# From simple to complex categories: How structure and label information guides the acquisition of category knowledge

Wai Keen Vong
School of Psychology
University of Adelaide
January, 2018

# Table of Contents

# List of Publications

The following is a full citation list of publications appearing in this thesis.

- Chapter 2: Vong, W.K, Perfors, A. & Navarro, D. J. (2016). The helpfulness of labels in semi-supervised learning depends on category structure. *Psychonomic Bulletin & Review*. 23(1), 230-238.
- Chapter 3: Vong, W.K, Hendrickson, A.T., Navarro, D. J. & Perfors, A. (unpublished manuscript). Learning structure with labeled examples.
- Chapter 4: Vong, W.K, Hendrickson, A.T., Perfors, A. & Navarro, D. J. (manuscript submitted for publication). Do additional features help or hurt category learning? The curse of dimensionality in human learners.

# From simple to complex categories: How structure and label information guides the acquisition of category knowledge

## Abstract

Categorization is a fundamental ability of human cognition, translating complex streams of information from the all of different senses into simpler, discrete categories. How do people acquire all of this category knowledge, particularly the kinds of rich, structured categories we interact with every day in the real-world? In this thesis, I explore how information from category structure and category labels influence how people learn categories, particular for the kinds of computational problems that are relevant to real-world category learning. The three learning problems this thesis covers are: semi-supervised learning, structure learning and category learning with many features. Each of these three learning problems presents a different kinds of learning challenge, and through a combination of behavioural experiments and computational modeling, this thesis illustrates how the interplay between structure and label information can explain how humans can acquire richer kinds of category knowledge.

# Declaration

I certify that this work contains no material which has been accepted for the award of any other degree or diploma in my name, in any university or other tertiary institution and, to the best of my knowledge and belief, contains no material previously published or written by another person, except where due reference has been made in the text. In addition, I certify that no part of this work will, in the future, be used in a submission in my name, for any other degree or diploma in any university or other tertiary institution without the prior approval of the University of Adelaide and where applicable, any partner institution responsible for the joint-award of this degree.

I acknowledge that copyright of published works contained within this thesis resides with the copyright holder(s) of those works.

I also give permission for the digital version of my thesis to be made available on the web, via the University's digital research repository, the Library Search and also through web search engines, unless permission has been granted by the University to restrict access for a period of time.

I acknowledge the support I have received for my research through the provision of an Australian Government Research Training Program Scholarship.

Wai Keen Vong
January 2018

# Acknowledgments

First and foremost, I owe a great deal of gratitude to each of my supervisors. To Amy, for initially taking the time to answer a curious undergraduate's e-mails about cognitive science with such enthusiasm. Looking back, so much of the development of my career as a researcher can be attributed to your advice and support at various critical junctures. To Dani, for helping distill my vague research ideas into actual scientific questions that other researchers would care about. The work presented here was greatly improved as a result of your input, from the genesis to the final product. And to Drew, for his unrelenting optimism and willingness to go deep into the trenches with me to collaborate on research projects together. Asking myself "What would Drew do?" has become a very useful tactic for getting myself unstuck on research problems.

I would also like give a special thanks to all of the other people at the Computational Cognitive Science Lab, which turned out to be an excellent place to spend some of my most formative years. To Lauren Kennedy, Steven Langsford and Keith Ransom for being part of the PhD journey together. And thanks to all of the others whose time at the CCS Lab overlapped with mine: Simon De Deyne, Dinis Gokaydin, Rachel Stephens, Anastasia Ejova, Luke Maurits, Sean Tauber, Wouter Voorspoels, Natalie May and Joey Ong. Also, thanks to Jessica O'Rielly, Emma Stewart, Heidi Long of the Active Vision Lab whom I all had the pleasure of sharing an office space with. Despite no windows in our basement offices, there was plenty of good company and cheers. Anna Ma-Wyatt also deserves a special thanks for filling in as my supervisor at the tail end of my thesis.

As part of my research, I was very fortunate to receive funding to travel to so many places and present my work at various conferences. Amy and Dani deserve another round of thanks here, for making sure I never needed to worry about funding for travel when the opportunity arose. In addition, the School of Psychology generously helped out with additional funding, as well as the Cognitive Science Society for a Student Travel Grant. I was also very fortunate to have been awarded the Marr Prize (Best Student Paper) by the Cognitive Science Society for the work presented in Chapter 4 of this thesis.

During my travels, I was able to visit a number of research labs to present my research. Many thanks to Patrick Shafto and the Cognitive and Data Science Lab at Rutgers University - Newark, Todd Gureckis and the Computation and Cognition Lab at New York University, Sam Gershman and the Computational Cognitive Neuroscience Lab at Harvard University and Steven Piantadosi and Celeste Kidd at the Computation and Language Lab and Rochester Baby Lab at the University of Rochester. Also thanks to the graduate students and postdocs from all of these labs for their company and conversations.

My research also led me to attend a number of workshops where I gained a lot of useful skills for research. The Bayesian Modeling for Cognitive Science workshop by Michael Lee and EJ Wagenmakers taught me a lot about Bayes. The Probabilistic Programming and Machine Learning Summer School hosted by Galois, where Frank Wood and his Machine Learning group at Oxford University and Noah Goodman's Computation and Cognition Lab from Stanford brought me up to speed with the world of probabilistic programming.

I am indebted and spoiled by the many good friendships outside of the lab. A particular thanks to Dennis Liu for his shared interest in all things intellectual, Jordan Gabriels for his commitment to the pomodoro technique, Guo Yu for being the most dependable friend, Stuart Lee for keeping me up to date on both R and all things cultural and George Stamatescu for his comedic wit. Other friends I would like to thank are Andrew Jia-Jun Ying, Asif Islam, Michael Le, Clifford Shum, Daniel Lee, Berny Chan, Chris Lee, Lily Guo, Hong Yeung, Julia Zhou, Monica Kei, Helen Lam, Jon Black-Tiong, Kate Bryson, Orddom Leav and many more.

Finally, to my family (Mum, Dad, Sal Ging and Wai Ho) for being the foundation for which all of this was possible.

# 1

# Introduction

The world is full of complexity and uncertainty, where people rarely encounter exactly the same situation twice. How have humans managed to survive and thrive in such difficult and diverse environments? One answer is that we have a number of cognitive abilities such as categorization, that allow us to flexibly adapt and generalize to situations humans have not encountered before. This ability allows humans to turn the blooming, buzzing confusion of everyday life into simpler and manageable chunks. Studying how people learn categories has been an influential area of cognitive science since its origin (Bruner, Goodnow, & Austin, 1956; Shepard, Hovland, & Jenkins, 1961) and remains an active area of inquiry today (Anderson, 1991; Goodman, Tenenbaum, Feldman, & Griffiths, 2008; Lake, Salakhutdinov, & Tenenbaum, 2015). It has had a wide-ranging impact on many related areas of cognition such as inductive reasoning (Osherson, Smith, Wilkie, Lopez, & Shafir, 1990), language

(Waxman & Markow, 1995) and perception (Goldstone & Hendrickson, 2010). In addition, research into category learning has led to the discovery of a number of deep and unifying connections with statistics and machine learning (Griffiths, Canini, Sanborn, & Navarro, 2007; Jäkel, Schölkopf, & Wichmann, 2008; Sanborn, Griffiths, & Navarro, 2010).

While our ability to categorize objects appears to be intuitive and effortless from the outside, there is a surprising amount of complexity to this problem below the surface. Objects can vary in many different ways: their orientation, their size, their shape or their colour. They can vary by whether they are occluded by other objects, and to the presence or absence of any number of varying features. This begs the question: how is it that people are able to learn and acquire knowledge about so many different kinds of categories?

To begin, it is illustrative to provide a concrete example to emphasize some of the challenges for learning to categorize between different objects. Consider a toddler who is playing outside and is learning to distinguish *cats* from *dogs*. At first glance, objects in both of these categories might appear to be similar in appearance and for a toddler, it might be unclear how they would begin to learn how to distinguish between these two categories. Should they be attending to the SHAPE or COLOUR of the objects? Or whether the object HAS SPOTS, or HAS WHISKERS? After some experience, they might begin to notice that objects that contain the feature HAS SPOTS are more likely to be dogs, and begin paying more attention to that particular feature than the COLOUR of the animal. However, this particular feature may not always be perfect, for example upon seeing a *Bengal cat* using the rule HAS SPOTS would cause the child to misclassify the animal as a dog instead of a cat. Thus, the computational problem of category learning involves figuring out how to combine information from both features and category labels to organize objects into meaningful categories. This process of learning categories requires the child to not only learn how to classify existing objects into categories, but to also *generalize* this knowledge to categorize new, previously unseen objects too.

One way that the difficulty of this learning problem could vary is to change the structure of the categories, or the kinds of features that are relevant for categorization. For example, what if the toddler was learning to categorize *dogs* and *horses* instead? In contrast to cats and dogs which are quite similar in appearance to each other, dogs and horses are noticeably different and a toddler might easily learn that they can be categorized on the basis of a salient feature such as SIZE. On the other hand, consider the task of learning to categorize between *dalmatians* and *border collies*. Given that they are both breeds of dogs, they share many similar features, so one could imagine that learning to categorize them would be much more difficult in comparison.

Another type of information that could also vary is through the labeled information provided to a learner. In a typical *supervised* category learning experiment, participants are presented with an object and are given feedback with the correct category label after each response. While this experimental design is common for studying category learning in the laboratory, there are many other ways in which children might receive category label information that could lead to different inferences. For example, labels might be explicitly provided by parents, with the intent to provide the most helpful examples to a learner. Another possibility is that the learner receives no category labels at all, and has to learn how to organize objects into categories on the basis of feature information alone. A third possibility might involve being presented with multiple labels for the same object, like hearing the words DOG and DALMATIAN for the same item, and figuring out how both labels relate to the same object. Thus, the different ways in which category labels are provided to a child might serve as an important signal for category learning.

## 1.1 Outline of this thesis

This thesis investigates how people learn categories, and addresses a number of key questions about how different kinds of information about category structure and category labels can shape our understanding of category learning. The main chapters in this thesis cover three different research questions related to how people learn categories, focusing on how people use information from category structure and category labels in tasks that are inspired by real-world category learning. Below, I provide a brief outline of the three research questions that I examine in this thesis.

1. How does information about category structure and category labels influence how people categorize objects into different categories? One of the computational problems a learner needs to solve is to determine how to sort a set of objects into different categories, which is particularly challenging because the learner may not know in advance how many categories to group the objects into. Past research has primarily examined this question from a unsupervised learning perspective, presenting participants with a set of stimuli, but without any category labels, and examining the effect of categorization behaviour with different category structures (Pothos et al., 2011). The focus of this chapter is to explore behaviour on this task in a semi-supervised context where some labeled examples are provided in conjunction with many additional unlabeled examples, and examining how the addition of labeled examples can interact with category structure and influence the kinds of categories people form.

2. How does information about category structure and category labels shape what kind of structured category representations people learn? A second learning problem is how humans are able to learn rich, structured category representations. For example, DOGS and CATS can be nested under the larger category of ANIMALS as part of a taxonomy. Alternatively, a learner could organize them under a different set of categories such as PETS that cut across these taxonomic category representations. This problem of structure learning builds on from the first problem, as not only does the learner need to learn how to categorize a set of objects, but also has to infer the correct manner in which they should be organized into a structure. While past work has focused on learning structure from unlabeled examples (Anderson, 1991; Love, Medin, & Gureckis, 2004; Kemp & Tenenbaum, 2008), in this chapter I explore how both category structure and category labels play a role in learning structured representations. I explore whether people can learn structure with labeled examples, whether there are any differences between learning different kinds of structures, and how people generalize and transfer this knowledge to new systems of categories.

3. How does the structure of categories affect how people learn complex real-world categories? Experiments studying human category learning typically use artificial stimuli with a limited number of features. Many real-world categories consist of a large number of features, and one open question is whether the learning categories with few features is the same as learning

> categories with many features. In this project, I look at how category structure affects learning categories with a large number of features. In addition, I compare human performance across a variety of category learning tasks to an ideal observer model of category learning to explain when and how learning is possible for complex categories with many features.

Each of these three research questions examines a important problem in category learning, and are all connected by two aspects: how varying the kind of information (category structure of category labels) given to a learner shapes the categories that are learned, and being motivated by the kinds of computational problems that are relevant to learning real-world categories. Before going into the details of each of these questions, the remainder of this introduction chapter is aimed at providing the relevant background and context of the study of human category learning.

I begin with a broad overview of human category learning, focusing on how information from category structure and category labels have driven our understanding of human category learning. In the section on category structure, I review early category learning through the seminal work of Shepard et al. (1961). This is followed by a discussion of the work by Rosch (1973) on family resemblance category structures, and the development of prototype and exemplar models of categorization. I then describe mixture models of categorization, and show how prototype and exemplar models can be unified under this framework. In particular, I present Anderson's (1991) Rational Model of Categorization an one example of a mixture model, and how such a model allows for greater flexibility in category representation than prototype and exemplar models. In the final section, I explore some attempts to measure the complexity of learning different kinds of categories based on their category structure.

The second part of this introduction focuses on the role of category labels. I look at the various ways people receive information from category labels, and how these differences can affect the inferences people make about the kinds of categories they learn. I review research in the area of unsupervised learning, where learners are only provided feature information without any corresponding

| $F_1$ | $F_2$ | $F_3$ | $F_4$ | Category |
|---|---|---|---|---|
| 1 | 1 | 1 | 0 | A |
| 1 | 0 | 1 | 0 | A |
| 1 | 0 | 1 | 1 | A |
| 1 | 1 | 0 | 1 | A |
| 0 | 1 | 1 | 1 | A |
| 1 | 1 | 0 | 0 | B |
| 0 | 1 | 1 | 0 | B |
| 0 | 0 | 0 | 1 | B |
| 0 | 0 | 0 | 0 | B |

**Table 1.1:** The category structure used in Experiment 2 of Medin and Schaffer (1978). There were nine different stimuli, each with four binary features, which were organized in two different categories labeled A and B. The structure of these categories follow a family resemblance structure, where members of category A typically had more features that take on a value of 1, while members in category B were more likely have features that take on feature values of 0.

category label information. The corresponding goal of unsupervised learning is to discover how to organize objects into categories in the absence of any label information. I contrast unsupervised learning with semi-supervised learning, where the learner has access to some category label information rather than none, and how this additional information might influence how people form categories. Finally, I explore how providing multiple category labels for each object helps with learning rich, structured category representations.

## 1.2 How category structure influences category learning

Why is category structure critical to understanding category learning? Many of the seminal ideas in category learning can be traced back to the study of different kinds of category structure. This in turn has led to the flourishing of different computational models that capture various phenomena of human category learning. The benefit of studying different category structures provides the experimenter with a set of empirical patterns (such as the difficulty of learning different structures), which constrain the space of plausible psychological models that can give rise to the same patterns of behaviour. Thus, exploring categories that are richer in structure in ways that resemble real-world categories should allow us to develop better theoretical accounts of human category learning.

Before I go any further, what exactly does the term *category structure* mean? In brief, a category structure refers to the arrangement of features that map onto different categories. One example of a category structure is shown in Table 1.1, where there are two categories to be learned (A and B)., Each example from both categories consists of $D$ different features ($f_1, f_2, \ldots, f_D$), where each feature takes on binary values (either 0 or 1).* These features are mapped to different visual features of the stimuli stimuli which are displayed to participants.

### 1.2.1 Early category learning

Early theories of category learning viewed categories as deterministic. This *classical* view posited that category membership was an *all or none* phenomenon: either an object was a member of a category or it was not (Bruner et al., 1956). From this viewpoint, category learning was the result of discovering the correct rule, or learning the set of necessary and sufficient features to achieve perfect classification of the stimulus set. This early, but influential work promoted the use of the supervised classification paradigm for studying category learning and the use of different category structures to provide empirical constraints on theories of category learning. The supervised classification design is straightforward to run experimentally, and has been the de facto method of studying how people learn categories (Shepard et al., 1961; Medin & Schaffer, 1978; Minda & Smith, 2001). It has also served as a benchmark for computational models of category learning and machine learning classifiers, leading to the development of numerous models that can both learn to correctly classify existing stimuli, but also generalize sensibly to new, unseen data. Generalization is a particularly im-

*There are also many instances of category structures where features can take on more than two discrete feature values, or take on continuous values instead.

portant aspect of category learning, as testing people with stimuli they were not trained on provides a strong test of whether participants merely memorized the examples during training, or learned some abstract rules or patterns about the categories that allow them to generalize appropriately.

One of the most influential and well-studied set of category structures from this period was the work of Shepard et al. (1961). They studied learning across six different types of category structure of varying difficulty, each of which contained eight different stimuli consisting of three binary features each, split into two categories consisting of four stimuli each. The Type I structure only required attending to a single feature dimension to correctly classify all of the items, while the Type VI structure required attending to all three. The results from this study showed that there were differences in the time it took participants to learn each of these different category structures, and provided evidence that people were not merely memorizing the link between each stimulus and its label (which would result in no difference in performance across the six category structures), but rather attempted to form an abstract rule to categorize each stimulus on the basis of shared category structure. In addition, these differences across these six category structures have served as a benchmark for computational models of category learning, as any model attempting to explain human category learning would need to be able to capture the learning curves for these six category structures.

### 1.2.2 Prototypes and exemplars

Beginning in the early 1970s, researchers began to move away from using category representations structured by logical definitions, and moved towards category representations that more closely resembled the structure of natural categories. During this time, a number of studies looking at people's knowledge of natural categories established that people's representations of real-world categories were not an "all-or-none" phenomenon (Rosch, 1973; Rosch & Mervis, 1975a). Rather, they could be explained by positing that people's representations of categories were more like *prototypes* instead, where prototypes are an idealized representation of a particular category. This research showed that certain members of a category are more likely to be viewed as “typical” members of a category than others. For example, *apple* is considered to be a much better member of the category of *fruits* than a *tomato*. In particular, the most typical item in a category is known as the *prototype*. Additionally, Rosch and Mervis (1975a) showed that prototypes naturally form from a category structure known as a *family resemblance* structure. In this kind of structure, different objects share a varying number of overlapping features, with the most prototypical members of each category as the ones that most closely resemble the other members of the same category, and the least resemblance towards members of other categories. Family resemblance structures predict how people fare on generalization or test items. Test stimuli that closely resemble category prototypes are easier and quicker to classify, while items that are equally similar to both categories tend to be much more difficult (Medin & Schaffer, 1978).

How might these typicality and generalization effects be explained by a computational model? As mentioned above, family resemblance category structures naturally arise from categories where the most prototypical members share most features with members of the same category, and the least with members of other categories. Prototype models (Reed, 1972; Posner & Keele, 1968) and exemplar models (Medin & Schaffer, 1978; Nosofsky, 1986) offer two complementary possibilities. Prototype models represent each category $k$ with an idealized prototype $p_k$ that captures the category's central tendency, which is calculated by taking the mean along each feature dimension across all observed items in the category. Then, the classification of a new stimulus $x_{n+1}$ is performed by calculating the similarity $s(x_{n+1}, p_k)$ between each of the category prototypes and the new stimulus.[†] Categorization decisions are then made by looking at the relative probability that the new stimulus belongs to each of the categories, and a decision is made using the Luce choice rule (Reed, 1972; Luce, 1959).

[†] The shape of the similarity function often takes the form of a Gaussian distribution (Nosofsky, 1986) or a decaying exponential function centered around the prototype (Shepard et al., 1987).

In contrast to prototype models which assume a single idealized representation for each category, a different set of researchers proposed that people's knowledge of categories is instead stored in memory as exemplars (Medin & Schaffer, 1978; Nosofsky, 1986). Exemplar models work by taking the summed similarity across all of the exemplars for each category to the target example, and then make a decision by applying the Luce choice rule (Medin & Schaffer, 1978; Nosofsky, 1984, 1986; Kruschke, 1992). Additionally, exemplar models often include additional parameters that encode *attentional weights* for each stimulus dimension, allowing such models to favour attending to only a single dimension or multiple dimensions for a given stimulus.

### 1.2.3 Mixture models of categorization

Despite the apparent differences of prototype and exemplar models, work by Ashby and Alfonso-Reese (1995) showed that by viewing the problem of category learning from a *computational* standpoint, both prototype and exemplar models can be viewed as different approaches to *density estimation*.[‡] The goal of density estimation is to infer the underlying probability distribution from which examples are being drawn from, where learning a prototype model is an example of *parametric* density estimation (where the distribution is known but the parameters are learned), while exemplar models are a form of *non-parametric* density estimation (where the distribution does not take on a simple, parametric form).

This unification of prototype and exemplar models showed that they could be viewed as two ends of a continuum, rather than as distinct theories. On one end categories could be represented by a single prototype, while on the other categories could be represented by all of their exemplars, but in both cases they were an attempt to solve the computational problem of density estimation. Framing categorization as a problem of density estimation hinted that it should also be possible to

[‡] This idea of categorization as density estimation also applies to decision-bound models of categorization (Ashby, 1992), although I focus less on these models in this thesis.

come up with computational models with category representations that flexibly interpolated between prototype and exemplar representations. Having a category representation with a mixture of sub-prototypes created from a subset of exemplars could capture important sub-structure *within* categories. A number of researchers have proposed such models of category learning, which include the Rational Model of Categorization (Anderson, 1991), SUSTAIN (Love et al., 2004) and the Varying Abstraction Model (Vanpaemel & Storms, 2008). While there are a number of subtle differences between these different mixture models, for the remaining of this section I focus on the Rational Model of Categorization (Anderson, 1991), as it will be relevant to the research presented in Chapter 2.

The Rational Model of Categorization (RMC) can flexibly interpolate between prototype and exemplar representations, which allows it to capture relevant sub-structure that prototype and exemplar models cannot. The RMC achieves this by having a prior distribution over cluster assignments known as the Chinese Restaurant Process prior. It is calculated in the following manner:

$$p(z_i = k|\mathrm{z}_{\mathrm{i}-1}) = \begin{cases} \frac{M_k}{i-1+a} & \text{if } M_k > 0 \quad \text{i.e. } k \text{ is old} \\ \frac{a}{i-1+a} & \text{if } M_k = 0 \quad \text{i.e. } k \text{ is new} \end{cases}$$

where $M_k$ is the number of items in each cluster $k$, and $\alpha$ is the dispersion parameter. Thus, for each new example the model encounters, it can either place it into one the existing clusters with some probability proportional to the number of items in each cluster, or create a new cluster to place it into, with a straightforward application of the Luce choice rule. The dispersion parameter $a$ controls the extent to which it favours placing items into new clusters, with lower values of $a$ favouring fewer clusters and higher values of $a$ favouring more clusters. From this perspective, prototype models can be viewed as placing all of the items of a category into a single cluster, while exemplar models can be viewed as placing each item into its own separate cluster.

One advantage of the Rational Model of Categorization is that it can flexibly grow as it observes more data, by creating new clusters to explain patterns in the data, assuming that having an additional cluster provides a better explanation for the observed data than not. As prototype and exemplar representations are limited to either a single category representation or storing every example, they do not have the representational power to flexibly trade-off between complexity and fit like the RMC. A second advantage of the Rational Model of Categorization is that it can perform both supervised learning and unsupervised learning, and discover how to organize items into different clusters through feature information alone. The ability to discover and represent structure in an unsupervised manner is another important aspect of human categorization (Love et al., 2004; Kemp & Tenenbaum, 2008; Shafto, Kemp, Mansinghka, & Tenenbaum, 2011), and will be discussed in more detail in Chapter 2. Finally, recent work has established a number of connections between the RMC and the Dirichlet Process Mixture Model used in statistics and machine learning (Griffiths, Canini, et al., 2007; Sanborn et al., 2010). Highlighting this connection showed that order effects in category learning could be captured by approximate inference techniques applied to this model, that other computational models could not.

### 1.2.4 Complexity and learnability of structure

So far, I have focused on how computational models of category learning have advanced to allow them to capture and represent the structure of categories in an increasingly detailed manner. This final section explores a related, but different question related to determining the subjective difficulty of learning different category structures. Is there an easy method to determine the subjective complexity for learning a particular kind of category structure? The six category structures from Shepard et al. (1961) showed that some types were easier to learn than others, but lacked a theoretical explanation for why this pattern of learning was observed. By examining the space of *boolean* concepts, J. Feldman (2000) showed that the subjective difficulty of learning different concepts could be predicted by the boolean complexity of the concept, i.e. the length of the shortest Boolean formula that is logically equivalent to the concept. However, while this measure of complexity predicts the difficulty of learning particular categories, it is not always the case that participants are using the same classification rule as predicted by these measures, suggesting that people often have difficulty discovering the simplest way to represent certain categories (Lafond, Lacouture, & Mineau, 2007). More recent work has attempted to generalize these findings, expanding the scope of measuring complexity from just boolean concepts to richer concepts and categories that can be expressed in a propositional or predicate logic (Kemp, 2012; Piantadosi, Tenenbaum, & Goodman, 2016).

While the subjective complexity of different kinds of category structures has been well-studied, a related, but slightly different measure of complexity is defined in terms of the number of features or the *dimensionality* of the categories, which has received less attention. Work by Rosch and Mervis (1975a) in feature-naming studies have shown that people's representation of real-world categories is sufficiently rich to list many different features from everyday categories. In comparison, the number of features used in artificial stimuli for category learning experiments is small relative to real-world categories, suggesting that some existing theories about how humans learn categories with few features may not adequately explain how humans learn categories with many features, like the kinds encountered in the real world. The few studies that have looked into how the number of features affects learning have provided inconsistent accounts about the effect of dimensionality. Some studies have shown that additional features is helpful for learning categories (Hoffman & Murphy, 2006; Minda & Smith, 2001), particularly when given additional prior information that highlighted that features were related (Hoffman, Harris, & Murphy, 2008). On the other hand, other research has argued that additional features are harmful for category learning, as it increases the number of features to search through (Edgell et al., 1996), or that participants focus on finding classification rules that rely on a single dimension (Ashby, Queller, & Berretty, 1999; Nosofsky, Gluck, Palmeri, McKinley, & Glauthier, 1994). I explore this question of category learning with a large number of features in more detail in Chapter 4 of this thesis.

The role of category structure has been integral to the study of human category learning. Through the development of standardized category structures, researchers have been able to examine effects such as the difficulty of learning different kinds of structures and to develop richer computational models of category learning that capture various facets of how people learn categories (Nosofsky, 1986; Kruschke, 1992; Love et al., 2004; Sanborn et al., 2010). While the supervised learning paradigm has been successful in examining the influence of category structure on learning, it is limited in its ability to manipulate the kind of labeled input given to learners. And yet, the way in which we acquire knowledge about categories is dependent not only on the structure of categories, but also the kinds of category labels too. In the next section of the introduction, I look at other kinds of computational problems related to the influence of category labels.

## 1.3 How category labels influence category learning

In the real world, there is considerable variation in how often people receive information from category labels, and the manner in which these labels are provided. In this section, I highlight how different ways in which learners receive category label information can alter the kinds of computational problems that need be solved. In particular, I look at three kinds of learning that are driven by differences in labeling: unsupervised learning, semi-supervised learning and multi-label category learning.

### 1.3.1 Unsupervised learning

In a standard supervised learning experiment, participants are presented with either category labels (during each trial) or category feedback (after each trial) to help them learn the categories. Despite the dominance of this experimental procedure, a number of researchers have criticized whether the supervised learning paradigm of learning to classify objects into two different categories mimics the equivalent real world process (Murphy, 2004; Pothos & Close, 2008). To counter this response, a number of other category learning paradigms have been used to examine learning in other kinds of ways. *Unsupervised learning* refers to a different kind of learning where no category labels nor feedback is provided to the learner. In the absence of any feedback, unsupervised learning occurs by spontaneously figuring out how to organize objects into categories on the basis of feature information alone. Common techniques in unsupervised learning include dimensionality reduction and clustering. Despite less attention for unsupervised learning, there are many instances of real-world learning where categories are learned in an unsupervised fashion. For example, the problem of phoneme learning requires children to figure out the boundaries for the different phoneme categories of their native language from hearing and listening to speech sounds from other people (N. Feldman, Griffiths, & Morgan, 2009; Lake, Vallabha, & McClelland, 2009). Since category labels are not provided for speech sounds, this is a challenging unsupervised category learning problem.

Is there experimental evidence in whether or not people can learn categories in an unsupervised manner? There have been two main experimental approaches to study unsupervised category learning. The first of these methods has been to adopt the supervised learning experiment paradigm and examine whether or not participants can correctly classify objects into the right categories (Love, 2002; Gureckis & Love, 2003). The major difference in the unsupervised version of these category learning tasks is that no feedback or category labels are provided, so participants must rely on information from stimulus features and the underlying category structure to solve this task. Because of the added difficulty, the kinds of categories people are taught in unsupervised category learning tasks are designed so that attending to the distributional information alone is sufficient to correctly classify them. One study by Love (2002) showed that participants were able to correctly learn to classify one-dimensional categories without any supervision at a level similar to participants in a supervised condition. However, the results also found that performance was worse when participants were prompted to form explicit rules compared to implicitly learning the categories, suggesting that unsupervised learning occurs passively. However, the tasks used in this particular study consisted of categories which only varied along one feature dimension, making the task comparatively easy despite the lack of feedback. Other work in unsupervised category learning has found that learning complex rules involving multiple features are much more difficult, with people sub-optimally preferring to categorize along a single feature dimension (Ashby et al., 1999).

One drawback of these unsupervised learning experiments is that they assume there exists a single, correct organization of objects into categories. However, as unsupervised learning involves no category labels there is no true organization, and there could be many different ways in which objects could be organized into sensible categories. The second approach to studying unsupervised learning is motivated by this intuition, conducting experiments where participants are asked to sort objects into different categories rather than classify them (Medin, Wattenmaker, & Hampson, 1987; Milton & Wills, 2004; Pothos & Chater, 2002; Pothos et al., 2011). Many of these unsupervised categorization experiments have examined people's sorting behaviour by asking participants to sort objects into a fixed number of categories (Medin, Wattenmaker, & Hampson, 1987; Regehr & Brooks, 1995; Milton & Wills, 2004). Similar to results found in unsupervised category learning studies, research in unsupervised categorization has found a strong preference for sorting on the basis of a single feature (Medin, Wattenmaker, & Hampson, 1987; Ahn & Medin, 1992). However, by tweaking the experimental design, researchers have been able to influence people to sort along multiple dimensions. Such methods include providing a causal connection between two different features (Medin, Wattenmaker, & Hampson, 1987), or using a category structure where it is equally intuitive to sort along a single dimension or multiple dimensions (Pothos & Close, 2007). More recently, researchers have begun to explore sorting behaviour in "free-sorting" paradigms, where participants are allowed to use as many categories as they wish (Pothos & Close, 2007; Pothos et al., 2011). One advantage of this experimental paradigm is that the learner needs to determine the number of categories to sort objects into, which more closely resembles one of the challenges in real-world category learning than sorting into a fixed number of categories.

What kinds of computational models can explain behaviour in unsupervised learning experiments? One approach has been to adapt existing supervised learning models to categorize on the basis of similarity alone (without category label information), such as the Unsupervised Generalized Context Model (Pothos & Bailey, 2009). A second approach involves using existing category learning models such as SUSTAIN (Love et al., 2004) and the Rational Model of Categorization (Anderson, 1991) that can perform both supervised and unsupervised learning out-of-the-box. By having a single model that can perform both kinds of learning, it suggests that both unsupervised and supervised learning should be viewed or subsumed under a single kind of learning process, rather than treating them as two distinct kinds of learning.

The evidence that people can learn in an unsupervised fashion suggests that category labels are not always required for learning, and these tasks also capture an essential aspect to real-world category learning. This is true for both understanding how people perform unsupervised classification and unsupervised sorting tasks.

### 1.3.2 Semi-supervised learning

While some kinds of categories are learned in a completely unsupervised manner such as phonemes, in many instances people receive occasional feedback or category label information, in a manner that cannot be described as completely unsupervised or completely supervised. This kind of learning is known as *semi-supervised learning*. Arguably, semi-supervised learning more closely resembles the kind of labeled input that people in the real-world receive when learning categories. Therefore, having a better understanding of how people perform semi-supervised learning should provide a more comprehensive understanding of real-world category learning.

Much of the research into how humans perform semi-supervised learning have employed classification tasks similar to the ones used to study supervised learning (Zhu, Rogers, Qian, & Kalish, 2007; Lake & McClelland, 2011; Kalish, Rogers, Lang, & Zhu, 2011; Gibson, Rogers, & Zhu, 2013). In these semi-supervised category learning tasks, participants are provided with labeled examples (like in supervised learning tasks), but also additional unlabeled examples (like unsupervised learning). One focus of human semi-supervised learning has been to establish whether varying additional unlabeled examples can influence people's categorization behaviour. Research has shown that additional unlabeled examples can lead to a shift in the classification boundary for simple one-dimensional categories, depending on how the distribution of unlabeled examples differs from the distribution of labeled examples (Zhu et al., 2007; Lake & McClelland, 2011; Kalish et al., 2011). However, similar to results in unsupervised learning, category learning with more complex two-dimensional classification rules often leads people to ignore the additional structure information from unlabeled examples and categorize on the basis of the labeled examples only (Vandist, De Schryver, & Rosseel, 2009; McDonnell, Jew, & Gureckis, 2012).

The existing approach to semi-supervised learning has framed the problem as first having labeled examples, and examining the effect of additional unlabeled examples. Yet, the opposite framing of semi-supervised learning is equally valid, assuming the learner has access to unlabeled examples, and examining the effect of additional labeled examples. One possible reason why category labels from labeled examples might be special is that labels are generated by humans to communicate information about the structure of the world, and taking into account the social context in which label information is provided is especially informative compared to feature information (Xu, 2002). Research from developmental psychology has shown that infants are more sensitive to information conveyed through category labels than other means such as random noises (Waxman & Markow, 1995), and that labels and language modulates perception and cognition at a deep level (Lupyan, 2012). However, the benefits of category labels still hotly disputed, as other researchers have argued that these results can be explained without privileging labeled examples and information, and can be explained by associative or attentional-driven accounts of learning instead (Colunga & Smith, 2005).

### 1.3.3 Multi-label category learning

In what other ways might people receive useful information about category labels in the real world? So far, I have covered how either no category labels (unsupervised learning) or some category labels (semi-supervised learning) can influence how people learn categories. A third possibility in which category labels are important is when people receive multiple, different category labels for the same object. For example, in the opening example I alluded to how objects be labeled as “dogs”, but also as “animals” or “pets”. How should a learner integrate additional category labels to an object which they already have an existing category label for, and how does the presence of multiple labels change people’s inferences about how categories should be represented in our minds?

One difficulty about incorporating new labeled information is that humans exhibit a bias for mutual exclusivity, preferring that novel labels be applied to novel objects (Markman & Wachtel, 1988), rather than applying multiple labels to the same object. Is there any evidence that people can overcome this bias and learn multiple labels for the same object? One area that has studied this particular learning problem is research in *cross-situational word learning*. Unlike category learning experiments where the goal is to learn how to classify objects into categories, cross-situational word learning tasks focus on learning the mapping between words and objects (Yu & Smith, 2007; Kachergis & Yu, 2017). On each trial, participants are shown different objects on the screen, along with different labels. Initially, the mapping between words and objects will appear to be random. However, as participants observe more trials, certain mappings between labels and objects will begin to appear much more plausible while others can be ruled out. While most of the experiments have focused on learning mutually exclusive one-to-one mappings (Yu & Smith, 2007), a number of studies have shown evidence that people can learn many-to-one mappings (where each object can have multiple labels), such as objects organized in a taxonomic hierarchy (Yurovsky & Yu, 2008; Gangwani, Kachergis, & Yu, 2010). On the other hand, Ichinco, Frank, and Saxe (2009) found that once participants had learned one set of mappings from category labels to the objects, they were unable to learn any additional mappings, arguing that participants were unable to overcome the mutual exclusivity bias due to blocking from the existing learned category label. One additional difficulty with overcoming this bias is that having an assumption of mutual exclusivity can often lead to faster more efficient learning (Hidaka, Torii, & Kachergis, 2017).

While the study of the mutual exclusivity bias has received considerable attention in cross-situational word learning studies, it is a problem that has been given less consideration among category learning researchers, which is surprising given that both of these experimental paradigms have a shared interest in the same kinds of phenomena involved in learning. However, this problem is interesting from a category learning perspective, as not only do people need to learn to map different labels onto the same object, but also learn that these labels will generalize differently to other objects. Is there any evidence that people can learn categories that override the mutual exclusivity assumption? In one study, Canini (2011) showed that participants were able to learn and reproduce a taxonomy composed of multiple labels for each object. Participants were taught different labels for objects organized into a taxonomy. During the test phase, they were then directly asked to reconstruct the taxonomy based on relations between each of the labels learned. Their results showed that people were able to reconstruct the hierarchy across a variety of different circumstances. However, one limitation of this work is that participants were directly told during the test phase to reconstruct a taxonomic-like structure, thus informing participants the kind of structure to recreate. On the other hand, in the real world children must somehow discover the correct structure themselves through label and feature information. In Chapter 3, I explore this question of learning multiple category labels and their role in learning conceptual structure.

## 1.4 Summary

In the following chapters of this thesis, I present three new studies that cover how information from both category structure and category labels influence the kinds of categories people can learn, across a variety of category learning tasks. Each chapter examines a different computational problem that is motivated by the kinds of structure and label information that learners might receive in the real world.

Across the three chapters, I investigate category learning through a combination of behavioural experiments and computational models. While the use of experiments alone can provide insights into human behaviour and learning, combining human behavioural data in conjunction with computational models offers a number of additional benefits which I briefly expand upon. First, computational models are built to help provide an explanation for *why* people are behaving the way they do, with different kinds of computational models providing explanations at different levels of understanding. These different levels can be categorized through Marr's (1982) three *levels of analysis*. At the *computational* level, computational models aim to explain behaviour in terms of the underlying computational problem the mind needs to solve a given task, with respect to the environmental and computational constraints. One example of a computational level category learning model is the Rational Model of Categorization (Anderson, 1991), which seeks to explain how categorization arises from rationally organizing a set of objects into clusters (Sanborn et al., 2010). The *algorithmic* level offers explanations of behaviour in terms of the cognitive processes involved such as similarity and choice. Both prototype and exemplar models are examples of computational models at this level. Finally, the *implementation* level seeks explanations in terms of how these computational models are physically instantiated, whether that be at the neural level or *in silico*.

In this thesis, my primary focus is the use of models at the *computational* level, by looking at which kinds of computational problems people are trying to solve when learning categories that are inspired by different learning scenarios in the real-world. A second benefit of computational models is that they force researchers to be explicit about the prior assumptions they make, in contrast to vaguely specified verbal theories which can often omit essential details that make them difficult to implement or cannot actually learn the task at hand. Having concrete and well-specified models allows researchers to make novel predictions, and then compare different kinds of computational models to determine which ones best explain and capture human behaviour.

In Chapter 2, I investigate the problem of semi-supervised categorization. How do people sort objects into categories when provided with a combination of both labeled and unlabeled examples? Past research in this area has focused primarily on unsupervised categorization where a learner is provided with only unlabeled examples. The addition of labeled examples turns this learning problem into one that more closely resembles the real-world problem a learner needs to solve. What effect do labeled examples have on people's ability to sort objects into different categories? Our results show that when the underlying category structure is sufficiently distinct, the addition of category labels to some examples has little to no effect on how people categorize. On the other hand, when the category structure is ambiguous, our results indicate that category labels drive sorting behaviour, and that this is dependent on which category labels people observe. To capture the behaviour in these semi-supervised categorization tasks, I extend a well-known model of category learning – the Rational Model of Categorization – that uses both information about category structure and category labels to determine how to cluster objects into different categories.

In Chapter 3, I look at how people learn different kinds of conceptual structure through information from both feature and label information. Past research has focused on either only learning multiple category labels through cross-situational word learning (Yurovsky & Yu, 2008; Gangwani et al., 2010; Ichinco et al., 2009) or making strong assumptions about the structure to be learned (Canini & Griffiths, 2011). Across two experiments, I explore how people learn either taxonomic or cross-cutting structures by presenting different feature and label information. I examine whether people can learn these structures, whether there exist any differences between learning different conceptual structures and how people generalize and transfer knowledge of conceptual structure.

In Chapter 4, I focus on how it is possible people can learn complex categories and overcome the curse of dimensionality. The curse of dimensionality is a problem that suggests learning in high-dimensional spaces or with a large number of features should be difficult, as the number of possible stimuli configurations blows up exponentially. This is relevant to understanding one problem of real-world categories, as such categories are rich with many different possible features human learners could attend to. And yet, humans are able learn these kinds of real-world categories, so it suggests that somehow we are able to overcome the curse of dimensionality. Across two experiments, I show how category structure plays a large role in determining what kinds of categories are learnable or not in high-dimensional spaces. Results show that family resemblance categories where each feature is predictive of membership, but noisily so, are learnable even with a large number of features. On the other hand, categories that are more rule-based and rely on a single feature to correctly classify objects become increasingly more difficult to learn with more and more features. Using an ideal observer model for this task, I examine when human performance matches this computational model and when it does not, showing that while it falls short of the ideal learner in many instances, it can still perform well in a limited number of circumstances.

Finally in Chapter 5, I conclude by summarizing my research findings within the broader context of category and concept learning, and how the ideas from my thesis may yield potentially interesting future research directions that help bridge the gap towards understanding real-world category learning.

# Statement of Authorship

TITLE OF PAPER: The helpfulness of labels in semi-supervised learning depends on category structure.

PUBLICATION STATUS: Published

PUBLICATION DETAILS: Vong, W.K, Perfors, A. & Navarro, D. J. (2016). The helpfulness of labels in semi-supervised learning depends on category structure. Psychonomic Bulletin & Review. 23(1), 230-238.

PRINCIPAL AUTHOR

NAME OF PRINCIPAL AUTHOR (CANDIDATE): Wai Keen Vong

CONTRIBUTION TO THE PAPER: Designed and ran experiments, performed data analysis, implemented computational model, wrote manuscript and acted as corresponding author.

OVERALL PERCENTAGE (%): 80%

SIGNATURE:

DATE: 15/1/2018

## CO-AUTHOR CONTRIBUTIONS

By signing the Statement of Authorship, each author certifies that:

1. the candidate's stated contribution to the publication is accurate (as detailed above);
2. permission is granted for the candidate in include the publication in the thesis; and
3. the sum of all co-author contributions is equal to 100% less the candidate's stated contribution.

NAME OF CO-AUTHOR: Amy Perfors

CONTRIBUTION TO THE PAPER: Supervised development of work, helped with experimental design, model development and editing of the manuscript

SIGNATURE:

DATE: 24/1/18

NAME OF CO-AUTHOR: Dani Navarro

CONTRIBUTION TO THE PAPER: Supervised development of work, helped with experimental design, model development and editing of the manuscript

SIGNATURE:

DATE: 24/1/18

# 2

# The helpfulness of category labels in semi-supervised learning depends on category structure

## 2.1 Introduction

Imagine you are walking through an art gallery with an artist friend. As you proceed, your friend occasionally stops to point out a particular painting and tell you the name of the artist who painted it, thereby providing you with labeled data for that painting. All around you are dozens of other paintings in various styles. Although your friend has not commented on these paintings, this unlabeled data might be very informative to you: it can refine your understanding of particular painters or what art styles exist. You might even detect groups of similar paintings that you suspect were painted by the same person, even though your friend has not commented on them.

Learning from both labeled and unlabeled data, as in this example, is referred to as semi-supervised learning (Zhu et al., 2007; Gibson et al., 2013) and has not been studied to the same extent as other categorization problems. Most research investigates supervised learning, where each example is paired with the appropriate category label (Medin & Schaffer, 1978; Nosofsky, 1986). Other research explores unsupervised learning, where people must learn categories based without any category labels or feedback (Love, 2002; Pothos & Chater, 2002; Pothos et al., 2011). Despite the attention focused on supervised and unsupervised learning, in real life the majority of situations involve mostly semi-supervised learning: a few labeled instances in conjunction with a large set of unlabeled experiences.

What do we know about human semi-supervised learning? Unfortunately, the literature is somewhat split about its effectiveness. One possibility is that receiving both unlabeled and labeled examples provides very little information over either source alone. Consistent with this, some studies have found that adding unlabeled data has no effect when labeled examples have already been provided (Vandist et al., 2009; McDonnell et al., 2012). Similarly, others have found that people are able to learn the structure of categories in an unsupervised manner, and only labels to map words onto existing category representations (Bloom, 2000a). Both of these areas of research suggest that semi-supervised learning is not very different from either supervised or unsupervised learning. However, there is evidence for the other possibility too: some studies have found that adding unlabeled data can affect category learning in both humans (Zhu et al., 2007; Lake & McClelland, 2011; Kalish et al., 2011; Gibson et al., 2013) and computers (Chapelle, Schölkopf, & Zien, 2006).

How can we reconcile these apparently contradictory findings? We begin by noting that the typical framing of semi-supervised learning tasks is somewhat puzzling. Although semi-supervised learning extends both supervised and unsupervised learning, papers on the topic almost invariably compare it to supervised learning. By adopting this perspective, researchers are led to ask whether unlabeled data provides any additional benefit to the learner over and above what can be learned from labeled data. This framing is implicit in the way our art gallery example was described: it was simply assumed that the labeled examples from the artist friend would be useful, and the question was whether the unlabeled paintings might be an additional source of information.

Yet this framing is easily reversed. Consider instead the following variant: As you walk around the art gallery, you see hundreds of examples of paintings. From this wealth of data you might form theories about art styles, pick out individual paintings, and so on. As you do so, your friend points to a few paintings and tells you that those are by Picasso and Monet. So far, the literature on semi-supervised learning has typically assumed that the labeled examples are distributed in a similar fashion as the unlabeled examples. However, this is not true in our art gallery example nor (often) in real life: there may have been many other painters such as Magritte, Pollock and Rembrandt which you only saw unlabeled examples of. More generally, the distribution from which the world generates raw (unlabeled) data need not be at all similar to the one from which a knowledgeable teacher chooses to select (labeled) examples, and a child learning language cannot assume that people are labeling all and only the relevant categories of objects. Indeed, what is relevant changes from context to context, and what is labeled is conditioned on many things (attention, conversational goal) other than providing the optimal category learning information. Thus by framing the problem of semi-supervised learning as one in which unlabeled data as the primary source of information, the focus now shifts to the evidentiary value of the additional labeled data.

Now the relevant question is: When and how might labeled examples be beneficial for category learning above and beyond having only unlabeled examples? One method to assess this would be to compare semi-supervised learning to performance in purely unsupervised learning. Traditionally, the problem of grouping objects into categories has been explored primarily from an unsupervised perspective (Medin, Wattenmaker, & Michalski, 1987; Pothos & Close, 2008; Pothos et al., 2011). However, one of the challenges of unsupervised categorization is the sheer combinatoric explosion of possible ways to sort a group of objects into categories. The number of ways to sort *n* items is given by the *n*th Bell number, which grows very rapidly as a function of *n*: even having only ten stimuli can result in over 100,000 possible different classifications (Medin & Ross, 1997). While one could argue that most of these classifications would be implausible *a priori*, it is also possible that a sufficiently large number would still be plausible, and would still make this a computationally demanding problem to be solved. Despite this search challenge, one can easily imagine circumstances where labeled instances need not be necessary. For instance, people might not need any labels to determine that a Picasso painting was not created by the same artist who created a Monet – the styles are so different that it is obvious just from the unlabeled data that there were two separate categories of artists. In such a situation, semi-supervised learning might not be noticeably different from unsupervised learning.

On the other hand, distinguishing the work of Klee from that of Kandinsky represents a much harder problem for the novice. In fact, when the training items are similar, unsupervised categorization is hard and people have difficulty in determining how many categories to sort objects into and to do so in a consistent manner (Pothos et al., 2011). We hypothesize that it is in precisely these kinds of relatively ambiguous situations where some additional labeled examples may be beneficial, where even just a few labeled examples can substantially reduce the difficult of this huge search problem.

But how would *just a few* labeled examples help so dramatically? One possibility is that labeled examples might serve as a cue to people about what dimensions to attend to. For example, if the labels suggest that there are multiple relevant dimensions, the presence of the labeled data may prompt people to switch from a unidimensional classification strategy to a multidimensional one. While people tend to exhibit a strong unidimensional bias in unsupervised learning (Ashby et al., 1999; Medin, Wattenmaker, & Michalski, 1987), some recent work has shown that the presence of a sufficient number of labeled examples can cause people to shift towards multi-dimensional classification strategies (Vandist et al., 2009; McDonnell et al., 2012). However, this is not the only possibility as to how labeled examples might drive categorization. A different set of labeled examples might guide the learner into classifying using only a single dimension instead. Thus, we also hypothesize that labeled examples serve as cues as to which classification strategies to pursue, and that different sets of labeled examples should lead to different classifications.

This paper investigates how and when a small number of labeled examples improves category learning outcomes based on unsupervised data. We test these predictions through an experiment in which people sort unlabeled multidimensional rectangle stimuli into categories. In some conditions, the true category structure is distinct, while in others it is ambiguous. Conditions also differ by whether the labels people are given pick out distinct categories or not. Consistent with our hypotheses, we find (a) that people rely on labels when the underlying category structure is ambiguous, and (b) in that case, people's classification strategies are affected by the labeled examples they receive. In addition, we develop a modified version of the Rational Model of Categorization (Anderson, 1991) and show that it naturally captures people's behaviour in this novel semi-supervised paradigm.

## 2.2 Method

Our experiment took the form of an unconstrained free-sorting task in a semi-supervised setting. Participants were shown 16 two-dimensional stimuli, a maximum of three of which were labeled (depending on the condition). They were asked to sort the objects into different categories any way they wished. Different conditions manipulated both the kinds of structures people saw as well as the labels associated with the stimuli. The goal of the experiment was to examine which (if any) of these settings promoted semi-supervised learning.

### 2.2.1 Participants

Data were analyzed from 504 participants (312 males) recruited from Amazon Mechanical Turk and paid either US$0.30 or US$0.50. An additional 34 did not complete the experiment and 52 were excluded for failing to properly respond to a check question (see below). The age of participants ranged from 18 to 69 (mean: 33.3) 56% of participants were from the USA, 39% were from India, and the remainder were from other countries.

### 2.2.2 Stimuli

The stimuli used in the experiment, shown in Figure 2.1, consisted of white rectangles with a black border. Inside each of the white rectangles was an inner gray rectangle along to the bottom-right corner. The stimuli varied along two continuous dimensions* corresponding to the height of the white rectangle (25 to 65 pixels high) and the length of the inner gray rectangle (10 to 50 pixels long). There were two different stimulus sets, one for each of the two stimulus structure conditions described below. Depending on the label condition, three of the stimuli might have been labeled with a nonsense word (*dax*, *fep* or *wug*), which appeared underneath the associated stimuli. A total of 16 different stimuli were used, all presented simultaneously on the screen.

### 2.2.3 Design

Participants were randomly assigned to conditions based on two between-subject manipulations. In the first, we varied the coherence of the underlying category structure. In the DISTINCT STRUCTURE condition, the stimuli consisted of three well-separated clusters that varied along both stimulus dimensions, as shown in the top row of Figure 2.2. The AMBIGUOUS STRUCTURE condition also consisted of three equally sized clusters, but they were much closer together in the stimulus space,

*The extent of the variation along the two stimulus dimensions was calibrated by applying multidimensional scaling to a set of pairwise similarity ratings to ensure both dimensions were equally salient.

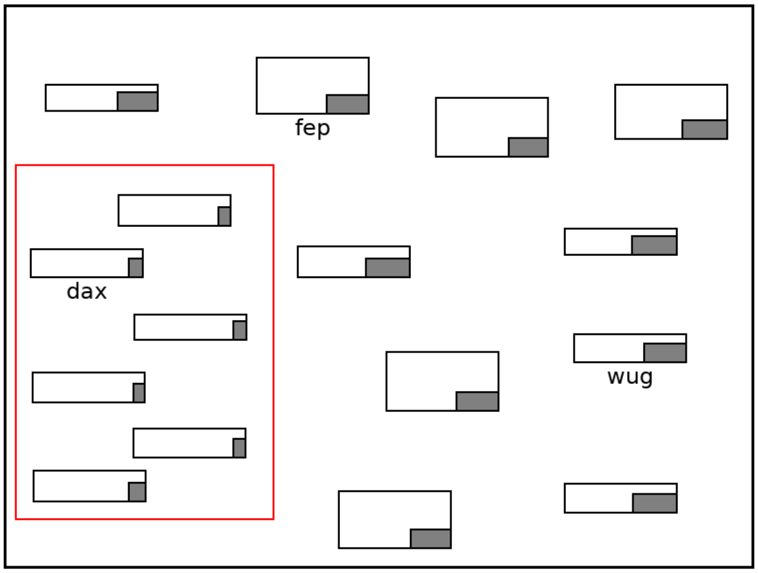


**Figure 2.1:** Screenshot from the task illustrating the stimuli and labels used in the experiment (this example is from the *Distinct Structure* and *Distinct Labels* condition). In all conditions, people were asked to sort the rectangles into categories by dragging them around the screen into clusters. In this screenshot the participant has already drawn one box around one of the categories they identified.

as in the bottom row of Figure 2.2. This made it difficult to distinguish the cluster boundaries from feature information alone. In all conditions the participants were not told how many categories there were: they were instructed to sort the stimuli into as many categories as they felt was necessary.

The second experimental manipulation, shown in the columns of Figure 2.2, varied the informativeness of the labels that were included. As a baseline, the NO LABEL condition was fully unsupervised with no labels at all. In the DISTINCT LABEL condition, people saw a helpful and informative set of labels: one labeled example from each of the three clusters. By contrast, in the AMBIGUOUS LABEL condition people saw potentially misleading labels: one came from the first cluster, two came from the second cluster, and none came from the third cluster. Of interest is how people's categorizations were affected by the informativeness of the label in combination with the structural coher-

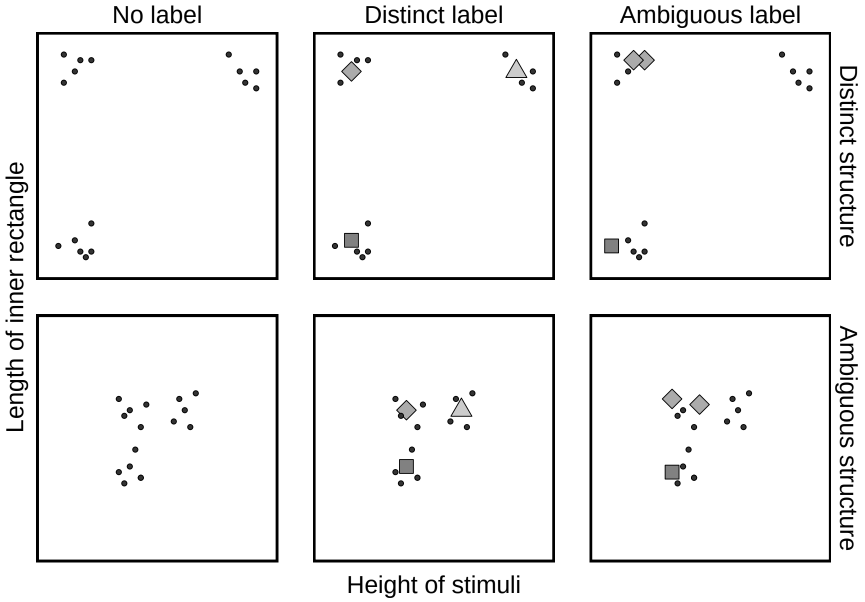


**Figure 2.2:** A visualization of the experimental design. The stimuli varied along two continuous dimensions (stimuli height and the length of the inner rectangle). The small black dots represent the unlabeled examples, while the larger stimuli represent the labeled examples, with each shape corresponding to a different category label (*dax*, *wug* or *fep*).

ence in the unsupervised data.

### 2.2.4 Procedure

The experiment was run online through a custom website. The cover story informed participants that archaeologists had discovered a number of unknown objects on a recent expedition, and needed help to sort them into different categories. In the labeled conditions participants were told that the archaeologists had discovered some of the names of the objects, which could be used as a guide on how they sorted the stimuli. In the no label condition the instructions simply recommended using the appearance of the objects to guide the sorting. In all conditions, no indication was given of how many different categories were present in the data.

In order to make sure that people understood the sorting task, before beginning the main task the participants completed a demonstration trial. This trial contained three squares and three triangles of different sizes, where they were asked to sort the shapes into separate piles that they thought should naturally go together. The position of objects in both the demonstration trial and main task were arranged to be non-overlapping and randomly ordered for each participant. The user interface required participants to first click and drag on stimuli until they were sorted into piles they thought should belong together, and then to draw boxes around each pile. If people were unhappy with the boxes they could revert to the click and drag stage until they were satisfied. If anyone failed to group any stimuli inside a box or assigned any stimuli to multiple boxes, a warning would appear and they could not submit their response. The demonstration trial also served as an exclusion criterion: 52 people failed to sort those stimuli in a sensible way (i.e., not by size or shape) and their data from the main experiment were therefore excluded from further analysis.

## 2.3 Results

Participants produced 216 unique sorts out of 504 solutions analyzed. This level of variability is commensurate with similar tasks in unsupervised categorization (e.g. Pothos et al., 2011). However, the extent of the variability was very different across conditions. To quantify this variability we use the adjusted Rand index (*adjR*), which measures the similarity between two classifications (Hubert & Arabie, 1985). It has a maximum of one when both classifications are the same, and drops to zero if the agreement between them is no more than would be expected by chance. The average *adjR* score among all pairs of participants in each condition is shown in Figure 2.3, and reveals two key findings.

The first finding was that people did indeed appear to find the ambiguous structure more ambiguous: responses in the DISTINCT STRUCTURE condition were more similar to one another

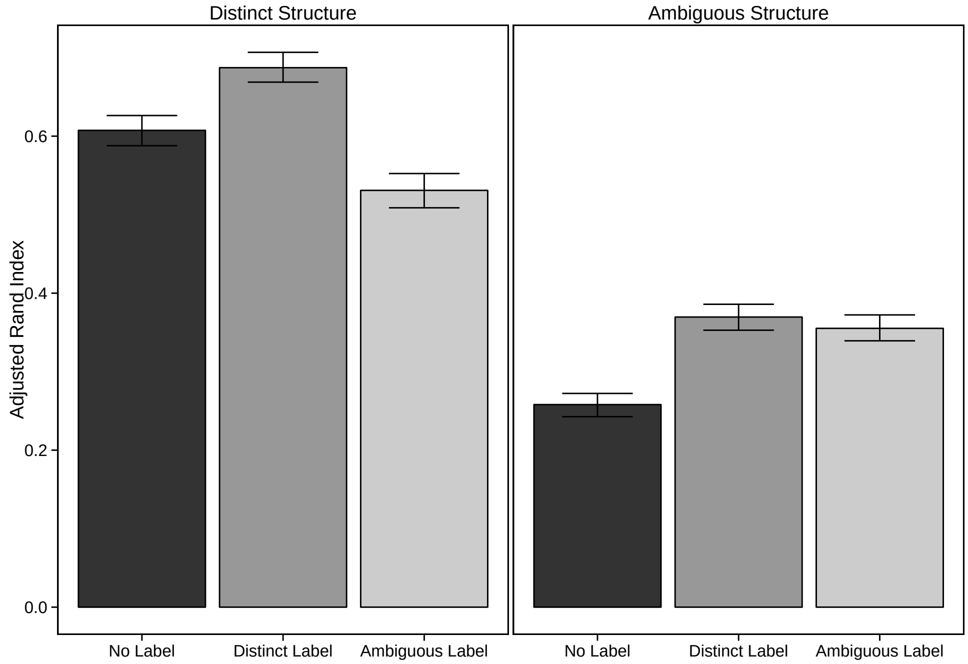


**Figure 2.3:** Agreement between participants within condition. Each bar plots the average similarity between solutions (i.e., adjusted Rand index) taken across all subjects in the same condition. Error bars are bootstrapped 95% confidence intervals.

than those in the AMBIGUOUS STRUCTURE condition. Consistent with this, a two-way ANOVA on structure × label revealed a significant main effect of structure ($F(1, 498) = 293.5, p < 0.001$).

The second finding, of more importance, is that the effect of labels was different in different contexts: while there was a significant main effect of label ($F(2, 498) = 14.2, p < 0.001$), there was also a significant interaction between the structure condition and label condition ($F(2, 498) = 10.9, p < 0.001$). In the AMBIGUOUS STRUCTURE condition, adding labels increased the degree of agreement among participants regardless of which label set was provided. However, in the DISTINCT STRUCTURE condition, the effect was more subtle. When the DISTINCT LABELS were provided, the labeled data were consistent with the structure of the unlabeled data, and the agreement among participants increased relative to the NO LABEL condition. But when the AMBIGUOUS LABELS were

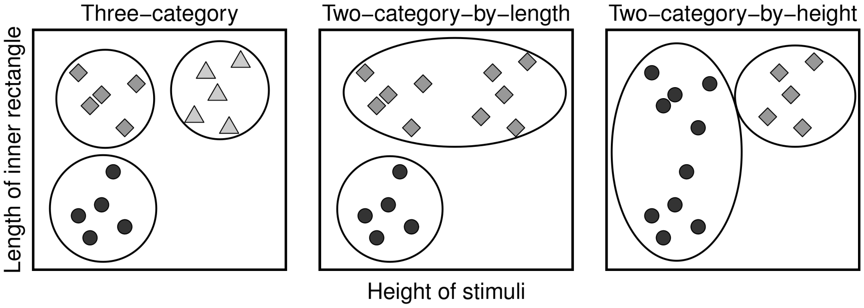


**Figure 2.4:** The three canonical classifications used to classify people's responses in the task. While this figure only depicts the canonical classifications for the *Ambiguous Structure* condition, the strategies are analogous for the *Distinct Structure* condition. The three-category strategy required attending to both stimulus dimensions when sorting. On the other hand, the two-category-by-length and two-category-by-height strategies only required attending to a single stimulus dimension corresponding to either the length of the inner rectangle or the height of the stimuli respectively.

provided, the structure among the labeled examples did not precisely match the structure of the unlabeled data. As a result, the agreement among participants dropped relative to the NO LABEL condition.

On close inspection it turns out that most answers were variants[†] of one of the three classification schemes shown in Figure 2.4, which we refer to as the three "canonical classifications" for the task. Participants almost always approximately (a) sorted into three categories using both stimulus dimensions, (b) sorted into two categories based on height, or (c) sorted into two categories based on length. We assigned people to one of the three classifications by calculating the *adjR* value between each person's sort and each of the three canonical classifications, and then selected the one that was highest as their classification strategy.[‡] The breakdown of classification type by condition is shown

[†] It was not unusual for a participant to classify *most* of the stimuli according to one of these schemes, with some of the boundary cases being different; situations like this meant that 187 of the 216 distinct unique sorts were only produced by a single participant.

[‡] We also ran analyses in which we grouped participants into an "other" strategy if their solutions were insufficiently similar to any of the canonical classifications (e.g. having *adjR* values below a certain threshold like 0.2 for all three canonical classifications). The qualitative pattern of results remains unchanged across

in the top row of Figure 2.5.

In the DISTINCT STRUCTURE condition the results were straightforward. The choice of labeling scheme had no effect on the classification strategy ($\chi^2(4) = 1.90, p = 0.75$) and participants tended to use the three category solution regardless of the nature of the labels. Even when one cluster of stimuli was given no labels at all, as in the AMBIGUOUS LABEL condition, people detected the unlabeled cluster and did not attempt to group those items with items in the labeled clusters. This suggests that if the category structure is coherent and obvious enough, labels make very little difference to people's categorizations.

For the AMBIGUOUS STRUCTURE condition the story is more complex, and there is a significant difference in classifications depending on the nature of the labels shown ($\chi^2(4) = 26.48, p < 0.001$).[§] In the NO LABEL condition, people were evenly split between the three classification schemes. This reflects the fact that the raw stimulus information was not sufficient for people to infer how to categorize the items. When labels were provided, participants relied on them heavily. In the DISTINCT LABEL condition people preferred the three category solution, since the labeling scheme explicitly picked out the three clusters. In the AMBIGUOUS LABEL condition, the labels ruled out the two-category-by-height strategy, but did not distinguish between the other two strategies. This is reflected in the data, with people split evenly between the three-category and two-category-by-length strategies.

Although the overall pattern of results is a complicated interaction between stimulus structure and labeling scheme, the interpretation of this interaction effect is simple. When the stimulus structure was unambiguous, providing additional labeled data had no influence on how people learned. In such cases semi-supervised learning looks the same as unsupervised learning. In contrast, when

different threshold values ranging from 0.1 to 0.5.

[§] Significant differences were also observed between each pair of label conditions (NO LABEL and DISTINCT LABEL: $\chi^2(2) = 10.47, p < .01$, NO LABEL and AMBIGUOUS LABELS: $\chi^2(2) = 15.61, p < .01$ and DISTINCT LABELS and AMBIGUOUS LABELS: $\chi^2(2) = 12.69, p < .01$).

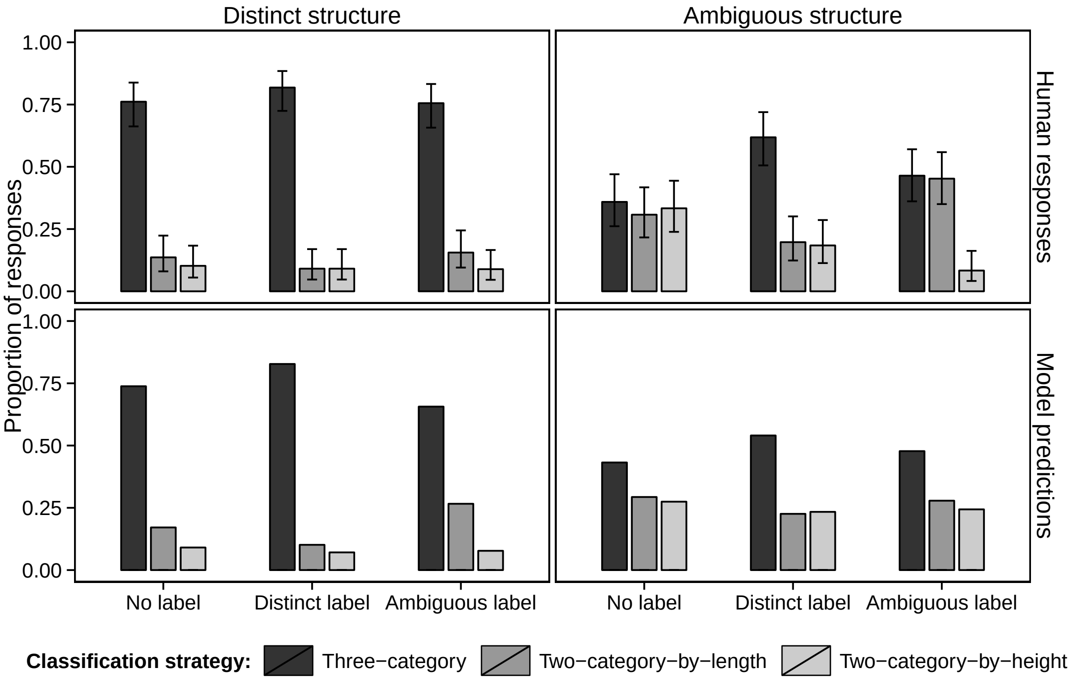


**Figure 2.5:** Comparison between the proportion of strategies used by humans and predicted by the Rational model across each of the experimental conditions. Error bars plot 95% confidence intervals for the human responses. People in the *Distinct Structure* mostly relied on unlabeled information, with labeled examples having little effect in their choice of classification strategy. In contrast, there was a strong effect in how labels were used by people in the *Ambiguous Structure* conditions. The rational model of categorization captures people's responses reasonably well in both conditions.

the stimulus structure was ambiguous, even a very small number of labeled examples had a big impact on how people learned, pushing people towards one solution or another depending on the information provided by the labels.

## 2.4 Model fitting

It appears that people produced sensible behavior in this task, but one question remains: can we account for this performance based on standard psychological theories of categorization, or is it necessary to postulate entirely different mechanisms or abilities? To address this question, we applied a modified version of Anderson's (1991) Rational Model of Categorization (RMC) to the task. The RMC is a Bayesian category learning model that has previously been applied to a variety of tasks in supervised learning (Anderson, 1991), unsupervised learning (Clapper & Bower, 2002; Pothos et al., 2011) and semi-supervised learning (Zhu et al., 2010). We chose to focus on the RMC because it lends itself well to the situation our participants were in: it assumes that stimuli belong to one of several categories, but does not know how many categories exist and so attempts to infer this from the data. However, there is no inherent reason why other successful category learning models such as SUSTAIN (Love et al., 2004) could not also be similarly adapted. The RMC learns the number of categories by relying on a simple sequential assignment method known as the Chinese restaurant process, which specifies the prior probability of a particular category (proportional to the number of items in that category) and the prior probability of a new category (a constant). For a detailed discussion of the RMC in the form we implemented it, see Sanborn et al. (2010).

It was necessary to modify the RMC slightly in order to apply to this task. A critical feature of the RMC is that category labels are viewed as an additional feature possessed by stimuli. From this perspective our task involves two continuous features (height and length) and one discrete one (label). A category is associated with a probability distribution over all three features. In Anderson's (1991) formulation, the number of possible values that a discrete feature can take is assumed to be known in advance. In our task this assumption is inappropriate, since the number of possible labels is not known to the learner. Fortunately this is easy to rectify: we assume that the distribution over labels is itself sampled from a Chinese restaurant process, consistent with the prior distribution over category assignments. Thus, labels of the same type would tend to belong to the same clusters, while items with unseen labels would be more likely to be assigned to new clusters.

Each run of the RMC outputs a set of category assignments for the observed stimuli (directly analogous to the responses we collected from participants). This output was compared to human responses by applying the same procedure that we applied to the human data: assigning each classification to one of the three canonical strategies based on the *adjR* index. Results for each condition reflect 5000 independent runs, with the order that the stimuli were presented to the model randomized between runs.

The output of the RMC, plotted in the bottom row of Figure 2.5, is qualitatively consistent with the pattern of responses produced by human subjects. For example, in the DISTINCT STRUCTURE conditions, the model predicted that the three category classification would be preferred regardless of the nature of the labels. It also predicted, similarly to people, greater variation in the strategies in the AMBIGUOUS STRUCTURE conditions. There were a few cases where the model predictions did not exactly match the responses given by people, most notably in the AMBIGUOUS STRUCTURE, AMBIGUOUS LABEL condition, where it did not rule out the two category by height classification like people did.[¶]

Overall, the correlation between the predictions of the modified RMC and the data from participants in the proportion of responses for each strategy was 0.92. This suggests that despite its imperfections, the RMC is able to roughly reproduce human performance for a novel semi-supervised task. Given that this is the first study that we are aware of that tries to compare semi-supervised learning to unsupervised learning (rather than to supervised learning) and where the number of labels is not known, it is reassuring to see that existing theory generalizes well to this situation.

[¶] The model's responses in this condition suggest that this result is driven primarily by runs where it did not observe the labeled instances necessary for correct classification until near the end of the run.

## 2.5 Discussion

Most of the literature on semi-supervised learning takes supervised learning as its starting point, and examines the extent to which additional unlabeled data shifts people's learned category representations relative to people only presented only with labeled data. The results in this area have been mixed, with studies finding that in some situations unlabeled data has an effect in semi-supervised learning (Zhu et al., 2007; Lake & McClelland, 2011; Kalish et al., 2011) and in others where it does not (McDonnell et al., 2012). Our work adopts a very different framing of the semi-supervised learning problem: instead of asking how semi-supervised learning differs from supervised learning, we ask how it differs from unsupervised learning. Instead of asking when unlabeled data have an influence on learning, we investigate when labeled data are helpful.

Our core results bear a superficial similarity to previous work, insofar as our key finding is that labeled data is sometimes helpful, and sometimes it has no effect on learning. However, our experimental manipulations make it clear when and why it happens. When the unlabeled data is informative enough that the category structure is unambiguous, people do not need labeled data to guide learning. As Bloom (2000a) suggests, semi-supervised learning appears indistinguishable from unsupervised learning in this scenario. In contrast, when the unlabeled data is ambiguous, labels become more powerful and have a large effect on the categories that people infer – in this case, the specific set of labels shown helps people determine which dimensions are relevant for classification. This includes whether to stick with a simpler unidimensional strategy or to switch to a more complex multi-dimensional classification strategy. Of course, ambiguous situations may not be the only kind of instance where labeled examples are useful. The results from Vandist et al. (2009) suggest that labeled examples can also help in learning complex Information-Integration categories – in that case, the categories are well-separated and not ambiguous but still require integrating information from multiple dimensions.

The historical prevalence of supervised learning as a topic of interest in cognitive science and machine learning has implicitly taken supervised learning to be the natural reference point against which semi-supervised learning should be assessed. In our view, this assumption also reflects an incomplete view of human semi-supervised learning. The category learning problems people – especially children – face in real life do not usually involve a few unlabeled examples in addition to many labeled ones. Rather, the world naturally presents people with a rich distribution of unlabeled data, which helpful teachers (such as parents) supplement by labeling.

Comparing semi-supervised learning to unsupervised learning sheds light on the critical role that labeled data plays in human learning. In particular, much of the difficulty in how humans learn categories is in the unsupervised aspects of determining how things should be grouped together. Here we argue that labels play a fundamental part in making sense of it, especially when the categories are ambiguous without them. It is an open question to what extent categories in the natural world are ambiguous in this way. Future work should investigate cases where labeled examples are informative in other ways, such as when objects belong to multiple cross-cutting categories (Shafto et al., 2011) or when items organized into taxonomies have multiple labels (Canini & Griffiths, 2011).

# Statement of Authorship

TITLE OF PAPER: Learning structure with labeled examples

PUBLICATION STATUS: Unpublished and Unsubmitted work written in a manuscript style

PUBLICATION DETAILS: Vong, W.K, Hendrickson, A.T., Navarro, D. J. & Perfors, A. (unpublished manuscript). Learning structure with labeled examples.

PRINCIPAL AUTHOR

NAME OF PRINCIPAL AUTHOR (CANDIDATE): Wai Keen Vong

CONTRIBUTION TO THE PAPER: Designed and ran experiments, performed data analysis, wrote manuscript and acted as corresponding author.

OVERALL PERCENTAGE (%): 75%

SIGNATURE:

DATE: 15/1/2018

CO-AUTHOR CONTRIBUTIONS

By signing the Statement of Authorship, each author certifies that:

1. the candidate's stated contribution to the publication is accurate (as detailed above);
2. permission is granted for the candidate in include the publication in the thesis; and
3. the sum of all co-author contributions is equal to 100% less the candidate's stated contribution.

NAME OF CO-AUTHOR: Andrew T. Hendrickson

CONTRIBUTION TO THE PAPER: Supervised development of work and helped with experimental design.

SIGNATURE:

DATE: 24/1/18

NAME OF CO-AUTHOR: Dani Navarro

CONTRIBUTION TO THE PAPER: Supervised development of work, helped with experimental design and editing of the manuscript signature:

date: 24/1/18

name of co-author: Amy Perfors

contribution to the paper: Supervised development of work, helped with experimental design and editing of the manuscript

signature:

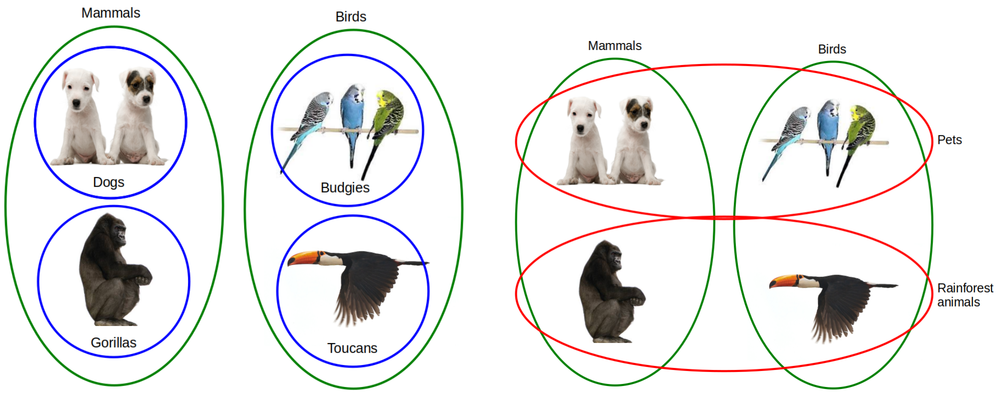
date: 24/1/18

# 3

# Learning structure with labeled examples

## 3.1 Introduction

One of the fundamental challenges for both children and scientists is determining how to organize the concepts and categories they have into structured representations and theories about the world. For example, many object kinds are organized into taxonomies, where categories at lower levels are also members of the categories above it. All dogs and gorillas are mammals, and all mammals are animals (see Figure 3.1a for a graphical representation). Another type of common structural representation is a cross-cutting structure, where categories can be organized in many ways depending on the context (Ross & Murphy, 1999; Nguyen & Murphy, 2003). Using the same example of animal categories, they can be grouped both based on their species (mammals vs. birds) or the environments



**(a)** A taxonomic structure

**(b)** A cross-cutting structure

**Figure 3.1:** Examples of two different kinds of structure. In both structures, the same objects can be categorized or grouped in different ways depending on the set of category labels that are presented with each object. One set of labels produces a taxonomic structure (where smaller categories are nested within larger categories), while another set of labels produces a cross-cutting structure (where categories overlap each other).

that they live in (pets vs. rain-forest animals). As illustrated in Figure 3.1b, these different sets of categories appear to cut across each other, hence the name "cross-cutting". How do people acquire these different kinds of structured representations about categories? In contrast to the standard category learning paradigm where each item belongs to a single category from a set of mutually exclusive categories (Nosofsky, 1986; Anderson, 1991), the problem of structure learning involves both learning all the labels for each of the categories a particular item can belong to, as well as learning the relations between each of the different categories.

While there are existing computational models of structure learning (see Kemp & Tenenbaum, 2008; Shafto et al., 2011), these models learn in an unsupervised fashion, using only feature information to determine the structure. To what extent do these computational models provide a plausible account of how a child might learn structure? One difference is that children receive vast amounts of labeled examples during development in addition to unlabeled examples, which might be beneficial for structure learning as these category labels provide cues to the organization of objects at multiple levels.

In this chapter, I examine the role that labeled examples play in learning two different kinds of structure: taxonomic structures and cross-cutting structures. For these two kinds of structures, I address three specific questions. First, can labeled examples facilitate structure learning? Second, are there any differences in the learnability of taxonomic and cross-cutting structures? And finally, can people transfer structure knowledge from one domain to another? Previous work has provided partial answers to these questions, which I review below, before presenting two experiments that investigate the problem of structure learning from labeled examples.

First, can labeled examples facilitate the learning of structure? While a number of studies have shown that people can learn structure from linguistic input (see Eaves & Shafto, 2014; Kemp, Goodman, & Tenenbaum, 2008), they have often relied on explicitly providing participants with the relevant relational information between categories to establish this structure. However, these kinds of overt, explicit relational cues are not always what children and people in the real world get, but rather weaker verbal cues such as category labels instead. This work suggests a natural follow-up question: is it possible to learn structure through labeled items alone? Some work addresses this question, but is limited by the range of structures investigated. For instance, a study by Canini and Griffiths (2011) showed that taxonomic structures could be learned through labeled examples. Participants received supervised training for categories organized in a taxonomy, and learned all of the category labels in the taxonomy. During the test phase, they asked participants to reconstruct the taxonomy by connecting the various category names with arrows that denoted relationships. Results found that 41% of participants could reconstruct the structure perfectly, while those who could not made relatively minor errors. One limitation of this particular study is that they only examined learning with a taxonomic structure, but not other kinds of conceptual structures, and it remains an open question whether this method of structure learning can generalize.

Second, are some structures easier or harder to learn than others? Examining the learnability of different structures would be informative in determining which inductive biases people bring to the problem of structure learning. This approach has been used in concept and category learning experiments, from the study of the six category types used in Shepard et al. (1961), to more recent work exploring the plausbility of different kinds of representational languages for concepts (Kemp, 2012; Piantadosi et al., 2016). In contrast, there are fewer accounts about which structures are easier or harder to learn. One study by Eaves and Shafto (2014) looked into this problem in a series of experiments; they presented participants with pairs of labels joined by an arrow representing a particular relation for a given structure (i.e. SED → VER). They examined different structures in separate experiments: a linear structure, a tree structure (similar to a taxonomy) and a more complex "orbital" structure. After participants had learned a particular structure, they were then asked to generate as many of the relations as they could. Their results showed that the relations for more difficult structures (tree and orbital) were harder to recall learn than the linear structure. However, they found that useful orderings of the relations, such as presenting the relations of a tree in a breadth-first order, speeded up learning for more difficult structures. Their results suggested that the number of relations between categories determined the difficulty of learning a particular relational structure. Would one observe similar results when learning structure from labeled examples, where the relations between categories are implicit instead?

Third, what is the role of knowledge transfer in structure learning? In the real world, children do not learn new categories and structure from scratch, but can leverage existing category knowledge as a starting point. This phenomenon is known as transfer learning or "learning-to-learn". One example of transfer learning is *selective attention*, where attending to relevant feature dimensions to learn past categories can be helpful for learning new categories (Nosofsky, 1986). A similar kind of phenomenon occurs by shape as a predictor for category membership, known as the *shape bias*, as many object categories are organized by their shape (Landau, Smith, & Jones, 1988; Kemp, Perfors, & Tenenbaum, 2007).

Does a similar kind of transfer effect occur when learning structure with labeled examples? For example, does knowing that one set of categories is a taxonomy help out when learning a new set of categories that is also a taxonomy? Like some of the other questions, researchers have explored the problem of structural transfer but only for tasks where participants are provided the relations explicitly. In a study by Kemp et al. (2008), participants first learned a structure by observing relations between pairs of entities. They were then asked to generate a set of relations for a new set of entities. The set of relations participants generated showed a strong similarity to the original structure, suggesting that participants had learned the original structure and generalized to a new domain. While the results from Kemp et al. (2008) showed a transfer of information, they did not look at any other potential benefits of transfer, including whether transfer of knowledge can speed up learning or whether this knowledge is specific to the particular structure learned. Some recent work has shown that a procedure known as inference-learning, where participants are provided with a category label and an incomplete stimulus and asked to fill in the missing features, led to better transfer of relational knowledge (Goldwater, Don, Krusche, & Livesey, 2018).

Across two experiments, I explore these three questions of learnability, differences in learnability and transfer to the acquisition of structural knowledge through labeled examples. Our results show that labeled examples can help facilitate the learning of different kinds of structure, although I do not observe any differences in learning either taxonomic or cross-cutting structures. In addition, while I find evidence supporting transfer from learning one structure to another, this was not specific structural-knowledge but related to other aspects of the task.

## 3.2 Experiment One

In Experiment One, participants performed a supervised category learning task using bug-like stimuli. The design of the stimuli allowed them to be categorized into either a taxonomic or cross-cutting structure based on the available feature information, but the set of category labels presented to participants identified the structure to use. This design allowed for the exploration of inductive biases in how people use both labeled information and feature information to learn structure, and whether there were differences in learnability across the taxonomic and cross-cutting structures. The results showed that most participants were able to learn either structure, although no observable differences were observed in learning between taxonomic or cross-cutting structures. A generalization task at the end of the experiment showed that the structure participants had learned affected how they generalized to categories containing new features.

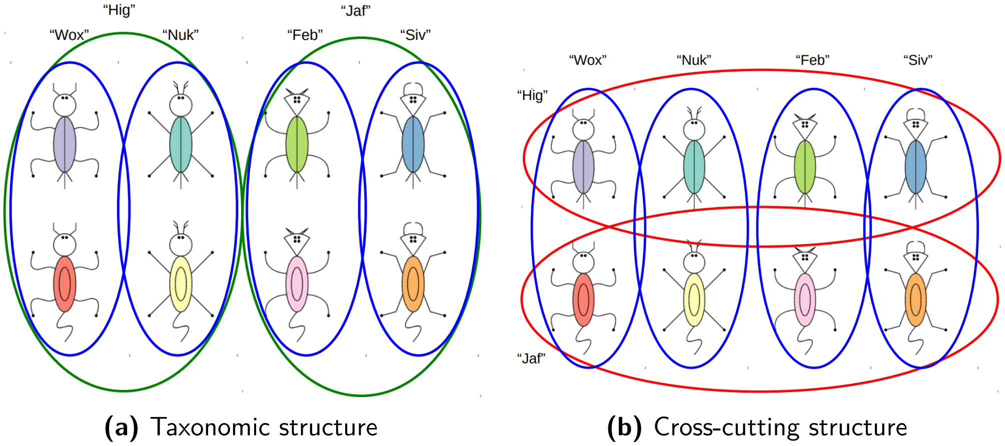
### 3.2.1 Method

#### Participants

148 participants (85 males) were recruited via Amazon Mechanical Turk. Participants ranged from ages 18 to 69 (mean: 32.5). They were paid US $4.25 for completing the experiment, which took roughly 20 minutes to complete. 9 participants failed to complete the task, and their data were excluded from further analyses.

#### Design

The goal of the task was to learn how to categorize eight different bug-like stimuli. Each stimulus belonged to two out of six possible categories, with each category associated with a different category label. The labels for each stimulus differed depending on whether participants were assigned to the taxonomic or cross-cutting structure conditions, as illustrated in Figure 3.2. In both of these structures, the six categories were divided into four small categories consisting of two different stimuli each, and two large categories consisting of four stimuli each that overlapped the smaller categories differently for each *Structure* condition. In the taxonomic category structure, the small cat-



**(a)** Taxonomic structure **(b)** Cross-cutting structure

**Figure 3.2:** The *Taxonomic* and *Cross-cutting* structures used in Experiment One. Both structures used the same eight stimuli and six category label names, but the organization between the stimuli and categories differed across *Structure* conditions. The *Taxonomic* structure consisted of two large categories of four stimuli each, and within each of these categories, two smaller categories of two stimuli each. On the other hand, the *Cross-cutting* structure used the same four smaller categories as the taxonomic structure, and then two large categories that overlapped with one stimuli from each of the four smaller categories.

egories were nested within the large categories, such that two small categories were contained within one large category. In the CROSS-CUTTING structure, the large and small categories overlapped each other in a way such that each one of the two stimuli in every small category belonged to one of the two large categories.

The category labels corresponded with different sets of features from the stimuli. Each of the stimuli consisted of seven discrete features, and the number of possible discrete feature values differed across the features. Thus, learning the categories required attending to different sets of features depending on the *Structure* condition. For example, learning the names of each of the small categories required attending to a specific subset of features, while the large categories required attending to a different subset of features. For the CROSS-CUTTING structure, the relevant sets of features for the large and small categories were independent, such that learning the small and large categories required paying attention to different sets of features. For the TAXONOMIC structure, the features

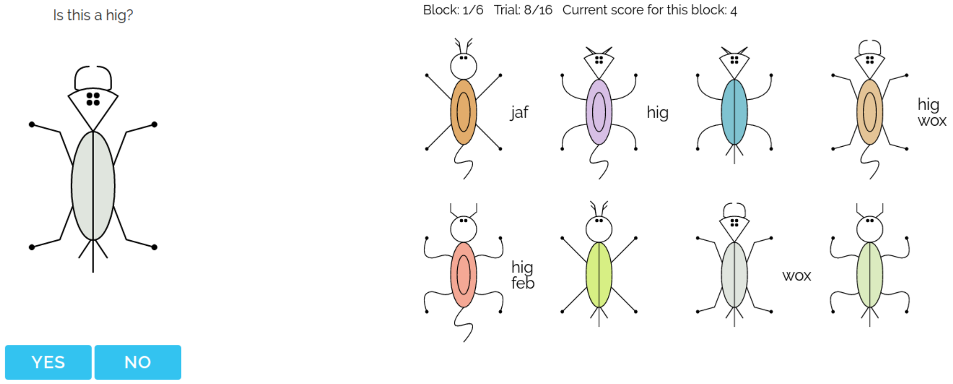


**Figure 3.3:** An example of a training trial in Experiment One. Participants were asked to judge whether the current stimuli was a member of a given category with either a Yes or No response. During training trials, participants were shown all of the stimuli on the right hand side, and depending on the block, some of the category labels were also presented.

that defined the large categories also shared features that defined the small categories.

### Procedure

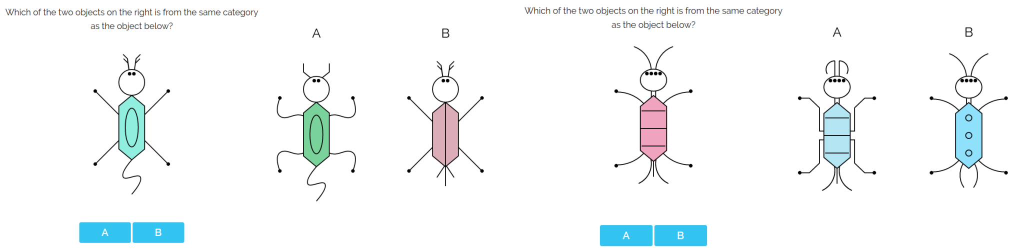
Train and Test Trials: The first part of the experiment involved subjects learning the categories through a mixture of training and test blocks. Each block consisted of 16 supervised trials, where each block covered both labels associated with each stimulus once in a randomized order. On each trial, participants were shown an object and asked "Is this an X?", and could respond either "Yes" or "No" as shown in Figure 3.3. On half of the trials, X was chosen to be one of the correct labels, otherwise the label X that was presented to participants was randomly chosen from one of the other categories of the same label type (small or large). In the training blocks, participants were given feedback after responding which informed participants both whether they were correct or not, and the correct category label regardless of whether or not they responded correctly. For example, if X corresponded to an incorrect label for the current stimuli and a participant responded "No", the feedback would be of the form "Correct! This is not an X, this is actually a Y", so that they would be pro-

| Block | Block Type | Number of Trials | Labels shown during trial |
|---|---|---|---|
| 1 | Training | 16 | Label shown after each trial |
| 2 | Training | 16 | All labels except current stimuli shown |
| 3 | Test | 16 | No labels shown |
| 4 | Training | 16 | All labels except current stimuli shown |
| 5 | Training | 16 | No labels shown until feedback after response |
| 6 | Test | 16 | No labels shown |
| 7 | Generalization | 32 | No labels shown |

**Table 3.1:** The type of blocks and the number of trials in each block for Experiment One. During training blocks, participants were presented with all of the stimuli and some of their associated labels to help with learning, while these were hidden in all test blocks. The generalization trials asked participants to select which stimuli participants thought should belong in the same category as another stimuli.

vided with the correct label for such trials. The structure of the test blocks was similar to the training blocks except participants received no feedback for their responses, and went onto the next trial immediately.

To reduce the memory load for participants, all of the stimuli were displayed on the right of the screen during the training blocks. In addition, labels were shown at particular times during the training blocks. For the first training block, participants initially only saw the stimuli, but a new label would appear alongside each stimulus after each trial consisting of the stimuli-label pair they just learned. For the second and third training blocks, the relevant labels for the stimuli on each trial were hidden (although participants could see the labels for all of the other stimuli), and only revealed after their response during the feedback screen. A similar process occurred for the final training block, except during the trial all of the labels were hidden, and were only revealed after responding. At anytime during the experiment, participants were also free to move these stimuli around, clicking two different stimuli would cause them to swap positions allowing participants to compare and contrast the different stimuli and their associated labels.


Which of the two objects on the right is from the same category as the object below?

Which of the two objects on the right is from the same category as the object below?

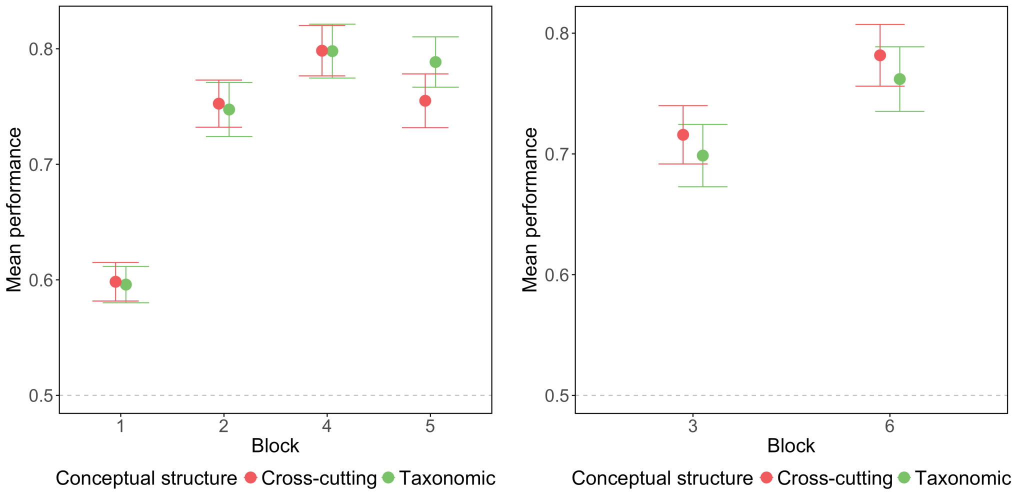
**(a)** First-order generalization trial

**(b)** Second-order generalization trial

**Figure 3.4:** Examples of first-order and second-order generalization trials in Experiment One. Participants were asked to select whether the stimuli labeled A or B belonged to the same category as the target stimuli on the left. Both of the generalization stimuli (A and B) differed from the target stimuli by the same number of features, but on features that either matched the categories they learned during training or not. While the first-order generalization trials consisted of stimuli with the same feature values from the experiment, the second-order generalization trials involved completely new feature values.

Generalization Trials: After completing the first part of the experiment, participants began the second part of the experiment, which was composed of 32 generalization trials that asked participants to judge which objects belonged to the same category. The 32 trials were split between an equal number of *first-order* generalization trials (same stimuli and features values during training) and *second-order* generalization trials (different stimuli and features values compared to training, but same distribution of feature values). Figure 3.4 illustrates both kinds of generalization trials. Using a completely different set of feature values, but the same category structure therefore allowed us to determine whether or not people were generalizing by the exact feature values they were trained on, or whether they generalized more broadly to the same features (legs, eyes etc.). The order of generalization trials was randomized for each participant, with first-order and second-order trials interspersed together.

In each generalization trial (as shown in Figure 3.4), participants were shown one target object on the left and then two objects on the right, and asked "Which of the two objects on the right is from the same category as the object below?" The two objects on the right were chosen such that they



**(a)** Training performance

**(b)** Test performance

**Figure 3.5:** Performance across training (left) and test (right) blocks in Experiment One, grouped by *Structure* condition. While both figures show an increase in performance over time for both training and test, only performance in the test blocks can be compared as there were differences from block to block in the training blocks. However, there is little to no difference in performance between the two *Structure* conditions.

matched the number of features to the target object on the left, but the relevant matching features differed. As the labels organized into the two different *Structure* conditions mapped onto different features, I hypothesized that participants would generalize along the same set of relevant features they were trained on, based off the underlying conceptual structure. I hypothesized that generalization across relevant features would be indicative that participants had learned a particular structure.

### 3.2.2 Results

Three questions motivated this first experiment: (1) Can people learn to map category labels onto the different structures? (i.e. are they learnable?) (2) Are there any differences in how people can learn taxonomic and cross-cutting structures? and (3) What effect does structure have in how people generalize or transfer their knowledge to new categories? I address each of these questions in order.

First, I explored whether people were able to learn the task. Figure 3.5 shows the mean performance across the different *Structure* conditions, separating out the training and test blocks.[*] The results from Figure 3.5 show an increase in performance across both training and test blocks, suggesting that people were learning across blocks, which I further quantify with additional analyses.

To determine whether people showed any learning across the six training and test blocks, I compared a Bayesian mixed effects model with *Block* as a continuous variable (across all six training and test blocks) to an intercept only model.[†]. The results showed that a model that included *Block* as a variable was strongly preferred ($BF > 10^{20} : 1$) over the intercept-only model, which suggests that people learned over the course of the experiment. However, it is unclear whether this analysis provides clear cut evidence for learning, as the number of labeled shown on the side of the experiment varied across the different training blocks.

As a follow-up analysis to control for this difference, I compared performance for the two test blocks only, where participants did not receive any label aids. The mean performance in the first and second test blocks was 71% and 76% respectively. I conducted a paired samples Bayesian t-test comparing the difference in performance across the two test blocks, with results showing strong evidence of a difference in performance, which was higher in the second test block compared to the first ($BF > 200 : 1$).

The second motivating question was whether there were any differences in learning between the two *Structure* conditions. I compared a Bayesian mixed effects model of *Structure* (coded as a discrete variable) and *Block* (coded as a continuous variable) to a model only containing *Block* as a variable. Similar to the analysis above, I compared performance in test blocks only to control for differences during training. Our results favoured the model with only *Block* as a variable ($BF >$

[*] Because the task differed slightly between training and test blocks, performance is shown separately. See below for further information.

[†] All mixed effects models in this paper assume a random intercept for each subject. I calculated Bayes factors using the default parameters (Rouder, Morey, Speckman, & Province, 2012; Liang, Paulo, Molina, Clyde, & Berger, 2012) of the BayesFactor package 0.9.12-2 (Morey & Rouder, 2015) in R 3.2.3.

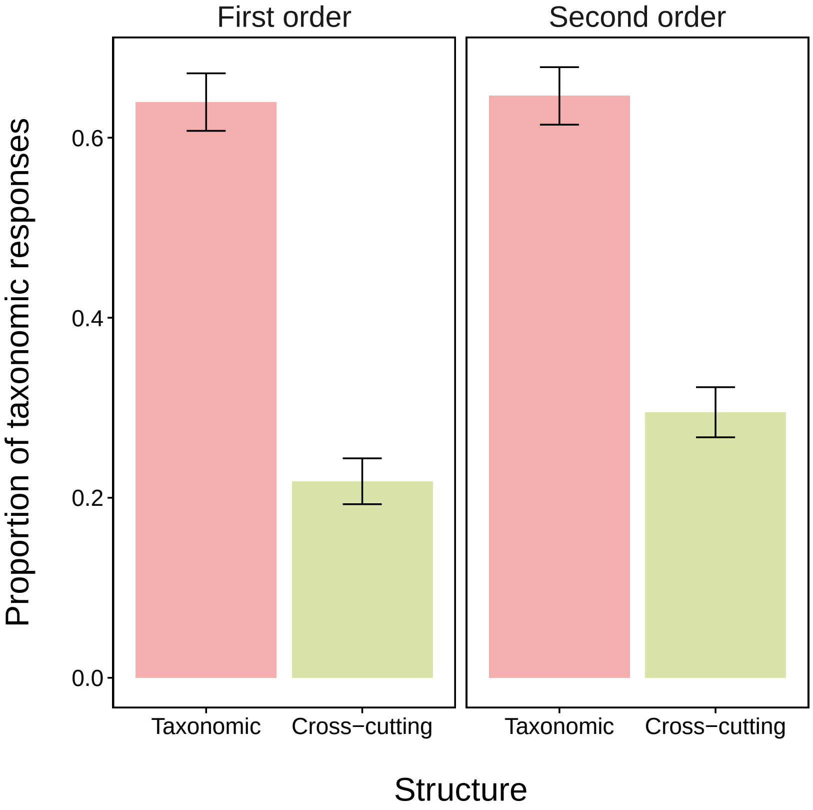


**Figure 3.6:** Proportion of generalization responses where participants selected the taxonomic response, grouped by *Structure* and type of generalization question. Across both first and second-order generalization trials, participants generalized according to the set of category labels based on the *Structure* they learned and not by the number of shared features between the target and generalization stimuli.

3.6 : 1), provided weak evidence that there was no difference in learning between a TAXONOMIC or CROSS-CUTTING structure,.

Finally, what effect did learning have on people's generalization to new objects and categories, and was this influenced by the *Structure* they had learned? While our earlier analyses showed no difference in learning performance between the two *Structure* conditions, would different sets of labels produce different kinds of generalization behaviour for the generalization trials at the end of the experiment?

In each of these trials, participants were shown a target stimuli and asked to pick one of two stimuli that either matched based on TAXONOMIC structure or the CROSS-CUTTING structure. In the first-order generalization trials, all of the feature values were the same as during training, while the second-order generalization trials involved completely new feature values to examine transfer behaviour. To analyze participant's generalization behaviour, I coded their responses as a percentage of responses which matched the TAXONOMIC response versus the CROSS-CUTTING response.

Figure 3.6 shows the percentage of taxonomic responses broken down by *Structure* and the *Generalization Type* (first-order or second-order), showing that in both the first and second-order generalization trials, participants preferred to generalize in the same manner as the structure they had learned during training. I compared a Bayesian mixed effects model that included *Structure* (coded as a discrete variable) to an intercept-only model on the proportion of taxonomic responses. Results strongly favoured the model with *Structure* and *Generalization Type* over an intercept-only model, indicating that there was a difference in how people generalized according to the *Structure* condition they were in. A follow-up analysis comparing a Bayesian mixed effect model of *Structure* and *Generalization Type* to a model with only *Structure* showed that there was no difference in people's responses across first and second-order generalization responses ($BF \approx 1.2 : 1$), which indicated that people were not memorizing the exact feature values when making generalization responses, but instead were using information from the structure they had learned to generalize in a sensible manner.

### 3.2.3 SUMMARY

The goal of Experiment One was to examine people's inductive biases in how they mapped labels to structured categories. While our results showed that people could learn mappings for different kinds of *Structure*, there was no difference in learning for the two structures that I examined. Despite this lack of difference during learning, when comparing behaviour on the generalization trials, I found that participants strongly preferred to select the stimuli that matched on structure to be in the same category, suggesting that learning affects generalization at multiple levels of abstraction.

However, there exists an alternative explanation for the pattern of generalization behaviour I observed in this task. Since the two *Structure* conditions were based on different sets of features to group the different bugs into categories, if participants had learned which features were relevant in either structure (without actually learning the relational structure itself) they would produce the same kinds of responses that I observed. Due to this potential confounding explanation, I ran a follow-up experiment that attempted to tease apart whether participants were actually learning the relational structure or only the relevant features for categorization.

## 3.3 Experiment Two

The purpose of Experiment Two was to address the potential confound from our results on the generalization task of Experiment One. I addressed this by modifying the experimental design, using a task with two phases where participants learned two different sets of categories. I hypothesized that if structure played an important role in learning, then we would expect to see better performance in the second phase when learning the same kind of structure. However, if structure is irrelevant, then no benefit in the transfer phase of the second phase would be observed. The results from Experiment Two showed a general learning benefit for the transfer task that was independent of structure, suggesting that other abstract learning mechanisms may be important. I explore some of these possibilities at the end of this section.

### Participants

397 participants (223 male) were recruited via Amazon Mechanical Turk.[‡] Participants ranged in age 18 to 71 (mean: 34.0), and were paid $3.00 USD for completing the task. 46 participants failed to complete the task, and there were no differences between which experiment conditions these incomplete participants were assigned to. Their data were excluded from any further analysis.

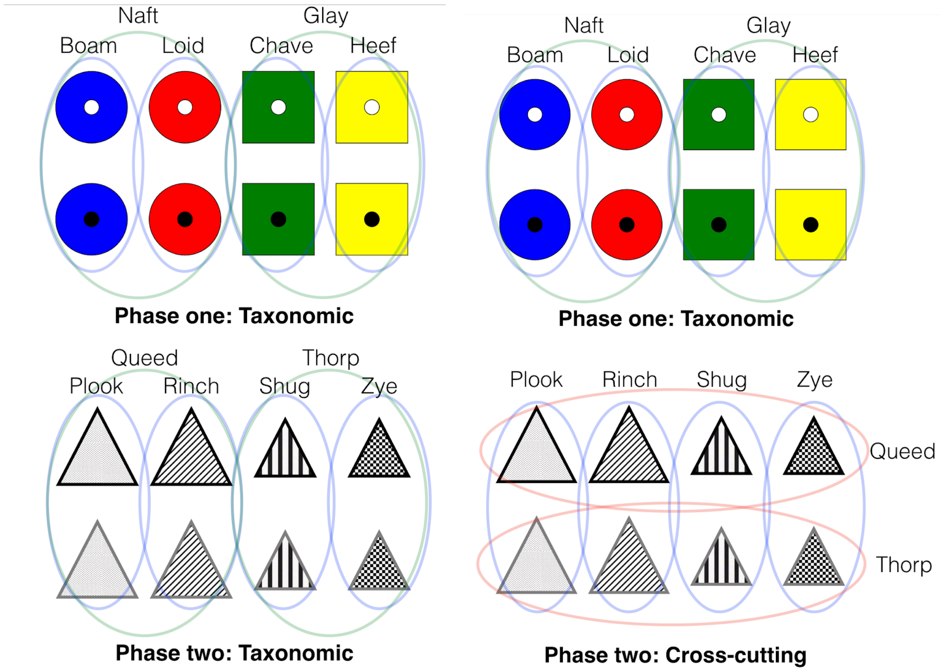
### Design

The structure of Experiment Two was split into two phases, where each phase required participants to learn a set of mappings from category labels to novel objects similar to Experiment One. The procedure in both phases was the same, but with different stimuli and category labels used in each phase. The same two structures were retained in Experiment Two (taxonomic and cross-cutting), but with fewer features to attend to.

In both phases, participants were presented with eight different stimuli and asked to learn all of the names for each stimulus. Exactly like Experiment One, the mappings between the stimuli and category labels in both phases either followed a taxonomic or cross-cutting organization, with the structure randomly selected for both phases. This resulted in a set of four different between-subjects conditions: taxonomic/taxonomic, taxonomic/cross-cutting, cross-cutting/taxonomic and cross-cutting/cross-cutting. Participants who learned the same structure in both phases were labeled as *Matching* structures, while those who learned different structures were labeled as *Non-matching*, as illustrated in Figure 3.7.

The stimuli in each phase consisted of three discrete features, which were different and randomly created for both phases. The eight stimuli were grouped into six different categories: two big categories and four smaller ones. There were two different sets of features for the small categories:

[‡] The larger sample size for Experiment Two compared to Experiment One was the result of doubling the number of conditions I wanted to test.

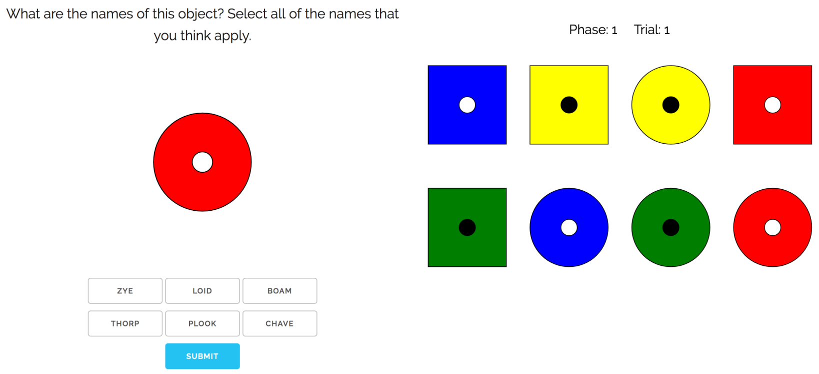


**(a)** Matching structures

**(b)** Non-matching structures

**Figure 3.7:** Examples of structures used in both phases in Experiment Two. On the left is a set of matching structure conditions, where participants learned a *Taxonomic* structure in both phases. On the right is a set of non-matching structures, where participants learned a *Taxonomic* structure in the first phase and a *Cross-cutting* structure in the second phase.

colour and pattern, and four sets of features for the large categories: shape, size, border colour and dot. In the first phase of the experiment, one of the small features was selected, along with two of the large features. The remaining features were then used as the features for the second phase. Each of the small features mapped to a different category label for each small category, while the two large features randomly mapped to the TAXONOMIC and CROSS-CUTTING structures respectively, with the exact mapping between labels and features were dependent on the *Structure* condition.



What are the names of this object? Select all of the names that you think apply.

**Figure 3.8:** Example of a training trial in Experiment Two. Participants selected all of the labels they thought applied to the current stimuli. Additionally, as a memory aid all of the stimuli were always presented on the right in a shuffled order.

### Procedure

On each trial, I presented participants with one of the eight stimuli and asked "What are the names of this object? Select all of the names that you think apply." Below the object were six buttons with the names of all the possible labels. Participants could select any of the six labels by clicking, or unselect them by clicking again. In addition, on the right side of the screen, participants were shown all eight stimuli in a random order on a 2 × 4 grid as a helpful guide. After selecting all the labels they believe applied to the current stimuli, participants submitted their responses and received feedback. This feedback indicated whether they were correct or not, and if they were incorrect they were also provided with the correct set of category labels. Figure 3.8 illustrates an example of a training trial in Experiment Two.

To determine whether participants had learned the task in either phase, I set a predefined learning criterion and examined their performance after each block to determine if they passed. The criterion was whether they had made less than two incorrect responses during the block (by either selecting an incorrect label or not selecting a correct label).[§] After reaching this criterion in the first phase, they would automatically begin the second phase of the experiment. Otherwise, they would remain in the first phase until they finished all six training blocks before proceeding to the second phase. The procedure for the second phase was identical to the first, but with new stimuli and category labels that differed based on the assigned *Structure* condition for each participant. Regardless of performance, all participants completed six blocks of training in the second phase.

### 3.3.1 Results

The aim of Experiment Two was to investigate what kinds of information people transferred between tasks, and in particular whether this included structural information. The results from this study did not show a benefit for matching structure conditions, but I did observe a general benefit of learning from the first phase to the second phase. I conclude this section outlining possible explanations and the impact of this result.

Before exploring the effects of transfer between the first and second phase I look at how well people were able to learn the task. I calculated the time to reach the learning criterion in both phases for all participants. I calculated the learning criterion by looking at the first block for each phase where participants' $F_1$ score was greater than or equal to 0.85, which corresponded to two or fewer mistakes in a single block. In the first phase, 266 (67%) participants reached the learning criterion (141 in the taxonomic, 125 in cross-cutting). In phase two, 262 (66%) people reached the criterion (131 in taxonomic, 131 in cross-cutting), which suggests that a majority of participants were able to learn all the labels associated with the different categories.[¶] Because I was interested in evaluating

[§] As participants could select multiple labels, measuring the number of correct responses provides an incomplete characterization of people's learning. Instead, I used the $F_1$ score as a measure of performance, where two or fewer incorrect responses corresponds to an $F_1$ score that is greater than 0.85, where $F_1 = 2 \cdot \frac{\text{precision} \cdot \text{recall}}{\text{precision} + \text{recall}}$.

[¶] An analysis of participants who did not reach the criterion suggests that they were able to learn one set of category labels (either the small or large categories) but not the other within the six blocks.

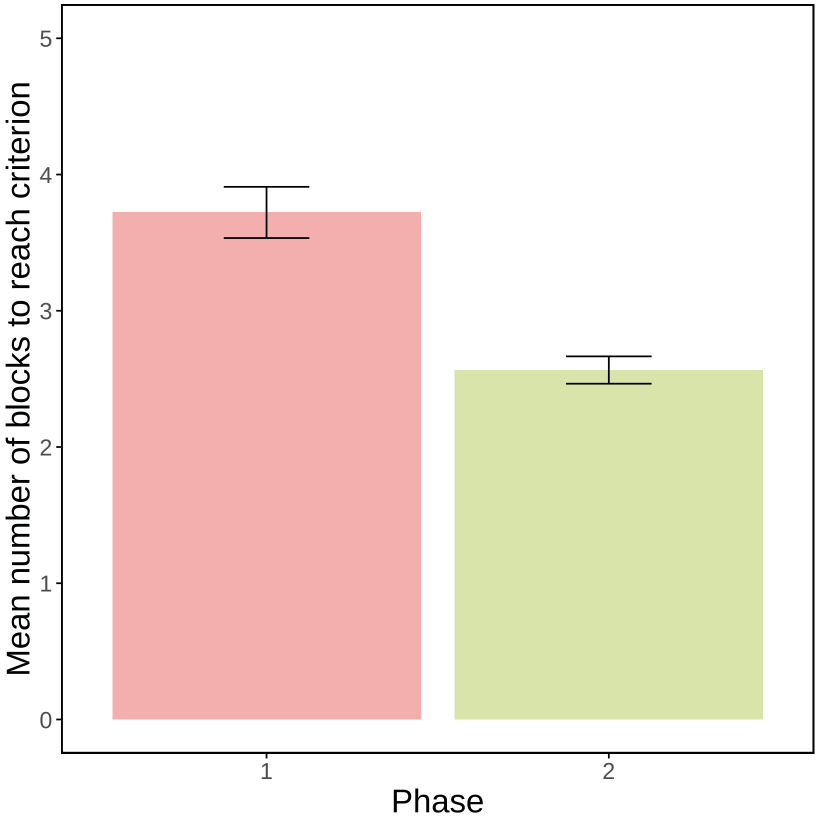


**Figure 3.9:** Mean number of blocks to criterion in both phases of the experiment collapsed across all conditions (for participants who learned the task in both phases). Overall, the results show that that participants reached criterion much faster in the second phase than the first phase, indicating there was a general benefit of transfer across the two tasks for people who learned the categories.

whether people would generalise knowledge to a new task, the rest of the results only consider the 229 participants who reached the learning criterion in both phases.

For the participants who learned both tasks, was there any effect of transfer of learning between the first and second phases? Do participants in the second phase reach criterion earlier, and does this depend on whether the *Structure* conditions in both phases were matching? I begin by addressing the first question. Figure 3.9 shows the average number of blocks it took for participants to reach

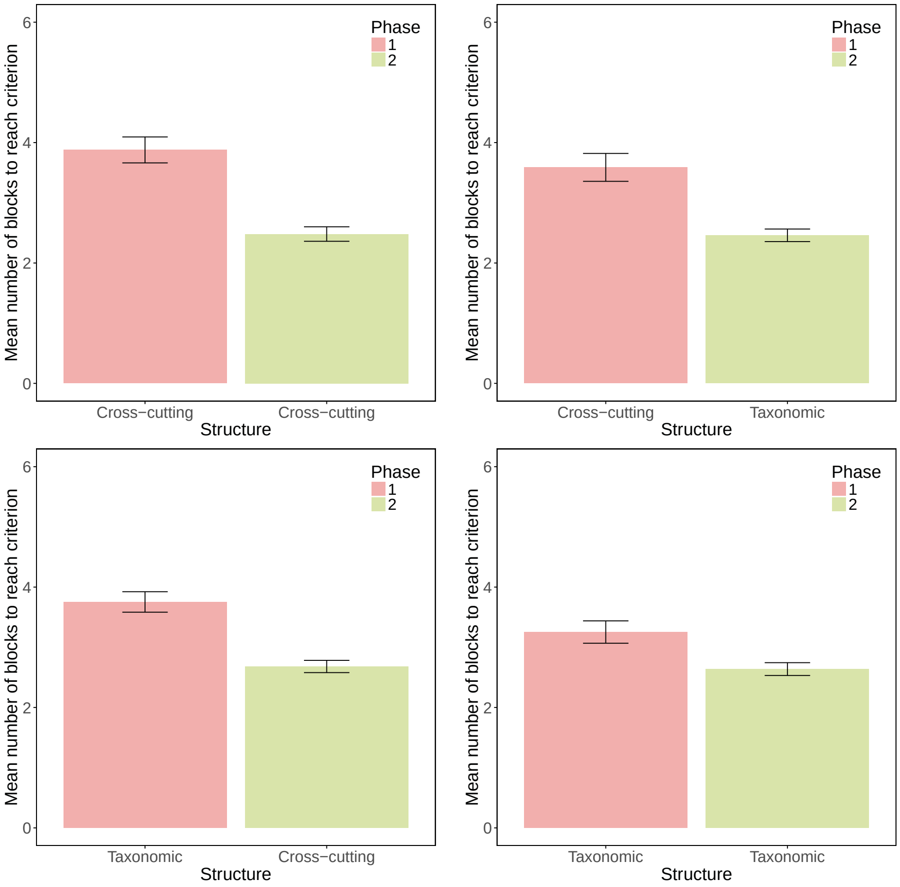


**Figure 3.10:** Mean number of blocks to reach criterion in both phases, broken down by each pair of *Structure* conditions. Participants across all conditions reached the learning criterion faster in the second phase, but there were no differences in the speed of learning across any of the conditions, suggesting that there was no benefit for matching structures.

the learning criterion in phase one and two, collapsed across all *Structure* conditions. The results show the mean time to reach criterion was 3.63 blocks for phase one, and 2.57 blocks in phase two. A paired samples Bayesian t-test comparing the time to reach criterion in each phase presented strong evidence of faster learning in the second phase ($BF > 7 \times 10^5 : 1$).

Were there further differences in transfer performance based on the kind of *Structure* participants learned in both phases? I was particularly interested in whether learning a set of MATCHING structures (i.e., TAXONOMIC/TAXONOMIC, or CROSS-CUTTING/CROSS-CUTTING) compared to non-matching structures NON-MATCHING structures (i.e., TAXONOMIC/CROSS-CUTTING or CROSS-CUTTING/TAXONOMIC) would lead to participants reaching the learning criterion earlier in phase two. Figure 3.10 illustrates the breakdown in mean time to reach criterion across the different conditions. While there were some differences in learning speed, they appeared to be minor. Indeed, a comparison between Bayesian mixed effects models agrees with this observation. I compared one model that included PHASE and MATCHING as discrete factors to another model with only PHASE as a factor for how well they accounted for the time to reach criterion. The results weakly favoured the simpler model with only PHASE as a factor ($BF = 7.35 : 1$), suggesting that learning a matching structure compared to a non-matching one had little to no effect on the speed of learning.

Overall, these results suggest that participants did not have an inductive bias to prefer structurally similar categories in the second phase. However, this combination of results does instead suggest a general transfer effect. What is the exact nature of this transfer effect and inductive bias, if it is not driven by structure? While there exist many other inductive biases, I consider one alternative explanation that explores how other parts of the task can be sufficiently abstracted to produce a general transfer effect.

One source of knowledge in the task that was independent of the structure was how many labels applied to each object in the task. Previous work suggests that people have a strong inductive bias towards mutual exclusivity, favouring a single label for each category (Markman & Wachtel, 1988). We can measure the strength of this inductive bias by examining the proportion of labels people selected on the first trial. Figure 3.11 highlights this inductive bias, with most participants selecting only a single label, and the remainder of participants preferring either two or three labels at most. However, as shown on the right of Figure 3.11, the proportion of selected labels from the first trial of the second phase shows a different set of responses. Here, the modal response was two responses, suggesting that participants shifted their inductive bias on the number of labels for each item, based

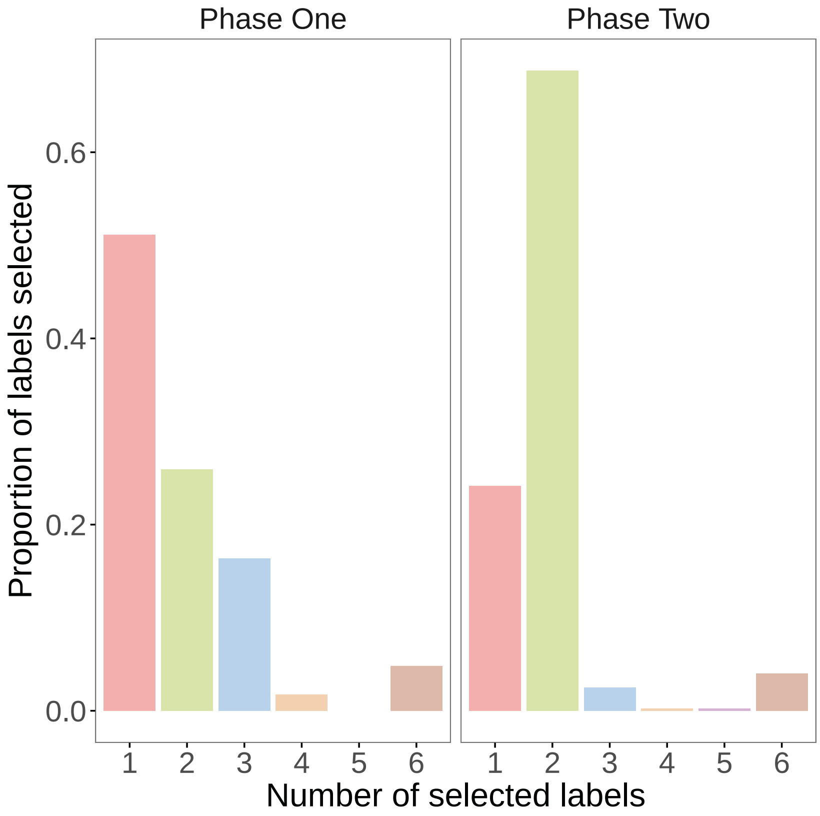


**Figure 3.11:** The proportion of selected labels (from one to six) on the first trial from both phases. In phase one, selecting a single label was the modal response, with the remainder of the participants favouring a small number of labels (two or three). However, in the first trial of phase two the distribution of responses shifts, with most participants selected two labels even before they had seen any feedback during this phase, suggesting a explanation for a general transfer effect.

off knowledge learned from phase one of the experiment.

Overall, the results from Experiment Two showed that there was a general effect of transfer from phase one to phase two. However, the additional benefit did not differ when participants had learned the same structure in both phases or not. A closer inspection into people's responses suggest that this benefit of transfer learning stemmed from learning useful meta-aspects about the task, such as the distribution of the number of labels for each object.

## 3.4 Discussion

In this work, I conducted two experiments investigating the problem of structure learning. The aims across both experiments were to explore three questions regarding this problem: the learnability of structure with labeled examples, differences across learning various structures, and how structure knowledge might transfer or generalize to new domains. The pattern of results obtained from both experiments provided satisfactory responses for some of these questions, while for others it was less clear. I discuss these results, relating them back to the literature and potential directions for future work in the rest of this chapter.

In both experiments, I found that participants were able to learn both taxonomic and cross-cutting structures. In Experiment One, the results showed that participants in both *Structure* conditions were able to answer questions during the test phase at above chance levels, and generalized to novel features based on the *Structure* they had learned. In Experiment Two a majority of participants in both *Structure* conditions were able to reach the learning criterion that was set as an indicator for having learned the structure. These results are consistent with previous work showing that labeled examples can guide structure learning (Canini & Griffiths, 2011), and I extend these results to cross-cutting structures. It also suggests that labeled examples may be useful for learning other kinds of structure.

One limitation of both experiments was that participants were never required to reconstruct the learned structure, unlike past experiments such as Kemp et al. (2008); Canini and Griffiths (2011). Rather, I asked participants to perform other tasks such as generalization that would reveal whether they had acquired structure knowledge implicitly. Unfortunately, one potential confounding explanation of the experimental results I observed is that participants may have been able to learn the categories and their labels, but not the relations between the categories that determined the relevant structure. One possibility of disentangling these competing explanations would be to ask participants to generate the relevant structure for future experiments. An alternative possibility would be to query participants with questions such as: "Are all X's also Y's?" as providing a correct answer to such questions would require participants to determine various relations between categories in an explicit fashion. Future research should attempt to determine whether learning all the category labels for an object is sufficient for encoding structural knowledge.

In exploring differences across learning various structures, I found no significant differences in the difficulty of learning a taxonomic from a cross-cutting structure. One explanation for this null result is from the experiments conducted in Eaves and Shafto (2014), arguing that the difficulty of learning structure was proportional to the number of relations between categories, and not from the number of categories. In the taxonomic and cross-cutting structures I used, the number of relations between categories were similar for both structures. Thus, one would expect no difference in difficulty from learning either structure according to this argument. However, this null effect is actually consistent with results from the developmental literature. Nguyen and Murphy (2003); Nguyen (2007) show that children learn both taxonomic and script categories at an early age (2-3 years), and can interchange between using either structure for different inferences including cross-classification. If we integrate these findings from the developmental literature with the results from both experiments, it suggests that lack of difference in the results is not particularly unexpected, but suggests that further exploration of this particular question should test out particular structures with a clearer gradient of difficulty.

The third purpose was investigating how well people could generalize and transfer knowledge of structure. In Experiment One, I evaluated generalization performance by asking participants which two objects belonged to the same category. While the results showed that participants responded in a way that matched the learned structure, participants may have only learned to attend to the relevant dimensions instead of the underlying structure. To correct for this, in Experiment Two I evaluated transfer by asking participants to complete a new category learning task with novel stimuli that was either consistent or not consistent with the original structure. The results from this experiment showed that participants who were able to learn the first structure were able to learn the second structure faster, but there was no difference in learning speed when structures were consistent or not. These results suggest that people transferred knowledge about particular aspects of the task independent of structure, which led to speeded learning regardless of the first learned structure.

Do there exist any computational models that can help explain the results from these two experiments? Most computational models of structure learning do so on the basis of feature information only (Kemp & Tenenbaum, 2008; Shafto et al., 2011; Heller & Ghahramani, 2005; Lake, Lawrence, & Tenenbaum, 2016). One structure learning model that does use labeled examples is the Tree-HDP model proposed by Canini and Griffiths (2011), although it was only designed learn taxonomic structures. A future challenge might be to extend the Tree-HDP model, or create new computational models that can learn a more flexible range of structures from the basis of different sets of labeled examples.

In summary, our results show that people can learn different kinds of structured categories through labeled examples, without telling participants the particular structure they were learning. There exist many future possibilities to test and examine people's representations of structural knowledge, as well developing computational models that can capture this kind of learning.

# Statement of Authorship

TITLE OF PAPER: Do additional features help or hurt category learning? The curse of dimensionality in human learners.

PUBLICATION STATUS: Submitted for Publication

PUBLICATION DETAILS: Vong, W.K, Hendrickson, A.T., Perfors, A. & Navarro, D. J. (manuscript submitted for publication). Do additional features help or hurt category learning? The curse of dimensionality in human learners.

PRINCIPAL AUTHOR

NAME OF PRINCIPAL AUTHOR (CANDIDATE): Wai Keen Vong

CONTRIBUTION TO THE PAPER: Designed and ran experiments, performed data analysis, implemented multiple computational models, wrote manuscript and acted as corresponding author.

OVERALL PERCENTAGE (%): 80%

SIGNATURE:

DATE: 15/1/2018

## CO-AUTHOR CONTRIBUTIONS

By signing the Statement of Authorship, each author certifies that:

1. the candidate's stated contribution to the publication is accurate (as detailed above);
2. permission is granted for the candidate in include the publication in the thesis; and
3. the sum of all co-author contributions is equal to 100% less the candidate's stated contribution.

NAME OF CO-AUTHOR: Andrew T. Hendrickson

CONTRIBUTION TO THE PAPER: Supervised development of work, helped with experimental design, development of computational models and editing of the manuscript

SIGNATURE:

DATE: 24/1/18

NAME OF CO-AUTHOR: Amy Perfors

CONTRIBUTION TO THE PAPER: Supervised development of work, helped with experimental design, development of computational models and editing of the manuscript

SIGNATURE:

DATE: 24/1/18

NAME OF CO-AUTHOR: Dani Navarro

CONTRIBUTION TO THE PAPER: Supervised development of work, helped with experimental design, development of computational models and editing of the manuscript

SIGNATURE:

DATE: 24/1/18

# 4

# Do additional features help or hurt category learning? The curse of dimensionality in human learners

## 4.1 Introduction

Despite the fact that category learning is logically difficult in many ways, people easily and naturally learn real-world categories. Quine (1960) identified one well-known problem, originating from the fact that the category referent for any particular word is under-determined and could be any from an infinite set of possibilities. This is an example of the problem of induction (e.g., Goodman, 1983), which concerns the difficulty in identifying how to generalize when there are a potentially infinite set of possible bases to do so. The problem of induction, in its different forms, has been widely studied within cognitive science. Proposed solutions often center around the existence of inductive biases, although the exact nature of these biases remains a debated issue (see, e.g., Markman, 1989; Landauer & Dumais, 1997; Griffiths, Kemp, & Tenenbaum, 2008; Chater, Clark, Goldsmith, & Perfors, 2015; Minda, 2015). In this paper, we focus on a less studied but related problem known as the curse of dimensionality. It is similar in that it is a fundamental problem of learnability, but different in that it relates to the specific problem of learning in high-dimensional spaces or with a large number of features. We show why the acquisition of real-world categories *should* be difficult due to the curse of dimensionality, but propose that the structure of real-world categories may alleviate the curse for humans, at least in many situations.

The curse of dimensionality has been well-studied within statistics (e.g., Bellman, 1961; Donoho, 2000) and machine learning (e.g., Verleysen & François, 2005; Keogh & Mueen, 2011), and has a number of interesting and widespread effects. Within computer science, the curse of dimensionality means that if the amount of data on which to train a model (e.g., a classifier) is fixed, then increasing dimensionality can lead to overfitting. This is because as the space grows larger, the examples themselves grow ever sparser; the only way to avoid the issue is to bring in exponentially more data for each additional dimension. In statistics and mathematics, the curse means that it is not possible to numerically optimize functions of many variables by exhaustively searching a discretized search space, thanks to the combinatorial explosion of parameters to explore.

A similar problem arises in the domain of category learning: as we consider categories with more and more features, the size of the possible feature space and number of examples required to fully learn a category grows extraordinarily quickly. For objects with $N$ independent binary features, there are $2^N$ possible examples and $2^{2^N}$ possible ways of grouping these objects into two categories. The size of possible categories grows at a double-exponential rate to the number of independent features of a category (Searcy & Shafto, 2016). As a result, even for moderate values for $N$, learning categories should be extremely difficult. If features can take on more than two values – as in most real-world categories – the problem grows even more acute. For instance, items with 16 possible features of five possible values each yields $1.5 \times 10^{11}$ possible exemplars.

Most real-world categories have a large number of available features for categorization (Rosch, 1973), which suggests that – in theory at least – the curse of dimensionality means that acquiring natural categories should be a difficult learning problem. Yet people, including children, can learn real-world categories with relative ease, often based on only a few exemplars. How do people accomplish this feat?

We know surprisingly little about the answer to this question. Most experimental work in category learning has not run into the problem of the curse of dimensionality, either because studies have used categories that people have already learned or because they tested categories using stimuli with only a few, highly salient features (e.g, Shepard et al., 1961; Medin & Schaffer, 1978; Nosofsky, 1986). Although this body of work has substantially contributed to our understanding of category learning, it remains an open question how learning is affected when there are a large number of features. The limited studies that have investigated category learning with varying numbers of features have yielded conflicting results, with some studies finding that additional features impair learning (Edgell et al., 1996), others finding that they facilitate learning (Hoffman & Murphy, 2006; Hoffman et al., 2008), and others finding that they have no effect on learning at all (Minda & Smith, 2001).

How can we resolve this apparent discrepancy? One possibility is that each of these studies differ in the kinds of category structures being learned. After all, the curse of dimensionality stems from having so many possible stimuli configurations in a high-dimensional space that it is difficult to learn which set of features people should use for classification. This should lead to the greatest inefficiency when most of the possible features are not predictive of category membership and only one or a few matter, as in Edgell et al. (1996). By contrast, if all features are predictive to some degree – especially if they are not perfectly correlated with each other – then additional features should be beneficial, or at least not harmful (Hoffman & Murphy, 2006; Hoffman et al., 2008; Minda & Smith, 2001). This possibility is especially interesting given the fact that most real-world categories have precisely this sort of family resemblance structure (Rosch & Mervis, 1975b; Murphy, 2002).

This hypothesis – that a family resemblance category structure may mitigate the impact of the curse of dimensionality, but that other kinds of category structures may not – appears superficially plausible, but to date no studies have tested it. The goal of the current paper is to examine this hypothesis by manipulating category structure and the number of features while holding other factors constant. Our results do indeed suggest that people do not succumb to the curse if the categories follow a family resemblance structure. However, if the categories are rule-based, the curse of dimensionality affects humans. We compare these results to an ideal observer model for this task that shows the theoretical accuracy one could achieve by combining information across all of the available features. Despite human performance falling well below the achievable limit given by this ideal observer model, our results show that there are situations in which people can still achieve high classification performance and overcome the curse of dimensionality.

## 4.2 Experiment One

We addressed two questions in this experiment. First, how does learning performance change as the number of features increase? Second, to what extent does the structure of the category influence this learning? We therefore systematically manipulated the number of features and the manner in which categories were structured within the context of a standard supervised learning task.

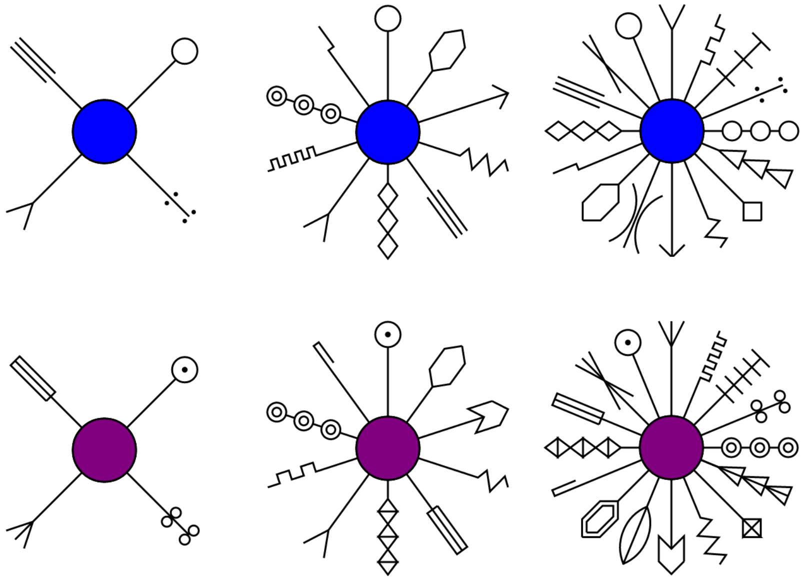


**Figure 4.1:** Example stimuli, displaying two instances from each of the three possible *Dimensionality* conditions (4, 10, and 16, from left to right). Features were binary and correspond to the legs of the amoebas. Together, the two 16-*Feature* examples show all possible feature values.

### 4.2.1 Method

#### Participants

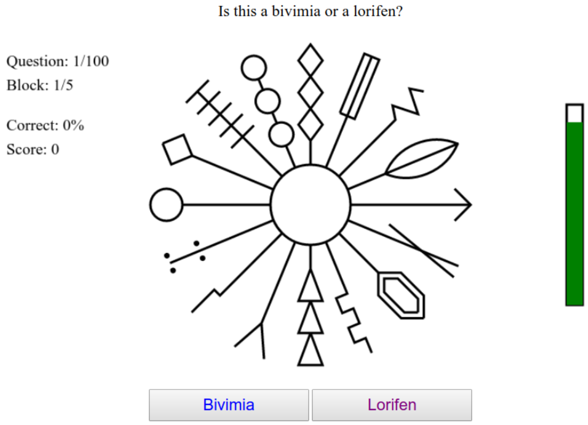
886 participants (496 male, 388 female, 2 other) were recruited via Amazon Mechanical Turk. This is a relatively high number of participants because we ran two experiments with slightly different methodologies (described below) but pooled the results since they were qualitatively identical. Participants ranged in age 18 to 76 (mean 34.2). They were paid US$2.00 for completion of the experiment, which took roughly 12 minutes. Data from an additional 42 participants were excluded from analysis, either from failure to complete the task (37 participants) or participating in a pilot version of the study (5 participants).

## Design

The experiment presented people with a supervised category learning problem, in which they were asked to classify an amoeba stimulus as either a bivimia or lorifen. Each amoeba consisted of a circular base with a set of binary features (legs). The full set of 16 unique pairs of features are shown on the two stimuli in the right column of Figure 4.1.

Nine experimental conditions were created by manipulating the *Dimensionality* of the stimuli and the *Structure* of the category in a 3 × 3 between-participants design; people were randomly assigned to each condition. The three levels of *Dimensionality* reflect the number of binary features present on the stimuli: 4-FEATURE ($N = 302$), 10-FEATURE ($N = 277$), or 16-FEATURE ($N = 307$). For the lower-dimensionality conditions, the set of displayed features were a randomly selected subset of the features used in the 16-FEATURE condition. The position of features on the amoeba were randomized differently for each participant.

The three category *Structures* were designed in the following way. In every condition there was one feature (chosen randomly) that was 90% predictive of the category label, such that 90% accuracy could be achieved by using that feature alone. However, the predictiveness of the other features differed as a function of *Structure* condition. In the RULE condition ($N = 294$), all other features were completely non-predictive (i.e., the value of that feature predicted a given label 50% of the time). As such, the best performance in the RULE condition would be achieved by identifying the single predictive feature and making categorization decisions using only it. By contrast, in the FAMILY resemblance condition ($N = 301$), all features were 90% predictive, and as a consequence the best possible performance was achievable by aggregating the information provided by all features. Finally, in the INTERMEDIATE condition ($N = 291$), the other features were 70% predictive. Thus, one feature was most diagnostic but it would be theoretically possible to achieve better performance by using all of the features in concert.



Is this a bivimia or a lorifen?

Question: 1/100
Block: 1/5

Correct: 0%
Score: 0

**Figure 4.2:** An example of a trial from the 16-*Feature* condition. Participants were asked to classify the amoeba as either a *bivimia* or a *lorifen*. In one version of the experiment, a green timer on the right was displayed to incentivize participants to respond faster for additional points, which they were given for correct answers only. Another version of the experiment was also run that did not include the timer.

## Procedure

The experiment consisted of five blocks of 20 learning trials each, resulting in a total of 100 trials. On each trial people were presented with an amoeba as shown in Figure 4.2 and were instructed to classify it as either a bivimia or a lorifen.[*] People received points for correct answers but did not lose points for incorrect ones. In one version of the experiment ($N = 439$) people were given as much time as they wanted to respond; in the other ($N = 447$), they were still given as much time as they liked but they saw a timer (the green bar on the right of Figure 4.2) that slowly decreased. In the version with the timer, they received more points for faster answers. There were no differences in performance between these two versions so the data was pooled and results reported are from the combined dataset.

[*] In all conditions the values for each feature were generated probabilistically and independently of one another subject to the constraint that they on average achieve the desired level of predictiveness. Moreover, the stimuli for each person were generated randomly according to the appropriate category structure, rather than pre-generating 100 specific stimuli and showing the same ones to everybody.

Participants were given feedback which was displayed for three seconds. It consisted of a correct or incorrect message, the number of points earned, the correct category label, and a change to the color of the circular base of the amoeba to indicate category membership (blue for bivimias and purple for lorifens). Before the next trial was displayed, a blank screen was shown for one second. At the end of each block of 20 trials, people were given a short summary of their performance, showing them their accuracy and points earned in the current block and all previous blocks.

### 4.2.2 Results

Participants learned well in all conditions, with accuracy increasing across training block (Figure 4.3). We quantified this effect through the use of Bayesian mixed effects model comparison in which we compared a baseline model that contains only a random intercept for each participant to a model that includes a continuous effect of *Block*.[†] The Bayes factor for this comparison ($BF > 10^{77} : 1$) overwhelmingly favors the model that includes an effect of *Block*. The posterior estimate of block shows a positive slope of 0.025 (95% CI is 0.023 to 0.028) indicating that average accuracy increased by about 2.5% for each block of training.

While it is reassuring that there is a general improvement in accuracy throughout training, one of our main questions was whether accuracy differed as a function of category *Structure*. Figure 4.3 suggests two things: first, that accuracy in the FAMILY structure is much higher than the INTERMEDIATE and RULE category structures; and second, that the learning rate may be identical across all category structures. To investigate the first issue, we evaluated whether there was an effect of *Structure* on accuracy. Indeed, a model with two predictors (*Structure*, coded as a three-level categorical variable, and *Block*) is strongly preferred over a model containing only *Block* ($BF > 10^{90} : 1$). Posterior estimates reveal that accuracy in the FAMILY condition is 0.16 higher (95% CI is 0.12 to 0.20)

[†] All mixed effects models in this paper assume a random intercept for each subject. Bayes factors were calculated using the default parameters (Rouder et al., 2012; Liang et al., 2012) of the BayesFactor package 0.9.12-2 (Morey & Rouder, 2015) in R 3.2.3.

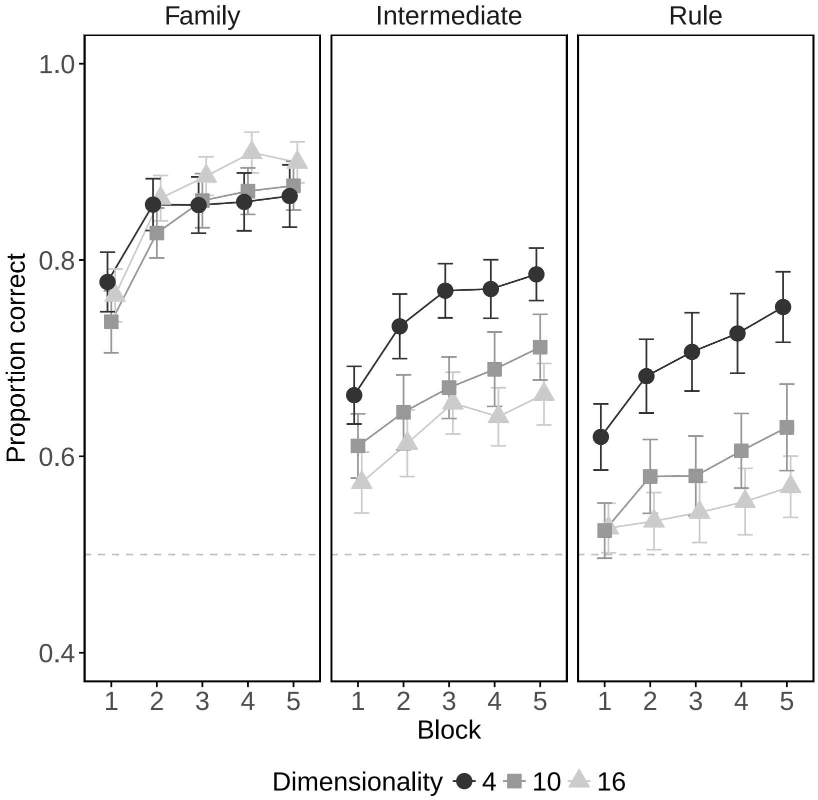


**Figure 4.3: Accuracy in Experiment One**. Human learning across the *Dimensionality* and *Structure* conditions. While learning within the *Family* resemblance categories was unaffected by the number of features, more features led to poorer performance in the *Rule* and *Intermediate* categories. Error bars depict 95% confidence intervals, and the dotted line reflects chance performance.

than the INTERMEDIATE structure, which is slightly higher than the RULE (0.08 more, 95% CI from 0.04 to 0.12). In order to investigate the second issue, we compared the *Structure* and *Block* model to a more complex model that also included an interaction between *Structure* and *Block*. The model without an interaction is strongly preferred ($BF = 40 : 1$), suggesting that the rate of learning across blocks is not different in the three *Structure* conditions.

Our second question was whether there is evidence for an effect of stimulus *Dimensionality* on performance. We found that there was: a model containing *Dimensionality* (coded as a three-

|  | Family | Intermediate | Rule |
|---|---|---|---|
| Bayes factor | 0.3:1 | $10^8 : 1$ | $10^{11} : 1$ |
| 4 vs. 10 | 0.06 (-0.05 to 0.06) | 0.08 (0.01 to 0.14) | 0.10 (0.03 to 0.19) |
| 10 vs. 16 | -0.02 (-0.08 to 0.03) | 0.04 (-0.03 to 0.10) | 0.03 (-0.04 to 0.11) |
| 4 vs. 16 | -0.02 (-0.08 to 0.04) | 0.11 (0.05 to 0.18) | 0.15 (0.07 to 0.22) |

**Table 4.1:** Bayes factors and parameter estimates for the post-hoc analyses of the effect of stimulus *Dimensionality* for each category *Structure* in Experiment One. The first row indicates the Bayes factor in favour of a model with Dimensionality and Block as predictors relative to a model with only Block; all other rows show the posterior estimates of the differences between Dimensionality conditions within that category structure. Results indicate that the effect of stimulus dimensionality was larger in the *Rule* and *Intermediate* than the *Family* category structure. The 95% confidence interval estimates are shown inside the brackets.

level categorical variable) and *Block* was strongly preferred over a model containing only *Block* ($BF > 10^7 : 1$). The posterior estimates of the effect of number of dimensions show that the only reliable difference was between the 4-feature and 16-feature conditions, with the 4-feature one being on average 0.08 more accurate (95% CI is 0.03 to 0.13); all other 95% confidence intervals of the difference span zero (4-feature vs. 10-feature and 10-feature vs. 16-feature).

Of course, we are less interested in whether dimensionality or category structure *alone* has an effect on learning, and most interested in whether there is an interaction: do more stimulus dimensions, as hypothesized, hurt learning in the rule-based structures but not in more family resemblance ones? To evaluate this, we compared a Bayesian mixed effects model containing *Block*, *Structure*, and *Dimensionality* alone to a model with these three variables plus an interaction term between *Structure* and *Dimensionality*. This shows strong evidence in favor of the model containing the interaction ($BF > 10^8 : 1$), indicating that additional features have different effects on learning in different category structures.

What differences drive this interaction? Figure 4.3 suggests that accuracy in the intermediate and rule category structure conditions decreases much more strongly as the number of stimulus features increases. In order to investigate this quantitatively, we conducted a post-hoc analysis of the effect of *Dimensionality* on accuracy within each category structure. The results, shown in Table 4.1, indicate the Bayes factor in favor of a mixed effects model containing *Dimensionality* and *Block* relative to a model containing only *Block*. The results show that for the INTERMEDIATE and RULE structures, the model with the higher Bayes factor inludes the *Dimensionality predictor*, suggesting that the number of features affects learning for these structures. However, for the FAMILY structure the preferred model based on its Bayes factor is one without *Dimensionality* as a predictor. This is consistent with our hypothesis that additional features should hurt learning much more strongly when categories do not follow a family resemblance structure.

### 4.2.3 Summary

Experiment One suggests that increasing the number of features has a differential impact depending on the underlying category structure. In the two conditions that contain a single highly predictive feature and other features that are less predictive (RULE and INTERMEDIATE), learning is clearly improved when there are fewer features overall. This is most evident in the final two columns of Table 4.1, which show 10-14% increases in overall accuracy for learning from four rather than 16 features in the INTERMEDIATE and RULE conditions. The same advantage does not occur in the FAMILY resemblance condition.

The fact that learning was not impaired in the FAMILY resemblance category structure may not be particularly surprising, given that all features were equally useful and there were no features that were less predictive. In that sense it is the lack of *advantage* for more features that is perhaps more surprising, especially since other studies have shown a learning advantage when there are additional features (e.g., Hoffman & Murphy, 2006; Hoffman et al., 2008). One possibility here is that performance in the FAMILY condition reflects a ceiling effect. Since all of the features were 90% predictive, it could be that the task was quite easy no matter how many features there were. We test this directly in Experiment Two by investigating only family resemblance structures, but manipulating the degree to which the features are predictive of the category label.

## 4.3 Experiment Two

This experiment explicitly tests whether additional features have an effect on category learning within family resemblance categories when the features are less predictive than in the previous experiment. If there is no effect of the number of features on learning, we can be more certain that the differences due to category structure found in Experiment One are actually due to category structure rather than to the informativeness of the features. We test this by systematically manipulating the predictiveness of the features. Does this affect the degree to which additional features affect learning?

### 4.3.1 Method

#### Participants

888 people (459 male, 425 female, 4 other) were recruited via Amazon Mechanical Turk. As before, the high number of participants reflects the fact that we ran two experiments with slightly different methodologies and pooled the results since they were qualitatively identical ($N = 436$ for the version without the timer, $N = 452$ for the version with it). Participants ranged in age from 19 to 74 (mean 34.6). They were paid US$2.00 for completion of the task, which took 12 minutes. Data from an additional 37 participants were excluded from analysis, either from failure to complete the task (32 participants) or participating in an earlier version of this study (5 participants).

### Design

The task and stimuli were identical to Experiment One, with participants randomly allocated in a 3 × 3 between-participants design. As before, we manipulated the *Dimensionality* by altering the number of features present in the stimuli to make three conditions: 4-FEATURE (*N*=286), 10-FEATURE (*N*=327), and 16-FEATURE (*N*=275). Unlike before, all the category structures were family resemblance structures, with all features being equally predictive of the category. This time we manipulated the degree of *Predictiveness* to make three conditions: 70% predictive (*N*=310), 80% predictive (*N*=258), and 90% predictive (*N*=320). The 90% condition was a replication of the FAMILY structure in Experiment One.

### Procedure

The procedure was identical to Experiment One. Similar to the previous experiment, in one version of the experiment ($N = 436$), there was no time limit for providing a response on each trial. In the other version of the experiment ($N = 452$), there was still no time limit, but they saw a timer that slowly decreased (see Figure 4.2), and they received more points for faster responses.

### 4.3.2 Results

How was learning affected by *Dimensionality* and *Predictiveness*? We evaluated this question by comparing Bayesian mixed effects models that included some combination of *Block* as a continuous variable and *Dimensionality* and *Predictiveness* as discrete variables. Reassuringly, we found that people did indeed learn over the course of training: a model including *Block* was strongly preferred over a model that only contained a random effect for each participant ($BF > 10^{74}$:1). As before, posterior estimates suggest that average accuracy increased by about 2.3% for each block of training (95% CI is 0.021 to 0.026).

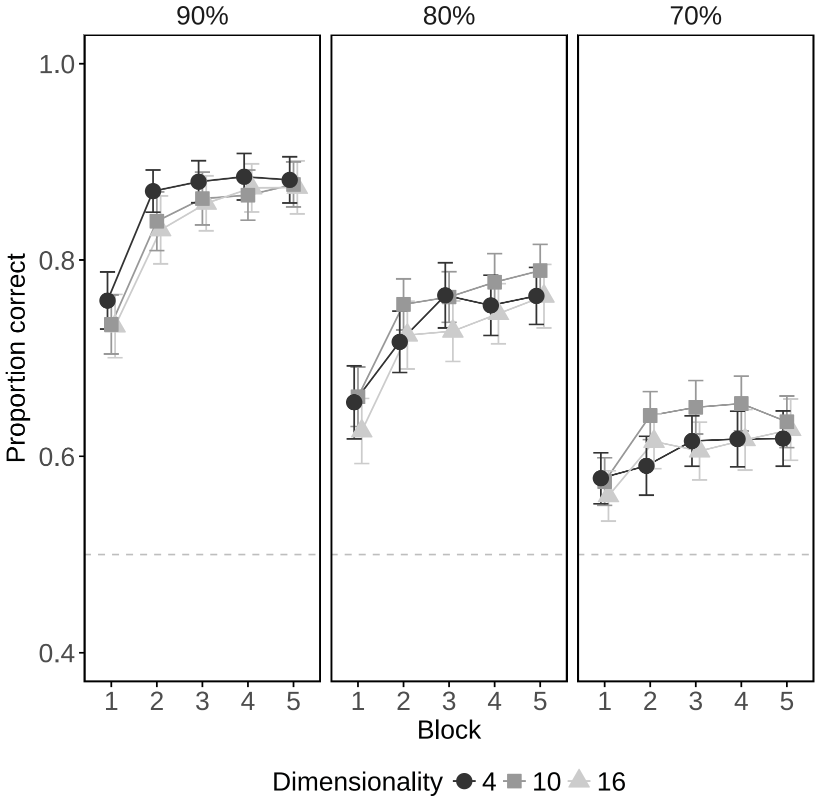


**Figure 4.4: Results from Experiment Two**. Mean accuracy across the three *Predictiveness* and *Dimensionality* conditions. While the mean performance decreased as the level of *Predictiveness* was reduced, within each *Predictiveness* condition there was no change based on the number of features. Error bars show 95% confidence intervals, and the dotted line reflects chance performance.

How did the *Predictiveness* of features affect learning? As Figure 4.4 shows, and as one would expect, overall learning was lower in categories with lower predictiveness. This is borne out in a Bayesian model comparison between a model with *Predictiveness* (coded as a three-level categorical variable) and *Block* as compared to a model with only *Block* as a predictor. The two-predictor model was strongly preferred ($BF > 10^{116} : 1$), and the posterior estimates indicate that accuracy was 11% higher for both the 90% to the 80% condition (CI is 0.07 to 0.14) as well as the 80% to the 70% condition (CI is 0.08 to 0.15).

Our main question, of course, was whether additional number of features had an impact on categorization accuracy. Figure 4.4 suggests that *Dimensionality* does not have an effect on learning, and a Bayesian mixed effects model comparison confirms this: a model containing only *Block* was preferred ($BF > 17 : 1$) over a model containing both *Dimensionality* and *Block*. This is further supported by post-hoc analyses that find strong preference for a model containing *Block* and *Predictiveness* predictors over all models containing *Dimensionality*.

### 4.3.3 Summary

Experiment Two provides strong evidence that the number of features does not affect learning when the category structure follows a family resemblance pattern, and that this cannot be attributed to a ceiling effect. Interestingly, learning in the 70% predictive family resemblance category structure is only slightly above chance ($M$ = 0.63 in the final block). Despite the fact that there was evident room for improvement, there was no benefit of increasing the number of features, suggesting that these results do not reflect a floor effect either. Taken in conjunction with Experiment One, these findings suggest that the curse of dimensionality affects people more as the category structure grows more rule-based. Indeed, in family resemblance categories, there appears to be no detrimental effect of additional features at all (but neither is there much benefit).

## 4.4 An ideal observer model for this task

In this section, we explore the predictions of an ideal observer model for this particular task, which we call the OPTIMAL model, that makes full use of the information from the task. The purpose of this analysis is to establish the achievable limit of classification performance for the various kinds of category structures in Experiments One and Two, and how human performance compares to the ideal limit.

### 4.4.1 Notation

We briefly describe the notation used to describe the OPTIMAL model. The input for each trial is a $D$-dimensional stimuli vector $\mathrm{x} = (x_1, x_2, \ldots, x_D)$, where $D$ is the dimensionality of the stimulus and each $x_i$ is a binary feature, i.e. $x_i \in \{0, 1\}$. The predicted category response $\hat{y} \in \{0, 1\}$ for trial $N$ is defined by the feature information from trial $N$ along with the representation learned by the model based on the previous $N - 1$ trials.

### 4.4.2 Model

Next, we describe the details of an ideal statistical learner. In our experiments, the stimuli were generated by following the principle of class-conditional independence (e.g., Anderson, 1990; Jarecki, Meder, & Nelson, 2013). As long as one knows the true category label $y$, then the probability of any particular feature value $x_i$ is completely independent of any other feature. As a consequence, every category can be represented in terms of a single feature vector $\vartheta = (\vartheta_1, \ldots, \vartheta_D)$ where $\vartheta_i = p(x_i|y)$ describes the probability that feature $i$ will have value $x_i$. Although class-conditional independence is not always satisfied in real life where feature correlations are possible (Malt & Smith, 1984), it is a reasonable simplification in many situations (Jarecki et al., 2013), and one that is appropriate to our experimental design. Moreover, because the category can be represented in terms of a single idealised vector $\vartheta$ that describes the central tendency of the category, it is broadly similar to standard prototype models (Posner & Keele, 1968).[‡]

[‡] Although we do not explicitly evaluate any exemplar models (Nosofsky, 1986) or mixture models (Sanborn et al., 2010; Love et al., 2004), we expect that their behavior would be very similar to the prototype model on these category structures. Exemplar models are perfectly capable of learning prototype-like category structures (Nosofsky, 1988), and as such we would not expect this experimental design to be predictive as regards to the prototype vs. exemplar distinction. Rather, we expect that the lessons implied for the optimal model would turn out to be similar for exemplar models and any other sufficiently rich statistical

Formally, we implement this statistical learning model using a naive Bayes classifier which makes the same assumption of class-conditional independence. In it, the posterior probability that novel object $x$ belongs to the category $y$ is given by:

$$p(y|\mathrm{x}) \propto \prod_{i=1}^{D} p(x_i|y)p(y) \tag{4.1}$$

where the marginal probability $p(x_i|y)$ is given by the posterior expected value of $\vartheta_i$ given the previously observed category members. Specifically, if the learner has observed $n_y$ previous exemplars that belong to category $y$, of which $n_{yi}$ were observed to have the feature $x_i$, then the model estimates the following probability:[§]

$$p(x_i|y) = E[\vartheta_i|n_{yi}, n_y] = \frac{n_{yi} + 1}{n_y + 2} \tag{4.2}$$

Applying a similar logic, the model learns the base rate of the category labels over time, and so the prior probability $p(y)$ of category $y$ is computed by applying a (smoothed) estimate of the observed base rate so far:

$$p(y) = \frac{n_y + 1}{n + 2} \tag{4.3}$$

Finally, as an ideal observer model, the statistical learning model is assumed to always choose the category label with highest posterior probability, and thus the response $\hat{y}$ is selected deterministically by applying the rule:

$$\hat{y} = \arg\max_{y} p(y|\mathrm{x}) \tag{4.4}$$

learning model.

[§] Formally, this expression arises if the learner places a uniform prior over an unknown Bernoulli probability $\vartheta_i$ and updates those beliefs via Bayes' rule. It is equivalent to the Laplace smoothing technique.

The OPTIMAL model is appealing for two reasons. Firstly, it serves as an ideal observer model for this experiment, insofar as it is a statistical learning model whose structure precisely captures the structure of the task (i.e., conditional independence) and learns the specific categories by applying Bayes rule. As such it can reasonably be claimed that the performance of this model represents the upper bound on what might be achievable in the learning task. Secondly, because of its connection to prototype models, it may be taken as a representative of a broad class of "family resemblance models" that have dominated the theory of category learning since the 1970s. This model is not intended to be a fully general model of human categorization, but rather to act as a gold standard to compare to human performance in these tasks across different levels of dimensionality, category structure, and feature predictiveness.

### 4.4.3 Model results

The OPTIMAL model was simulated 10,000 times in each of the experimental conditions from both experiments, where each simulation mimicked a 100-trial experiment. On each trial, a new stimulus was generated in exactly the same manner as the experiment. The model then made a prediction of the category label of the current stimulus, and then received feedback which it would use to update its category representation.

We begin by looking at the predictions of the ideal observer model across both tasks. Figure 4.5 (top row) shows that for Experiment One, the model predicts that performance is highest in the FAMILY category structure, followed by the INTERMEDIATE structure, with the worst performance in the RULE structure. Similarly, as Figure 4.6 (top row) reveals, the model predicts the best performance in the 90% condition and the least in the 70% condition.

However, despite human performance matching the qualitative trends as predicted by the OPTIMAL model across *Structure* for Experiment One, and *Predictiveness* for Experiment Two, the predictions of the OPTIMAL model deviate from human performance in a number of ways. This is

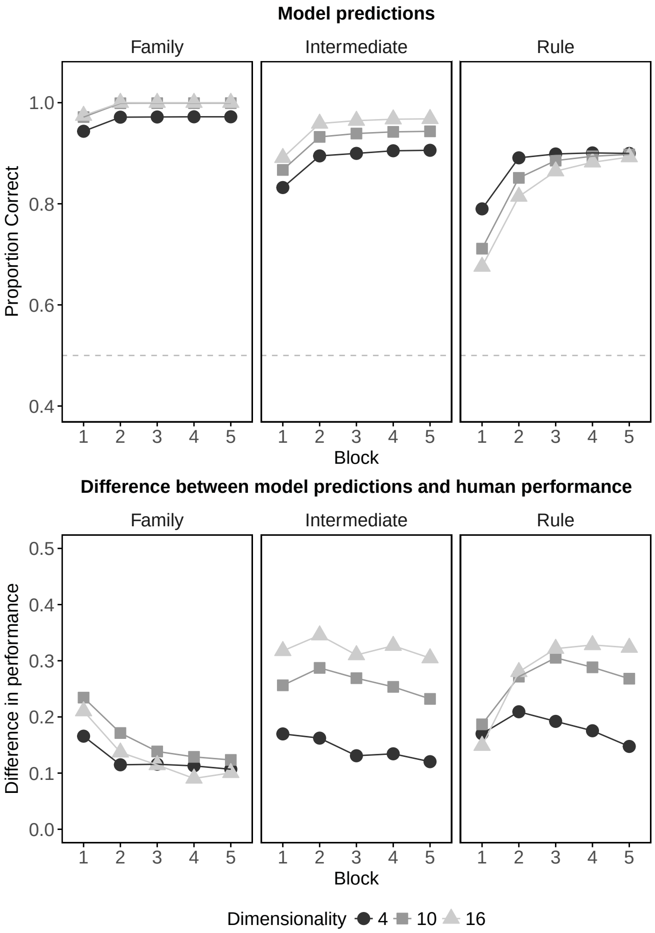


**Figure 4.5: Comparison of *Optimal* and human performance in Experiment One.** Each panel on the top row shows human performance (circles) compared to performance of the *Optimal* model (triangles) in one of the nine conditions we ran in Experiment One. The model captures the fact that performance in the *Family* condition is higher than the *Intermediate* condition, which is higher than the *Rule* condition. However, it deviates from human behaviour in two notable ways: its performance is far higher than human performance across all conditions, and it predicts that performance should improve with more features in the *Family* and *Intermediate* conditions where the additional features are more informative. The panels on the bottom row show how the *Optimal* model deviates from human performance. While human performance tracked the model's predictions for the *Family* condition, for the *Intermediate* and *Rule* conditions the model's performance was much higher, especially as the number of features increased.

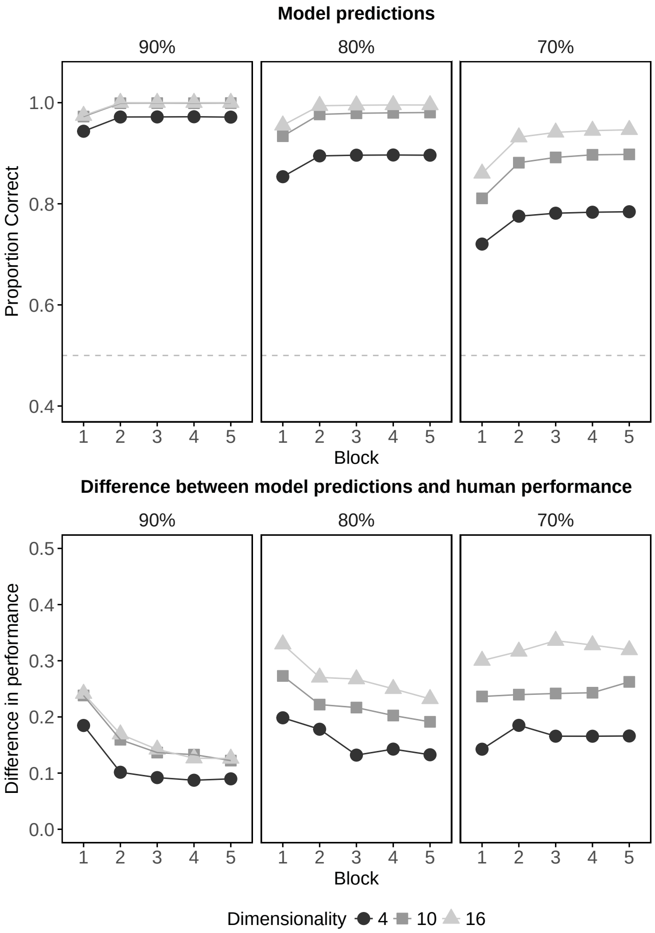


**Figure 4.6: Comparison of *Optimal* model and human performance in Experiment Two.** Each panel on the top row shows human performance (circles) compared to performance of the model (triangles) in one of the nine conditions we ran in Experiment Two. Similar to the model prediction results of Experiment One, while the ideal observer model matches human behaviour in showing that lower *Predictiveness* results in lower performance, its performance far exceeds human performance, especially as the number of features increases. The panels on the bottom show how the predictions of the *Optimal* model deviates from human performance. Again, we find that as the *Predictiveness* decreases the model systematically overpredicts performance for categories with a higher number of features.

illustrated in the panels on the bottom rows of Figures 4.5 and 4.6, where the difference between the model predictions and human performance are plotted. First, the OPTIMAL model consistently predicts that performance should be far higher than what we empirically observe, as the difference between the model predictions and human performance is above zero for every condition. Second, for all of the conditions where additional features are helpful for categorization, the model shows increasing performance for additional features. On the other hand, for humans the effect of additional features hurts performance when most of the features are weakly predictive (like the INTERMEDIATE structure from Experiment One), or that performance stays the same (as observed for all of the family resemblance structures). In addition, for the RULE condition in Experiment One where additional features are not informative, humans show decreasing performance for more features, whereas the ideal observer model predicts that by the end of the experiment should result in equal performance across the *Dimensionality* conditions. This analysis highlights that human behaviour is far from the maximum performance that is achievable as demonstrated by the ideal observer model. We discuss possibilities on how computational models could explain this pattern of results in the discussion section.

## 4.5 Discussion

The term "curse of dimensionality" has been applied to a range of problems in machine learning, statistics and engineering, all of which share the common property that the space of possible solutions to an inference problem grows extraordinarily rapidly as the dimensionality increases. The same phenomenon applies to human category learning, and our goal in this paper has been to explore how the curse plays out for human learners.

At an empirical level we observed a clear pattern in which the curse of dimensionality is strongly mediated by the structure of the categories that need to be learned. Rule-based categories, in which only a small number of features are relevant for predicting category membership, are heavily affected by dimensionality because the search problem (identifying the predictive features) becomes harder as more features are added. In contrast, the number of features does not appear to affect learning for family resemblance categories in which all features are somewhat predictive of category membership. Comparing our results to an ideal observer model that could attain the highest possible performance, we find numerous instances in which people's behaviour deviates from the model's predictions and suggests that people may be employing a different strategy to learn the categories.

One of the main conclusions from this work is that it is not very meaningful to discuss the effect of dimensionality without considering what kind of category structure is being learned. In fact, the role of category structure may help explain away apparent differences in the literature. For instance, previous research indicating that additional features hurt performance used features that were not predictive of category membership (Edgell et al., 1996), consistent with our RULE condition in Experiment One. Other studies that found that adding features did not have any effect used family resemblance categories, consistent with the FAMILY conditions in both of our experiments (Hoffman and Murphy (2006) Experiments 1 and 2, Minda and Smith (2001) Experiment 4, Hoffman et al. (2008) Experiment One). There were only two results we were not able to replicate, both reflecting an improvement in learning with more features (Hoffman and Murphy (2006) Experiment 3 and Minda and Smith (2001) Experiment One). However, in both of those studies, other aspects of category structure covaried with the number of features, providing an alternate explanation for results that differed from ours. For instance, in Experiment One of Minda and Smith (2001), the categories with fewer features were less structured, and when this confound was addressed in Experiment 4 of that paper, the effect went away.

### 4.5.1 Broader implications for human learning

The fact that human learning deviates systematically from the ideal observer model is theoretically interesting and highlights an important difference between real world learning and many category learning experiments. Both our experiments had features with *class-conditional independence*, in which the stimulus features are conditionally independent of one another as long as one knows the category to which the stimulus belongs. This assumption does not hold in general, but in some situations it might provide a reasonable first approximation. Indeed, people do appear to assume class-conditional independence, at least at first, in some category-learning tasks (Jarecki, Meder, & Nelson, in press).

However, from a computational modeling perspective, it is important to recognize the limitations that this assumption imposes: the reason that our ideal observer model is able to perform *better* on family resemblance categories as the number of features increases is that it exploits the fact that every additional feature conveys independent information about the category. When class-conditional independence holds, family resemblance categories become easier to learn as the dimensionality increases. This does *not* match the pattern we observed in our data, in which people's performance on family resemblance categories was the same regardless of the number of features in the stimuli.

One possible explanation is that is that the class-conditional independence structure in our experiments does not match the true relationship between features and categories for most categories the world. If that is the case, it would imply that perhaps people are adaptive by relying on a single or small set of features when learning categories. Depending on the relationship among features and between them and category labels, in real life one might end up making better categorization decisions by using a limited number of predictive features rather than attempting to process all information inherent in the stimuli. In other words, human learners might differ from our optimal statistical learning model not because of the limits of human cognition but because human cognition is shaped by an environment in which class conditional independence is a poor assumption, and that human learners are better described by other kinds of inductive biases.

The question of what inductive biases are required to explain how humans are affected by the curse of dimensionality in some cases and not others is beyond the scope of this paper, but we can speculate about possible answers. One possibility is an inductive bias for sparsity (Gershman, Cohen, & Niv, 2010), which assumes that only one (or a limited) number of features is relevant for categorization. Thus, the relevant features for this task could be learned through selective attention, a process where attentional weights for particular features increase or decrease based on their ability to make correct classification decisions. This kind of approach has been successfully employed by a number of existing models of categorization to explain other patterns in human category learning (Nosofsky, 1986; Kruschke, 1992), and is a potential future avenue of exploration for a more complete explanation of how people learn categories with many features.

A second, alternative approach towards lifting the curse dimensionality is to reduce the number of features that are represented or encoded in the first place. Such methods focus on reducing the number of dimensions via manifold learning (Tenenbaum, 1997) or using structured representations (Kemp & Tenenbaum, 2009; Tenenbaum, Kemp, Griffiths, & Goodman, 2011; Lake et al., 2015). These kinds of approaches have been pursued with considerable success in semantic representation (Griffiths, Steyvers, & Tenenbaum, 2007; Landauer & Dumais, 1997). It is, of course, possible that human learning is versatile enough to incorporate the fundamental insights from both exploiting limited memory and attentional capacities and reducing the effective dimensionality of incoming stimuli. Pursuing these issues further is a matter for future work.

# 5
# Discussion

In this thesis, I explored how the interaction between information from category structure and category labels can influence the acquisition of category knowledge. Each chapter explored a different computational problem to be solved, motivated by the ways in which structure and label information vary for real-world categories. This final chapter is devoted to summarising and connecting the work between each of these three chapters to the broader literature on category learning. I conclude by discussing some future directions related to how other kinds of structure and label information might shed light on other important aspects of real-world category learning.

As a quick summary for the reader, I briefly review the results from each chapter:

- Chapter 2 focused on the computational problem of semi-supervised learning. Suppose a learner is given a set of objects, where some of the objects are provided with category labels but others are not. How should a learner sort these objects into meaningful categories? This research question captures one aspect of category learning in the real world, namely that the input a child receives about the world is mostly unsupervised, with some small amount of supervision. The experimental results in this chapter revealed that when the category structure was sufficiently distinct, participants sorted the objects into the same kinds of categories regardless of what category labels were presented. However, when given an ambiguous category structure instead, participants relied more heavily on the category label information as a guide to how the objects should be organized into categories. I found that the results across the different structure and label conditions could be predicted by a modified version of Anderson's Rational Model of Categorization that could accommodate for both information from labeled and unlabeled examples.

- In Chapter 3, I examined the computational problem of learning structure from labeled examples. Many real-world categories can be organized into different structures, such as trees, or rings, or cliques, but this computational problem has mostly been considered a problem of unsupervised learning. In this chapter, I looked at how labeled information might provide additional useful constraints to help people learn different kinds of structure. Across two experiments, participants were presented with objects that were either organized as a taxonomic or cross-cutting structure in a supervised category learning paradigm, and were asked to learn all of the category labels. The results from the first experiment showed that participants could learn both kinds of structures, and what they had learned influenced how they generalized to new objects. The second experiment showed that while structure knowledge didn't transfer across tasks, other aspects allowed them to learn faster regardless.

- Finally, in Chapter 4, I looked at the curse of dimensionality in category learning. The computational problem in this section explored how people could learn categories with many features, in a way that more closely resembled the complexity of real-world categories. Across two experiments, I looked at categories with different category structures and a varying number of features. I showed that when the categories followed a family resemblance structure, people were unaffected by the presence of additional features in learning. However, when the categories were rule-based, they succumbed to the curse and performed worse with additional features. I compared and contrasted these results to an ideal observer model for this category learning task, showing how far human performance deviates from the ideal standard across the different conditions.

Across these three chapters, there are a number of connections that are worth emphasizing and relating to the wider literature. I consider three overarching themes related to how structure and label information influence learning in this thesis: labels as cues for category structure, insights from realistic category structures, and structure as a determinant of learnability.

## 5.1 Theme 1: Labels as a cue for structure

The first major theme of this thesis is that category labels provide cues to structure. In Chapters 2 and 3, the goal for participants was to learn the underlying latent structure that governed how objects were organized into different categories. In Chapter 2, the categories were organized as different clusters in a two-dimensional space, while in Chapter 3, the categories were organized into taxonomies and cross-cutting structures where categories were based on different sets of features. In both cases, the set of category labels given to participants were designed to be informative for the task in different ways, either by labelling particular stimuli (as in Chapter 2), or giving them multiple labels (as in Chapter 3). Traditionally, the learning problems explored in these two chapters — clustering and structure learning — have been studied from an unsupervised learning perspective (Sanborn et al., 2010; Kemp & Tenenbaum, 2008; Shafto et al., 2011). However, using only feature information as the basis to determine the underlying structure of a set of categories is computationally infeasible due to a combinatorial explosion in the number of possible structures. In this thesis, I argue that category labels can play an important role in dealing with the complexity by providing additional constraints for this problem. How do the results from these two chapters align with existing theories of the role of labels in concept and category learning? I describe two different theories of category labels using the framework in Lupyan and Lewis (2017): *labels-as-mappings* and *labels-as-cues*.

The labels-as-mappings theory suggests that the role of labels is to map already-learned categories in the mind to a category label (Bloom, 2000b; Xu & Tenenbaum, 2007b). This theory argues that people's knowledge of categories is first acquired in an unsupervised manner through nonlinguistic means, and that the purpose of category labels is merely a problem of discovering which labels to map onto which categories. Thus, the challenge of the labels-as-mappings view is figuring out exactly which category labels to map to which existing categories in the mind.

In contrast, rather than presupposing that humans learn categories in an unsupervised fashion and then map linguistic tokens to these categories, the labels-as-cues theory argues that labels act as cues to help the learner *construct* the correct meaning of a category. Thus, the use of category labels is combined with information from category structure to figure out the problem of mapping labels to objects and generalizing the meaning of a label simultaneously. One piece of supporting evidence for this theory is from cross-linguistic studies, where research has argued there may not be any common vocabulary that is common across all languages (Borin, 2012), suggesting that there are few, if any categories that are truly acquired without any kind of language.

The results from Chapters 2 and 3 provide partial evidence to both of these theories. In Chapter 2, participants in the DISTINCT structure condition did not need to rely on any label information to sort the stimuli into categories, suggesting that people may have merely mapped the labels onto the existing categories (at least for the conditions where they were presented with category labels). On the other hand, participants in the AMBIGUOUS structure condition categorized the stimuli in a way that was heavily influenced by what kinds of labels they saw, indicating that they may have used the labels as a cue to determine the categories in the task. In Chapter 3, despite using the same stimuli for both *Structure* conditions, people categorized and generalized on the basis of what they had learned, favouring the labels-as-cues hypothesis. However, another interpretation is that people nay have had an inductive bias towards particular types of category organization (such as mutually exclusive ones), but that this inductive bias was swamped by sufficient data suggesting otherwise.

Are there computational models that capture these different effects of how category labels affect category learning? I presented a modified version of the RMC that accounted for the results across most conditions in Chapter 2, although its ability to handle some of the label conditions in the AMBIGUOUS structure condition was less impressive, suggesting there is room for future improvement. While I did not build a computational model to help explain the behavioural results in Chapter 3, one starting point may be to borrow some of the ideas from the semi-supervised RMC in Chapter 2. However, instead of assuming that the categories are mutually exclusive as the Chinese Restaurant Process prior does, one could use a different prior that allowed for overlapping categories. Finally, while the labels-as-mappings theory seems to be the default viewpoint in word learning (Barrett, 1986; Bloom, 2000b; Xu & Tenenbaum, 2007b), the evidence presented in this thesis suggests that there is a lot of interesting future research to be done on how people use category labels as cues to construct categories instead.

## 5.2 Theme 2: Insights from realistic category structures

The second theme of this thesis is studying richer category structures similar to the kinds of structure one might encounter with different real-world categories. The kinds of category structures and stimuli used in category learning research have focused on a limited number of simple, but well-calibrated structures. The research presented in these two chapters look at whether the kinds of empirical results from category learning experiments involving simple structures can generalize to more complex kinds of category structures.

In Chapters 3 and 4, I studied categories that were designed to resemble some of the ways in which real-world categories exhibit interesting structure. In Chapter 3, I examined category learning in two kinds of conceptual structure: taxonomic and cross-cutting structures. Both taxonomic and cross-cutting structures have stimuli that can belong to multiple categories, rather than assuming mutual exclusivity. This makes these categories more difficult to learn, but also more closely matches the kinds of systems of categories that exist in the real world. In addition, many different experiments examining people's knowledge of real-world categories point to evidence of rich, semantic knowledge that is structured in these complex ways (Osherson et al., 1990; Ross & Murphy, 1999). Yet, little is known about how people learn and acquire category knowledge of such systems. In the experiments conducted in Chapter 3, people were able to learn these structures from labeled examples and generalize in an appropriate manner. In Chapter 4, I investigated whether people could learn category structures that consisted of many more features than previously studied. Past empirical work has shown that people possess knowledge of many different features from real-world categories (Rosch, Mervis, Gray, Johnson, & Boyes-Braem, 1976; Malt & Smith, 1982). Again, little previous work had tackled the question of whether people could even learn categories with many features. My results show that the main predictor of being able to learn categories with many features is the underlying structure of the categories.

What insights do the results of Chapters 3 and 4 contain about the study of more realistic category structures? Despite the fact that both of these studies investigated learning with rather different and complicated categories, there was strong evidence that people were able to learn them within the timeframe of the experiment. This suggests that studying the dynamics of how people learn richer category structures is experimentally feasible, in line with other recent work in this area (Kemp, Tenenbaum, Niyogi, & Griffiths, 2010; Canini & Griffiths, 2011; Piantadosi et al., 2016). One exciting result were the transfer effects I observed in Chapter 3, where people who had learned the structure in the first phase were able to more quickly learn the structure from the second phase. This ability of *learning-to-learn*, where people are able to leverage the information previously learned to help speed up learning has been argued to be one of the hallmarks of human intelligence (Lake, Ullman, Tenenbaum, & Gershman, 2017).

## 5.3 Theme 3: Category structure as a determinant of learnability

The third and final theme of this thesis is how category structure determines the learnability of categories. This is not a novel insight from this thesis by any means, but rather adds growing support to an active area of inquiry in category learning (J. Feldman, 2000; Chater & Vitányi, 2003; Pothos & Chater, 2002; Kemp, 2012). In category learning, a number of researchers have proposed different methods to predict the learnability of categories through different measures of category structure, through representational languages from boolean algebra (J. Feldman, 2000) to first-order predicate logic (Kemp, 2012; Piantadosi et al., 2016). In Chapters 2 and 4, I find that the learnability of categories is strongly related to category structure, albeit in slightly different ways.

In Chapter 2, one of the main experimental manipulations was to use two different kinds of category structure, a DISTINCT structure where clusters of stimuli were well separated in feature space, and an AMBIGUOUS structure where clusters of stimuli were closer together in feature space. By measuring how consistent the responses were from participants within each condition, I found that there was far greater consistency in the DISTINCT structure condition compared to the AMBIGUOUS structure condition. In Chapter 4, I explored a range of category structures, varying both the number of features in the categories (4, 10 or 16 features), and the kind of category structure, ranging from family resemblance categories where most features were correlated with one another, to rule-based categories where only a single feature was relevant for categorization. Here, the results showed that performance can be explained by the structure of the categories, with performance higher for family resemblance structures than rule-based categories, and also for family resemblance structures where each of the features were more predictive of category membership.

To what extent are the results from Chapters 2 and 4 consistent with previous research using simplicity to predict learnability? The results from both chapters seem to intuitively map onto existing ideas of complexity. However, it is complicated to make a direct comparison to existing theories as the category structures used in both tasks do not easily map onto the kinds of representational languages used to compare learnability. For example, work by Pothos et al. (2011) has argued that a one predictor of unconstrained sorting tasks (such as the experiment in Chapter 2) were stimulus sets that had *low* within-cluster variance, rather than how well separated different clusters were. Yet, the sets of stimuli used in the different structure conditions in Experiment 2 differed mostly by the amount of cluster separation from each other, rather than the within category variance, driving the main difference of results I observed. One additional limitation of the findings from Pothos et al. (2011) is that they only measured behaviour in unsupervised categorization tasks, yet our results showed that consistency of sorting behaviour varied depending on the set of category labels provided to participants. One promising research problem in semi-supervised learning may be to explore how structure and label information interact in these sorting tasks can be used to predict the intuitiveness of sorting for different kinds of stimulus sets.

For similar reasons, it is not easy to predict learnability of the categories used in Chapter 4 either, as the features corresponded to the categories in a probabilistic, rather than deterministic manner. Regardless, we can observe a clear pattern of behaviour in the two experiments, where performance across conditions in both experiments can be predicted by some combination of the set features that are high in predictiveness, as well as the overall number of features.

This suggests that there are ample opportunities to try and expand the scope of existing approaches examining learnability of categories. For example, is it possible to create some measure of learnability that can predict the learnability of categories that are unsupervised, semi-supervised and fully supervised? One reason researchers have been able to make successful theories about learnability is the wealth of data from conducting the same experiments on the same category structures, and we should do the same for richer and more realistic datasets so that the same methods can be applied.

## 5.4 Future directions

In the closing section of this thesis, I discuss some other promising avenues for incorporating structure and label information into category learning, particularly in ways that solve other computational problems related to the challenge of learning real-world categories. For the role of structure, I consider ideas such as feature learning and compositionality. For the role of labeled information, I look at some possibilities offered by active learning and teaching.

### 5.4.1 Future directions for category structure

How might we hope to better understand the role that category structure plays in category learning? The standard approach to studying category structure is to explicitly define the feature values for each feature dimension, and then to use these values to generate visual stimuli that are presented to participants. This was the approach I used across all of the studies presented in this thesis. These inputs, whether they are discrete or continuous features, are then also used as inputs to computational models of category learning to generate predictions. One of the major advantages to specifying features in this manner is that it allows the researcher to easily control and vary the category structure and to test the hypotheses that they are interested in.

However, one significant disadvantage of this approach is that in the real-world, the various features of categories are not automatically extracted, but must be learned in conjunction with the categories themselves. Real-world categories are not defined by a finite set of pre-existing features, and that learning both the relevant features and categories represents a significantly more challenging kind of learning problem. Past work has shown that people are sensitive to correlations across features, and can learn to differentiate features that are relevant for categorization to those that are not (Schyns & Rodet, 1997; Austerweil & Griffiths, 2011).

Connecting to the work in Chapter 4 on learning categories with many features, how might we approach the problem of learning visual categories from images rather than a large set of feature values? Recent computational models using deep learning suggest one possibility: instead of going directly from a set of fixed inputs to a category response, deep learning models transform the raw input through a number of successive layers, that gradually encode more abstract information about the stimuli (LeCun, Bengio, & Hinton, 2015). Recent work from a number of cognitive scientists have used pre-trained deep networks to establish how the learned features from these models are predictive of phenomena such as category typicality (Lake et al., 2015) and similarity (Peterson, Abbott, & Griffiths, 2016). However, these studies have typically used pre-trained networks trained on existing image datasets, rather than examining how humans and deep learning models might learn the same kinds of feature representations from scratch.

While deep learning models are remarkable at the kinds of pattern recognition for low-level features from raw input, a number of researchers have argued that they are unable to capture the kinds of learning exhibited at more abstract levels, which display aspects of relational, compositional and causal knowledge (Lake et al., 2017). Thus, one of the major open questions in category learning is to discover ways to explore how we can integrate models of deep learning with other computational models that can handle structured representations of category knowledge.

### 5.4.2 Future directions for category labels

Similarly, what other aspects of category label information are worth studying to capture the different kinds of learning exhibited in the real world? In this thesis, I focused on how labeled information influenced learning in two ways: when some labels are provided in Chapter 2 in the case of semi-supervised learning, and when multiple labels were provided in Chapter 3 for structure learning. Yet, there are many other ways in which labels convey useful information, particularly in how that information is obtained. I consider two other ways in which the method of obtaining category labels may be relevant to real-world category learning: active learning and pedagogy.

In all of the experiments I conducted in this thesis, participants were presented the stimuli and label information in a *passive* manner, receiving information without any requiring any volition of their own. In contrast, in the real world people can also decide to gather information for themselves and learn categories in an *active* manner, searching out for the relevant kinds of information themselves to make decisions (Bruner, 1961; Nelson, 2005; Markant & Gureckis, 2014). Active learning has been argued to lead to faster and more successful learning than passive observation, as it allows the learner to hone in on the true hypothesis through more efficient queries that more effectively reduce uncertainty than random passive samples of data (Markant & Gureckis, 2014; Markant, Settles, & Gureckis, 2016). In addition, active learning is an important part of real-world category learning as learners in the real world will interact with their environments and asking the right kinds of questions in an active manner is an integral part of learning.

A second way in which labeled information can be provided learners is *pedagogically*. In a pedagogical context, a teacher explicitly reasons about how to select information that is most informative to the learner in inferring the correct hypothesis, and the learner is sensitive to how this information was generated (Shafto & Goodman, 2008; Shafto, Goodman, & Griffiths, 2014). There is a growing body of evidence that people are sensitive to pedagogical cues (Xu & Tenenbaum, 2007a; Bonawitz et al., 2011; Shafto et al., 2014; Eaves Jr, Feldman, Griffiths, & Shafto, 2016). Similar to active learning, the importance of teaching is relevant to learning in the real world as children receive much of their information in a pedagogical manner. Thus, discovering how to model these kinds of phenomena more accurately is an important avenue for the future.

## 5.5 Conclusion

So much of our knowledge about the world is wrapped in concepts and categories. How do we acquire all of this knowledge about categories? This thesis has argued that understanding how people learn categories is shaped by the combination of two factors: category structure and category labels. Through a combination of behavioural experiments and computational models, I have examined how these two factors interact and shape learning in a number of tasks inspired by the challenges learning in more realistic scenarios.

# A

# Details of the Semi-supervised Rational Model of Categorization

As discussed in Chapter 2, our version of the RMC does not assume the learner knows the number of possible labels in advance, and – much like the number of categories itself – uses a Chinese restaurant process to capture this uncertainty (see Navarro, Griffiths, Steyvers, & Lee, 2006, for an overview). This implies that the probability of observing the $j$th label for a stimulus belonging to the $k$th category is

$$P(\text{label } j \mid \text{category } k) = \frac{n_{jk}}{n_{.k} + l}$$

where $l$ is a parameter that governs the learner's willingness to tolerate differently labeled items within the same category (fixed at $l = 1$ in all simulations). In this expression $n_{jk}$ denotes the number of times the label $j$ has been observed in cluster $k$, and $n_{.k}$ is the total number of labeled examples assigned to category $k$. Relatedly, the probability of observing a new label for an item in category $k$ is

$$P(\text{new label} \mid \text{category } k) = \frac{l}{n_{.k} + l}$$

For unlabeled data, a complete solution would be to have the model treat the label as missing data, and to try to infer those labels via Bayesian inference by sampling from the posterior in the same way that the model infers the category assignment. In our applications we adopt a simplification in which the model simply computes the expected prior probability of the label that should have been assigned to that observation, integrating over all possible values for that label. For an item assigned to category $k$, this is given by:

$$\begin{aligned}
P(\text{unlabeled} \mid \text{category } k) &= E_{P(\text{label} \mid \text{category } k)}[P(\text{label} \mid \text{category } k)] \\
&= P(\text{new label} \mid \text{category } k)^2 + \textstyle\sum_j P(\text{label } j \mid \text{category } k)^2 \\
&= \left(\tfrac{l}{n_{.k}+l}\right)^2 + \textstyle\sum_j \left(\tfrac{n_{jk}}{n_{.k}+l}\right)^2 \\
&= \tfrac{l^2+\sum_j {n_{jk}}^2}{(n_{.k}+l)^2}
\end{aligned}$$

This approach is an approximation (sufficient for our purposes) that uses the prior to integrate out the learner's uncertainty about the identity of the missing labels, and is considerably simpler than the full Bayesian solution that would use the full joint posterior distribution over all unobserved quantities to achieve the same end.

In all other respects, including parameter values, the model we used is identical to the version of the RMC described by Sanborn et al. (2010), in which we used Markov chain Monte Carlo (MCMC) methods to approximate Bayesian inference within the model. For each condition, the adapted RMC was run 5000 times, randomizing the order of the stimuli presented to the model each time.

# B

# Details of limited capacity computational models for Chapter 4

As the ideal observer model described in Chapter 4 fails to fully match human behaviour in the category learning tasks we studied, it suggests that there may be other computational models that can better explain our results. In this appendix section, I evaluate a number of other computational models that vary in their capacity to attend and learn. I discuss two additional models known as the RULE model and the LIMITED model that vary in how many features they attend to and how they use their category representation to make judgments. The notation for both of these additional models is the same as used to describe the OPTIMAL or ideal observer model.

## B.1 A RULE BASED MODEL

The first model we consider is the simplest and most limited in capacity. It is a rule-based categorization model (which we call RULE) that classifies objects into categories using only a single feature. On each trial, the model considers a single hypothesis $h_{ia}$ for the rule that defines how it makes categorization decisions. All of the rules in the hypothesis space take the following form: `if` $x_i = a$ `then` $\hat{y} = 0$`, otherwise` $\hat{y} = 1$, such that each rule learns to use one feature ($a$) to predict the category outcome ($\hat{y}$). The space of hypotheses is defined by the set of features. As an example, a particular hypothesis the model might use is: `If the third feature takes value 0` (i.e. $x_3 = 0$)`, then respond bivimia` ($\hat{y} = 0$)`, otherwise respond lorifen` ($\hat{y} = 1$)`.` Relative to other rule-based models in the literature (e.g., Nosofsky, Palmeri, & McKinley, 1994; Goodman et al., 2008) this is fairly simplistic because it can never adapt and use more than a single feature to make a category judgment. This was deliberate because we wanted to consider a model at one end of the spectrum, capturing the important intuition that the learner attends to and uses only one feature at a time.

The RULE model learns by updating the utility $u$ of every hypothesis in the hypothesis space. All utility values are initialized to 0.5 at the start of the learning process and are bounded between 0 and 1. At the end of each trial, the model updates the utility of the currently-considered hypothesis only, and it does so by assuming that the utility is proportional to the number of correct decisions that the rule has led to on those trials where the learner was considering that rule. Formally, this utility function is denoted:

$$u(h_{ia}) = \frac{1 + (\text{correct predictions with } h_{ia})}{2 + (\text{trials with } h_{ia})} \tag{B.1}$$

At the end of every trial, the model updates the utility of the current hypothesis. If it makes the correct prediction, the hypothesis is retained for the next trial, otherwise it is discarded and a new hypothesis is selected from the set of hypotheses with probability proportional to the utility, as in Equation B.2:

$$p(h_{ia}) = \frac{u(h_{ia})}{\sum_{x,y} u(h_{xy})} \tag{B.2}$$

## B.2 A LIMITED CAPACITY STATISTICAL LEARNER

The OPTIMAL and RULE model differ in several respects, and if one of them learns in a more human-like fashion than the other we would like to know *why* this is the case. The OPTIMAL statistical model employs a category representation that closely mirrors a probabilistic prototype, whereas the RULE model represents categories using simple decision rules. The OPTIMAL model updates its category representations using all the information available to it, whereas the RULE model only updates its beliefs about the one specific rule it is currently considering. Finally, the OPTIMAL model makes its categorization decisions by always choosing the most likely category, whereas the rule based model – though also deterministic – is entirely capable of following a particular rule to make a correct decision and then immediately discarding that rule.

Given these differences, we developed a LIMITED model variant of the OPTIMAL statistical learning model that retains the prototype-style representation but is limited in both how many features to use when making decisions and the ability to learn from their observations.*

As in the OPTIMAL model, the category label selected on a given trial by the the LIMITED model is dictated by Equation 4.4. However, instead of multiplying across all features when making a decision as in Equation 4.1, a single feature $f$ drives decision making:

$$p(y|\mathrm{x}) \propto p(x_f|y)p(y) \tag{B.3}$$

Learning is also limited in this model, which we implemented by applying Equation 4.2 to only a *single* feature on each trial. This limitation captures the same qualitative principle that underpins the single-hypothesis belief updating procedure used by the RULE model. Highlighting this connection, the updating process of the LIMITED model shifts its attention across features using a utility-based

*We also considered models that were limited in only one way (i.e., either how many features to use in making decisions or in learning, but not both) but the model reported here, which was limited in both, had the best performance. Thus, for ease of exposition and overall clarity we report only this particular model.

rule that is almost identical to Equations B.1 and B.2 for the RULE model:

$$u(f_i) = \frac{1 + (\text{correct predictions with } f_i)}{2 + (\text{trials with } f_i)} \tag{B.4}$$

$$p(f_i) = \frac{u(f_i)}{\sum_x u(f_x)} \tag{B.5}$$

The LIMITED model is thus very restricted: it absorbs information only from a single attended feature, and this is the only feature that contributes to the categorization decision. The differences between this model and the RULE model are fairly modest.

## B.3 Model results

Similar to the results obtained for the OPTIMAL model, we simulated the RULE and LIMITED models 10,000 times for each of the experimental conditions in both experiments, where each simulation mimicked a 100-trial experiment. On each trial, a new stimulus was generated in exactly the same manner as the experiment, the model then made a prediction of the category label, and then received feedback and updated its category representations. The full set of results comparing human performance to all three models are shown in Figures B.1 and B.2.

How do the predictions of the RULE and LIMITED model differ from the OPTIMAL model? Similar to the analyses in the main body of this Chapter, I begin by looking at aggregate performance across participants. Similar to the OPTIMAL model, both the RULE and LIMITED models produce

|  | Experiment 1 | Experiment 2 |
|---|---|---|
| Optimal | 0.228 | 0.217 |
| Limited | 0.038 | 0.028 |
| Rule | 0.029 | 0.037 |
| Random | 0.237 | 0.251 |

**Table B.1: Error estimates between model predictions and human performance.** Errors are root-mean-squared error (RMSE) between the aggregate model predictions of accuracy and the aggregate human accuracy. In both experiments, the *Limited* and *Rule* models fit far better than the *Optimal* model and a *Random* model that guesses on each trial, although which of the two has the best fit changes from Experiment 1 to 2.

the main qualitative patterns shown in the participant data across both experiments. However, where the OPTIMAL model failed to capture the quantitative aspects of human performance, both the RULE and LIMITED models do a better job. In particular, when categories are not family resemblance structures, more features should hurt performance, and when the category structure is family resemblance based, additional features should make no difference. Since this behavior precisely matches the qualitative pattern shown by people, it is perhaps no surprise that the RULE and LIMITED models quantitatively fit human performance far better (see Table B.1).

Perhaps somewhat disconcertingly, neither the quantitative nor the qualitative fits give compelling reason to prefer either the RULE or LIMITED model over each other. That said, it is important to realise that this analysis so far reflects *aggregate* data: how well each model predicts the overall population average amongst our participants. Yet we know that population averages may be highly misleading when the goal is to infer what kinds of individual processes give rise to the behavior in question. For instance, if a population was actually made of two very different kinds of people, the aggregate might bear no resemblance to the behavior of any of the individuals involved.

For this reason we also calculated which model best fit each of the individuals in each of the experiments (as reflected in the RMSE between each individual's accuracy and the model predictions from the same experimental condition as the individual), as shown in Figure B.3. It is evident that although there is substantial variation across people, in the majority of conditions most people are best fit by the LIMITED model. The only exception is that in the most difficult conditions a substantial number are best fit by a model that guesses randomly, suggesting these participants were unable to learn the task. The OPTIMAL model describes performance the least well of the three theoretically motivated models.

While both the RULE and LIMITED models better account for the performance of humans at the aggregate and individual levels much better than the OPTIMAL model, when interpreting them as process models they make a number of strong assumptions that are weaker than what humans seem to be doing. For example, in the LIMITED model it only updates and decides using a single feature on any given trial, but other research in category learning suggests that people can combine information from more than a single feature to make judgments (Gluck, Shohamy, & Myers, 2002). Thus, this reduces the plausibility that the LIMITED model is a complete process-level description of people's behaviour in this task, and that further work is necessary.

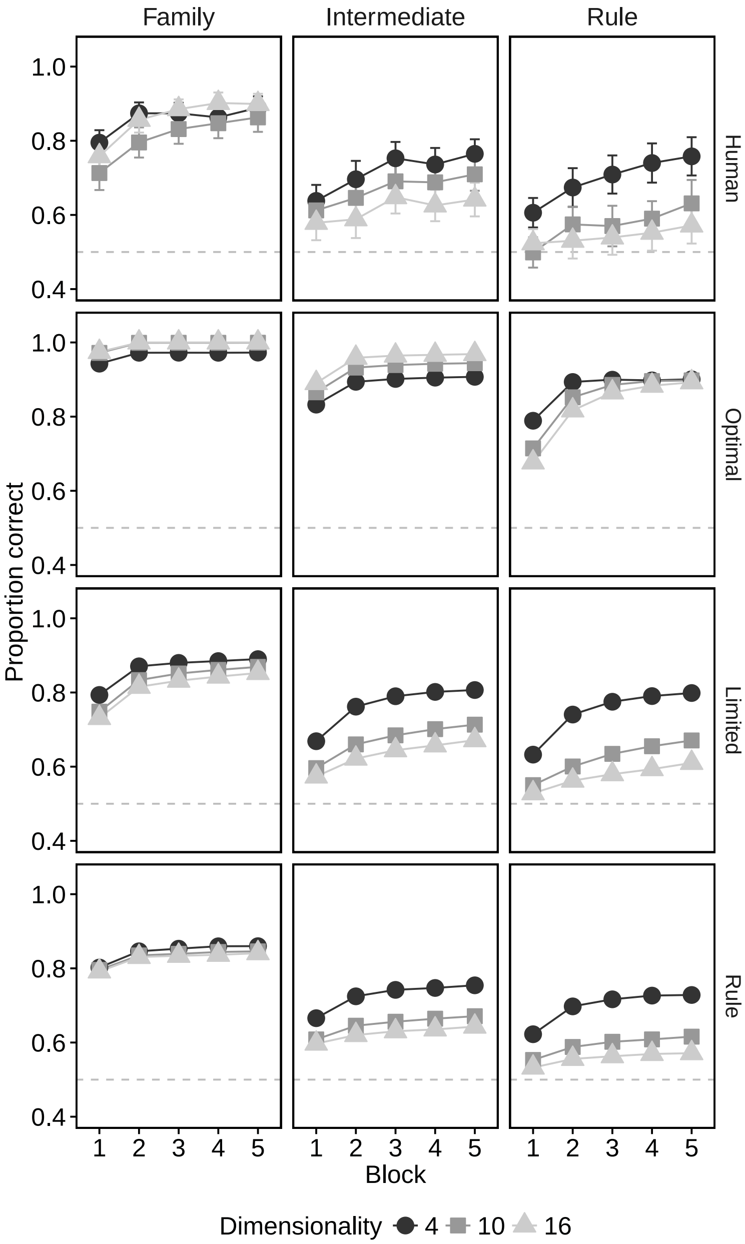


**Figure B.1: Comparison of model and human performance in Experiment 1.** The top row shows human performance and subsequent rows show the predictions made by each model. The columns correspond to the different *Structure* conditions. All of the models capture the fact that people perform better in the *Family* condition. However, only the *Rule* and *Limited* models predict performance in the other two category structures: namely, that accuracy diminishes with additional features.

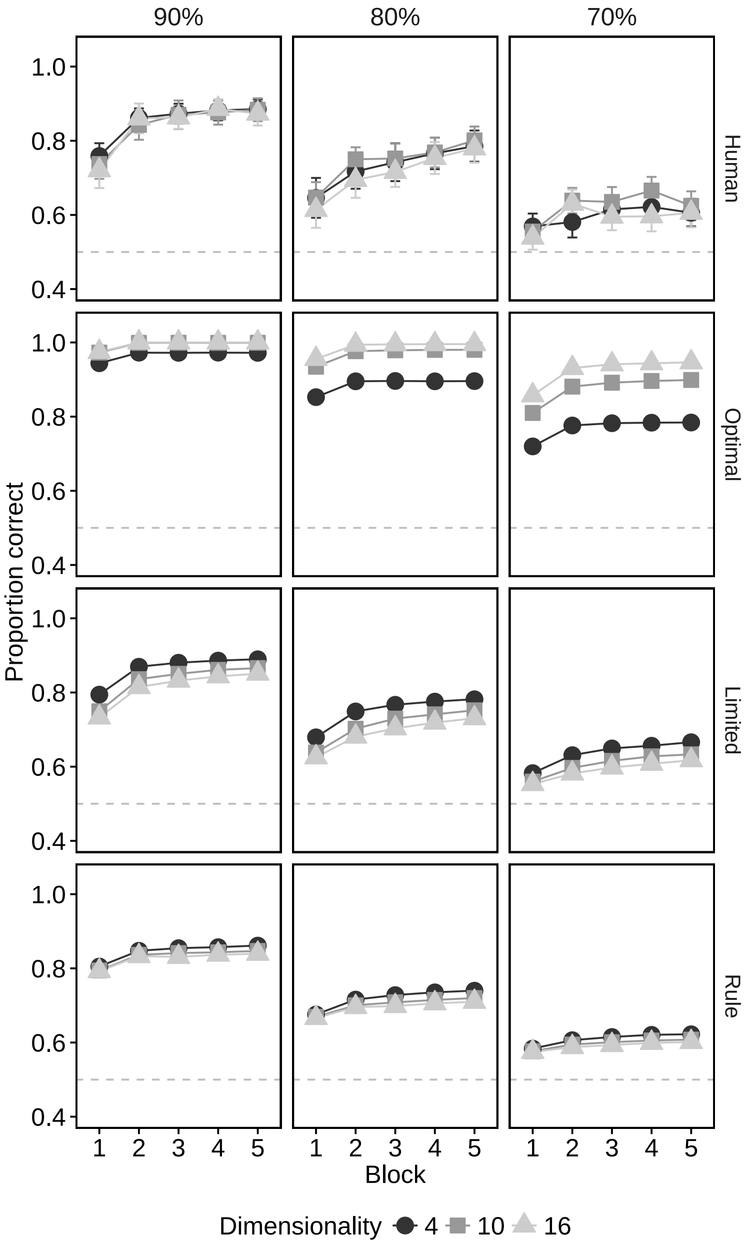
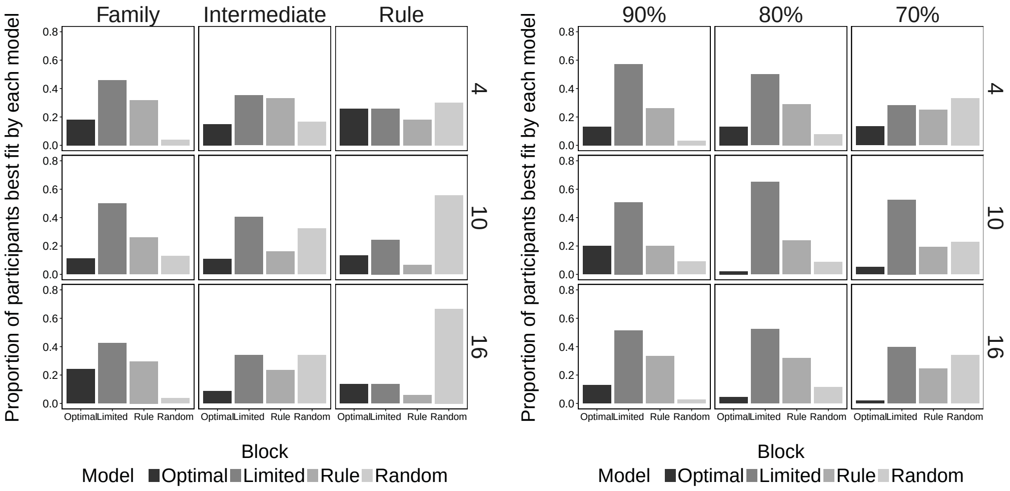


**Figure B.2: Comparison of model and human performance in Experiment 2.** The top row shows human performance and subsequent rows show the predictions made by each model. The columns indicate the *Predictiveness* of the features in that condition. All of the models capture the fact that human performance is better in the 90% condition. However, the *Optimal* model incorrectly predicts that performance should improve with additional features, whereas the other models and humans show similar accuracy regardless of how many features there are.



**(a)** Experiment 1

**(b)** Experiment 2

**Figure B.3: Proportion of individuals who are best fit by each of the models in both experiments.** Most of the time, most people's pattern of performance most closely matched the predictions of the *Limited* model. The main exception is in the more difficult conditions like the *Rule*-based category structure, in which a model consistent with random guessing did the best.

# References

Ahn, W.-K., & Medin, D. L. (1992). A two-stage model of category construction. *Cognitive Science*, *16*(1), 81–121.

Anderson, J. R. (1990). *The adaptive character of thought*. Hillsdale, NJ: Erlbaum.

Anderson, J. R. (1991). The adaptive nature of human categorization. *Psychological Review*, *98*(3), 409–429.

Ashby, F. G. (1992). *Multidimensional models of categorization.* Lawrence Erlbaum Associates, Inc.

Ashby, F. G., & Alfonso-Reese, L. A. (1995). Categorization as probability density estimation. *Journal of mathematical psychology*, *39*(2), 216–233.

Ashby, F. G., Queller, S., & Berretty, P. M. (1999). On the dominance of unidimensional rules in unsupervised categorization. *Perception & Psychophysics*, *61*(6), 1178–1199.

Austerweil, J. L., & Griffiths, T. L. (2011). A rational model of the effects of distributional information on feature learning. *Cognitive psychology*, *63*(4), 173–209.

Barrett, M. D. (1986). Early semantic representations and early word-usage. In *The development of word meaning* (pp. 39–67). Springer.

Bellman, R. (1961). *Adaptive control processes: A guided tour*. Princeton, NJ: Princeton University Press.

Bloom, P. (2000a). *How children learn the meaning of words*. Cambridge, MA: MIT Press.

Bloom, P. (2000b). *How children learn the meanings of words.* The MIT Press.

Bonawitz, E., Shafto, P., Gweon, H., Goodman, N. D., Spelke, E., & Schulz, L. (2011). The double-edged sword of pedagogy: Instruction limits spontaneous exploration and discovery. *Cognition*, *120*(3), 322–330.

Borin, L. (2012). Core vocabulary: A useful but mystical concept in some kinds of linguistics. In *Shall we play the festschrift game?* (pp. 53–65). Springer.

Bruner, J. S. (1961). The act of discovery. *Harvard educational review*.

Bruner, J. S., Goodnow, J. J., & Austin, G. A. (1956). *A study of thinking*. New York: John Wiley and Sons.

Canini, K. R. (2011). *Nonparametric hierarchical bayesian models of categorization* (Unpublished doctoral dissertation). University of California, Berkeley.

Canini, K. R., & Griffiths, T. L. (2011). A nonparametric bayesian model of multi-level category learning. In *Proceedings of the 25th AAAI Conference on Artificial Intelli-*

*gence.*

Chapelle, O., Schölkopf, B., & Zien, A. (2006). *Semi-supervised learning* (Vol. 2). MIT press Cambridge.

Chater, N., Clark, A., Goldsmith, J., & Perfors, A. (2015). *Empiricism and language learnability.* Oxford, England: Oxford University Press.

Chater, N., & Vitányi, P. (2003). Simplicity: a unifying principle in cognitive science? *Trends in cognitive sciences*, *7*(1), 19–22.

Clapper, J. P., & Bower, G. H. (2002). Adaptive categorization in unsupervised learning. *Journal of Experimental Psychology: Learning, Memory, and Cognition*, *28*(5), 908.

Colunga, E., & Smith, L. B. (2005). From the lexicon to expectations about kinds: a role for associative learning. *Psychological review*, *112*(2), 347.

Donoho, D. L. (2000). High-dimensional data analysis: The curses and blessings of dimensionality. In *American Mathematical Society Conference on Math Challenges of the 21st Century.*

Eaves, B. S., & Shafto, P. (2014). Order effects in learning relational structures. In *Proceedings of the 36th Annual Conference of the Cognitive Science Society.*

Eaves Jr, B. S., Feldman, N. H., Griffiths, T. L., & Shafto, P. (2016). Infant-directed speech is consistent with teaching. *Psychological review*, *123*(6), 758.

Edgell, S. E., Castellan Jr, N. J., Roe, R. M., Barnes, J. M., Ng, P. C., Bright, R. D., & Ford, L. A. (1996). Irrelevant information in probabilistic categorization. *Journal of Experimental Psychology: Learning, Memory, and Cognition*, *22*(6), 1463–1481.

Feldman, J. (2000). Minimization of boolean complexity in human concept learning. *Nature*, *407*(6804), 630–633.

Feldman, N., Griffiths, T. L., & Morgan, J. L. (2009). The influence of categories on perception: explaining the perceptual magnet effect as optimal statistical inference. *Psychological review*, *116*(4), 752.

Gangwani, T., Kachergis, G., & Yu, C. (2010). Simultaneous cross-situational learning of category and object names. In *Proceedings of the 32nd Annual Conference of the Cognitive Science Society* (Vol. 32).

Gershman, S. J., Cohen, J. D., & Niv, Y. (2010). Learning to selectively attend. In S. Ohlsson & R. Catrambone (Eds.), *Proceedings of the 32nd Annual Conference of the Cognitive Science Society* (pp. 1270–1275). Austin, TX: Cognitive Science Society.

Gibson, B. R., Rogers, T. T., & Zhu, X. (2013). Human semi-supervised learning. *Topics in Cognitive Science*, *5*(1), 132–172.

Gluck, M. A., Shohamy, D., & Myers, C. (2002). How do people solve the "weather prediction" task?: Individual variability in strategies for probabilistic category learning. *Learning & Memory*, *9*(6), 408–418.

Goldstone, R. L., & Hendrickson, A. T. (2010). Categorical perception. *Wiley Interdisci-*

*plinary Reviews: Cognitive Science*, *1*(1), 69–78.

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.

Goodman, N. (1983). *Fact, fiction, and forecast* (4th ed.). Cambridge, MA: Harvard University Press.

Goodman, N., Tenenbaum, J. B., Feldman, J., & Griffiths, T. L. (2008). A rational analysis of rule-based concept learning. *Cognitive Science*, *32*(1), 108–154.

Griffiths, T. L., Canini, K. R., Sanborn, A. N., & Navarro, D. J. (2007). Unifying rational models of categorization via the hierarchical dirichlet process. In *Proceedings of the cognitive science society* (Vol. 29).

Griffiths, T. L., Kemp, C., & Tenenbaum, J. B. (2008). Bayesian models of cognition. In R. Sun (Ed.), *The Cambridge handbook of computational psychology* (pp. 59–100). Cambridge, MA: Cambridge University Press.

Griffiths, T. L., Steyvers, M., & Tenenbaum, J. B. (2007). Topics in semantic representation. *Psychological Review*, *114*(2), 211–244.

Gureckis, T. M., & Love, B. C. (2003). Towards a unified account of supervised and unsupervised category learning. *Journal of Experimental & Theoretical Artificial Intelligence*, *15*(1), 1–24.

Heller, K. A., & Ghahramani, Z. (2005). Bayesian hierarchical clustering. In *Proceedings of the 22nd International Conference on Machine learning* (pp. 297–304).

Hidaka, S., Torii, T., & Kachergis, G. (2017). Leveraging mutual exclusivity for faster cross-situational word learning: A theoretical analysis. In *Proceedings of the 39th annual conference of the cognitive science society.*

Hoffman, A. B., Harris, H. D., & Murphy, G. L. (2008). Prior knowledge enhances the category dimensionality effect. *Memory & Cognition*, *36*(2), 256–270.

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.

Hubert, L., & Arabie, P. (1985). Comparing partitions. *Journal of Classification*, *2*(1), 193–218.

Ichinco, D., Frank, M. C., & Saxe, R. (2009). Cross-situational word learning respects mutual exclusivity. In *Proceedings of the 31st Annual Conference of the Cognitive Science Society* (Vol. 31).

Jäkel, F., Schölkopf, B., & Wichmann, F. A. (2008). Generalization and similarity in exemplar models of categorization: Insights from machine learning. *Psychonomic Bulletin & Review*, *15*(2), 256–271.

Jarecki, J., Meder, B., & Nelson, J. (in press). Naive and robust: Class-conditional independence in human classification learning. *Cognitive Science*.

Jarecki, J., Meder, B., & Nelson, J. D. (2013). The assumption of class-conditional independence in category learning. In M. Knauff, M. Pauen, N. Sebanz, & I. Wachsmuth (Eds.), *Proceedings of the 35th Annual Conference of the Cognitive Science Society* (pp. 2650–2655). Austin, TX: Cognitive Science Society.

Kachergis, G., & Yu, C. (2017). Observing and modeling developing knowledge and uncertainty during cross-situational word learning. *IEEE Transactions on Cognitive and Developmental Systems*.

Kalish, C. W., Rogers, T. T., Lang, J., & Zhu, X. (2011). Can semi-supervised learning explain incorrect beliefs about categories? *Cognition*, *120*(1), 106–118.

Kemp, C. (2012). Exploring the conceptual universe. *Psychological Review*, *119*(4), 685–722.

Kemp, C., Goodman, N. D., & Tenenbaum, J. B. (2008). Theory acquisition and the language of thought. In *Proceedings of the 30th Annual Conference of the Cognitive Science Society.*

Kemp, C., Perfors, A., & Tenenbaum, J. B. (2007). Learning overhypotheses with hierarchical Bayesian models. *Developmental Science*, *10*(3), 307–321.

Kemp, C., & Tenenbaum, J. B. (2008). The discovery of structural form. *Proceedings of the National Academy of Sciences*, *105*(31), 10687–10692.

Kemp, C., & Tenenbaum, J. B. (2009). Structured statistical models of inductive reasoning. *Psychological Review*, *116*(1), 20–58.

Kemp, C., Tenenbaum, J. B., Niyogi, S., & Griffiths, T. L. (2010). A probabilistic model of theory formation. *Cognition*, *114*(2), 165–196.

Keogh, E., & Mueen, A. (2011). The curse of dimensionality. In C. Sammut & G. Webb (Eds.), *Encyclopedia of machine learning.* New York, NY: Springer.

Kruschke, J. K. (1992). Alcove: an exemplar-based connectionist model of category learning. *Psychological review*, *99*(1), 22.

Lafond, D., Lacouture, Y., & Mineau, G. (2007). Complexity minimization in rule-based category learning: Revising the catalog of boolean concepts and evidence for non-minimal rules. *Journal of Mathematical Psychology*, *51*(2), 57–74.

Lake, B., Lawrence, N. D., & Tenenbaum, J. B. (2016). The emergence of organizing structure in conceptual representation. *arXiv preprint arXiv:1611.09384*.

Lake, B., & McClelland, J. (2011). Estimating the strength of unlabeled information during semi-supervised learning. In *Proceedings of the 33rd Annual Conference of the Cognitive Science Society* (pp. 1400–1405).

Lake, B., Salakhutdinov, R., & Tenenbaum, J. B. (2015). Human-level concept learning through probabilistic program induction. *Science*, *350*(6266), 1332–1338.

Lake, B., Ullman, T. D., Tenenbaum, J. B., & Gershman, S. J. (2017). Building machines that learn and think like people. *Behavioral and Brain Sciences*, *40*.

Lake, B., Vallabha, G. K., & McClelland, J. L. (2009). Modeling unsupervised perceptual category learning. *IEEE Transactions on Autonomous Mental Development*, *1*(1), 35–43.

Landau, B., Smith, L. B., & Jones, S. S. (1988). The importance of shape in early lexical learning. *Cognitive Development*, *3*(3), 299–321.

Landauer, T., & Dumais, S. (1997). A solution to Plato's problem: The Latent Semantic Analysis theory of acquisition, induction, and representation of knowledge. *Psychological Review*, *104*(2), 211-240.

LeCun, Y., Bengio, Y., & Hinton, G. (2015). Deep learning. *Nature*, *521*(7553), 436–444.

Liang, F., Paulo, R., Molina, G., Clyde, M. A., & Berger, J. O. (2012). Mixtures of g priors for Bayesian variable selection. *Journal of the American Statistical Association*, *103*(481), 410–423.

Love, B. C. (2002). Comparing supervised and unsupervised category learning. *Psychonomic Bulletin & Review*, *9*(4), 829–835.

Love, B. C., Medin, D. L., & Gureckis, T. M. (2004). Sustain: a network model of category learning. *Psychological review*, *111*(2), 309.

Luce, R. D. (1959). Individual choice behavior.

Lupyan, G. (2012). Linguistically modulated perception and cognition: the label-feedback hypothesis. *Frontiers in psychology*, *3*.

Lupyan, G., & Lewis, M. (2017). From words-as-mappings to words-as-cues: the role of language in semantic knowledge. *Language, Cognition and Neuroscience*, 1–19.

Malt, B. C., & Smith, E. E. (1982). The role of familiarity in determining typicality. *Memory & Cognition*, *10*(1), 69–75.

Malt, B. C., & Smith, E. E. (1984). Correlated properties in natural categories. *Journal of Verbal Learning and Verbal Behavior*, *23*(2), 250–269.

Markant, D. B., & Gureckis, T. M. (2014). Is it better to select or to receive? learning via active and passive hypothesis testing. *Journal of Experimental Psychology: General*, *143*(1), 94.

Markant, D. B., Settles, B., & Gureckis, T. M. (2016). Self-directed learning favors local, rather than global, uncertainty. *Cognitive science*, *40*(1), 100–120.

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

Markman, E., & Wachtel, G. F. (1988). Children's use of mutual exclusivity to constrain the meanings of words. *Cognitive Psychology*, *20*(2), 121–157.

Marr, D. (1982). *Vision: A computational investigation into the human representation and processing of visual information*. Henry Holt & Company.

McDonnell, J. V., Jew, C. A., & Gureckis, T. M. (2012). Sparse category labels obstruct generalization of category membership. In *Proceedings of the 34th Annual Conference of the Cognitive Science Society* (pp. 749–754).

Medin, D. L., & Ross, B. H. (1997). *Cognitive psychology.* (Second ed.). Harcourt Brace

Jovanovich.

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

Medin, D. L., Wattenmaker, W. D., & Hampson, S. E. (1987). Family resemblance, conceptual cohesiveness, and category construction. *Cognitive psychology*, *19*(2), 242–279.

Medin, D. L., Wattenmaker, W. D., & Michalski, R. S. (1987). Constraints and preferences in inductive learning: An experimental study of human and machine performance. *Cognitive Science*, *11*(3), 299–339.

Milton, F., & Wills, A. (2004). The influence of stimulus properties on category construction. *Journal of Experimental Psychology: Learning, Memory, and Cognition*, *30*(2), 407.

Minda, J. P. (2015). *The psychology of thinking: Reasoning, decision-making and problem-solving*. New York, NY: Sage.

Minda, J. P., & Smith, J. D. (2001). Prototypes in category learning: The effects of category size, category structure, and stimulus complexity. *Journal of Experimental Psychology: Learning, Memory, and Cognition*, *27*(3), 775–799.

Morey, R. D., & Rouder, J. N. (2015). BayesFactor: Computation of Bayes Factors for Common Designs [Computer software manual]. Retrieved from `http://CRAN.R-project.org/package=BayesFactor` (R package version 0.9.12-2)

Murphy, G. (2002). *The big book of concepts*. Cambridge, MA: MIT Press.

Murphy, G. (2004). *The big book of concepts*. MIT press.

Navarro, D. J., Griffiths, T. L., Steyvers, M., & Lee, M. D. (2006). Modeling individual differences using Dirichlet processes. *Journal of Mathematical Psychology*, *50*(2), 101–122.

Nelson, J. D. (2005). Finding useful questions: On bayesian diagnosticity, probability, impact, and information gain. *Psychological review*, *112*(4).

Nguyen, S. P. (2007). Cross-classification and category representation in children's concepts. *Developmental Psychology*, *43*(3), 719.

Nguyen, S. P., & Murphy, G. L. (2003). An apple is more than just a fruit: Cross-classification in children's concepts. *Child Development*, *74*(6), 1783–1806.

Nosofsky, R. M. (1984). Choice, similarity, and the context theory of classification. *Journal of Experimental Psychology: Learning, memory, and cognition*, *10*(1), 104.

Nosofsky, R. M. (1986). Attention, similarity, and the identification–categorization relationship. *Journal of Experimental Psychology: General*, *115*(1), 39.

Nosofsky, R. M. (1988). Exemplar-based accounts of relations between classification, recognition, and typicality. *Journal of Experimental Psychology: Learning, Memory, and Cognition*, *14*(4), 700–708.

Nosofsky, R. M., Gluck, M. A., Palmeri, T. J., McKinley, S. C., & Glauthier, P. (1994).

Comparing modes of rule-based classification learning: A replication and extension of shepard, hovland, and jenkins (1961). *Memory & cognition*, *22*(3), 352–369.

Nosofsky, R. M., Palmeri, T. J., & McKinley, S. C. (1994). Rule-plus-exception model of classification learning. *Psychological Review*, *101*(1), 53–79.

Osherson, D. N., Smith, E. E., Wilkie, O., Lopez, A., & Shafir, E. (1990). Category-based induction. *Psychological review*, *97*(2), 185.

Peterson, J. C., Abbott, J. T., & Griffiths, T. L. (2016). Adapting deep network features to capture psychological representations. *arXiv preprint arXiv:1608.02164*.

Piantadosi, S. T., Tenenbaum, J. B., & Goodman, N. D. (2016). The logical primitives of thought: Empirical foundations for compositional cognitive models. *Psychological Review*, *123*(4), 392.

Posner, M. I., & Keele, S. W. (1968). On the genesis of abstract ideas. *Journal of Experimental Psychology*, *77*(3), 353–363.

Pothos, E. M., & Bailey, T. M. (2009). Predicting category intuitiveness with the rational model, the simplicity model, and the generalized context model. *Journal of Experimental Psychology: Learning, Memory, and Cognition*, *35*(4), 1062.

Pothos, E. M., & Chater, N. (2002). A simplicity principle in unsupervised human categorization. *Cognitive Science*, *26*(3), 303–343.

Pothos, E. M., & Close, J. (2007). One or two dimensions in spontaneous classification: A simplicity approach. *Cognition*, *107*(2), 581–602.

Pothos, E. M., & Close, J. (2008). One or two dimensions in spontaneous classification: A simplicity approach. *Cognition*, *107*(2), 581–602.

Pothos, E. M., Perlman, A., Bailey, T. M., Kurtz, K., Edwards, D. J., Hines, P., & McDonnell, J. V. (2011). Measuring category intuitiveness in unconstrained categorization tasks. *Cognition*, *121*(1), 83–100.

Quine, W. (1960). *Word and object*. Cambridge, MA: MIT Press.

Reed, S. K. (1972). Pattern recognition and categorization. *Cognitive psychology*, *3*(3), 382–407.

Regehr, G., & Brooks, L. R. (1995). Category organization in free classification: The organizing effect of an array of stimuli. *Journal of Experimental Psychology: Learning, Memory, and Cognition*, *21*(2), 347.

Rosch, E. (1973). Natural categories. *Cognitive Psychology*, *4*(3), 328–350.

Rosch, E., & Mervis, C. (1975b). Family resemblances: Studies in the internal structure of categories. *Cognitive Psychology*, *7*(4), 573–605.

Rosch, E., & Mervis, C. B. (1975a). Family resemblances: Studies in the internal structure of categories. *Cognitive Psychology*, *7*(4), 573–605.

Rosch, E., Mervis, C. B., Gray, W. D., Johnson, D. M., & Boyes-Braem, P. (1976). Basic objects in natural categories. *Cognitive psychology*, *8*(3), 382–439.

Ross, B. H., & Murphy, G. L. (1999). Food for thought: Cross-classification and category organization in a complex real-world domain. *Cognitive Psychology*, *38*(4), 495–553.

Rouder, J. N., Morey, R. D., Speckman, P. L., & Province, J. M. (2012). Default Bayes factors for ANOVA designs. *Journal of Mathematical Psychology*, *56*(5), 356–374.

Sanborn, A. N., Griffiths, T. L., & Navarro, D. J. (2010). Rational approximations to rational models: alternative algorithms for category learning. *Psychological Review*, *117*(4), 1144–1167.

Schyns, P. G., & Rodet, L. (1997). Categorization creates functional features. *Journal of Experimental Psychology-Learning Memory and Cognition*, *23*(3), 681–696.

Searcy, S. R., & Shafto, P. (2016). Cooperative inference: Features, objects, and collections. *Psychological Review*, *123*(5), 510-533.

Shafto, P., & Goodman, N. (2008). Teaching games: Statistical sampling assumptions for learning in pedagogical situations. In *Proceedings of the 30th annual conference of the cognitive science society* (pp. 1632–1637).

Shafto, P., Goodman, N. D., & Griffiths, T. L. (2014). A rational account of pedagogical reasoning: Teaching by, and learning from, examples. *Cognitive psychology*, *71*, 55–89.

Shafto, P., Kemp, C., Mansinghka, V., & Tenenbaum, J. B. (2011). A probabilistic model of cross-categorization. *Cognition*, *120*(1), 1–25.

Shepard, R. N., Hovland, C. I., & Jenkins, H. M. (1961). Learning and memorization of classifications. *Psychological Monographs: General and Applied*, *75*(13), 1–42.

Shepard, R. N., et al. (1987). Toward a universal law of generalization for psychological science. *Science*, *237*(4820), 1317–1323.

Tenenbaum, J. B. (1997). Mapping a manifold of perceptual observations. *Advances in Neural Information Processing Systems*, *10*, 682–688.

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.

Vandist, K., De Schryver, M., & Rosseel, Y. (2009). Semisupervised category learning: The impact of feedback in learning the information-integration task. *Attention, Perception, & Psychophysics*, *71*(2), 328–341.

Vanpaemel, W., & Storms, G. (2008). In search of abstraction: The varying abstraction model of categorization. *Psychonomic bulletin & review*, *15*(4), 732–749.

Verleysen, M., & François, D. (2005). The curse of dimensionality in data mining and time series. In *Proceedings of the 8th International Workshop on Artificial Neural Networks* (pp. 758–770).

Waxman, S. R., & Markow, D. B. (1995). Words as invitations to form categories: Evidence from 12-to 13-month-old infants. *Cognitive psychology*, *29*(3), 257–302.

Xu, F. (2002). The role of language in acquiring object kind concepts in infancy. *Cognition*, *85*(3), 223–250.

Xu, F., & Tenenbaum, J. B. (2007a). Sensitivity to sampling in bayesian word learning. *Developmental science*, *10*(3), 288–297.

Xu, F., & Tenenbaum, J. B. (2007b). Word learning as Bayesian inference. *Psychological Review*, *114*(2), 245–272.

Yu, C., & Smith, L. B. (2007). Rapid word learning under uncertainty via cross-situational statistics. *Psychological Science*, *18*(5), 414–420.

Yurovsky, D., & Yu, C. (2008). Mutual exclusivity in cross-situational statistical learning. In *Proceedings of the 30th Annual Conference of the Cognitive Science Society* (Vol. 30).

Zhu, X., Gibson, B. R., Jun, K.-S., Rogers, T. T., Harrison, J., & Kalish, C. (2010). Cognitive models of test-item effects in human category learning. In *Proceedings of the 27th International Conference on Machine Learning* (pp. 1247–1254).

Zhu, X., Rogers, T., Qian, R., & Kalish, C. (2007). Humans perform semi-supervised classification too. In *Proceedings of the National Conference on Artificial Intelligence* (pp. 864–870).