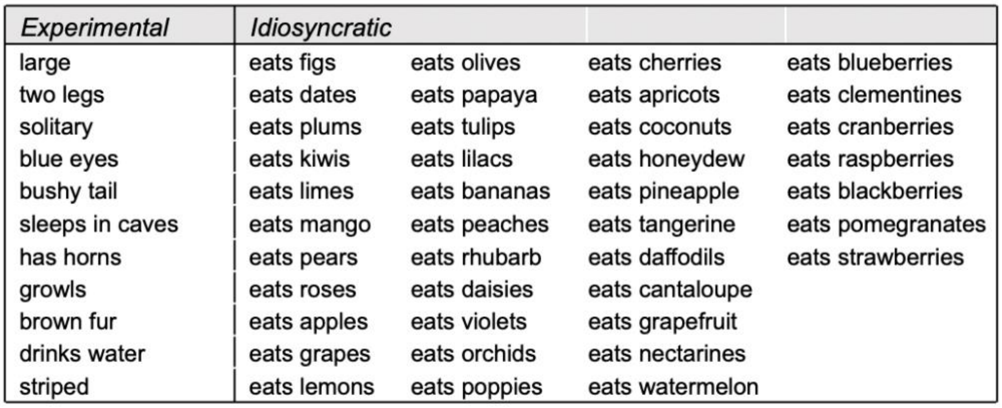
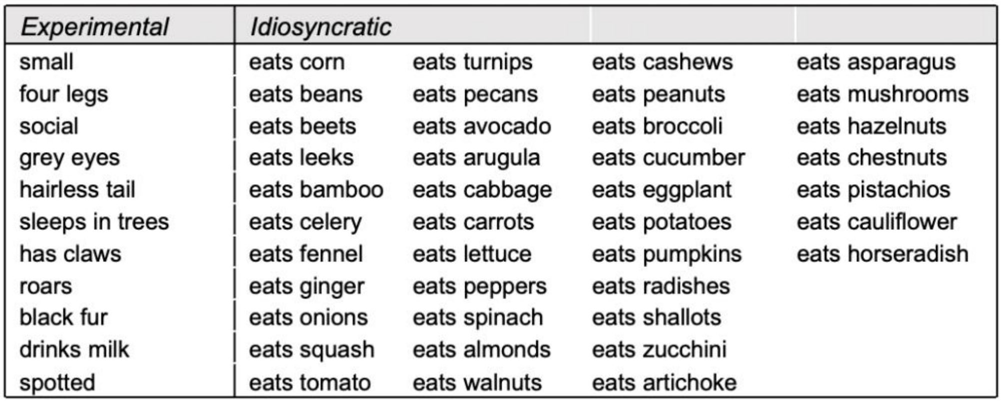
Supplementary Materials: Verbal Category Stimuli

Feature List A



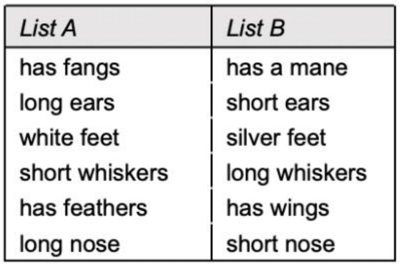
Experimental	*Idiosyncratic*			
large	eats figs	eats olives	eats cherries	eats blueberries
two legs	eats dates	eats papaya	eats apricots	eats clementines
solitary	eats plums	eats tulips	eats coconuts	eats cranberries
blue eyes	eats kiwis	eats lilacs	eats honeydew	eats raspberries
bushy tail	eats limes	eats bananas	eats pineapple	eats blackberries
sleeps in caves	eats mango	eats peaches	eats tangerine	eats pomegranates
has horns	eats pears	eats rhubarb	eats daffodils	eats strawberries
growls	eats roses	eats daisies	eats cantaloupe	
brown fur	eats apples	eats violets	eats grapefruit	
drinks water	eats grapes	eats orchids	eats nectarines	
striped	eats lemons	eats poppies	eats watermelon	

Feature List B



Experimental	*Idiosyncratic*			
small	eats corn	eats turnips	eats cashews	eats asparagus
four legs	eats beans	eats pecans	eats peanuts	eats mushrooms
social	eats beets	eats avocado	eats broccoli	eats hazelnuts
grey eyes	eats leeks	eats arugula	eats cucumber	eats chestnuts
hairless tail	eats bamboo	eats cabbage	eats eggplant	eats pistachios
sleeps in trees	eats celery	eats carrots	eats potatoes	eats cauliflower
has claws	eats fennel	eats lettuce	eats pumpkins	eats horseradish
roars	eats ginger	eats peppers	eats radishes	
black fur	eats onions	eats spinach	eats shallots	
drinks milk	eats squash	eats almonds	eats zucchini	
spotted	eats tomato	eats walnuts	eats artichoke	

Catch Features



List A	*List B*
has fangs	has a mane
long ears	short ears
white feet	silver feet
short whiskers	long whiskers
has feathers	has wings
long nose	short nose

Supplementary Table 1: Full list of verbal features used in Exp. 1. For each participant, feature lists A and B were randomly assigned to the Modular and Random structures. The first three features in each list were assigned to the core nodes, and the other eight features were randomly assigned to the peripheral nodes. Idiosyncratic features were included to add variability across exemplars and were only seen once or twice by each participant. Each catch feature only appeared once in a catch trial.

Supplementary Material: Additional Models

Experiment 1: Exemplar Model

Methods

The generalized context model (GCM). We instantiated an exemplar theory of category learning in an adaptation of GCM. Though our equations employ the set of parameters used by Nosofsky et al. (2018), our model embodies the classic exemplar-based representations of Medin & Schaffer's (1978) original context model. The probability (or the match value) of category J given test stimulus i was calculated as the summed similarity of stimulus i to all exemplars j in category J:

$$P(C_J|i) = \sum_{j \in J} s_{ij}$$

where i and j are binary vectors with 11 values representing the presence (1) or absence (0) of each of the 11 category features, and s represents the similarity of the current stimulus i to each stored exemplar j. This similarity measure (s_{ij}) is an exponential decay function of the distance between i and j (d_{ij}) in multidimensional feature space where distance is determined by a (weighted) Minkowski power model:

$$s_{ij} = \exp\,(-c \cdot d_{ij})$$

$$d_{ij} = \left[\sum w_m \left|x_{im} - x_{jm}\right|^r\right]^{1/r}$$

where x_{im} and x_{jm} are the values for items *i* and *j* on feature *m*, w_m is the attention weight given to dimension *m*, *r* determines the distance metric used, and *c* is a sensitivity parameter. Following Nosofsky et al. (2018), we set *r*=2 which results in a Euclidean distance metric and set *w*=1 for all features. We did not fit attention weights for individual features and subjects because our counterbalancing ensures that any differences across core or across peripheral features should wash out.

The sensitivity parameter c determines the rate at which similarity declines with distance; any positive value of c means that exemplars that are in close proximity with the test stimulus will have the greatest influence on its match value. Nosofsky et al. (2018) reported an optimized sensitivity parameter value of ~1.8; we thus tested this parameter value specifically in addition to a wider range of values. The above equations were used to calculate the match of each test stimulus to its corresponding category in the analyses described below. The *p-store* parameter determines the probability with which each exemplar is stored into memory. Only exemplars that are stored contribute to category judgments. For example, a *p-store* value of 0.75 means that each exemplar has an 75% chance of being stored; reducing this parameter results in lower overall accuracy. Because the GCM often achieved ceiling performance, we lowered the *p-store* to 0.75 to better reveal differences between task conditions and to provide variation across model iterations.

Model assessment. In the missing feature task, humans were given feedback on each trial such that by the end of each trial they could encode a correct 6-feature exemplar (the five probed features plus the correct missing feature). In our exemplar model simulations, we therefore considered the stored exemplars of category J to be the correct 6-feature exemplars of category J across all trials. On each trial, three test stimuli were compared with the stored exemplars, corresponding to the three missing feature options for that trial. The correct stimulus i_0 contained six features: the five probe features plus the correct missing feature. Each of the two incorrect exemplars, i_1 and i_2, contained the five probe features plus one of the incorrect features. We calculated P_0, P_1, and P_2 as described above, where category J is always the correct category (if the current correct test stimulus is from a Modular category, we only compare that stimulus with Modular exemplars). That is, the match value of correct stimulus i_0 on trial t is calculated as the summed similarity of i_0 with all j exemplars from trial 1 to 144. We calculated the match of incorrect stimuli i_1 and i_2 in the same manner, and compared match values to determine the accuracy on each trial.

Much like other researchers focusing on inference learning and internal category structure, here we adapted the context model to assess individual categories, that is, to quantify the match of a current test stimulus to its corresponding category (Medin et al., 1982; Wattenmaker, 1991, 1993). Thus, the only meaningful way our GCM implementation differs from previous accounts (e.g., Nosofsky et al., 2018) relates to which values are compared at model assessment. In standard classification tasks, a single test item is compared to two or more categories to determine which category provides the best match. In our inference task, we compared multiple test items to a single category to determine which of the test items is the best match.

We assessed model behavior in two different ways. In a deterministic response mapping, the test probe with the highest category match was considered to be the model's chosen response, and each trial was assigned an accuracy value. In order to simulate human data, we iterated the model across many permutations of the task, with variations across iterations enabled by setting *p-store*<1. It has been argued that deterministic response mapping may be more appropriate when the GCM is used to simulate individual subject's data (Nosofsky & Zaki, 2002). Our deterministic response mapping was as follows: If P_0 was greater than both P_1 and P_2, $acc = 1$; if P_0 was equal to the greater value amongst unequal P_1 and P_2, $acc = 1/2$; if all values were equal, $acc = 1/3$; and if either P_1 or P_2 was greater than P_0, $acc = 0$. Mean accuracy was calculated separately for Random core and Modular core trials. Within each of 1000 simulations, trial order was randomized and mean accuracy was calculated within each condition.

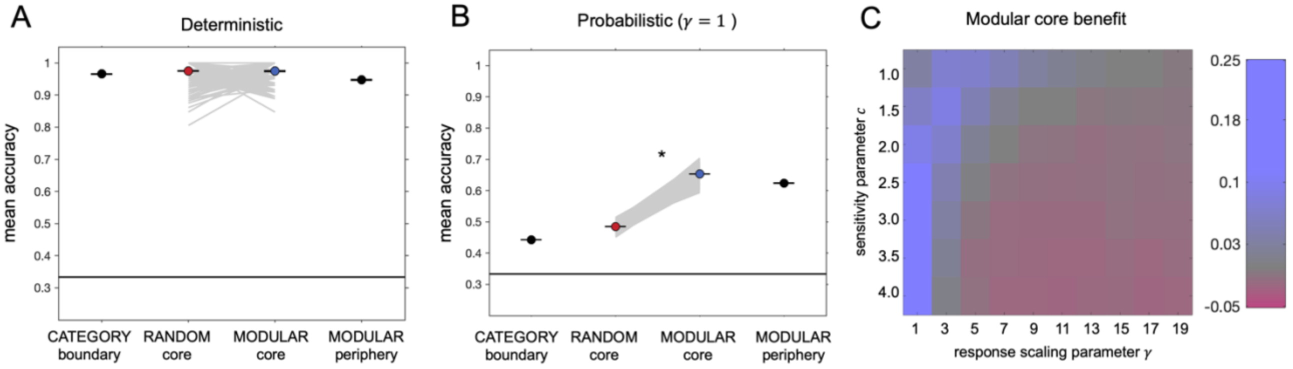
In a probabilistic response mapping, the model responds by probability matching based on the relative similarities of the test probes (P_0, P_1, and P_2), where the trial response *Resp* is calculated as follows:

$$Acc = \frac{{P_0}^{\gamma}}{{P_0}^{\gamma} + {P_1}^{\gamma} + {P_2}^{\gamma}}$$

where γ is a response-scaling parameter that influences the degree of determinism in the model's response on each trial. When $\gamma = 1$ the response is solely based on probability matching, whereas when $\gamma > 1$ responses become more deterministic. As γ approaches infinity, Acc becomes identical to the deterministic response mapping described above.

Results

Mean exemplar model accuracy for the GCM is shown in Supp. Fig. 1. Using a deterministic response mapping and $c = 1.8$, the exemplar model successfully learned the category boundary (*M*=96.6%, *p*<0.0001) and the Modular peripheral structure (*M*=94.8%, *p*<0.0001), and also performed above chance on Random core (*M*=97.3%, *p*<0.0001) and Modular core (*M*=97.5%, *p*<0.0001) trials. No difference between Random and Modular core trials was observed (*t*(999)=0.28, *p*>0.7), as shown in Supp. Fig. 1A. However, when responses were purely probabilistic (y=1), a Modular core benefit emerged (*t*(999)=249.5, *p*<0.0001; Supp. Fig. 1B). Across a range of c and γ values, the GCM revealed either a Modular core benefit or a Random core benefit (Supp. Fig. 1C). The fact that the GCM can replicate the Modular core benefit under certain parameter conditions suggests that an exemplar-based model of category representation is able to account for our human behavioral data, in addition to the neural network model reported in the main manuscript.



Supplementary Figure 1: Exemplar model simulations of the missing feature task in Experiment 1. (A) Using a deterministic response mapping, the GCM performed above chance in all structure conditions, but revealed no difference between Random and Modular core structure learning (*p*>0.7). (B) Using a probabilistic response mapping, a Modular core benefit was observed (*t*(999)=249.5, *p*<0.0001) (C) A visualization of the parameters under which the GCM reveals a significant Modular core (blue) or Random core (red) benefit. Deeper colors indicate a larger difference in accuracy between conditions.

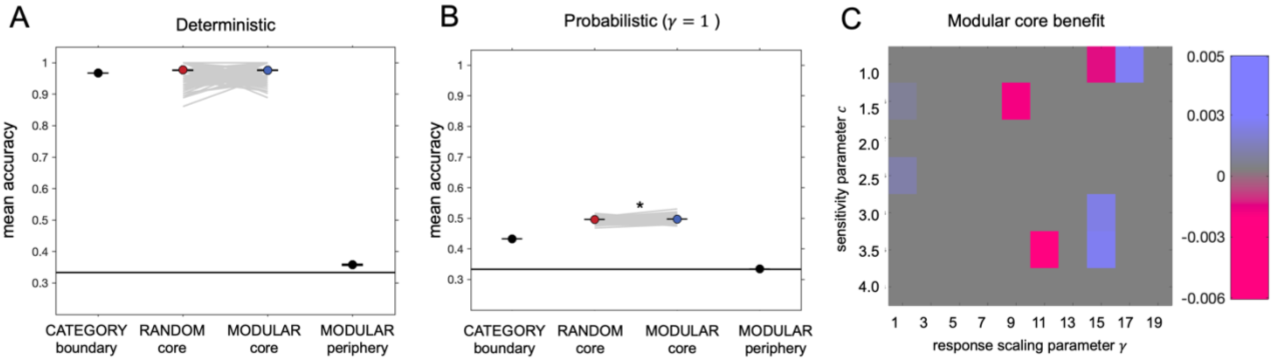
Experiment 1: Prototype Model

Methods

Our average distance prototype model was nearly identical to our exemplar model, except instead of comparing the test stimulus to each observed exemplar j, it was compared to the centroid of all stored exemplars in category J. The centroid was represented as the mean across exemplars, resulting in a vector with 11 feature values between 0 and 1. Because we wanted to target the nature of the category representation specifically, all of the parameters from the GCM were used in the prototype model simulations.

Results

Using a deterministic response mapping (Supp. Fig. 2A), the prototype model successfully discriminated between categories (*M*=96.8%, *p*<0.0001), and learned both Random core (*M*=97.7%, *p*<0.0001) and Modular core (*M*=97.6%, *p*<0.0001) structure successfully. Surprisingly, the model performed significantly above chance on Modular peripheral structure trials (*M*=35.8%, t(999)=5.76, *p*<0.0001). This pattern arises when Modular exemplars from the two distinct modules do not equally contribute to the prototype; this happens due to chance when *pstore*<1 and as a result of trial order. No difference between Modular core and Random core accuracy was observed (*t*(999)=0.61, *p*>0.5). However, as was the case with the GCM, the prototype model was able to replicate the Modular core benefit under probabilistic response conditions (*t*(999)=4.0, *p*<0.0001), suggesting that a prototype category representation can also in theory account for the human behavioral effects, in addition to the neural network model and exemplar model (Supp. Fig. 2B). However, relative to the exemplar model, the behavior of the prototype model is more erratic across different parameter conditions (Supp. Fig. 2C).



Supplementary Figure 2: Prototype model simulations of the missing feature task in Experiment 1. (A) Using a deterministic response mapping, the prototype model performed above chance in all structure conditions, but revealed no difference between Random and Modular core structure learning (*p*>0.7). (B) Using a probabilistic response mapping, a Modular core benefit was observed (*t*(999)=4.0, *p*<0.0001). (C) A visualization of the parameters under which the prototype model reveals a significant Modular core (blue) or Random core (red) benefit. Deeper colors indicate a larger difference in accuracy between conditions.

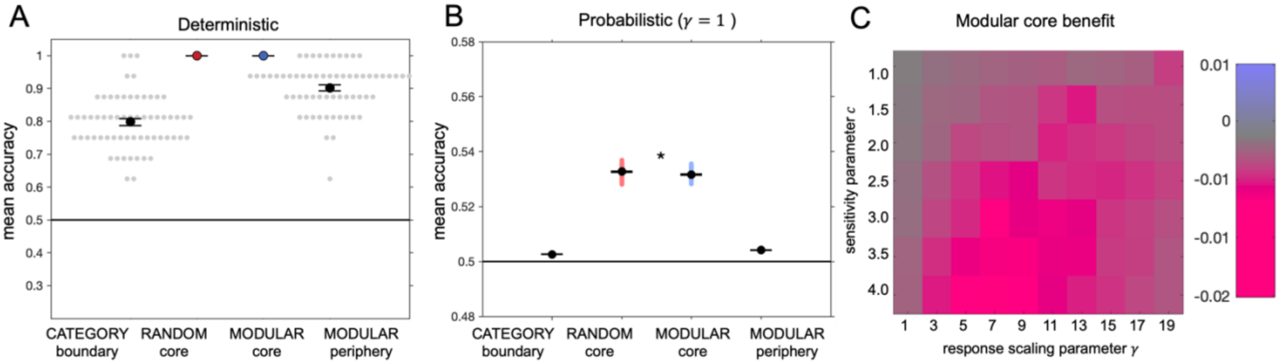
Experiment 3: Exemplar Model

Methods

The exemplar model used in Experiment 3 is nearly identical to the GCM model described above in Exp. 1. The exemplars and test stimuli are 11-feature binary vectors, except now the stored category exemplars correspond to the set of 2-feature stimuli presented during the SL task, and the test stimuli correspond to the exemplars presented in the 2AFC task. On each trial, category match P for the correct and incorrect 2AFC test stimuli was calculated based on the summed similarity between each test stimulus and the 550 category exemplars. Deterministic and probabilistic response mappings were implemented as in Exp. 1. Mean accuracy was calculated within each of 62 simulations within each category structure, corresponding to the 62 Lattice and 62 Modular paths used in the behavioral experiment.

Results

Under deterministic response conditions, the GCM was able to learn the core and peripheral structure of Lattice and Modular categories (Supp. Fig. 3A). The model performed significantly above chance on both the Lattice (*M*=88%; *SD*=6.8%; p<0.0001) and Modular (*M*=97%; *SD*=3.9%; *p*<0.0001) peripheral trials, and achieved ceiling performance for Lattice and Modular core trials (*M*=100%). Probabilistic response mapping revealed a more interesting pattern of results (Supp. Fig. 3B). Here we did find a significant difference between Lattice (*M*=53.3%) and Modular (*M*=53.2) core structure learning, but in the opposite direction to the human and neural network model results: the exemplar model learned the core structure significantly better in Lattice relative to Modular categories (*t*(122)=3.32, *p*=0.001). This difference was robust across a range of parameter conditions (Supp. Fig. 3C).



Supplementary Figure 3: Exemplar model simulations of the 2AFC task in Experiment 3. (A) Under deterministic response mapping, the exemplar model successfully learned the core and peripheral structure of both the Lattice and Modular categories, but no difference in accuracy was observed between the core (B) Using a probabilistic response mapping, the exemplar model revealed increased performance on Lattice core vs. Modular core trials, contrasting with the pattern of human performance. (C) A visualization of the parameters under which the prototype model reveals a significant Modular core (blue) or Random core (red) benefit. Deeper colors indicate a larger difference in accuracy between conditions. The exemplar model did not reveal a Modular core benefit under any parameter conditions.

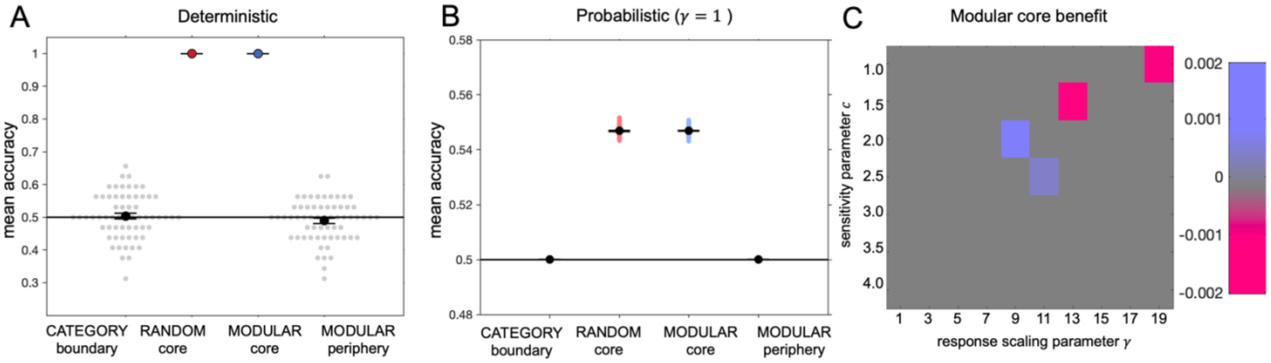
Experiment 3: Prototype Model

Methods

The prototype model was almost identical to the exemplar model, except instead of comparing the stimulus to each observed exemplar j, it was compared to the centroid of all stored exemplars in category J.

Results

The prototype model also was at ceiling at learning Lattice (100%) and Modular (100%) core structure under deterministic response conditions, with no difference between conditions (Supp. Fig. 4A). Unlike the exemplar model, the prototype model was not able to learn the peripheral structure of either the Lattice (*M*=50.3%, *p*<0.7) or Modular (*M*=48.9%, *p*=0.19) category. Using probabilistic response mapping, the same pattern of results emerged (Supp. Fig. 4B). While accuracy for Modular (*M*=54.7%, *p*<0.0001) and Lattice (*M*=54.7%, *p*<0.0001) core trials was brought off of ceiling, there was no difference between the two conditions (*t*(122)=0.17, *p*>0.8). However, as in Exp. 1, the prototype model's behavior was erratic across a range of parameter values (Supp. Fig. 4C).



Supplementary Figure 4: Prototype model simulations of the 2AFC task in Experiment 3. In both deterministic (A) and probabilistic (B) response mapping conditions, the prototype model was able to learn the Modular and Lattice core structure, but not able to learn the Modular and Lattice peripheral structure. No difference between Modular core and Lattice core accuracy was observed. (C) A visualization of the parameters under which the prototype model reveals a significant Modular core (blue) or Random core (red) benefit. Deeper colors indicate a larger difference in accuracy between conditions.

References

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.

Medin, D. L., & Schaffer, M. M. (1978). Context theory of classification learning. *Psychological review*, *85*(3), 207.

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.

Nosofsky, R. M., & Zaki, S. R. (2002). Exemplar and prototype models revisited: response strategies, selective attention, and stimulus generalization. *Journal of Experimental Psychology: Learning, Memory, and Cognition*, *28*(5), 924.

Wattenmaker, W. D. (1991). Learning modes, feature correlations, and memory-based categorization. *Journal of Experimental Psychology: Learning, Memory, and Cognition*, *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.