# The hierarchical structure is optimal given human cognition.

Takao Noguchi, Brad Love
Department of Experimental Psychology, University College London

Abstract

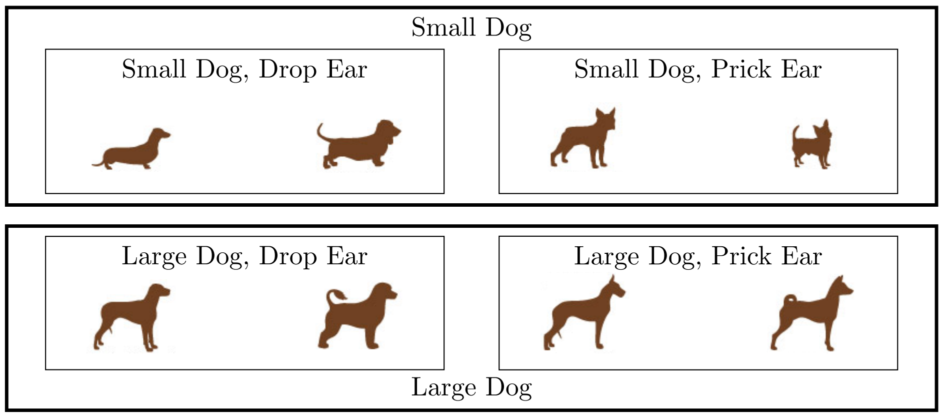
Environments surrounding people are often hierarchically structured: a single object may fit into a series of progressively more general categories (e.g., a terrier, dog, and animal). Given the prevalence of this hierarchical structure, previous studies often assumed that knowledge representation is also hierarchical, although empirical evidence does not provide a clear support. In this contribution, we demonstrate that the hierarchical structure enables people to make the most accurate inference and should be preferred, even when knowledge representation is flat. This optimality is most pronounced when knowledge is represented with clusters of observed objects. Further, the cluster representation naturally leads to the effects of learning order, which provides possible mechanisms behind previous empirical findings: the age of acquisition effects and the basic level advantage.

Even informal observation of everyday categorization reveals that many objects fit into a number of categories. A single object might be called a terrier, dog, mammal or animal. This hierarchical structure of categories — a sequence of progressively larger categories — has been suggested as a universal property of category structure across cultures (Atran, 1998; Berlin, 1992). Given the wide-spread adaptation of the hierarchical structure, it is tempting to assume that knowledge is hierarchically represented. This assumption of hierarchical representation, indeed, often underlies theoretical propositions (e.g., Tenenbaum, Kemp, Griffiths, & Goodman, 2011). In this study, however, we demonstrate that the hierarchical structure is optimal even when knowledge representation is flat, without the depth represented in the hierarchical structure.

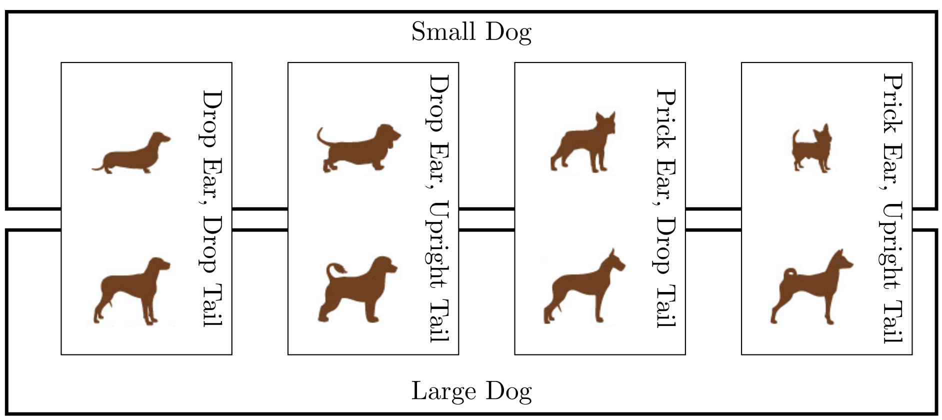
## Hierarchical structure of categories

To begin, let us illustrate the hierarchical structure. An example hierarchical structure is illustrated in Figure 1(a). Formally, the hierarchical structure satisfies the two characteristics: maximization of feature information and containment relation. First, the hierarchical

Word count: 2875.
Last updated: October 30, 2015.




(a) Hierarchical structure, where dogs share the maximum number of feature values at both levels, and categories at the lower level are nested within categories at the higher level.




(b) Non-hierarchical structure, where dogs share the maximum number of feature values at each level, but categories at the lower level are orthogonal to categories at the higher level.

*Figure 1*. Example structures with dogs. A rectangle encloses dogs in a category, with a thicker rectangle indicating a higher level category.

structure maximizes the information which objects, $X$, provide about a category $Y^{(i)}$:

$$I\left(Y^{(i)};X\right), \tag{1}$$

which is mutual information between category labels and objects. This feature information is maximized when all the objects in a category share values on the largest possible number of features.

The hierarchical structure also satisfies the containment relation, which maximizes the following measure:

$$\sum_{i=1}^{k-1} I\left(Y^{(i)};Y^{(i+1)}\right), \tag{2}$$

which is a sum of mutual information between multiple levels of categories. This containment measure is maximized when category labels form a partially ordered set: when all the objects in one category at the lower level fit into the same category at the higher level (see Figure 1 for examples).

Previous studies have demonstrated that people prefer the hierarchical structure. When people are asked to categorize an object in any manner people prefer, for example, people are very likely to provide hierarchical structure of categories (e.g., Rosch, Mervis, Gray, Johnson, & Boyes-Braem, 1976). Given the wide-spread preference for hierarchical structure of categories, it has been argued that knowledge is likely represented with hierarchy (e.g., Markman, 1989; Markman & Callanan, 1984). Markman (1989), for example, argues that people learn to hierarchically represent knowledge as they accumulate knowledge (see also Inhelder & Piaget, 1964; Vygotsky, 1962).

This assumption of hierarchical representation has been underlying theoretical propositions in cognitive science. For example, a mathematical model has been proposed to explain how people could learn to hierarchically represent knowledge (Kemp & Tenenbaum, 2008), and an inference drawn with this model correlates well with inference people make (Kemp & Tenenbaum, 2009). The assumption of hierarchical representation also underlies theories in many other domains within cognitive science (e.g., inference (Osherson, Smith, Wilkie, López, & Shafir, 1990), memory (Bower, Clark, Lesgold, & Winzenz, 1969; Glass & Holyoak, 1975), reasoning (Collins & Michalski, 1989; Shastri & Ajjanagadde, 1993), and word learning (Xu & Tenenbaum, 2007)).

The assumption of hierarchical representation, however, is not well supported with empirical evidence (see Murphy & Lassaline, 1997, for review). Sloman (1998), for example, report that when reasoning and making inferences with the hierarchical structure, people often neglect the containment relation, that all the objects in one category at the lower level fit into the same category at the higher level.

Thus, it has been left unexplained why and how people prefer the hierarchical structure, when knowledge is not hierarchically represented. In this study, we aim to provide answer to this open question. In particular, we demonstrate that even when knowledge representation is flat — when representation does not distinguish lower and higher levels — the hierarchical structure is optimal given human cognition. Models of human cognition is discussed below.

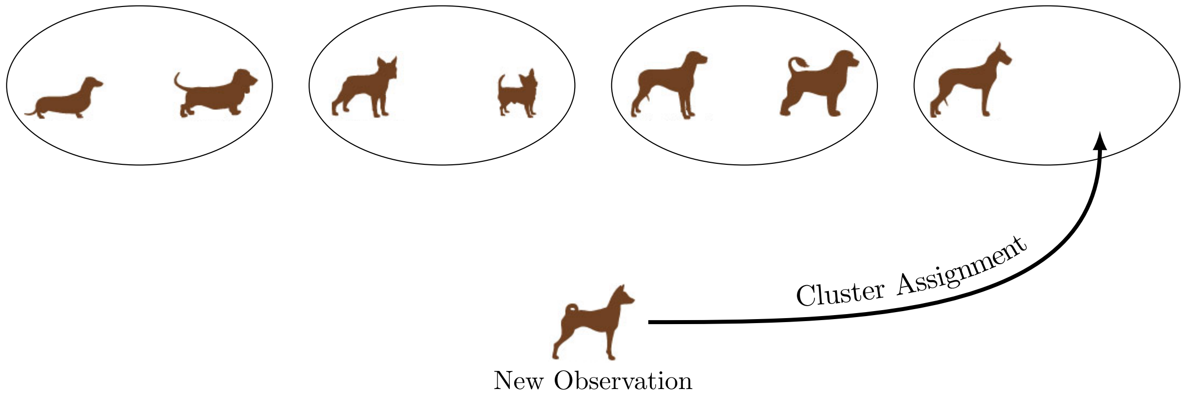



*Figure 2*. Example clusters. An ellipse encloses dogs in a cluster. A new observation is assigned to a cluster which contains most similar dogs.

## Models of human cognition

Cognitive models often assume that people represent knowledge as clusters of the observed objects (e.g., Anderson, 1991; Love, Medin, & Gureckis, 2004; Sanborn, Griffiths, & Navarro, 2010). Every time a new object is observed, the new object is assigned to an existing cluster with similar objects (see Figure 2 for an example). If none of the existing clusters contains sufficiently similar objects, a new cluster is created to host the new object. Then using the knowledge represented as clusters, an inference is drawn from the observed objects in the cluster. If a new object is assigned to a cluster predominantly with terriers, for example, the new object is considered to be likely to be a terrier.

Formally, the probability that object $x$ is inferred as an instance of category $w$ at level $i$ is expressed as follows:

$$p(y^{(i)} = w \mid x) = \sum_{k \in \mathbb{Z}} p(y^{(i)} = w \mid z = k)\, p(z = k \mid x). \tag{3}$$

Here, $\mathbb{Z}$ is a set of all the possible clusters. The probability of category label, $p(y^{(i)} = w \mid z = k)$, signifies drawing inference with clusters, and the probability of a cluster, $p(z = k \mid x)$, signifies knowledge representation. We first discuss how an inference can be improved with clusters, and then address how the improvement depends on knowledge representation.

**Inference.** First, an inference can be more accurate when the probability of category label given a cluster is optimized. In particular, $p(y^{(i)} = w \mid z = k)$ approaches 1, when all the objects in cluster $k$ fit into category $w$. Thus, an inference can be most accurate when the cluster structure follows the category structure. Thus as objects which share values tend to be assigned to the same cluster, an inference is more accurate when objects in the same category share values on more features. When objects in the same category share values on the maximum number of dimensions, the structure maximizes the feature information (Equation 1).

Also to allow an accurate inference across multiple levels of categories, objects in the same cluster have to fit into the same categories at both lower and higher levels. With the cluster structure illustrated in Figure 2, for example, an accurate inference can be made for

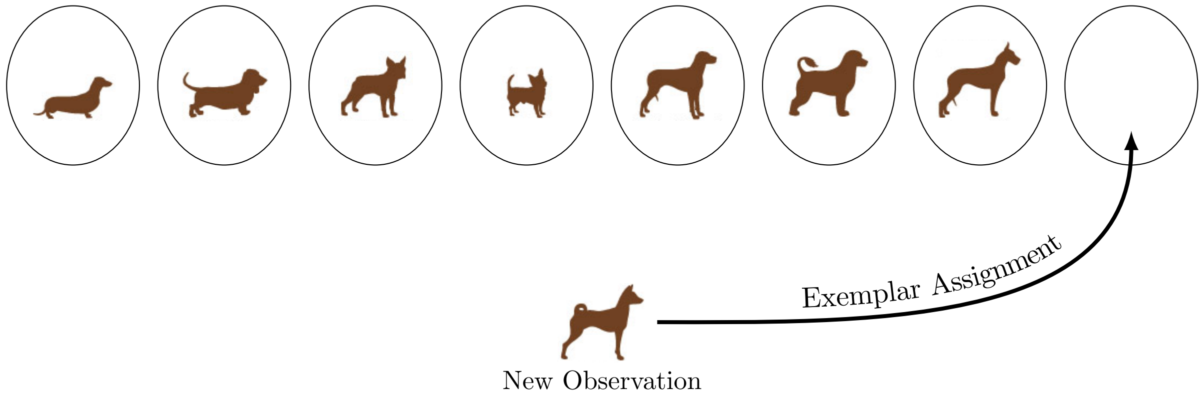



*Figure 3*. Exemplar representation, where each object is assigned to its own cluster.

the small/large dog categories at the higher level, and the prick/drop ear categories at the lower level (see Figure 1(a)). An inference, however, cannot be accurate when the category structure at the higher level is orthogonal to the category structure at the lower level: for example, when small/large dogs are distinguished at the higher level, but this distinction are neglected at the lower level (see Figure 1(b)). When all the objects in a category at the lower level fit into a same category at the higher level category, the category structure maximizes the containment measure (Equation 2).

**Cluster structure.** Therefore, the hierarchical structure is expected to be optimal, such that a human learner is most likely to make an accurate inference across multiple levels of categories. This optimality is, however, depends on how clusters are structured.

Previous research indicates that also psychologically plausible is the exemplar representation (Nosofsky, 1986, 1991) (see Figure 3 for an illustration). With the exemplar representation, each cluster contains a single object, and an inference is based on the most similar among the observed objects. If an object is most similar to an observed object labeled as a terrier, for example, the object is inferred to be a terrier. Thus, an inference is more accurate when similar objects fit into the same category: when the feature information is maximized. As the cluster structure cannot follow the category structure, however, the exemplar representation is insensitive to the containment relation: whether objects in a category at the lower level fit into the same category at the higher level.

The simulation we report below examined whether the maximization of the feature information and the containment relation improves inference accuracy. The exemplar and cluster representations are simulated with the same model but with different parameter values.

## Simulation 1

### Methods

We have computed the feature information (Equation 1) and the containment measure (Equation 2) for all the possible category structures with eight objects. We coded the three features of eight dogs in Figure 1 into a binary and constrained the possible category

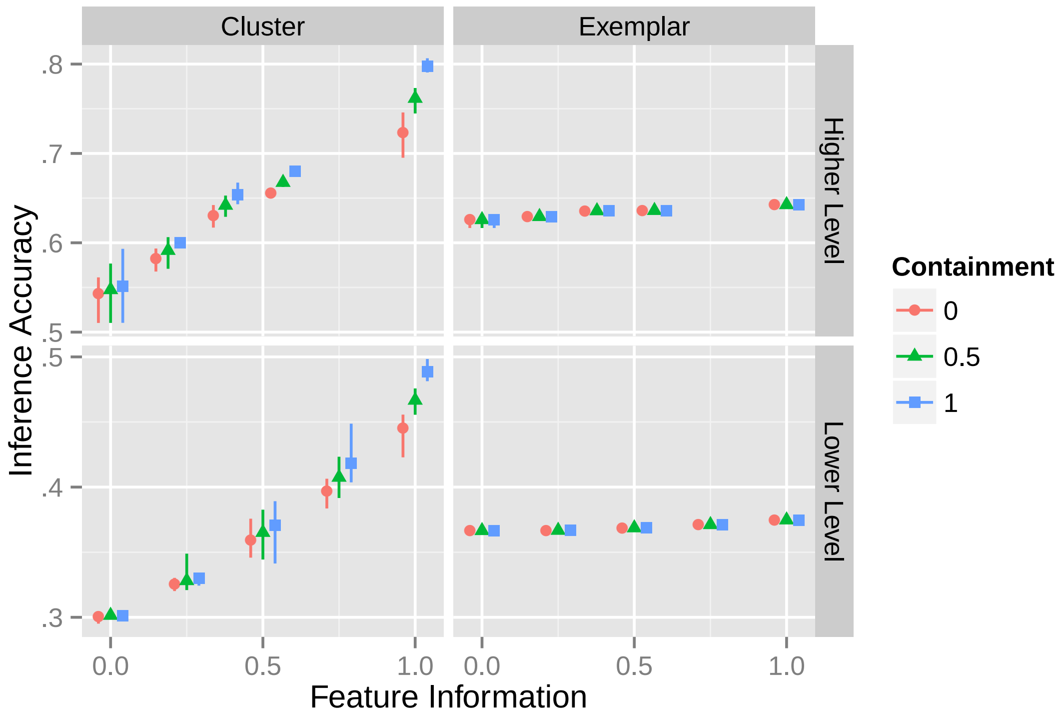



*Figure 4*. The average inference accuracy for each category structure. The levels of categories are shown in rows: the top panels illustrate the higher level, and the bottom panels illustrate the lower level. The two representations are shown in columns: the left panels illustrate the cluster representation, and the right panels illustrate the exemplar representation. Each panel summarizes the accuracy for each combination of feature information and containment measure. A dot represents mean, and error bar is 95% empirical interval. Each dot is jittered along the horizontal axis for the illustration purposes.

structure, such that the degree of branching is two: the higher level has two categories with four dogs each, and the lower level contains four categories with two dogs each. Since the ordering of features is exchangeable, the number of possible category structure totals to 3,675.

For each category structure, the model is trained for 10 blocks. Each block involves 16 trials, the eight objects with a category label at either lower or higher level, in a random order. In a trial with one level (e.g., the lower level), a category label at the other level (e.g., the higher level) is treated as missing. After the training blocks, the average inference accuracy was calculated for each level of categories. This accuracy was further mean-averaged across the $10^4$ simulations for each category structure (Please refer to Appendix A for more details of the model and the simulation).

## Results and Discussion

For illustration purposes, we scaled the feature information, so that it ranges from 0 to 1 on each category level. The simulation results are summarized in Figure 4. As predicted for the cluster representation, the inference is more accurate with the feature information and the containment measure: The highest accuracy is achieved with the largest feature information and the largest containment measure.

With the exemplar representation, however, the inference is only slightly more accurate with the feature information. This is because the identical objects are repeatedly presented in the simulation, and hence, an inference is often based on the identical object, as opposed to the most similar object among the observed ones. Thus, the similarity between objects tend to have little impact on inference. Here, the highest accuracy is achieved with the largest feature information but the containment measure does not appear to have an impact.

To confirm, we fitted linear regressions on the inference accuracy, which when appropriate, allow each level to have varying intercepts and slopes. The estimated slopes confirm the above observations: both feature information and containment measure increases the inference accuracy to a greater extent with the cluster representation than with the exemplar representation: the interaction effect is at $\beta = 0.15$ (95% CI [0.15, 0.16]) for the feature information, and at $\beta = 0.02$ (95% CI [0.02, 0.03]) for the containment measure. Further, the containment relation has a zero impact on the inference accuracy with the exemplar representation: $\beta = 0.00$ (95% CI [0.00, 0.00]).

Thus as expected, the simulation results show that the hierarchical structure is optimal given human cognition. This optimality is, however, more pronounced with the cluster representation than with the exemplar representation. While the cluster representation shows improved inference accuracy with both feature information and containment relation, the exemplar representation shows improved inference accuracy only with feature information.

## Simulation 2

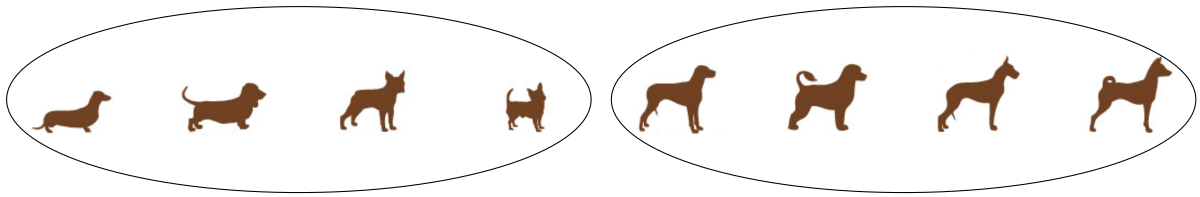
The above simulation results confirm that the hierarchical structure is optimal given the cluster representation. The cluster representation, however, also predicts effects of learning order on inference accuracy. The predicted effects were tested in Simulation 2.

The predicted effects of learning order are illustrated in Figure 5. As the cluster assignment of an object considers its category label as well as the object's features, objects from the same category tend to be perceived similar and fit into the same cluster. As a result, the cluster structure at first tends to follow the category structure learned at first. As subsequent learning at another category level builds upon this initial cluster structure, the cluster structure cannot as closely follow the category structure learned later. Consequently, an inference tends to be more accurate at the category level learned at first.

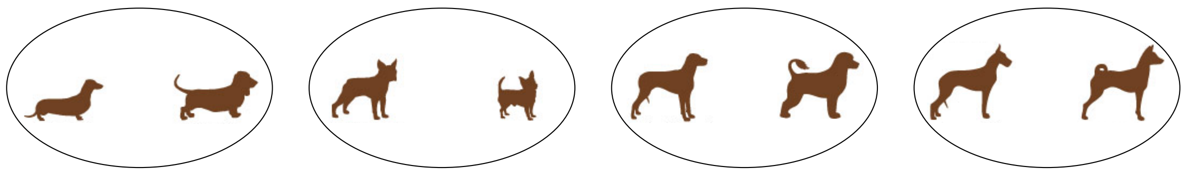
With the exemplar representation, however, each object is assigned to its own cluster, regardless of learning order. Thus, we expect effects of learning order with the cluster representation but not with the exemplar representation.

### Methods

Simulation 2 employs the almost identical methods to Simulation 1, except that we only tested the hierarchical structures with two possible learning orders: lower-higher or higher-lower. Also, a training block in Simulation 2 presents the eight objects in a random



(a) When categories at the higher level are learned first, the cluster structure mimics the category structure at the higher level, allowing an accurate inference at the higher level.



(b) When categories at the lower level are learned first, the cluster structure mimics the category structure at the lower level, allowing an accurate inference at the lower level.

*Figure 5*. Illustrative cluster structures for each learning order. An ellipse encloses dogs in a cluster.

order with a category label at either lower or higher level. Then, the learning order of lower-higher involves 10 training blocks with the lower level first, followed by 10 training blocks with the higher level categories. This order is reversed for the learning order of higher-lower.

**Results and Discussion**

Figure 6 illustrates the average inference accuracy for each learning order. This figure shows that with the cluster representation, an inference tends to be more accurate at the level learned at first. The top left panel in Figure 6, for example, shows that when the categories at the higher level are learned at first, an inference is more accurate at the higher level than when the same categories are learned later. With the exemplar representation, in contrast, the inference accuracy does not appear influenced by the learning order.

To confirm, we fit mixed-effect linear regressions to the inference accuracy. The estimated slopes confirm the above observations: the learning order has a larger impact with the cluster representation than with the exemplar representation, as indicated by the interaction effect ($\beta = 0.08$, 95% CI [0.08, 0.08]). With the cluster representation, the learning order has different effect at each level ($\beta = 0.08$, 95% CI [0.08, 0.08]). On the other hand with the exemplar representation, the learning order does not have an observable effect ($\beta = 0.00$, 95% CI [0.00, 0.00]).

These results confirm our predictions: with the cluster representation, learning is built upon the existing cluster representation. The cluster representation tends to follow the category structure learned at first. This cluster structure carries over to the subsequent learning, and as a result, an inference tends to be more accurate at the level which is learned first.

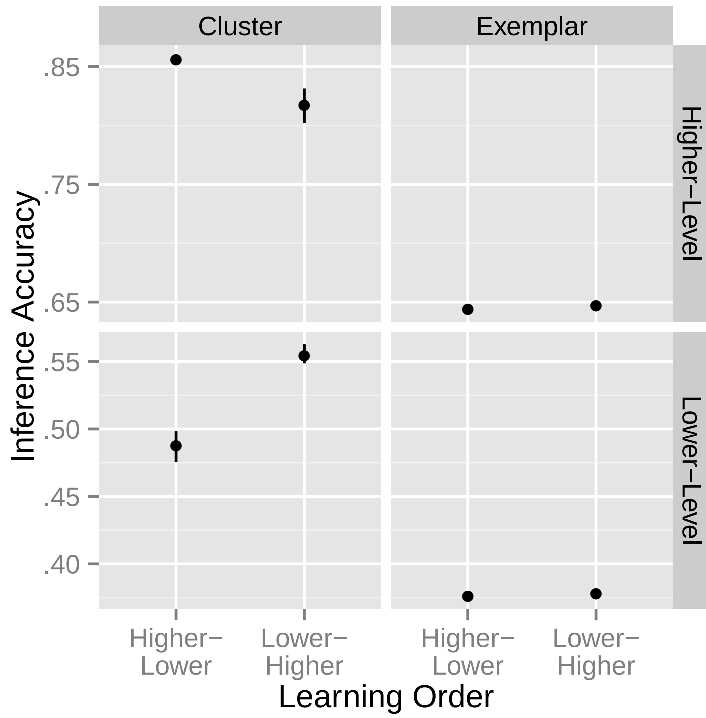



*Figure 6*. The average inference accuracy for learning order. The levels of categories are shown in rows: the top panels illustrate the higher level, and the bottom panels illustrate the lower level. The two representations are shown in columns: the left panels illustrate the cluster representation, and the right panels illustrate the exemplar representation. A dot represents mean, and error bar is 95% empirical interval.

## General Discussion

Previous studies within cognitive science have shown widespread preference for hierarchical structure (e.g., Rosch et al., 1976). This prevailing preference of hierarchical structure has led researchers to assume that knowledge is hierarchically represented (e.g., Markman, 1989; Markman & Callanan, 1984). This assumption of hierarchical representation has underlay theories in many domains within cognitive science.

Although we do not argue against the possibility that knowledge is hierarchically represented, our simulation results suggest that how people categorize objects may not necessarily correspond to how knowledge is psychologically represented. Our results indicate that the preference for the hierarchical structure is not necessarily due to the hierarchical representation. Even with the cluster representation, which does not distinguish higher and lower levels, the hierarchical structure is optimal and should be preferred. Thus, this study offers an alternative interpretation of the wide adaptation of hierarchical structure: the adaptation can be seen as evidence for the cluster representation over the exemplar representation.

In addition, the results from Simulation 2 highlight a mechanism behind the age-of-acquisition effects (Gerhand & Barry, 1998; Morrison & Ellis, 1995): words that are acquired earlier in childhood are processed more accurately than words that are acquired later in life. Our results show that a category structure learned at first tends to shape the knowledge representation, allowing a more accurate inference.

This initial representation underlies subsequent learning, which further explains the basic level advantage (Mervis & Rosch, 1981; Rosch et al., 1976), where an inference people make is generally more accurate at the basic level (e.g., dog) than at a more general (e.g., mammal) or more specific level (e.g., terrier). Our results provide the explanation that the basic level advantage may be a consequence of learning order. Indeed, previous research has demonstrated that categories at the basic level are learned earlier at the childhood than at other levels (Berlin, Breedlove, & Raven, 1973; Brown, 1958; Horton & Markman, 1980; Mervis & Crisafi, 1982).

In addition, the simulation method we used complements machine teaching (Zhu, 2013). Machine teaching is a procedure to identify the optimal teaching objects, which enables learners to achieve the most accurate inference. In this machine teaching, the procedure is conditioned on the learning model and the category structure. In contrast, our simulation is conditioned on the learning model and identified the category structure which enables the learners to achieve the most accurate inference.

Further, our simulation procedure has potential for practical applications: our simulation identifies the structure of objects which is best for human cognition. Thus for example, a retailer could use the same procedure to identify the best product categories to present to consumers. Such categories would best-fit consumers' cognition and may help consumers in finding the products which best suit their needs.

## References

Anderson, J. R. (1991). The adaptive nature of human categorization. *Psychological Review*, *98*, 409–429.

Atran, S. (1998). Folk biology and the anthropology of science: cognitive universals and cultural particulars. *Behavioral and Brain Sciences*, *21*, 547–609.

Berlin, B. (1992). *Ethnobiological classification: principles of categorization of plants and animals in traditional societies*. New Jersey: Princeton University Press.

Berlin, B., Breedlove, D. E., & Raven, P. H. (1973). General principles of classification and nomenclature in folk biology. *American Anthropologist*, *75*, 214–242.

Bower, G. H., Clark, M. C., Lesgold, A. M., & Winzenz, D. (1969). Hierarchical retrieval schemes in recall of categorized word lists. *Journal of Verbal Learning and Verbal Behavior*, *8*, 323–343.

Brown, R. (1958). How shall a thing be called? *Psychological Review*, *65*, 14–21.

Collins, A. & Michalski, R. (1989). The logic of plausible reasoning: a core theory. *Cognitive Science*, *13*, 1–49.

Gerhand, S. & Barry, C. (1998). Word frequency effects in oral reading are not merely age-of-acquisition effects in disguise. *Journal of Experimental Psychology: Learning, Memory, & Cognition*, *24*, 267–283.

Glass, A. L. & Holyoak, K. J. (1975). Alternative conceptions of semantic theory. *Cognition*, *3*, 313–339.

Horton, M. S. & Markman, E. M. (1980). Developmental differences in the acquisition of basic and superordinate categories. *Child Development*, *51*, 708–719.

Inhelder, B. & Piaget, J. (1964). *The early growth of logic in the child : classification and seriation*. London: Routledge and Kegan Paul.

Kemp, C. & Tenenbaum, J. B. (2008). The discovery of structural form. *Proceedings of the National Academy of Sciences of the United States of America*, *105*, 10687–10692.

Kemp, C. & Tenenbaum, J. B. (2009). Structured statistical models of inductive reasoning. *Psychological Review*, *116*, 20–58.

Love, B. C., Medin, D. L., & Gureckis, T. M. (2004). SUSTAIN: A network model of category learning. *Psychological Review*, *111*, 309–332.

Markman, E. M. (1989). *Categorization and naming in children*. Massachusetts: MIT Press.

Markman, E. M. & Callanan, M. A. (1984). An analysis of hierarchical classification. In R. J. Sternberg (Ed.), *Advances in the psychology of human intelligence* (Vol. 2, pp. 325–365). New Jersey: Lawrence Erlbaum Associates, Inc.

Mervis, C. B. & Crisafi, M. A. (1982). Order of acquisition of subordinate-, basic-, and superordinate-level categories. *Child Development*, *53*, 258–266.

Mervis, C. B. & Rosch, E. (1981). Categorization of natural objects. *Annual Review of Psychology*, *32*, 89–115.

Morrison, C. M. & Ellis, A. W. (1995). Roles of word frequency and age of acquisition in word naming and lexical decision. *Journal of Experimental Psychology: Learning, Memory, & Cognition*, *21*, 116–133.

Murphy, G. L. & Lassaline, M. E. (1997). Hierarchical structure in concepts and the basic level of categorization. In K. Lamberts & D. Shanks (Eds.), *Knowledge, concepts and categories* (pp. 93–132). East Sussex: Psychology Press.

Nosofsky, R. M. (1986). Attention, similarity, and the identification-categorization relationship. *Journal of Experimental Psychology: General*, *115*, 39–57.

Nosofsky, R. M. (1991). Relation between the rational model and the context model of categorization. *Psychological Science*, *2*, 416–421.

Osherson, D. N., Smith, E. E., Wilkie, O., López, A., & Shafir, E. (1990). Category-based induction. *Psychological Review*, *97*, 185–200.

Rosch, E., Mervis, C. B., Gray, W. D., Johnson, D. M., & Boyes-Braem, P. (1976). Basic objects in natural categories. *Cognitive Psychology*, *8*, 382–439.

Sanborn, A. N., Griffiths, T. L., & Navarro, D. J. (2010). Rational approximations to rational models: alternative algorithms for category learning. *Psychological Review*, *117*, 1144–1167.

Shastri, L. & Ajjanagadde, V. (1993). From simple associations to systematic reasoning: a connectionist representation of rules, variables and dynamic bindings using temporal synchrony. *Behavioral and Brain Sciences*, *16*, 417–451.

Sloman, S. A. (1998). Categorical inference is not a tree: the myth of inheritance hierarchies. *Cognitive Psychology*, *35*, 1–33.

Tenenbaum, J. B., Kemp, C., Griffiths, T. L., & Goodman, N. D. (2011). How to grow a mind: statistics, structure, and abstraction. *Science*, *331*, 1279–1285.

Vygotsky, L. (1962). *Thought and language*. Massachusetts: MIT Prss.

Xu, F. & Tenenbaum, J. B. (2007). Word learning as Bayesian inference. *Psychological Review*, *114*, 245–272.

Zhu, X. (2013). Machine teaching for bayesian learners in the exponential family. In C. J. C. Burges, L. Bottou, M. Welling, Z. Ghahramani, & K. Q. Weinberger (Eds.), *Advances in Neural Information Processing Systems 26.*

## Appendix
### Details of Simulation with Rational Model of Cognition

Suppose a learner has observed $n-1$ objects $\{x_1, x_2, \ldots, x_{n-1}\}$ with corresponding category labels $\{y_1, y_2, \ldots, y_{n-1}\}$. Each of these objects fits into a cluster. The cluster label for the $i$th object is denoted as $z_i$.

**Drawing an inference**

Then, the probability that the $n$th object fits into category $w$ is expressed as follows:

$$\begin{aligned} p(y_n = w \mid x_n) &= \sum_{k \in \mathbb{Z}} p(z_n = k \mid x_n)\, p(y_n = w \mid z_n = k) \\ &= \sum_{k \in \mathbb{Z}} \frac{p(z_n = k)\, p(x_n \mid z_n = k)}{p(x_n)}\, p(y_n = w \mid z_n = k) \\ &= \sum_{k \in \mathbb{Z}} \frac{p(z_n = k)\, p(x_n \mid z_n = k)}{\sum_{s \in \mathbb{Z}} p(z_n = s)\, p(x_n \mid z_n = s)}\, p(y_n = w \mid z_n = k). \end{aligned} \tag{4}$$

Here, $\mathbb{Z}$ is a set of all the possible clusters to which the $n$th object can be assigned. The three terms in Equation 4 are described below in turn.

First, the probability that the $n$th object fits into cluster $k$ is given by:

$$p(z_n = k) = \begin{cases} \dfrac{c\, m_k}{(1-c) + c\,(n-1)} & \text{if} \quad m_k > 0 \\[2ex] \dfrac{(1-c)}{(1-c) + c\,(n-1)} & \text{if} \quad m_k = 0 \end{cases} \tag{5}$$

where $c$ is a parameter called the coupling probability and $m_k$ is the number of objects assigned into cluster $k$.

Following Anderson (1991) and Sanborn et al. (2010), we assume that an object has independent dimensions. Therefore,

$$p(x_n \mid z_n = k) = \prod_{d \in D} p(x_{n,d} \mid z_n = k), \tag{6}$$

where $D$ is a set of dimensions in which an object is described. The above term is computed with

$$p(x_{n,d} = v \mid z = k) = \frac{B_{v,d} + \beta_c}{m_k + J_d\, \beta_c}, \tag{7}$$

where $B_{v,d}$ is the number of objects in cluster $k$ with value of $v$ on dimension $d$, and $J_d$ is the number of values which an object can take on dimension $d$. Parameter $\beta_c$ determines knowledge representation, as discussed below.

Similarly, the probability that the $n$th object has category label $w$, given a cluster, is given by:

$$p(y_n = w \mid z = k) = \frac{B_w + \beta_l}{B_. + J\,\beta_l}, \tag{8}$$

where $B_w$ is the number of observed objects with category label $w$ in cluster $k$, $B_.$ is the number of object in cluster $k$, and $J$ is the number of category labels. Also, parameter $\beta_l$ determines knowledge representation, as discussed below.

## Learning

The learning is equivalent to assigning an object into a cluster. The probability that an object is assigned to cluster $k$ is computed as

$$p(z_n = k \mid x_n,\, y_n) \propto p(z_n = k)\; p(x_n \mid z_n = k)\; p(y_n \mid z_n = k). \tag{9}$$

This is computed with Equations 5, 6 and 8. Additionally, this cluster assignment is conducted using the sequential Monte Carlo with one particle, which produces behavior much like human (Sanborn et al., 2010).

## Knowledge representation

With the above specification, knowledge representation is determined by values for parameter $\beta$. For the cluster representation, we used $\beta_f = 1.0$ and $\beta_l = 0.5$. These values are among the best fitting values to human performance (Sanborn et al., 2010).

For the exemplar representation, we used small values of $\beta$ ($\beta_f = 0.001$ and $\beta_l = 0.001$) during the training, ensuring that a cluster can only contain identical objects. Inference based on such small $\beta$ is, however, often deterministic, and to allow more probabilistic inference, we used a larger value for $\beta$ ($\beta_f = 1.0$ and $\beta_l = 5.0$) during the testing.

For both representations, the coupling probability $c$ is at 0.5 during the training blocks and at 1.0 during the testing blocks. These coupling probabilities prevent a new cluster from being created and ensure that an inference is not made with random guessing during the testing.