynn et al., 2020; Schapiro et al., 2013). Karuza et al. (2019) use novel visual stimuli in a series of behavioral experiments to reveal that humans are sensitive to graph structures containing a range of different numbers and sizes of communities, or modules. A recent behavioral study (Kakaei et al., 2021) suggests that temporal community structure accelerates recognition learning of novel visual objects and affects the order in which objects are learned. Using a sequential motor task in humans, Lynn et al. (2020) utilize random walks across different graph structures to reveal overall faster responses on a modular graph relative to a lattice graph. Additional behavioral work suggests that humans can extract higher-order structure from the environment and leverage this learned structure to aid subsequent learning (Mark et al., 2020). The formation of potential "cognitive maps" similarly involves the extraction and representation of environmental structure that can be applied across many domains (see below). Thus, humans appear sensitive to higher-order structure, and this has implications for real-world learning in domains such as navigation, language acquisition, event learning, and knowledge accumulation (Karuza et al., 2016; Lynn & Bassett, 2020; Schapiro et al., 2013). There also appears to be a common thread through behavioral work in humans revealing a learning benefit for modular graphs, with clusters of reliably associated nodes. Our present findings extend this phenomenon to a semantic context: Humans find it easier to learn categories characterized by a modular structure containing sets of reliably co-occurring semantic features. Furthermore, empirical support for the claim that structure is encoded within learned representations is relevant within any domain concerned with cognitive or neural representations.

Like humans, our neural network models revealed a learning benefit for Modular categories. These findings complement previous computational investigations of semantic learning. The semantic models of Rogers and McClelland (2004) were able to leverage coherent covariation among features to progressively differentiate semantic categories during learning in a trajectory that corresponds with the development of semantic memory in human infants. For example, animals share many features with each other (e.g., *can move*, *has skin*) that are not shared with plants. The model is presented with features of different animals and plants (e.g., *bird*, *tree*) and learns internal representations of these items across simulated developmental timescales. The model's learning algorithm drives items with shared features to share connection weights, which leads to increased similarity among animal representations (e.g., BIRD, FISH) and increased similarity among plant representations (e.g., TREE, FLOWER). The lack of shared features between animals and plants results in differentiation between these semantic classes. In other words, the coherent covariation among features enables the formation of semantic representations in the Rogers and McClelland model that reflect the real-world semantic hierarchy. Saxe et al. (2019) formalized a notion of category coherence that describes more precisely how deep linear neural networks extract categories from noise as a function of feature co-occurrence probabilities within subsets of items. In our current studies we show that

coherence can quickly govern representation building in a novel domain, in both humans and neural network models, and that this learning can reflect the structure *within* a category. Increased coherence within the Modular category input drove similarity of feature representations within modules and differentiation between modules, which led to better overall understanding of the Modular category in the models and in people.

There is a close relationship between coherent covariation and hierarchical organization which deserves some further attention. Saxe et al. (2019) explain that: "coherent categories consist of large subsets of items possessing, with high probability, large subsets of features that tend not to co-occur in other categories." A hierarchy, on the other hand, is an explicit structural representation that describes a layered organization of categories, superordinate categories, subordinate categories, and so on. The concept of coherence is more general and allows for more fuzziness than a strict hierarchy would allow, but categories with high coherence can also be described by a hierarchy. Indeed, the Rogers and McClelland (2004) neural network model of semantic learning leveraged coherent covariation among semantic features to learn representations that reflected the hierarchical organization of semantic space. While our neural network model could represent the structure of Modular categories without explicit hierarchical representations, our results could be recast as an effect of hierarchical structure on learning. The two categories can also be described as differing in "complexity," with the peripheral Modular components being simpler than the periphery of the Random or Lattice categories. Complexity, hierarchy, and coherent covariation usually track closely, but can in theory be disentangled, and future work could explore their unique impacts on category learning.

The potential role of additional cognitive factors in this structure learning process also remains to be explored. For example, it is possible that distinct category structures result in different patterns of attentional allocation across features, influencing how categories are learned. While our neural network model mirrored human performance simply through direct sensitivity to feature structure and without an attentional mechanism, it will be important for future work to investigate the role of attention and other possible mediating factors in these effects of structure on learning. Another open question is how internal category structure influences categorization. While our own experiments used inference tasks to boost within-category learning, our results generate predictions regarding classification tasks. Specifically, since Modular structures promote the efficient clustering and differentiation of feature representations, we might predict increased classification performance between two Modular categories, as opposed to two non-Modular categories. Further exploring the influence of internal category structure on inference as well as classification performance will be a fruitful direction of future work.

Implications for Models of Category Learning

A crucial question in theories of conceptual and category knowledge relates to the content and structure of the underlying representations. What information is contained within a category representation, and how is it stored? Broadly speaking, the main theoretical divide is between exemplar and abstraction theories of representation. In exemplar theories, a category representation consists of stored exemplar traces. In abstraction theories, a category is represented in terms of an abstracted, centralized category representation in which information from individual exemplars has been integrated. The most well-known abstraction theories claim that these centralized representations are "prototypes," which reflect the category's central tendency. When exemplar and prototype models are compared, especially in the context of feature correlation learning, exemplar theories receive more empirical and computational support (e.g., Medin & Schaffer, 1978). However, the range of abstraction models is vast. Whereas prototype models abstract and encode a category's central tendency, other models abstract and encode a category's feature correlations, thus retaining its feature-based structure. This flavor of abstraction model is instantiated in neural network models such as the one described by Rumelhart & McClelland, in which feature associations are encoded via the fine-tuning of weights during gradient descent learning. These relational abstraction models are predicted to behave similarly to exemplar-style models in the context of feature correlation learning (Medin & Schaffer, 1978). However, they differ in their assumptions of how feature correlation sensitivity arises: relational abstraction models abstract feature correlations during encoding and they are embedded within the category representation; exemplar models do not store feature correlations but rather abstract them at the time of retrieval. Dissociating the predictions generated by these two disparate theories of category representation is theoretically and empirically difficult (Barsalou, 1990).

Here our aim was to provide a relational abstraction model of category learning that can explain patterns of category structure learning in humans. Relational abstraction models, like the neural network model used here, encode feature correlations into the learned category representations. If feature correlations are embedded into the new category representation during encoding, this feature-based structure will likely influence subsequent processing. We incorporated these hypotheses into a design such that we could observe whether correlated structure in one set of features influenced learning of a completely different set of features, and predicted that both humans and a neural network model would reveal such an influence. Indeed, across three behavioral experiments we observed increased accuracy for core features in categories containing increased clustering of correlated features in their periphery. A neural network model revealed similar patterns of behavior. Taken together, these empirical and computational results show that a relational abstraction model can indeed capture this kind of human category learning behavior. While we cannot claim that this model outperforms exemplar-style models based on GCM simulations applied to Experiment 1, Experiment 3 provides intriguing evidence that, at least in some cases, a relational abstraction model may indeed outperform exemplar models in the context of category structure learning (see the online supplemental materials).

It is important to note that a relational abstraction theory of category representation need not be implemented in a neural network model, and not all neural network models instantiate a relational abstraction theory. Addressing the former point, early theorists supporting relational abstraction models argued that, during category learning, separate memory representations are created for all individual properties and property conjunctions (Hayes-Roth & Hayes-Roth, 1977). Much like how a modal prototype model tracks the frequency of individual features, the "property-set model" tracks the frequency of feature combinations. Feature associations are explicitly encoded and no network organization is invoked. Conversely,

neural network models can be designed to implement a wide range of theoretical models (Gluck & Bower, 1988). What the model represents is determined by its architecture and given inputs, and various decisions can result in an exemplar or abstraction model of category representation. Our current model's ability to reflect human category learning lies in its use of distributed representations—its input nodes correspond to localist feature units, but their weighted connections to the hidden layer result in a distributed code in which any extractable, useful information can be represented (e.g., features, categories, feature associations).

Exemplar-style network models do not share all of these characteristics. For example, REMERGE is a neural network model built to demonstrate how inference and categorization can be accomplished by orthogonalized neural codes in the hippocampus (Kumaran & McClelland, 2012). Its architecture contains a layer of localist feature units and a layer of localist conjunctive units; recurrent processing between these layers enables co-activation of multiple conjunctive units that code for indirectly related experiences. In the case of categorization, each training exemplar is assigned one conjunctive unit, and the features of individual test items will activate these units in proportion to their featural overlap (i.e., neither feature nor category representations are distributed). The retrieval dynamics in REMERGE carry out a function closely analogous to the similarity-based computation implemented in GCM (Kumaran & McClelland, 2012, Appendix). Indeed, the model shares the core characteristic of GCM: storing individual exemplars separately.

As another example, ALCOVE is a neural network model that pairs an exemplar theory of category representation with an error-driven learning mechanism (Kruschke, 1992). Similar to REMERGE, both features and categories are represented in localist rather than distributed codes. ALCOVE is also similar to GCM, but error-driven weight changes between feature units and exemplar units allows for increased attention to specific feature dimensions, and weight changes between exemplar and category units enables exemplars to differentially contribute to category decisions. When the attention-learning rate is high, ALCOVE is sensitive to correlated feature dimensions, consistent with our own GCM simulations (online supplemental materials). Additionally, ALCOVE's error-driven learning rule enables it to model forms of base-rate neglect in humans that non-network context models cannot (Nosofsky et al., 1992). However, the architecture of this model does not enable the encoding of feature associations, preserving an exemplar-based representational system.

Pushing away from explicit exemplar-style representations, SUSTAIN is another neural network model of category learning containing a powerful dimensional attention mechanism. It does not represent exemplars in a distributed code, but does not explicitly store representations of individual items either. Rather, SUSTAIN represents categories as "clusters" within a multidimensional feature space, with new clusters formed only when existing clusters lead to large prediction errors. Love et al. (2004) specifically intended to capture category substructure, and SUSTAIN would thus likely be able to capture the peripheral structures of our Modular and non-Modular categories. Like ALCOVE, attentional tuning within each dimension results in sensitivity to correlated feature dimensions, which can influence category decisions. However, SUSTAIN's ability to predict our Modular core benefit is unclear. Love et al. (2004) report that SUSTAIN prefers intercorrelated, versus non-intercorrelated, dimensions, but once the model learns one cluster of correlated features it finds it more difficult to learn another. This might lead to easier learning of core features in the non-Modular relative to Modular categories, the opposite of what we observed.

Many other neural network-style models can be considered to fall within the relational abstraction class of category learning models, and these would likely reflect our observed patterns of human behavior. The neural network model we employed here is only one instance of a multilayer network paired with error-driven learning mechanisms. For example, the divergent autoencoder (DIVA) model is a feedforward network employing error-driven learning and feature input nodes—its architecture enables the encoding of feature associations and would therefore fall within the relational abstraction class of category models (Kurtz, 2007). Here our general claim is that humans abstract and encode feature associations during category learning, and therefore category learning models that also encode feature associations should reflect human learning patterns. This argument applies to multilayer neural network models such as McClelland and Rumelhart (1985), Rogers and McClelland (2004), and Hinton (1986), as well as DIVA (Kurtz, 2007). We believe the empirically observed benefits for Modular categories likely require the kind of distributed representations that emerge in these models. This has implications for theories and models of category knowledge as well as the neural representations that we might expect to emerge during category learning. Further exploration of the computational landscape will be able to further reveal which learning and representational assumptions best explain human category learning and semantic inference.

Potential Neural Substrates

Based on prior neuroimaging investigations of SL, structure learning, and category learning, we believe it is likely that our effects are underpinned by the rapid development of distributed representations in the hippocampus. There is a growing literature indicating that the hippocampus tracks information across experiences in the service of category learning (Mack et al., 2018). Evidence from neuroimaging and studies with hippocampal amnesics similarly suggests that the hippocampus is recruited for rapid SL (Bornstein & Daw, 2012; Covington et al., 2018; Harrison et al., 2006; Schapiro et al., 2012, 2014, 2016; Strange et al., 2005; Turk-Browne et al., 2009, 2010), building overlapping representations that reflect statistically strong associations between stimuli (Schapiro et al., 2012). This ability to extract commonalities across experiences is in tension with the original CLS theory, which posited that the hippocampus houses sparse, separated neural patterns which are only subsequently consolidated as distributed representations in neocortex (McClelland et al., 1995). However, we have previously proposed that the hippocampus itself contains complementary learning systems due to different representational formats in subfields CA1 versus CA3 and Dentate Gyrus: A neural network model incorporating properties of these subfields demonstrates how sparse, separated patterns in CA3 and Dentate Gyrus that support episodic memory can coexist with distributed representations in CA1 that underlie rapid statistical learning (Schapiro et al., 2017). Our current model simulations suggest that distributed representations may be necessary to explain our category learning effects, leading us to hypothesize that these structured representations might emerge in the CA1 subfield of the hippocampus. Indeed, the models used here are functionally equivalent to

the monosynaptic pathway of our hippocampus model—the pathway connecting entorhinal cortex with region CA1.

Hippocampal involvement in the rapid extraction of category structure would be consistent with our previous finding that the hippocampus is sensitive to graph community structure. Schapiro et al. (2016) exposed humans to a sequence of abstract visual stimuli whose order was determined by random walks across a modular graph. Each module, or community, corresponded to a group of stimuli that appeared in close temporal proximity within the experiment. After exposure, multivoxel hippocampal patterns evoked by stimuli within the same community were more similar than those from different communities, revealing sensitivity to high-level graph structure. Additional evidence of structure learning and representation in the hippocampus comes from research on "cognitive maps." Tolman (1948) introduced cognitive maps as abstract, domain-general structures that represent relations between entities. These structured representations are powerful because they can enable inference of relationships between entities that have not been directly observed, leading to generalization and improved learning performance (Behrens et al., 2018; Whittington et al., 2020). Cognitive maps are also a powerful theoretical framework because they can be used to represent relations between varied kinds of entities, such as places (Epstein et al., 2017; Hafting et al., 2005; Javadi et al., 2017; Killian & Buffalo, 2018; O'Keefe & Nadel, 1978; Stachenfeld et al., 2017), objects (Mark et al., 2020; Whittington et al., 2020), people (Park et al., 2021; Tavares et al., 2015), concepts (Constantinescu et al., 2016), events (Hassabis et al., 2007), and other abstract spaces (Behrens et al., 2018; Franklin & Frank, 2018; Schuck et al., 2016; Theves et al., 2019). The medial temporal lobe and hippocampus specifically have been implicated in representing or otherwise processing structural elements of these cognitive maps (Constantinescu et al., 2016; Hafting et al., 2005; Hassabis et al., 2007; Javadi et al., 2017; Killian & Buffalo, 2018; Morton et al., 2020; O'Keefe & Nadel, 1978; Park et al., 2021; Stachenfeld et al., 2017; Tavares et al., 2015; Theves et al., 2019). Evidence thus suggests that graph or map-like structures rely at least in part on hippocampal representations, reinforcing the hypothesis that hippocampal representations underlie our observed effects of graph structure on category learning.

While the rapid learning capabilities of the hippocampus would facilitate structure learning in novel domains like the ones employed here, we would expect a consolidated form of this knowledge to rely primarily on neocortical areas that are known to support long-term conceptual knowledge, like the anterior temporal lobes (Patterson et al., 2007; Peelen & Caramazza, 2012; Ralph et al., 2010, 2017).

Relational Structure Within Real-World Concepts

We have demonstrated how correlational structure within a category's features—above and beyond the identity of the features themselves—can influence category learning more broadly. We interpret these findings to be consistent with a relational abstraction theory of category representation, in which feature correlations are encoded into the category representation itself, as opposed to an exemplar theory, in which a category is represented as the set of distinct exemplar traces. Considering theories of real-world conceptual knowledge, abstraction theories of representation are arguably more plausible than exemplar theories of representation; in fact, Murphy (2016) argues that an exemplar theory of concepts has not yet been put forward. An exemplar theory of concepts would have trouble representing hierarchical semantic knowledge (e.g., that a ladybug is a beetle, which is an animal, which is a living thing) as well as representing knowledge that is not tied to experience with exemplars (e.g., that whales have four-chambered hearts; Murphy, 2016). Furthermore, even if it were plausible to represent concrete or "entity" concepts (e.g., *ladybug, apple*) in terms of exemplars, it is unclear how that would work in the context of intangible concepts (e.g., *truth, democracy*).

Relational abstraction theories of representation are also consistent with theories of relational semantic knowledge. While some entity concepts can more easily be defined in terms of particular features (e.g., *ladybug*), other "relational" concepts are more easily defined in terms of particular relations between entities in the world (e.g., bridge, friend; Asmuth & Gentner, 2017; Gentner & Kurtz, 2005). The meaning of a relational concept comprises a certain relation between elements, rather than specific sensorimotor features. For example, a *bridge* is a structure connecting two entities—shape, material, and size features can be irrelevant for categorization. Indeed, a *bridge* might not be a physical object at all, but rather an abstract connection between two separate ideas. Relationality has consequences for how concepts are processed. For example, it is easier to imagine an ideal example of a relational concept relative to an entity concept (Goldwater et al., 2011), and relational concepts may be more semantically mutable across contexts than entity concepts (Asmuth & Gentner, 2017). Further, relational structure within different concepts or domains drives analogical reasoning (Gentner et al., 1993; Holyoak, 2012). Relationality is associated with conceptual abstractness, since both terms characterize concepts that are perhaps more appropriately or efficiently represented in terms of intangible, rather than concrete, features.

Similar ideas have been explored in the context of category learning. Feature relations play an important role in behavioral studies of "category coherence" and "abstract coherent categories" (Erickson et al., 2005; Murphy & Medin, 1985; Rehder & Ross, 2001; Spalding & Murphy, 1996; Wisniewski, 1995). Rather than referring to statistical regularities between any features, this body of work considers "coherence" to reflect known semantic, thematic, or causal relations between features that participants bring with them into an experimental setting. In other words, coherent categories are those whose feature combinations make sense in light of prior knowledge (e.g., *has wings, can fly*). Prior knowledge can affect interpretations of features during category processing and learning (Spalding & Murphy, 1996; Spalding & Ross, 2000; Wisniewski & Medin, 1994) and can make it easier to integrate features during category learning (Murphy & Allopenna, 1994). Incidentally, inference tasks promote learning of abstract coherent categories (Erickson et al., 2005) as well as relational discovery (Goldwater et al., 2018). Furthermore, similarity-based category learning models (e.g., exemplar and prototype models) cannot account for the acquisition of abstract, coherent categories (Erickson et al., 2005). In sum, it is easier to learn a new category when it contains sets of features that are known to be associated, or correlated, with each other in the real world. Correlated features in turn play an important role in relational category learning (Goldwater et al., 2018).

The novel insects learned in our current experiments did not contain rich semantic, thematic, and causal relations—indeed, our feature combinations could be considered *incoherent* with respect to prior knowledge (e.g., *antennae, claws*). The structure embedded within

each of our insect categories relied on feature correlations that were not imported from the real world but were learned online in the context of the experiment. Nevertheless, we replicated the finding that feature correlations benefit learning. While this is consistent with the category learning research summarized above, we need not invoke prior knowledge to explain our observed effect. Feature correlations and category coherence benefit learning even in the absence of prior knowledge. Our findings thus suggest how, without any prior semantic knowledge, feature correlations can aid concept learning *de novo* and may provide the scaffolding for relational meaning. More generally, representing concepts in terms of specific features as well as feature relations—and understanding how these feature relations are learned—may provide traction on understanding "relational" or "abstract" concepts that tend to be empirically elusive.

Representation of Abstract Structure

Our behavioral data reveal an effect of category structure on category learning, and our simulations suggest that the encoding of feature-based structure underlies this effect. However, we do not know whether the representations built during category learning are representations of structure *per se*. The effects of graph structure we observe here might be a result of feature-, exemplar-, or category-specific representations becoming more or less similar to each other during learning depending on their patterns of co-occurrence. Indeed, this is how the neural network models work. Relatedly, Saxe et al. (2019) offer a mathematical explanation of how the clustering or coherence of an environment influences learning in the context of a deep linear network model. It is possible, though, that category learning could sometimes result in, or depend on, abstracted representations of structure that exist independently from the feature representations themselves.

One influential framework that invokes explicit abstract structural representations is the structured statistical framework of Kemp, Tenenbaum, and colleagues (Kemp & Tenenbaum, 2008, 2009; Tenenbaum et al., 2011). This approach pairs representations of structural forms (e.g., hierarchical trees, directed graphs) with a domain-general statistical inference mechanism to model forms of inductive reasoning such as property induction. In these models, background knowledge of a domain is captured in an appropriate structural form in addition to a stochastic process that reflects how properties are likely to be distributed within the domain. For example, biological properties are best represented in a hierarchical tree paired with a "diffusion" process, whereas disease properties are best represented in a directed graph paired with a "transmission" process (Kemp & Tenenbaum, 2009). The combined structure and stochastic process generates a prior over which Bayesian statistical inference is performed. It is argued that the same statistical inference engine used for property inference can be used to learn the correct structure for a given domain (Kemp & Tenenbaum, 2008). In the context of our experiments, the structural statistical model would be tasked with (a) finding optimal structural forms for the Modular and non-Modular categories, (b) choosing the appropriate stochastic process(es), and (c) performing statistical inference over these priors and the observed category exemplars. However, it is important to note that our structural classifications (i.e., Modular, Random, Lattice; Figure 1) reflect the distances between *features* rather than the distances between exemplars; the structured statistical framework operates on exemplar-based structures. For example, the feature-based structure of our Modular categories has three clusters (i.e., core, mod1, mod2), but if the structure was redrawn based on exemplar distances it would only have two clusters (i.e., exemplars from mod1, exemplars from mod2). The structured statistical model would undoubtedly be able to learn this two-cluster structure. However, since core features are always present in all exemplars in all category structures—and thus do not relate to the exemplar-based structural forms—it seems unlikely that the statistical inference mechanism would more easily learn the core features in our Modular versus non-Modular categories. This approach could potentially be adapted, however, to address the phenomena about internal feature structure observed here.

If abstract representations of structure do emerge during learning, they can be empirically tested using a "structure transfer" paradigm. Mark et al. (2020) implemented such a paradigm to determine whether representations of structure are formed and whether they aid future learning. The researchers exposed participants to an environment defined either by a lattice or modular graph structure; visual images of real-world objects were assigned to different graph nodes, and the presence of graph edges indicated a possible temporal transition between these objects. After learning this initial environment, participants were introduced to a second environment structured according to the same modular graph but defined by a completely new set of visual objects. The researchers observed that participants who were initially exposed to a modular environment found it easier to learn a second modular environment, even though the environments had no visual objects in common. This suggests that participants not only learned the structure of the environment, but transformed it away from the environment's specific features and represented it in an abstract form that could then be applied to future learning environments (Mark et al., 2020). In the category learning context, future work could test whether learning a category characterized by a certain graph structure makes it easier to learn a second category with the same structure, even when the categories share no features in common. This would help clarify to what extent the brain learns representations of the category structure separate from the associations between particular category features.

Conclusions

We presented three human behavioral experiments and corresponding model simulations to test the influence of category structure on category learning. Our results provided support for an abstraction theory of representation in which feature correlations are encoded into the learned representation. We also found that category structure influenced how easily important category features were learned, and specifically that humans find it easier to learn categories containing sets of reliably co-occurring features. It is likely that clusters of reliably co-occurring features benefit structure learning more generally and that this learning is underpinned by rapidly formed distributed representations.

References

Ahn, W. K., & Kim, N. S. (2000). The causal status effect in categorization: An overview. In D.L. Medin (Ed.),*The psychology of learning and motivation* (pp. 23–65). Academic Press.

Anderson, A. L., Ross, B. H., & Chin-Parker, S. (2002). A further investigation of category learning by inference. *Memory & Cognition*, *30*(1), 119–128. https://doi.org/10.3758/BF03195271

Asmuth, J., & Gentner, D. (2017). Relational categories are more mutable than entity categories. *Quarterly Journal of Experimental Psychology*, *70*(10), 2007–2025. https://doi.org/10.1080/17470218.2016.1219752

Barsalou, L. W. (1990). On the indistinguishability of exemplar memory and abstraction in category representation. In T. Scrull & R. Wyer (Ed.), *Advances in social cognition, volume III: Content and process specificity in the effects of prior experiences* (pp. 61–88). Lawrence Erlbaum Associates.

Behrens, T. E., Muller, T. H., Whittington, J. C., Mark, S., Baram, A. B., Stachenfeld, K. L., & Kurth-Nelson, Z. (2018). What is a cognitive map? Organizing knowledge for flexible behavior. *Neuron*, *100*(2), 490–509. https://doi.org/10.1016/j.neuron.2018.10.002

Bornstein, A. M., & Daw, N. D. (2012). Dissociating hippocampal and striatal contributions to sequential prediction learning. *European Journal of Neuroscience*, *35*(7), 1011–1023. https://doi.org/10.1111/j.1460-9568.2011.07920.x

Bowman, C. R., Iwashita, T., & Zeithamova, D. (2020). Tracking prototype and exemplar representations in the brain across learning. *eLife*, *9*. Article e59360. https://doi.org/10.7554/eLife.59360

Bowman, C. R., & Zeithamova, D. (2021). *Coherent category training enhances generalization and increases reliance on prototype representations*. PsyArxiv. https://doi.org/https://psyarxiv.com/ct7ad

Carey, S. (1985). *Conceptual change in childhood* (Vol. 460). MIT Press.

Chin-Parker, S., & Ross, B. H. (2002). The effect of category learning on sensitivity to within-category correlations. *Memory & Cognition*, *30*(3), 353–362. https://doi.org/10.3758/BF03194936

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), 216–226. https://doi.org/10.1037/0278-7393.30.1.216

Constantinescu, A. O., O'Reilly, J. X., & Behrens, T. E. (2016). Organizing conceptual knowledge in humans with a gridlike code. *Science*, *352*(6292), 1464–1468. https://doi.org/10.1126/science.aaf0941

Covington, N. V., Brown-Schmidt, S., & Duff, M. C. (2018). The necessity of the hippocampus for statistical learning. *Journal of Cognitive Neuroscience*, *30*(5), 680–697. https://doi.org/10.1162/jocn_a_01228

Cree, G. S., McRae, K., & McNorgan, C. (1999). An attractor model of lexical conceptual processing: Simulating semantic priming. *Cognitive Science*, *23*(3), 371–414. https://doi.org/10.1207/s15516709cog2303_4

Devraj, A., Zhang, Q., & Griffiths, T. (2021, July 26–29). The dynamics of exemplar and prototype representations depend on environmental statistics. In *Proceedings of the 43rd Annual Meeting of the Cognitive Science Society*. Virtual.

Epstein, R. A., Patai, E. Z., Julian, J. B., & Spiers, H. J. (2017). The cognitive map in humans: Spatial navigation and beyond. *Nature Neuroscience*, *20*(11), 1504–1513. https://doi.org/10.1038/nn.4656

Erickson, J. E., Chin-Parker, S., & Ross, B. H. (2005). Inference and classification learning of abstract coherent categories. *Journal of Experimental Psychology: Learning, Memory, and Cognition*, *31*(1), 86–99. https://doi.org/10.1037/0278-7393.31.1.86

Franklin, N. T., & Frank, M. J. (2018). Compositional clustering in task structure learning. *PLoS Computational Biology*, *14*(4), Article e1006116. https://doi.org/10.1371/journal.pcbi.1006116

Franks, J. J., & Bransford, J. D. (1971). Abstraction of visual patterns. *Journal of Experimental Psychology*, *90*(1), 65–74. https://doi.org/10.1037/h0031349

Frost, R., Armstrong, B. C., & Christiansen, M. H. (2019). Statistical learning research: A critical review and possible new directions. *Psychological Bulletin*, *145*(12), 1128–1153. https://doi.org/10.1037/bul0000210

Gentner, D., & Kurtz, K. J. (2005). Relational categories. In W.-K. Ahn, R. L. Goldstone, B. C. Love, A. B. Markman, & P. Wolff (Eds.), *Categorization inside and outside the laboratory: Essays in honor of Douglas L. Medin* (pp. 151–175). American Psychological Association. https://doi.org/10.1037/11156-009

Gentner, D., Rattermann, M. J., & Forbus, K. D. (1993). The roles of similarity in transfer: Separating retrievability from inferential soundness. *Cognitive Psychology*, *25*(4), 524–575. https://doi.org/10.1006/cogp.1993.1013

Gluck, M. A., & Bower, G. H. (1988). Evaluating an adaptive network model of human learning. *Journal of Memory and Language*, *27*(2), 166–195. https://doi.org/10.1016/0749-596X(88)90072-1

Goldwater, M. B., Don, H. J., Krusche, M. J., & Livesey, E. J. (2018). Relational discovery in category learning. *Journal of Experimental Psychology: General*, *147*(1), 1–35. https://doi.org/10.1037/xge0000387

Goldwater, M. B., Markman, A. B., & Stilwell, C. H. (2011). The empirical case for role-governed categories. *Cognition*, *118*(3), 359–376. https://doi.org/10.1016/j.cognition.2010.10.009

Gopnik, A. (1988). Conceptual and semantic development as theory change: The case of object permanence. *Mind & Language*, *3*(3), 197–216. https://doi.org/10.1111/j.1468-0017.1988.tb00143.x

Hafting, T., Fyhn, M., Molden, S., Moser, M. B., & Moser, E. I. (2005). Microstructure of a spatial map in the entorhinal cortex. *Nature*, *436*(7052), 801–806. https://doi.org/10.1038/nature03721

Hampton, J. A. (1979). Polymorphous concepts in semantic memory. *Journal of Verbal Learning and Verbal Behavior*, *18*(4), 441–461. https://doi.org/https://doi.org/10.1016/S0022-5371(79)90246-9

Harrison, L. M., Duggins, A., & Friston, K. J. (2006). Encoding uncertainty in the hippocampus. *Neural Networks*, *19*(5), 535–546. https://doi.org/10.1016/j.neunet.2005.11.002

Hassabis, D., Kumaran, D., Vann, S. D., & Maguire, E. A. (2007). Patients with hippocampal amnesia cannot imagine new experiences. *Proceedings of the National Academy of Sciences*, *104*(5), 1726–1731. https://doi.org/10.1073/pnas.0610561104

Hayes, B. K., Taplin, J. E., & Munro, K. I. (1996). Prior knowledge and sensitivity to feature correlations in category acquisition. *Australian Journal of Psychology*, *48*(1), 27–34. https://doi.org/10.1080/00049539608259502

Hayes-Roth, B., & Hayes-Roth, F. (1977). Concept learning and the recognition and classification of exemplars. *Journal of Verbal Learning and Verbal Behavior*, *16*(3), 321–338. https://doi.org/10.1016/S0022-5371(77)80054-6

Hinton, G. E. (1986, August 15–17). Learning distributed representations of concepts. In *Proceedings of the Eighth Annual Conference of the Cognitive Science Society* (Vol. 1, p. 12). Amherst, Massachusetts.

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

Hoffman, A. B., & Murphy, G. L. (2006). Category dimensionality and feature knowledge: When more features are learned as easily as fewer. *Journal of Experimental Psychology: Learning, Memory, and Cognition*, *32*(2), 301–315. https://doi.org/10.1037/0278-7393.32.3.301

Holyoak, K. J. (2012). Analogy and relational reasoning. In K. J. Holyoak & R. G. Morrison (Eds.), *The Oxford handbook of thinking and reasoning* (pp. 234–259). Oxford University Press.

Javadi, A. H., Emo, B., Howard, L. R., Zisch, F. E., Yu, Y., Knight, R., Pinelo Silva, J., & Spiers, H. J. (2017). Hippocampal and prefrontal processing of network topology to simulate the future. *Nature Communications*, *8*(1), Article 14652. https://doi.org/10.1038/ncomms14652

Kahn, A. E., Karuza, E. A., Vettel, J. M., & Bassett, D. S. (2018). Network constraints on learnability of probabilistic motor sequences. *Nature Human Behaviour*, *2*(12), 936–947. https://doi.org/10.1038/s41562-018-0463-8

Kakaei, E., Aleshin, S., & Braun, J. (2021). Visual object recognition is facilitated by temporal community structure. *Learning & Memory*, *28*(5), 148–152. https://doi.org/10.1101/lm.053306.120

Karuza, E. A., Kahn, A. E., & Bassett, D. S. (2019). Human sensitivity to community structure is robust to topological variation. *Complexity*, *2019*, Article 8379321. https://doi.org/10.1155/2019/8379321

Karuza, E. A., Kahn, A. E., Thompson-Schill, S. L., & Bassett, D. S. (2017). Process reveals structure: How a network is traversed mediates expectations about its architecture. *Scientific Reports*, *7*(1), Article 12733. https://doi.org/10.1038/s41598-017-12876-5

Karuza, E. A., Thompson-Schill, S. L., & Bassett, D. S. (2016). Local patterns to global architectures: Influences of network topology on human learning. *Trends in Cognitive Sciences*, *20*(8), 629–640. https://doi.org/10.1016/j.tics.2016.06.003

Keil, F. C. (1992). *Concepts, kinds, and cognitive development*. MIT Press.

Kemp, C., & Tenenbaum, J. B. (2008). The discovery of structural form. *Proceedings of the National Academy of Sciences*, *105*(31), 10687–10692. https://doi.org/10.1073/pnas.0802631105

Kemp, C., & Tenenbaum, J. B. (2009). Structured statistical models of inductive reasoning. *Psychological Review*, *116*(1), 20–58. https://doi.org/https://doi.org/10.1037/a0014282

Killian, N. J., & Buffalo, E. A. (2018). Grid cells map the visual world. *Nature Neuroscience*, *21*(2), 161–162. https://doi.org/10.1038/s41593-017-0062-4

Komatsu, L. K. (1992). Recent views of conceptual structure. *Psychological Bulletin*, *112*(3), 500–526. https://doi.org/10.1037/0033-2909.112.3.500

Kruschke, J. K. (1992). ALCOVE: An exemplar-based connectionist model of category learning. *Psychological Review*, *99*(1), 22–44. https://doi.org/10.1037/0033-295X.99.1.22

Kumaran, D., & McClelland, J. L. (2012). Generalization through the recurrent interaction of episodic memories: A model of the hippocampal system. *Psychological Review*, *119*(3), 573–616. https://doi.org/10.1037/a0028681

Kurtz, K. J. (2007). The divergent autoencoder (DIVA) model of category learning. *Psychonomic Bulletin & Review*, *14*(4), 560–576. https://doi.org/https://doi.org/10.3758/BF03196806

Lassaline, M. E., & Murphy, G. L. (1996). Induction and category coherence. *Psychonomic Bulletin & Review*, *3*(21), 95–99. https://doi.org/https://doi.org/10.3758/BF03210747

Love, B. C., Medin, D. L., & Gureckis, T. M. (2004). SUSTAIN: A network model of category learning. *Psychological Review*, *111*(2), 309–332. https://doi.org/10.1037/0033-295X.111.2.309

Lynn, C. W., & Bassett, D. S. (2020). How humans learn and represent networks. *Proceedings of the National Academy of Sciences*, *117*(47), 29407–29415. https://doi.org/10.1073/pnas.1912328117

Lynn, C. W., Kahn, A. E., Nyema, N., & Bassett, D. S. (2020). Abstract representations of events arise from mental errors in learning and memory. *Nature Communications*, *11*(1), Article 2313. https://doi.org/10.1038/s41467-020-15146-7

Mack, M. L., Love, B. C., & Preston, A. R. (2018). Building concepts one episode at a time: The hippocampus and concept formation. *Neuroscience Letters*, *680*, 31–38. https://doi.org/10.1016/j.neulet.2017.07.061

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

Mark, S., Moran, R., Parr, T., Kennerley, S. W., & Behrens, T. E. (2020). Transferring structural knowledge across cognitive maps in humans and models. *Nature Communications*, *11*(1), Article 4783. https://doi.org/10.1038/s41467-020-18254-6

Markman, A. B., & Ross, B. H. (2003). Category use and category learning. *Psychological Bulletin*, *129*(4), 592–613. https://doi.org/10.1037/0033-2909.129.4.592

Martin, R. C., & Caramazza, A. (1980). Classification in well-defined and ill-defined categories: Evidence for common processing strategies. *Journal of Experimental Psychology: General*, *109*(3), 320–353. https://doi.org/https://doi.org/10.1037/0096-3445.109.3.320

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), Article 419. https://doi.org/10.1037/0033-295X.102.3.419

McClelland, J. L., & Rogers, T. (2003). The parallel distributed processing approach to semantic cognition. *Nature Reviews Neuroscience*, *4*(4), 310–322. https://doi.org/10.1038/nrn1076

McClelland, J. L., & Rumelhart, D. E. (1985). Distributed memory and the representation of general and specific information. *Journal of Experimental Psychology: General*, *114*(2), 159–188. https://doi.org/10.1037/0096-3445.114.2.159

McRae, K., Cree, G. S., Westmacott, R., & Sa, V. R. D. (1999). Further evidence for feature correlations in semantic memory. *Canadian Journal of Experimental Psychology*, *53*(4), 360–373. https://doi.org/10.1037/h0087323

McRae, K., De Sa, V. R., & Seidenberg, M. S. (1997). On the nature and scope of featural representations of word meaning. *Journal of Experimental Psychology: General*, *126*(2), 99–130. https://doi.org/10.1037/0096-3445.126.2.99

Medin, D. L., Altom, M. W., Edelson, S. M., & Freko, D. (1982). Correlated symptoms and simulated medical classification. *Journal of Experimental Psychology: Learning, Memory, and Cognition*, *8*(1), 37–50. https://doi.org/10.1037/0278-7393.8.1.37

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

Medin, D. L., & Shoben, E. J. (1988). Context and structure in conceptual combination. *Cognitive Psychology*, *20*(2), 158–190. https://doi.org/10.1016/0010-0285(88)90018-7

Medin, D. L., Wattenmaker, W. D., & Hampson, S. E. (1987). Family resemblance, conceptual cohesiveness, and category construction. *Cognitive Psychology*, *19*(2), 242–279. https://doi.org/10.1016/0010-0285(87)90012-0

Mikolov, T., Chen, K., Corrado, G., & Dean, J. (2013). *Efficient estimation of word representations in vector space*. arXiv: 1301.3781. https://arxiv.org/abs/1301.3781

Morton, N. W., Schlichting, M. L., & Preston, A. R. (2020). Representations of common event structure in medial temporal lobe and frontoparietal cortex support efficient inference. *Proceedings of the National Academy of Sciences*, *117*(47), 29338–29345. https://doi.org/10.1073/pnas.1912338117

Murphy, G. L. (2016). Is there an exemplar theory of concepts? *Psychonomic Bulletin & Review*, *23*(4), 1035–1042. https://doi.org/10.3758/s13423-015-0834-3

Murphy, G. L., & Allopenna, P. D. (1994). The locus of knowledge effects in concept learning. *Journal of Experimental Psychology: Learning, Memory, and Cognition*, *20*(4), 904–919. https://doi.org/10.1037/0278-7393.20.4.904

Murphy, G. L., & Medin, D. L. (1985). The role of theories in conceptual coherence. *Psychological Review*, *92*(3), 289–316. https://doi.org/10.1037/0033-295X.92.3.289

Neumann, P. G. (1974). An attribute frequency model for the abstraction of prototypes. *Memory & Cognition*, *2*(2), 241–248. https://doi.org/https://doi.org/10.3758/BF03208990

Nosofsky, R. M. (1984). Choice, similarity, and the context theory of classification. *Journal of Experimental Psychology: Learning, Memory, and Cognition*, *10*(1), 104–114. https://doi.org/10.1037/0278-7393.10.1.104

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

Nosofsky, R. M. (2011). The generalized context model: An exemplar model of classification. In E. M. Pothos & A. J. Wills (Eds.), *Formal approaches in categorization* (pp. 18–39). Cambridge University Press. https://doi.org/10.1017/CBO9780511921322.002

Nosofsky, R. M., Kruschke, J. K., & McKinley, S. C. (1992). Combining exemplar-based category representations and connectionist learning rules. *Journal of Experimental Psychology: Learning, Memory, and Cognition*, *18*(2), 211–233. https://doi.org/https://doi.org/10.1037//0278-7393.18.2.211

Nosofsky, R. M., Palmeri, T. J., & McKinley, S. C. (1994). Rule-plus-exception model of classification learning. *Psychological Review*, *101*(1), 53–79. https://doi.org/https://doi.org/10.1037/0033-295X.101.1.53

Nosofsky, R. M., Sanders, C. A., & McDaniel, M. A. (2018). Tests of an exemplar-memory model of classification learning in a high-dimensional natural-science category domain. *Journal of Experimental Psychology: General*, *147*(3), 328–353. https://doi.org/10.1037/xge0000369

O'Keefe, J., & Nadel, L. (1978). *The hippocampus as a cognitive map*. Oxford University Press.

O'Reilly, R. C. (1996). *The Leabra model of neural interactions and learning in the neocortex* [Doctoral dissertation]. Carnegie Mellon University.

O'Reilly, R. C., Munakata, Y., Frank, M. J., Hazy, T. E., & Contributors. (2020). *Computational cognitive neuroscience* (4th ed.) Wiki Book. https://github.com/CompCogNeuro/ed4

Park, S. A., Miller, D. S., & Boorman, E. D. (2021). *Inferences on a multidimensional social hierarchy use a grid-like code*. bioRxiv, 2020-05. https://doi.org/10.1101/2020.05.29.124651

Patterson, K., Nestor, P. J., & Rogers, T. T. (2007). Where do you know what you know? The representation of semantic knowledge in the human brain. *Nature Reviews Neuroscience*, *8*(12), 976–987. https://doi.org/10.1038/nrn2277

Peelen, M. V., & Caramazza, A. (2012). Conceptual object representations in human anterior temporal cortex. *The Journal of Neuroscience*, *32*(45), 15728–15736. https://doi.org/10.1523/JNEUROSCI.1953-12.2012

Pudhiyidath, A., Roome, H. E., Coughlin, C., Nguyen, K. V., & Preston, A. R. (2020). Developmental differences in temporal schema acquisition impact reasoning decisions. *Cognitive Neuropsychology*, *37*(1-2), 25–45. https://doi.org/https://doi.org/10.1080/02643294.2019.1667316

Ralph, M. A. L., Jefferies, E., Patterson, K., & Rogers, T. T. (2017). The neural and computational bases of semantic cognition. *Nature Reviews Neuroscience*, *18*(1), 42–55. https://doi.org/10.1038/nrn.2016.150

Ralph, M. A. L., Sage, K., Jones, R. W., & Mayberry, E. J. (2010). Coherent concepts are computed in the anterior temporal lobes. *Proceedings of the National Academy of Sciences*, *107*(6), 2717–2722. https://doi.org/10.1073/pnas.0907307107

Rehder, B., & Ross, B. H. (2001). Abstract coherent categories. *Journal of Experimental Psychology: Learning, Memory, and Cognition*, *27*(5), 1261–1275. https://doi.org/10.1037/0278-7393.27.5.1261

Reitman, J. S., & Bower, G. H. (1973). Storage and later recognition of exemplars of concepts. *Cognitive Psychology*, *4*(2), 194–206. https://doi.org/10.1016/0010-0285(73)90011-X

Rips, L. J., Smith, E. E., & Medin, D. L. (2012). Concepts and categories: Memory, meaning, and metaphysics. In K. J. Holyoak & R. G. Morrison (Eds.), *The Oxford handbook of thinking and reasoning* (pp. 177–209). Oxford University Press. https://doi.org/10.1093/oxfordhb/9780199734689.013.0011

Rogers, T. T., & McClelland, J. L. (2004). *Semantic cognition: A parallel distributed processing approach*. MIT Press.

Rogers, T. T., & McClelland, J. L. (2008). Précis of semantic cognition: A parallel distributed processing approach. *Behavioral and Brain Sciences*, *31*(6), 689–714. https://doi.org/10.1017/S0140525X0800589X

Rosch, E. (1973). Natural categories. *Cognitive Psychology*, *4*(3), 328–350. https://doi.org/10.1016/0010-0285(73)90017-0

Rosch, E. (1975). Cognitive representations of semantic categories. *Journal of Experimental Psychology: General*, *104*(3), 192–233. https://doi.org/10.1037/0096-3445.104.3.192

Rosch, E., & Mervis, C. B. (1975). Family resemblances: Studies in the internal structure of categories. *Cognitive Psychology*, *7*(4), 573–605. https://doi.org/10.1016/0010-0285(75)90024-9

Rubinov, M., & Sporns, O. (2010). Complex network measures of brain connectivity: Uses and interpretations. *NeuroImage*, *52*(3), 1059–1069. https://doi.org/https://doi.org/10.1016/j.neuroimage.2009.10.003

Rubinov, M., & Sporns, O. (2011). Weight-conserving characterization of complex functional brain networks. *Neuroimage*, *56*(4), 2068–2079. https://doi.org/10.1016/j.neuroimage.2011.03.069

Rumelhart, D. E. (1990). Brain style computation: Learning and generalization. In S. F. Zornetzer, J. L. Davis & C. Lau (Eds.), *An introduction to neural and electronic networks* (pp. 405–420). Academic Press.

Saffran, J. R., Aslin, R. N., & Newport, E. L. (1996). Statistical learning by 8-month-old infants. *Science*, *274*(5294), 1926–1928. https://doi.org/10.1126/science.274.5294.1926

Saxe, A. M., McClelland, J. L., & Ganguli, S. (2019). A mathematical theory of semantic development in deep neural networks. *Proceedings of the National Academy of Sciences*, *116*(23), 11537–11546. https://doi.org/10.1073/pnas.1820226116

Schapiro, A. C., Gregory, E., Landau, B., McCloskey, M., & Turk-Browne, N. B. (2014). The necessity of the medial temporal lobe for statistical learning. *Journal of Cognitive Neuroscience*, *26*(8), 1736–1747. https://doi.org/10.1162/jocn_a_00578

Schapiro, A. C., Kustner, L. V., & Turk-Browne, N. B. (2012). Shaping of object representations in the human medial temporal lobe based on temporal regularities. *Current Biology*, *22*(17), 1622–1627. https://doi.org/10.1016/j.cub.2012.06.056

Schapiro, A. C., Rogers, T. T., Cordova, N. I., Turk-Browne, N. B., & Botvinick, M. M. (2013). Neural representations of events arise from temporal community structure. *Nature Neuroscience*, *16*(4), 486–492. https://doi.org/10.1038/nn.3331

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), Article 20160049. https://doi.org/10.1098/rstb.2016.0049

Schapiro, A. C., Turk-Browne, N. B., Norman, K. A., & Botvinick, M. M. (2016). Statistical learning of temporal community structure in the hippocampus. *Hippocampus*, *26*(1), 3–8. https://doi.org/10.1002/hipo.22523

Schuck, N. W., Cai, M. B., Wilson, R. C., & Niv, Y. (2016). Human orbitofrontal cortex represents a cognitive map of state space. *Neuron*, *91*(6), 1402–1412. https://doi.org/10.1016/j.neuron.2016.08.019

Sloman, S. A., Love, B. C., & Ahn, W. K. (1998). Feature centrality and conceptual coherence. *Cognitive Science*, *22*(2), 189–228. https://doi.org/10.1207/s15516709cog2202_2

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), 1411–1436. https://doi.org/10.1037/0278-7393.24.6.1411

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), 800–811. https://doi.org/10.1037/0278-7393.28.4.800

Solomon, K. O., Medin, D. L., & Lynch, E. (1999). Concepts do more than categorize. *Trends in Cognitive Sciences*, *3*(3), 99–105. https://doi.org/10.1016/S1364-6613(99)01288-7

Solomon, S. H., Medaglia, J. D., & Thompson-Schill, S. L. (2019). Implementing a concept network model. *Behavior Research Methods*, *51*(4), 1717–1736. https://doi.org/10.3758/s13428-019-01217-1

Spalding, T. L., & Murphy, G. L. (1996). Effects of background knowledge on category construction. *Journal of Experimental Psychology: Learning, Memory, and Cognition*, *22*(2), 525–538. https://doi.org/10.1037/0278-7393.22.2.525

Spalding, T. L., & Ross, B. H. (2000). Concept learning and feature interpretation. *Memory & Cognition*, *28*(3), 439–451. https://doi.org/10.3758/BF03198559

Stachenfeld, K. L., Botvinick, M. M., & Gershman, S. J. (2017). The hippocampus as a predictive map. *Nature Neuroscience*, *20*(11), 1643–1653. https://doi.org/10.1038/nn.4650

Strange, B. A., Duggins, A., Penny, W., Dolan, R. J., & Friston, K. J. (2005). Information theory, novelty and hippocampal responses: Unpredicted or unpredictable? *Neural Networks*, *18*(3), 225–230. https://doi.org/10.1016/j.neunet.2004.12.004

Tavares, R. M., Mendelsohn, A., Grossman, Y., Williams, C. H., Shapiro, M., Trope, Y., & Schiller, D. (2015). A map for social navigation in the human brain. *Neuron*, *87*(1), 231–243. https://doi.org/10.1016/j.neuron.2015.06.011

Tenenbaum, J. B., Kemp, C., Griffiths, T. L., & Goodman, N. D. (2011). How to grow a mind: Statistics, structure, and abstraction. *Science*, *331*(6022), 1279–1285. https://doi.org/10.1126/science.1192788

Theves, S., Fernandez, G., & Doeller, C. F. (2019). Hippocampus encodes distances in multidimensional feature space. *Current Biology*, *29*(7), 1226–1231.e3. https://doi.org/10.1016/j.cub.2019.02.035

Tolman, E. C. (1948). Cognitive maps in rats and men. *Psychological Review*, *55*(4), 189–208. https://doi.org/10.1037/h0061626

Turk-Browne, N. B., Scholl, B. J., Chun, M. M., & Johnson, M. K. (2009). Neural evidence of statistical learning: Efficient detection of visual regularities without awareness. *Journal of Cognitive Neuroscience*, *21*(10), 1934–1945. https://doi.org/10.1162/jocn.2009.21131

Turk-Browne, N. B., Scholl, B. J., Johnson, M. K., & Chun, M. M. (2010). Implicit perceptual anticipation triggered by statistical learning. *The Journal of Neuroscience*, *30*(33), 11177–11187. https://doi.org/10.1523/JNEUROSCI.0858-10.2010

Tyler, L. K., & Moss, H. E. (2001). Towards a distributed account of conceptual knowledge. *Trends in Cognitive Sciences*, *5*(6), 244–252. https://doi.org/10.1016/S1364-6613(00)01651-X

Tyler, L. K., Moss, H. E., Durrant-Peatfield, M. R., & Levy, J. P. (2000). Conceptual structure and the structure of concepts: A distributed account of category-specific deficits. *Brain and Language*, *75*(2), 195–231. https://doi.org/10.1006/brln.2000.2353

Unger, L., Savic, O., & Sloutsky, V. M. (2020, July 29–August 1). Simple mechanisms, rich structure: Statistical co-occurrence regularities in language shape the development of semantic knowledge. *Proceedings of the Annual Meeting of the Cognitive Science Society*, Virtual.

Ward, T. B., & Scott, J. (1987). Analytic and holistic modes of learning family-resemblance concepts. *Memory & Cognition*, *15*(1), 42–54. https://doi.org/https://doi.org/10.3758/BF03197711

Wattenmaker, W. D. (1991). Learning modes, feature correlations, and memory-based categorization. *Journal of Experimental Psychology: Learning, Memory, and Cognition*, *17*(5), 908–923. https://doi.org/10.1037/0278-7393.17.5.908

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

Watts, D. J., & Strogatz, S. H. (1998). Collective dynamics of 'small-world' networks. *Nature*, *393*(6684), 440–442. https://doi.org/10.1038/30918

Whittington, J. C., Muller, T. H., Mark, S., Chen, G., Barry, C., Burgess, N., & Behrens, T. E. (2020). The Tolman–Eichenbaum machine: Unifying space and relational memory through generalization in the hippocampal formation. *Cell*, *183*(5), 1249–1263.e23. https://doi.org/10.1016/j.cell.2020.10.024

Wisniewski, E. J. (1995). Prior knowledge and functionally relevant features in concept learning. *Journal of Experimental Psychology: Learning, Memory, and Cognition*, *21*(2), 449–468. https://doi.org/10.1037/0278-7393.21.2.449

Wisniewski, E. J., & Gentner, D. (1991). On the combinatorial semantics of noun pairs: Minor and major adjustments to meaning. In G.B. Simpson (Ed.) *Advances in psychology* (Vol. 77, pp. 241–284). North-Holland.

Wisniewski, E. J., & Medin, D. L. (1994). On the interaction of theory and data in concept learning. *Cognitive Science*, *18*(2), 221–281. https://doi.org/10.1207/s15516709cog1802_2

Wittgenstein, L. (2010). *Philosophical investigations*. John Wiley & Sons.

Yamauchi, T., & Markman, A. B. (1998). Category learning by inference and classification. *Journal of Memory and Language*, *39*(1), 124–148. https://doi.org/10.1006/jmla.1998.2566

Received March 16, 2022
Revision received February 2, 2023
Accepted March 27, 2023 ■