Learning the internal structure of novel categories

Sarah H. Solomon & Anna C. Schapiro

Department of Psychology, University of Pennsylvania

Penn Computational Cognitive Neuroscience Lab

Introduction

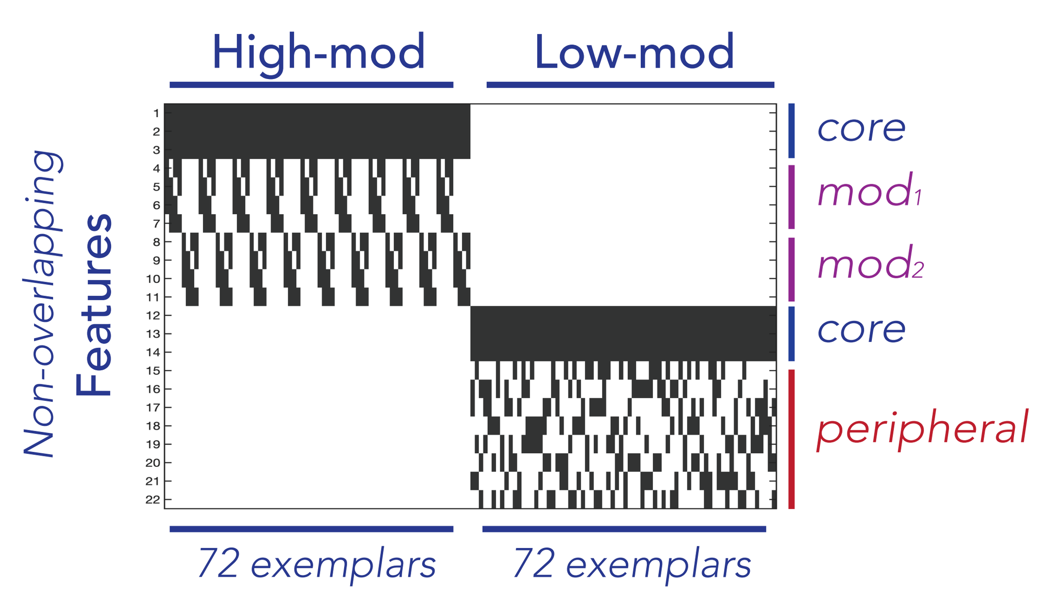
The concepts that compose our world are richly structured. Whereas the structure of semantic space as a whole enables us to differentiate semantic categories[1,2] the **internal structure of concepts** enables generalization to possible but yet unseen category exemplars.

A concept's internal structure can be characterized as the patterns of feature relationships across its exemplars. This structure can be represented as networks in which nodes represent conceptual features and edges represent their co-occurrences. This structure can vary across concepts in meaningful ways[3].

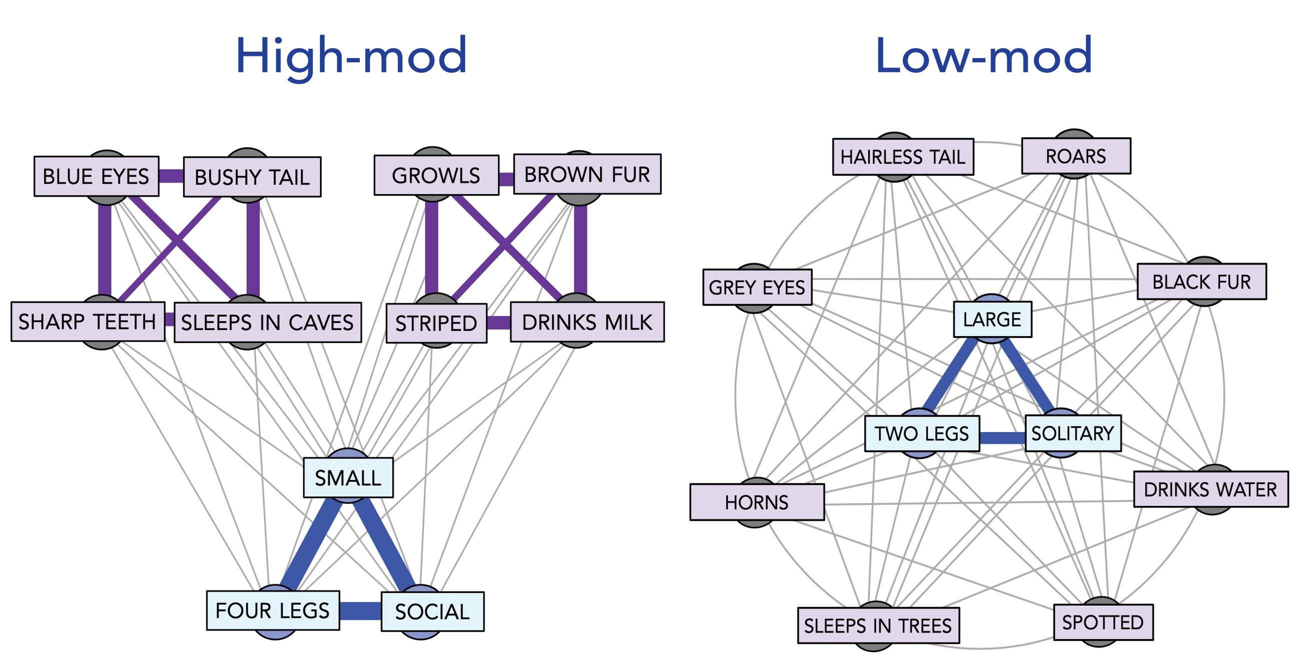
How is this internal category structure learned?

Category Structure Design

We defined one high modularity and one low modularity category by specifying patterns of feature co-occurrence.

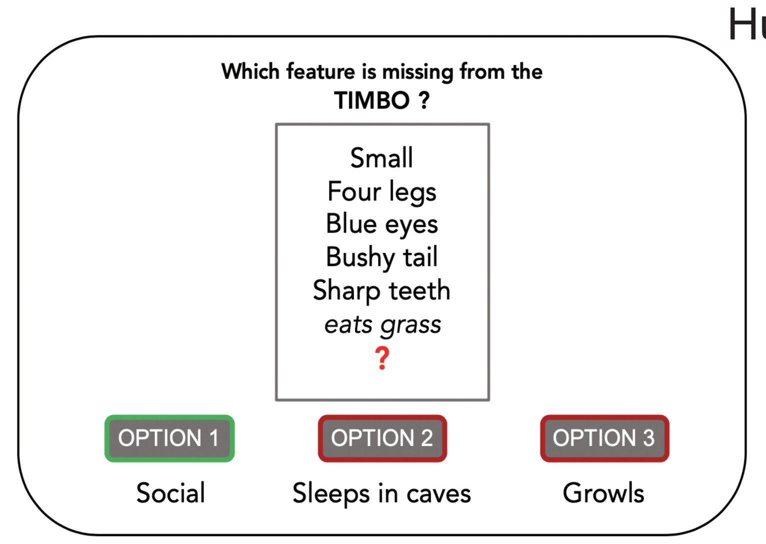


Exemplar features are described by graphs, where a connection between two features indicates that they can co-occur.



Two novel animal categories were designed such that one exhibited high-modularity and the other exhibited low-modularity.

Behavioral Task



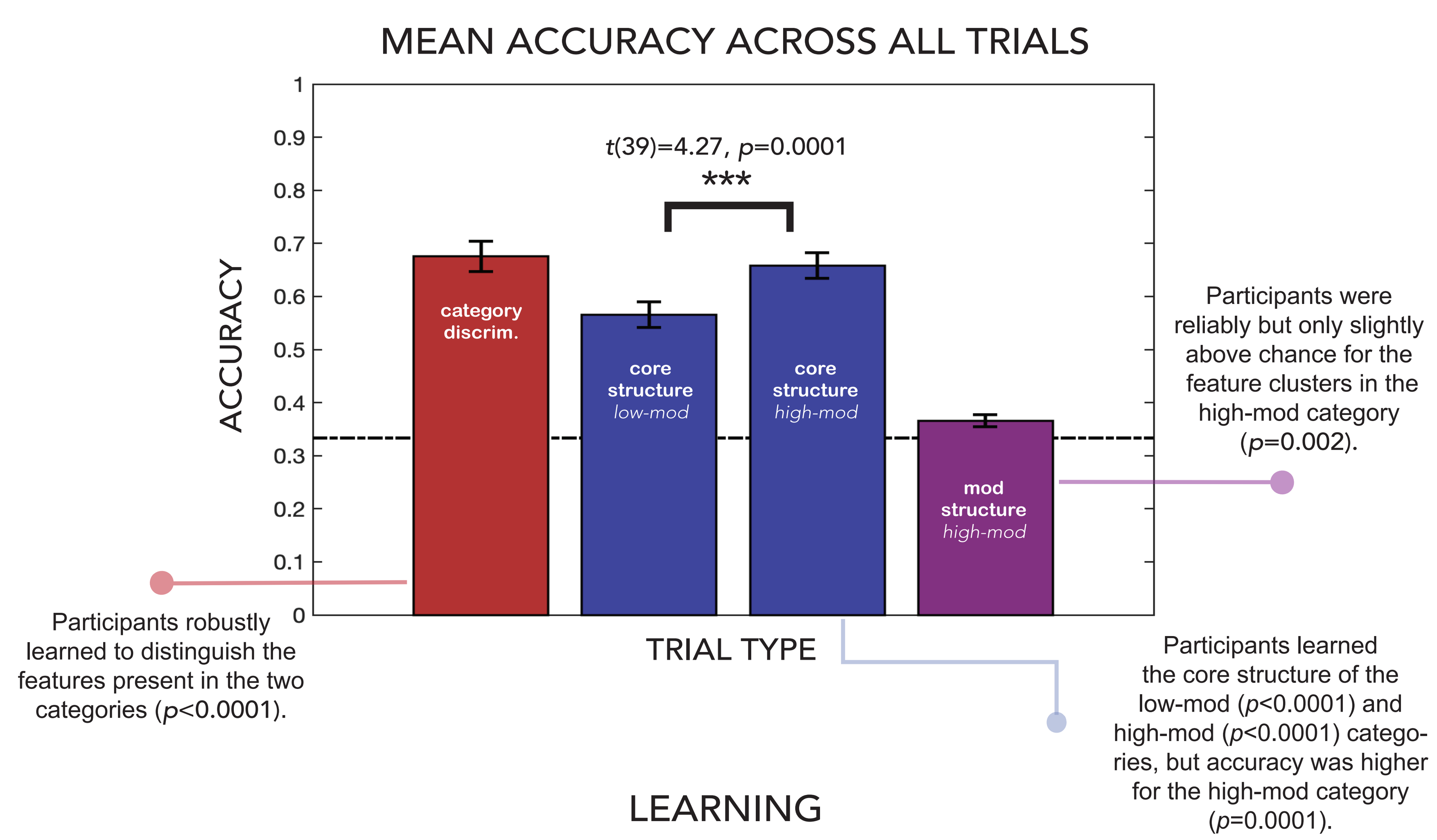
Human subjects (N=40) on MTurk completed a **missing-feature task**

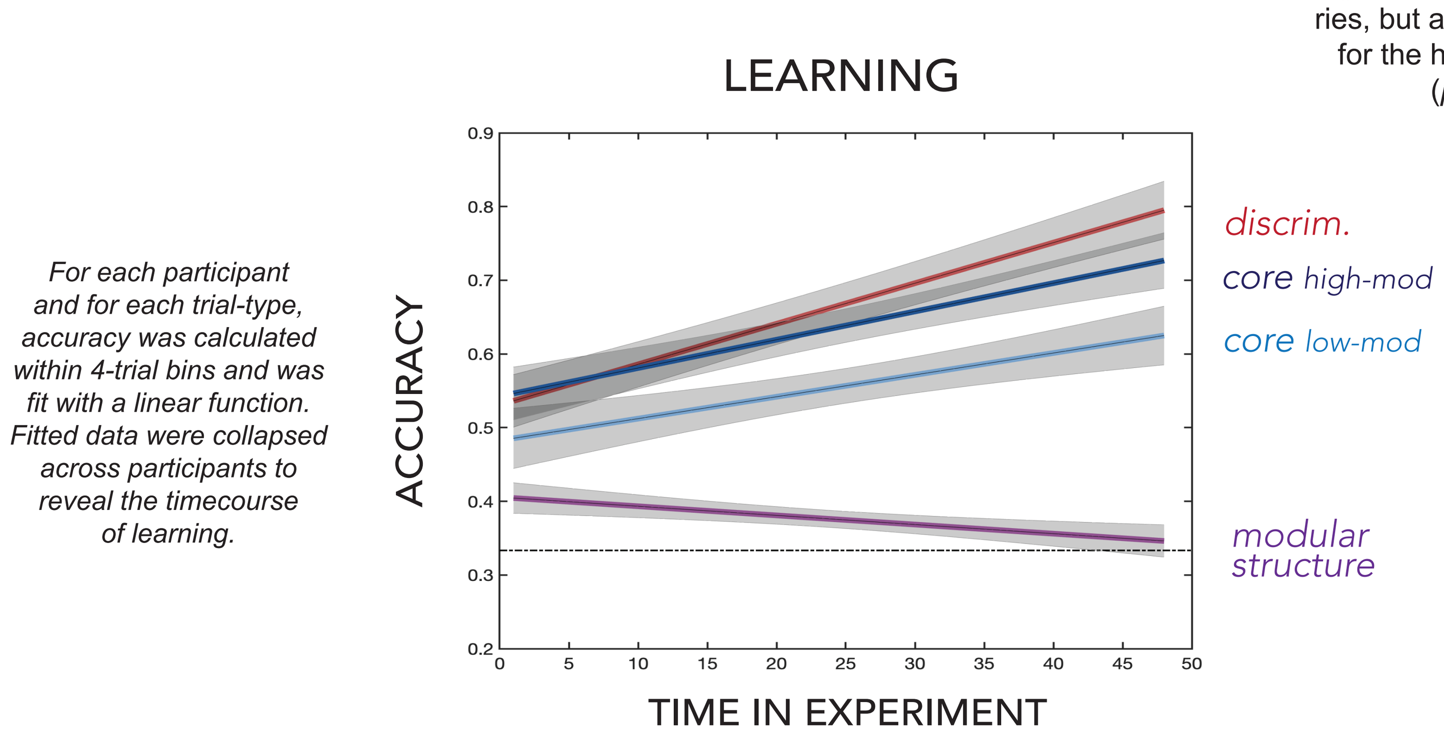
- 144 trials
- One feature was missing on each trial (**core**-feat, **mod**-feat, or **peripheral**-feat)
- Trial order pseudo-randomized with clusters of ~5 trials per category
- Feedback given after response, and trial repeated until correct feature was selected

A missing-feature task tested different kinds of structure knowledge: category distinctions, core structure, and modular structure.

Human Behavior Results

We calculated mean accuracy across subjects for each of the four kinds of structure-types: category discrimination, core structure, and modular structure.





Despite core structure being identical in the two categories, core features were easier to learn in the high-mod category.

Modeling Methods

Training: autoencoder

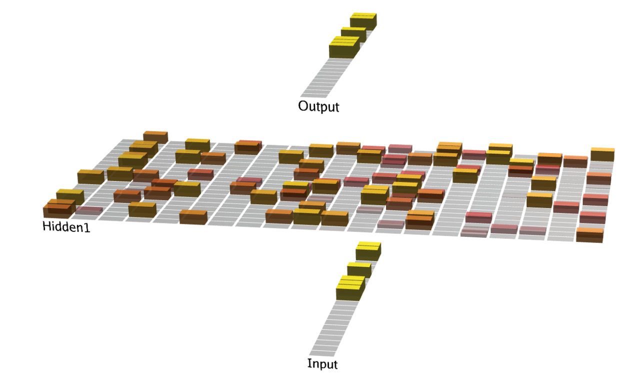
- 144 input patterns for the 144 behavioral trials
- Each input pattern is a complete exemplar *(-eats)*
- Replicates input pattern on output layer

Testing: pattern completion *(after every 8 training trials)*

- 144 input patterns for the 144 behavioral trials
- Each input pattern corresponds to shown features on each behavioral trial *(1 feature missing)*
- Activity on output layer reveals whether additional correct features activated

Architecture

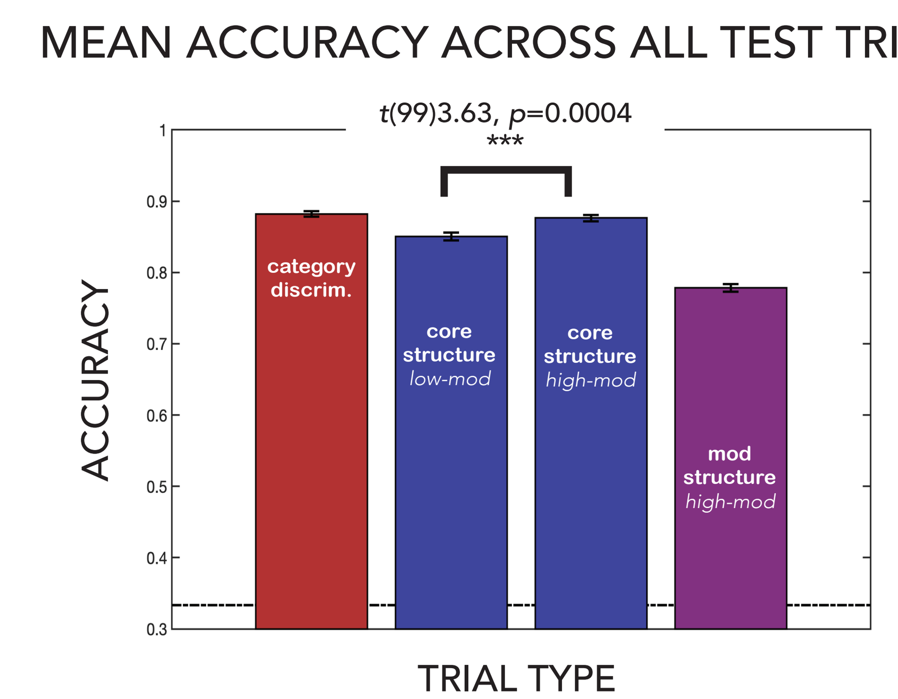
1. Input layer: 22 units *(features)*
2. Hidden layer: 100 units
3. Output layer: 22 units *(features)*

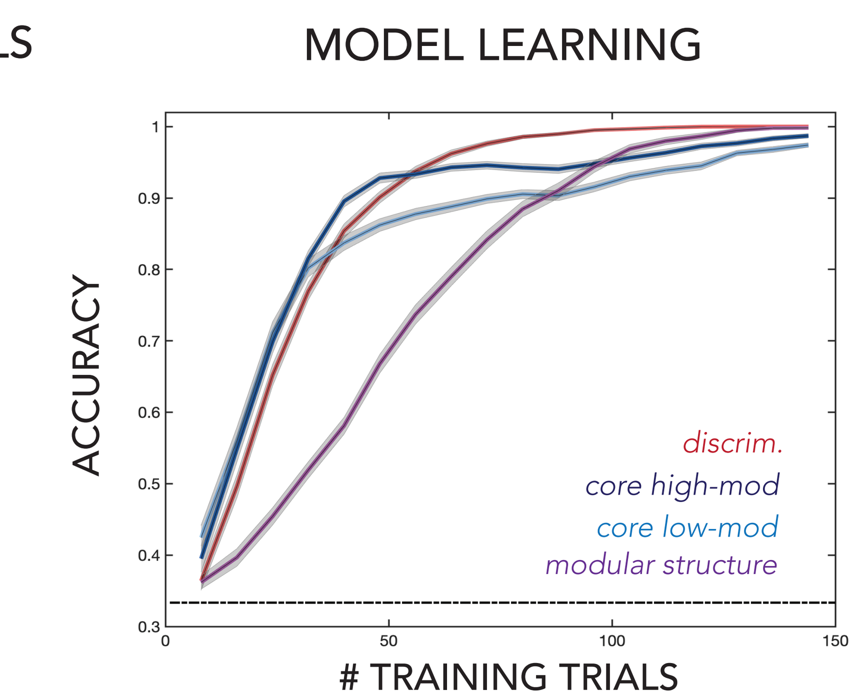


A neural network model learned these structured categories in a paradigm analogous to the missing-feature task used in human behavior.

Model Results

Accuracy was calculated over 100 runs of the model.





The model's behavior mirrored human behavior, such that core-structure was more easily learned in the high-mod category.

Conclusions

When learning the internal structure of two carefully designed novel categories, both humans and models found it easier to learn core *(always present)* features in the category that contained additional feature clusters *(high-modularity)*, even though this core structure was identical in the other category that did not contain feature clusters *(low-modularity)*. This could be because the high-mod category has more consistent pairs of co-occurring features that predict the core feature. More generally, these results suggest that learning individual components of category structure is influenced by the global structure of that category.

References:

1. Rogers, T. T., & McClelland, J. L. (2004). Semantic cognition: A parallel distributed processing approach. MIT press.
2. 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.
3. Solomon, S. H., Medaglia, J. D., & Thompson-Schill, S. L. (2019). Implementing a concept network model. Behavior research methods.

Contact: sarahsol@sas.upenn.edu