

# NO REPRESENTATION WITHOUT TAXATION: THE COSTS AND BENEFITS OF LEARNING TO CONCEPTUALIZE THE ENVIRONMENT

Michael Ramscar and Melody Dye

ramscar@stanford.edu, pkipsy@stanford.edu  
Department of Psychology, Stanford, CA 94305, USA

## ABSTRACT

How do the *ways* in which we learn influence our cognitive representations of *what* we learn? We show that in language learning tasks, the process of learning to conceptualize and categorize perceptual input shapes how that input gets represented in mind. In representation, there seems to be a give and take between veridicality and completeness, on the one hand, and discrimination and accurate categorization, on the other. Learning to better discriminate objects into categories by learning to value salient features makes people less likely to notice or remember the same objects' other features. Gains in response-discrimination between categories thus come at a cost to within category discrimination. This is a natural consequence of error-driven learning, a mechanism underlying most contemporary learning models. We present an exposition of error-driven learning, outline its implications for cognitive representations, and test these predictions, showing that the patterns of human learning are consistent with our analysis. We suggest that the mechanisms of human learning obey a simple principle: *there can be no representation without taxation*. We describe the implications of this principle for our conception of categorization and analogy.

## INTRODUCTION

While we perceive the world through our senses, we do not experience it in terms of raw sense data; we experience it in terms of *concepts*. We experience a world of objects and events – pages, screens, cars, people etc – and not the raw patterns of activity that external stimuli produce in the retinal cells of our eyes.

Most, if not all, cognitive activities appear to involve a process of converting the mass of

data we receive from our senses into 'meaningful' concepts. Learning imposes discontinuities on the continuous dimensions of inputs, so that raw sense data is grouped into larger representational wholes, which satisfy the informational requirements of various cognitive activities, such as reasoning about or communicating about the environment. We call this kind of discrimination-learning *categorization*: the process of taking a set of undifferentiated perceptual inputs and generating or tuning responses to those inputs. Categorization is an important aspect of cognition, and much effort has been invested in attempting to account for how the 'stuff of experience' is represented, manipulated and combined in the mind, and how it relates to language. Our research addresses an important question this process raises: How do the *ways* in which cognitive representations are developed and learned influence *what* gets learned and represented?

In this paper, we explore the hypothesis that different types of learning produce correspondingly different cognitive representations. We show that learning to conceptualize or categorize perceptual input has consequences for the representation of the input itself. We show that in both non-symbolic categorization (such as learning to predict environmental affordances and events) and symbolic categorization (such as in language-learning), improved response-discrimination appears to come at a cost to the representation of the original input. Thus, learning to better discriminate objects into categories based on their highly-salient features seems to make people less likely to notice or remember the same objects' less-salient features. Learners home in on cues that are highly predictive of a given category and simultaneously discard—or

'learn to ignore'—other probabilistic information that is less informative. Gains in response-discrimination between categories thus come at a cost to within category discrimination.

In this discussion, we use "cognitive representation" to describe a probabilistic understanding of the set(s) of cues that reliably predict a category label (i.e., which features predict which word). This is a departure from many traditional categorization models, which attempt a *descriptive* characterization of the mental representations of categories. What we take to be important, from the point of view of processing, is simply to give a consistent account of the principles by which representations of conceptual 'content' are meaningfully related in use. Spoken words provide a unique basis for considering the relationship between learning and representation. Because of some of the intrinsic characteristics of verbal labels, and the ways in which they are used and perceived, significant differences in discrimination-learning occur depending on the order in which objects and their features and labels are encountered in learning (Ramscar, Yarlett, Dye, Denny & Thorpe, *in submission*). These differences in learning are reflected in the different cognitive representations they produce.

In what follows, we show how taking the predictive structure of learning seriously can shed new light on the nature of symbolic representations, how it can offer new insights into many of the phenomena associated with analogical reasoning, a "high-level" cognitive process that appears to be driven by underlying symbolic structure (c.f., Gentner, 1983), and how it can provide a satisfactory account of the relationship between analogy and "straight-forward" categorization processes. The basic principle of *no representation without taxation* that underlies all these results is, we argue, a fundamental law of learning.

### Symbolic Representation

People use symbols (such as words, signs or pictures) and arrangements of symbols to communicate about the world. In our analysis of the symbolic learning that makes this possible,

we do not presuppose that symbolic thought is necessarily the same thing as "symbolic computation" (where symbolic computation is equated with a particular algorithmic—usually procedural—approach to computer programming). One might conceive of many ways in which symbolic thought could be implemented in mind, and the model embodied within the "symbolic approach" to cognitive science is just one of these.

"Symbolic" approaches to cognition typically characterize mental representations in terms of rules that define relationships between classes of entities (such as an "if X then Y"; see e.g., Fodor, 1998). This approach requires in turn that type/token relationships for these classes be defined. Defining what constitutes X's and Y's allows individual entities to be bound the appropriate part of structures like "if X then Y," allowing them to describe relationships in the world. If these definitions of classes are symbolic, this in turn requires that the symbols that comprise the definitions be defined (i.e., if the definition of X is "all X's have Z," one needs to define Z).

The problem with this approach is that definitions of symbolic categories that invoke further symbols soon take on the theoretical characteristics of a Russian Doll; they are inherently regressive. The symbol "dog," is a token of the type "noun," just as "spaniel" is a token of "dog," or "Fido" is a token of "spaniel." In order to explain (or generalize) the relationship between "Fido," "Spot" and "spaniel," it is not sufficient to say, "Fido is a spaniel" and "Spot is a spaniel." One must say *why* Fido and Spot are spaniels (as opposed to anything else).

The logic of the symbolic approach to cognition doesn't just require that class memberships be established in order to understand relations between classes. To the extent that symbolic representations are supposed to be interpretable, it also requires that one have an account of how one goes from classes to individuals. If symbolic representations are conceived of as "compositional" (i.e., so that sentences in natural language have structural meanings that are derived from the structure of

sentence, which in turn affects the specific meanings of words out of which the sentence is composed, see e.g., Fodor, 1998), one needs an account of how one can somehow extract the relevant individual tokens of meaning from descriptions that only mention types (i.e., one needs to be able to say which aspects of the meanings of “cat” “sat” and “mat” are relevant to the meaning of “the cat sat on the mat”).

Neither desideratum is satisfied by any existing symbolic approach (Fodor, 1998; Murphy, 2002). Indeed, there are good reasons to believe that they cannot be satisfied: the kinds of things that people represent and think about symbolically do not fall into discrete classes of X’s, Y’s or Z’s. Symbolic categories do not possess discrete boundaries (i.e., there are no fixed criteria for establishing whether an entity is an X or a Y), and entities are often assigned to multiple symbolic classes (i.e., they are sometimes X’s; sometimes Ys). As a result of these (and many other) factors, symbolic type/token relationships appear to be inherently underdetermined (see e.g., Wittgenstein, 1953; Quine, 1960; Fodor, 1998). This is a problem for all current symbolic approaches, and those few theorists who have addressed it seriously have been forced to conclude that although there must be a solution, it is both innate and largely inscrutable (Fodor 1998).

Alternative approaches to characterizing thought and language, especially those that take an associative (or connectionist) approach, are often termed “sub-symbolic,” to distinguish them from “symbolic” models (Fodor, 1998). However, to the extent that we think of thought as being symbolic (and it seems natural to do so, especially with regards language), and to the extent that associative, connectionist and “symbolic” approaches seek to explain the nature of thinking, differentiating between “sub-symbolic” and “symbolic” computational paradigms without a clear idea of what symbolic thought actually is runs the risk of missing the point altogether.

A far more important consideration than any distinction between the “symbolic” and “sub-symbolic” is that, overwhelmingly, theories of cognition of all persuasions tend to assume

that symbolic thought is referential. That is, they subscribe to the idea that symbols both represent, and in an—important sense—point to, meanings, so that symbols and their meanings share a bi-directional relationship. Symbols (and language) are typically seen as abstract representations that either exemplify (stand for) or refer (or point) to meanings (referents). These meanings can in turn be defined by reference to things in the world (i.e., symbols can be defined by reference to objects and events they refer to). So, for example, the symbol “dog” may be defined by reference to a class of things in the world—dogs).

The problems with this approach are largely the same as those for type/token definitions, and have been laid out exhaustively (Wittgenstein, 1953; Quine, 1960; Fodor, 1998; Murphy, 2002). What we focus on here is the referential presupposition of a bi-directional relationship between symbols and their meanings. Crucially, this assumption is at odds with the idea that symbols are abstract representations, because abstraction is not a bi-directional process. Abstraction involves reducing the information content of a representation, such that only information relevant to a particular purpose is retained (Wittgenstein, 1953). As such, abstraction is an inherently directed process: one can abstract from a larger body of information to an abstract representation of it (such as reading a research paper and summarizing it in an abstract), but one cannot reverse the process. However, the idea of “reverse abstraction” makes no sense, since one cannot recover information once it has been discarded as part of the process of abstraction (just as one cannot take the abstract of a research article and get from it the detailed methods and analysis sections that one has never seen). Given that abstraction is a directed process, and symbols serve as abstractions in communication and thought, then it follows that communication and thought may well be directed in a way that respects the basic principles of abstraction.

Accordingly, in what follows we take an explicitly predictive approach to symbolic thought. This approach is explicitly not refer-

ential. On the contrary, we treat symbols as abstractions in a literal sense, and given that abstraction is directed, we take the view that symbolic processing must be directed as well. Prediction is by its very nature directed. A prediction follows from the cues (or clues) that lead to an expectation. Correspondingly, we argue that the relationship between symbols and the concepts underlying their use is not bi-directional, and that symbolic processing is a process of predicting symbols.

### Symbolic Learning

In considering how symbols are represented and used, we begin by examining how they are learned. In what follows, we conceive of learning as a process by which information is acquired about the probabilistic relationships between important regularities in the environment (such as objects or events) and the cues that allow those regularities to be predicted (e.g., Rescorla & Wagner, 1972).

Crucially, the learning process is driven by discrepancies between what is expected and what is actually observed in experience (termed *error-driven learning*). The predictive values of cues produce expectations, and any difference between the value of what is expected and what is observed produces further learning. The predictive value of a given cue is strengthened when relevant events (such as events, objects or labels) are under-predicted by that cue, and weakened when they are over-predicted (Rescorla & Wagner, 1972). As a result, cues compete for relevance, and the outcome of this competition is shaped both by positive evidence about co-occurrences between cues and predicted events, and negative evidence about non-occurrences of predicted events. This process produces patterns of learning that are very different from what would be expected if learning were shaped by positive evidence alone (Rescorla, 1988).

This view of learning can be applied to symbolic thought by thinking of symbols (i.e., words) as both potentially important cues (predictors) and outcomes (things to be predicted). For example, the word “chair” might be pre-

dicted by, or serve to predict, the features that are associated with the things we call chairs (both when chairs and “chair” are present, or when they are being thought of). Thus word learning can take two forms:

- (i) *cues are labels and outcomes are features;*
- (ii) *cues are features and outcomes are labels.*

In (i), which we term LF-learning, learning is a process of acquiring information that allows the prediction of a feature or set of features given a label, whereas in (ii), which we term FL-learning, learning is a process of acquiring information that allows the prediction of a label from a given feature or set of features.

Many theories of symbolic cognition emphasize the importance of relations between things in our understanding the world (Gentner, 1983; Fodor, 1998; Goldstone, Medin & Gentner, 2001; Penn, Holyoak & Povinelli, 2008). Despite the widespread belief that associative models are unstructured (e.g., Fodor, 1998), the opposite is true. Treated properly, associative models are inherently structured. Though they are often referred to as models of association, all contemporary models of learning are *predictive*. Learning discovers probabilistic cue structures that share temporal, predictive relationships with other things (e.g., objects, events or labels) in the environment (see also Elman, 1991). Prediction is fundamentally relational, and LF- and FL-learning describe the two possible ways that these relations can be structured in symbolic learning.

In FL learning, the set of cues being learned from is generally larger than the set of outcomes being learned about, whereas in FL learning, the set of outcomes is generally larger than the set of cues. As we will now show, these set-size differences in the number of cues and outcomes that are being learned about in each these two forms of word learning result in asymmetrical cognitive representations and different levels of discrimination learning.

### The structure of labels and the world

Symbolic labels are relatively discrete, and possess little cue-structure, whereas objects and events in the world are far less dis-

## The costs and benefits of learning to conceptualize the environment

crete, and possess much denser cue-structure. (By cue-structure we mean the number of salient and discriminable cues they present.) Consider say, a *pan*. A pan presents a learner with many discriminable features; shape, color, size, etc. Now consider the label ‘pan.’ A native English speaker can parse this word into a sequence of phonemes [ $p^h an$ ], but otherwise will be largely unable to discriminate many further features within these. This is not to say that there are no other discriminable aspects of speech (i.e., emphasis, volume, or pitch contour), but rather to say that ordinarily, the phonetic level dominates semantic categorization. Other features, such as pitch contour, do not covary or *compete* with phonemes in the same way that size and shape do in an object. Moreover, because phonemes occur in sequence rather than simultaneously, there is no direct competition between them as cues. Labels essentially provide a learner with only a single cue, i.e. the label—‘pan’—itself.

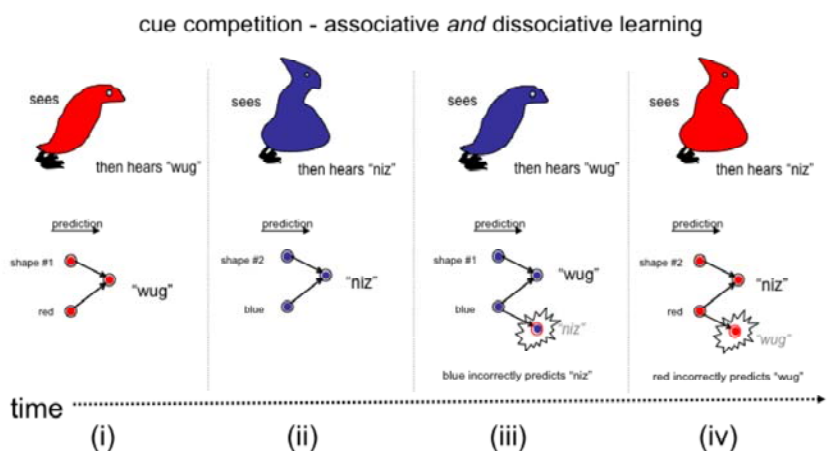
In LF-learning, because labels serve as cues and since individual labels have little cue-structure, learning involves predicting a set of features from a single cue (the label). LF-learning has a one-to-many form: one cue to many features. In contrast, FL-learning, where objects or events serve as cues, involves predicting a single response (a label) from a set of

cues. Thus FL-learning, has a many-to-one form: from many semantic features to a label.

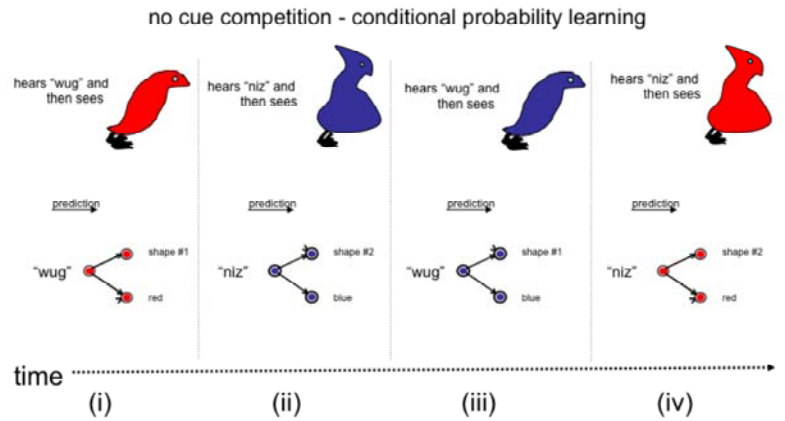
### Cue-competition in learning

When many cues are present simultaneously, they can compete for relevance in predicting an event. If a cue successfully predicts an event over time (positive evidence), the association between that cue and the event will increase. Conversely, if a cue unsuccessfully predicts an event—i.e., the event does not follow (negative evidence)—the association between the cue and the event will decrease.

In LF-learning, a single cue will be predictive of each of the features encountered in an object or event. Because no other cues are available to compete for associative value, there can be no loss of potential associative value to other cues over the course of learning trials. By contrast, in FL-learning, because many cues are available to compete for relevance, learning will separate the highly salient cues from the less salient cues, favoring cues with a high degree of positive evidence and disfavoring those with a high degree of negative evidence. FL-learning and LF-learning thus differ significantly in terms of cue-competition; the dense cue-structure of FL-learning fosters cue-competition, while the sparse cue-structure of LF-learning inhibits it.



**Figure 1.** Cue competition in FL-learning. The top panels depict the temporal sequence of events: an object is shown and then a word is heard over three trials. The lower panels depict the relationship between the various cues and labels in word learning.



**Figure 2.** When labels predict features (LF-learning), the absence of cue competition results a situation where the outcome of learning is simply a representation of the probability of the features given the label.

### Cue-structure and symbolic learning

To see how these factors affect symbolic learning, consider a world containing two kinds of animals: wugs and nizes. These animals have two main features: their shape and their color. Wugs are wug-shaped and are either blue or red. Likewise, nizes are nize-shaped and are either blue or red. Figure 1 represents FL-learning in this scenario.

At (i), a learner encounters an object with two features, shape-1 and red, and hears ‘wug’. The learner acquires information about two equally predictive relations, shape-1  $\Rightarrow$  ‘wug’ and red  $\Rightarrow$  ‘wug’. At (ii), the learner gets two new cues then a new label, and forms two new predictive relations, shape-2  $\Rightarrow$  ‘niz’ and blue  $\Rightarrow$  ‘niz’. At (iii), the learner encounters two previously seen cues, shape-1 and blue. Given what the learner already knows—shape-1  $\Rightarrow$  ‘wug’ and blue  $\Rightarrow$  ‘niz’—she expects ‘wug’ and ‘niz,’ but only ‘wug’ occurs. As a result: (1) the associative value of the relation shape-1  $\Rightarrow$  ‘wug’ increases; and (2) negative evidence—the non-occurrence of ‘niz’—causes a loss of associative value in blue  $\Rightarrow$  ‘niz.’ Crucially, as the value of blue  $\Rightarrow$  ‘niz’ decreases, its value *relative* to shape-2  $\Rightarrow$  ‘niz’ decreases. At (iv), a similar situation occurs. The learner encounters shape-2 and red and expects ‘niz’ and ‘wug’. When only ‘niz’ is heard, the value of shape-2  $\Rightarrow$  ‘niz’ increases, and red  $\Rightarrow$  ‘wug’ loses value. FL-learning is thus *competitive*:

losses can cause cue values to change *relative* to other cues: since one cue’s loss can be another’s gain, associative value can shift from one cue to another.

Now consider LF-learning (Figure 2). At (i), a learner encounters the label ‘wug’ and then an object with the features shape-1 and red. She thus learns about two equally valuable predictive relations ‘wug’  $\Rightarrow$  shape-1 and ‘wug’  $\Rightarrow$  red. Similarly, at (ii), the learner acquires two further equally valued relations ‘niz’  $\Rightarrow$  shape-2 and ‘niz’  $\Rightarrow$  blue. Now, at (iii), the learner hears ‘wug’ and expects red and shape-1. However, shape-1 occurs and blue occurs. This has three consequences: (1) positive evidence induces an increase in the associative value of ‘wug’  $\Rightarrow$  shape-1; (2) ‘wug’  $\Rightarrow$  blue becomes a new predictive relation; (3) negative evidence decreases the value of ‘wug’  $\Rightarrow$  red. However, since ‘wug’ is the only cue, this loss of associative value is *not* relative to any other cues (likewise at iv). LF-learning is thus *non-competitive*, and simply results in learning the “raw” probabilities of features given labels.

Both FL and LF-learning capture probabilistic information predictive relationships in the environment. However, there are fundamental differences between the two. In FL-learning predictive power, not frequency or simple probability, determines cue values; LF-learning is probabilistic in far more simple

terms. Given this, it seems that the sequencing of labels and features ought to have a marked affect on learning. We call this the **Feature-Label-Order (FLO) hypothesis**.

We formally tested the FLO hypothesis in simulations using a prominent learning model (Rescorla & Wagner, 1972; our analysis is compatible with many other error-driven learning models, and can be applied to learning other environmental regularities). The Rescorla-Wagner model states how the associative values ( $V$ ) of cue(s)  $i$  predicting an event  $j$  change over discrete training trials (where  $n$  indexes the current trial):<sup>1</sup>

$$\Delta V_{ij}^n = \alpha_i \beta_j (\lambda_j - V_{TOTAL}) \quad (1)$$

If there is a discrepancy between  $\lambda_j$  (the total possible associative value of an event) and  $V_{TOTAL}$  (the sum of current cue values), the saliency of the set of cues  $\alpha$  and the learning rate of the event  $\beta$  will be multiplied against that discrepancy. The resulting amount will then be added or subtracted from the associative strength of any cues present on that trial.

The associative strength between a set of cues  $i$  and an event  $j$  will increase in a negatively accelerated fashion over time, as learning gradually reduces the discrepancy between what is predicted and what is observed.

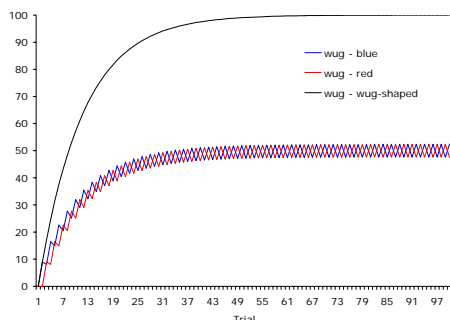


Figure 3. The development of cue values in a simulation of the LF-learning depicted in Figure 2.<sup>2</sup>

<sup>1</sup>  $V_{ij}$  is the change in associative strength on a learning trial  $n$ .  $\alpha$  denotes the saliency of  $i$ , and  $\beta$  the learning rate for  $j$ .

<sup>2</sup> In the simulations either a *niz* or a *wug* is encountered in each trial, and each species / color is equally frequent / salient in the environment,

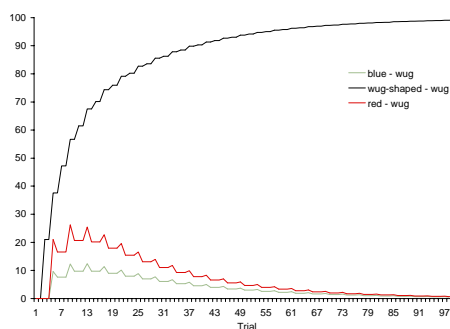


Figure 4. The development of cue values in a simulation of the FL-learning depicted in Figure 1.

### Discrimination and interference

Two Rescorla-Wagner simulations formally illustrate the differences in the representations learned after LF- and FL-training. As Figure 3 shows, LF-learning results in a representation of the probability of each feature given the label; e.g., the learned associative value of ‘wug’ $\Rightarrow$ red is about half of the associative value of ‘wug’ $\Rightarrow$ wug-shaped, because ‘wug’ predicts red successfully only 50% of the times and wug-shaped successfully 100% of the time. In FL-learning (Figure 4), the representations learned reflect the *value* of cues: the relationship ‘wug’ $\Rightarrow$ wug-shaped is very reliable, and is highly valued relative to cues that cause prediction error. In this case the association ‘wug’ $\Rightarrow$ red is effectively unlearned.

It is important to note that in LF-learning, the lack of discrimination produced by learning can lead to problems of interference in predicting events (or responses to them); LF-learning tends to produce representations in which a number of competing predictions are all highly probable. To illustrate this, we return to our wug / niz example, in which there were equal numbers of wugs and nizzes: *red* cued “wug” 50% of the time and “niz” 50% of the time. In this scenario, if a child trained LF on the animals saw a red wug and was asked what it was called, there is 100% probability that wug-shaped=wug and only 50% probability that red = niz. ‘Wug,’ is the obvious answer. Imagine, however, that there were fifty times as many blue wugs as blue nizzes in the population, and fifty times as many red nizzes

as red wugs. In this scenario, the color red will cue “wug” about 98% of the time and “niz” less than 2% of the time, simply based on frequency of occurrence. For a child trying to name a red wug, there’s again a near 100% probability that wug-shaped = wug, but now there’s also over 98% probability that red = niz. There will thus be a large degree of uncertainty regarding the correct answer. We call this *response interference*. The problem here is that tracking the frequencies of successful predictions does not pick out the cues that actually best *discriminate* a prediction from others. While both FL and LF-learning can lead to response-discrimination in an ideal world, LF-learning may fail to discriminate events when their frequencies vary; and in the real world, these frequencies inevitably do vary.

This analysis has been confirmed in further Rescorla-Wagner simulations. In these, the frequencies of categories were manipulated in the way described above, so that a non-discriminating feature was shared by high frequency and low frequency exemplars from different categories. When trained to predict labels from exemplars (FL-training) under these conditions, the model learned to discriminate and categorize well. Given LF-training, the model did not learn to discriminate, and categorized poorly.

Participants in several learning experiments demonstrated the same pattern of results. When trained to predict labels from exemplars (FL-learning), participants discriminated and categorized well. However, participants trained to predict exemplars from labels (LF-learning) discriminated and categorized poorly, even though logically, exactly the same information was available to them. Both groups saw the same labels and the same exemplars for the same amount of time. Simply manipulating the predictive / temporal relationship between the labels and exemplars in training resulted in very different patterns of performance (Ramscar et al, *in submission*). Similar asymmetries occur when categories are learned in either inference or classification tasks (Markman, & Ross, 2003). It is likely that the principles we

describe underlie these effects (see also Love, Medin, & Gureckis, 2004).

### **No Representation Without Taxation**

It is worth restating the role of expectation and error-driven learning in the formation of cognitive representations. FL-learning is—crucially—not confined to simply recording information about the probability of labels given cues. FL-training leads to accurate classification because the configuration of cues in such training produces cue competition, leading a learner to develop representations that ignore the actual (often misleading) probability of the cues present and focus solely on those cues that are reliably predictive. FL-training produces representations that trade completeness for discrimination. By contrast, LF-training produces representations that provide a more *veridical* picture of the structure of the world (i.e., the actual cue probabilities), yet are of less value when it comes to categorization and discrimination. In both cases, gains in information in one dimension come at the cost of a loss of information in another. This trade-off—gains at one level of representation are brought about by sacrifices at another level—appears to be an inherent feature of error-driven discrimination-learning. We call this the principle of *no representation without taxation*. Our previous findings confirm the advantages of FL-learning in categorization. However, they do not show that this resulted from any loss in the completeness of the representations of items at the cue-level. Experiment 1 was designed to examine this question. If FL-learning is advantageous in category learning because it distorts input representations, it follows that if LF-learning produces *less* distortion in what is learned,<sup>3</sup> we should expect that items learned about in LF-training will be represented more completely in memory.

### **Experiment 1**

#### ***Participants, Method and Materials***

To first examine the no representation without taxation hypothesis—i.e., that information

---

<sup>3</sup> Without cue competition cues will not “lose” value, so there will be no ‘distortion’ in the learned representation.

gains at one level of representation come at a cost to another level—we conducted an experiment. We wanted to see whether the increased accuracy in category discrimination tasks that FL-learning brings about result in less accurate memory for the items that participants learn from in training.

24 Stanford Undergraduates learned the names of six “species of aliens” (the six families of *fribbles* in Figure 5). The six families were divided into three family pairs, with each pair sharing the same body type (Figure 4 shows one pair of families, with each row corresponding to a separate family). Each participant learned one family pair in FL-configuration, one family pair in LF-configuration, and one pair split so that one family was learned LF and one FL. The families assigned to each configuration were counterbalanced across participants.

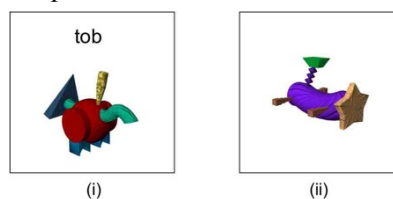


Figure 5. Exemplars of two *fribble* categories<sup>4</sup> used in training in Experiment 1. The configuration of the diagnostic features for each category is circled.

Training comprised 2 identical blocks presenting 18 exemplars of each of the six categories in a pseudo-randomized order (i.e., no two exemplars from a family pair were presented adjacently). To enforce LF or FL relationships as our participants studied “species of aliens” we minimized their ability to strategize (word learning is rarely a conscious process). All four categories were trained simultaneously, exemplars of each category were presented in a non-predictable sequence, and each exemplar was presented for only 175ms to inhibit participants’ ability to search for features. FL training trials comprised 1000ms presentation of a label (“this is a wug”), followed by a blank screen for 150 ms, followed by 175ms exposure to the exemplar. LF training trials comprised 175 ms

<sup>4</sup> Created by the Tarr lab at Brown University.

exemplar, 150 ms blank screen and 1000ms label (“that was a wug”). A 1000ms blank screen separated all trials.



Category Verification Exemplar Recognition

Figure 6. Category verification trials (i) paired *fribbles* with correct and incorrect labels. Exemplar recognition trials (ii) included *fribbles* seen in training, and new *fribbles* from the same families.

Participants were tested in two ways. A first test examined their ability to correctly discriminate between the categories they had learned about. Participants were presented with both an exemplar they had been exposed to in training and a label on-screen, and asked to respond “old” if the exemplar-label pairing was one they had learned, and “new” if it was not one they had learned. There were 10 “old” and 10 “new” tests per category. The “new” trials presented pairings not seen in training, and two exemplars were mismatched with each of the 5 alternative labels, yielding 120 test trials. The second test examined participants’ ability to discriminate the actual *fribbles* they had learned from novel exemplars. Participants were presented with 8 exemplars from each family that had been seen previously and 8 exemplars of each family that they had not seen previously, and asked to discriminate between them. This test yielded 96 test trials.

Category Verification Test

	Hit	False Alarm	d'
FL-trained	0.73	0.43	1.30
LF-trained	0.66	0.45	0.67

Exemplar Recognition Test

	Hit	False Alarm	d'
FL-trained	0.52	0.46	0.23
LF-trained	0.62	0.36	1.04

Table 1. Mean hit, false alarm, and *d'* rates by test and by training-type.

Results and Discussion

Table 1 presents rates for hits, false alarms and the signal detection measure *d'* for each task and training type. A 2 (training) x 2 (test)

repeated measures ANOVA revealed an interaction between the way that participants learned the categories and their performance in the tests ( $F(1,21) = 4.695, p < .05$ ). Post hoc paired t-tests revealed that participants were more accurate in verifying exemplar-label pairings when trained FL than LF ( $t(21) = 2.09, p < 0.05$ ). When participants were asked to *recognize* training items, the opposite was true: they were more accurate for items which had been trained LF rather than FL ( $t(21) = -1.9, p < 0.05$ ). As expected, improved learning about the categorization task appears to have come at the expense of accurate memory for the items this information was learned from.

### The Redistribution Of Wealth

Thus far, our discussion of learning has focused on learning to label objects in the world, and on classifying objects into labeled categories. In reality, of course, people learn far more about their environments than just the labels languages assign to things. We learn, for example, that objects of a certain form might be called “chairs,” but also learn that we can sit on chairs, that we can stand on them to reach high objects, and that should we find ourselves in a brawl in a wild west saloon, that chairs will shatter over the back of a black-hatted aggressor in an aesthetically pleasing, fight-ending manner. We also learn to extend these inferences by analogy, such that we might sit on a box when there is no chair to be had, or we might strike a saloon brawler with a bar stool when a chair is out of reach. To the extent that these kinds of predictions we make about the environment are more or less discrete, there is reason to believe that FLO-like effects, and the representational principles that accompany them, would be discernable in other learning tasks.

We noted at the outset that, in general, cognitive scientists have struggled with the concept of concept (see also Murphy, 2002). People use words, and words somehow relate to things in the world in relatively reliable ways, but these relations are often less than systematic. This had led to a certain amount of conceptual confusion when cognitive scientists

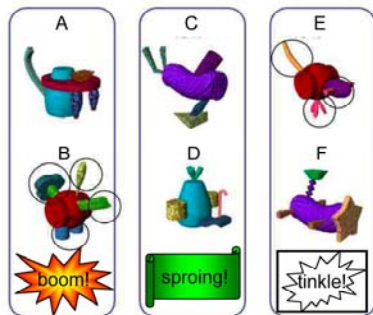
have come to think about interesting phenomena such as metaphor and analogy. The problem can be stated simply: metaphor and analogy are typically defined as ways of thinking that go beyond the literal; i.e., beyond the mappings between symbols and concepts. Given that the mappings between symbols and concepts are anything but concrete, this leaves the problem of distinguishing analogical, metaphorical and “categorical” processes hanging; there may *seem* to be a difference, but it’s hard to say what it is.

Our approach to symbolic thought offers a probabilistic solution to this problem: the difference between analogical and “categorical” processes is simply a matter of degree. In “categorical inferences,” two objects might both cue a very similar set of predictions, including a label. On the other hand, in “analogical inferences,” two objects might cue only a few, or even just one common prediction. For example, take two chairs: they will both cue the label “chair,” and numerous other inferences as well (we can sit on chairs, stand on them to reach high objects, etc.). Now consider a chair and a box: they cue some common predictions (sitting), but there are many inferences that boxes cue (storage), that chairs don’t, and vice versa; and of course, boxes and chairs predict different labels (“chair” and “box”).

The principles we have outlined also allow us to describe the effect that learning to conceptualize the world in this way ought to have on cognitive representations, and to consider the role of similarity in these conceptualizations. Learning to “represent” the world through LF-learning involves discovering the conditional probabilities of features given events and labels. In this case, similarity will be determined by the probability of the features that any two objects share; objects will be similar to the degree that they have salient (i.e., frequent) features in common. However, as learners come to conceptualize—i.e. discover the predictive structure of—their environments, they learn the predictive value of cues. As we have shown above, when learned in a system, predictive values can differ greatly

from conditional probabilities. In this case, we would expect similarity to be governed by predictive value, not frequency (see also Medin, Goldstone & Gentner, 1993).

Several points of interest arise out of this, which can be illustrated by considering the *fribble* families in Figure 7. The families are grouped into pairs that predict a distinctive sound (i.e. “boom” is predicted by families A and B in the left panel). If body types predict different noises, the resulting errors will cause them to lose value as cues. Assuming that similarity is a product of the alignment of shared dimensions (Markman, 1996), decreasing the value of a shared dimension should result in a decrease in similarity (in essence, two items will cease to have this dimension in common), while increasing the value of a shared dimension should result in an increase in similarity. What *actually* happens, however, will depend on cue competition, and the conceptual system being learned. As we will show, this leads to some surprising predictions.



**Figure 7.** Examples of the fribble families used in Experiment 2, with their predictive relations.

First, consider *fribble* families B and E: once their shared feature (body type) loses value in learning, the exemplars of families B and E should lose value still further when the predictive relations in Figure 7 are learned.

When families of *fribbles* begin with no obvious features in common and share no predictive structure, e.g., A and E, the effect of unlearning one dimension will neither increase nor decrease the number of alignable features; thus learning should have no effect on the similarity of these *fribble* families.

However, the diminishing of the predictive strength of the bodies (as in B and E) will make available associative value that can be acquired by other cues. This might act to increase the similarity of families that predict the same event. For example, families A and B both have pairs of features that project downwards, families C and D both have multiple posterior dorsal projections and single anterior dorsal projections, and families E and F both have small anterior lateral projections. Given that two families of *fribbles* predict each sound, these commonalities will be a more frequent cue than any other feature of the two *fribble* families, and thus the value of these features should increase relative to these other features, increasing the overall similarity of each of these pairs of families.

Finally, and perhaps most surprisingly, is the effect the corollary of this last point will have on exemplars of the **same** *fribble* family. Because each family of *fribbles* possesses a feature that generates error (body type), each will lose associative strength for that feature. This value will be distributed to other features as described above (with more value going to dimensions shared by families that predict the same event). However, in this case, the net effect of learning will be to *reduce* the overall number of alignable features in members of, say, A, and increase the value of dimensions that shared with members of B at the expense of features specific to A. Thus exactly the same processes that increases the similarity of A in relation to B may produce a within category *reduction* of similarity in A.

## Experiment 2

### *Participants, Method and Materials*

To see whether learning the conceptual system depicted in Figure 7 (Figure 5 provides a sense of within family variance) would bring about the redistribution of associative value described above, we taught it to 34 Stanford undergraduates using either an FL- or LF-configuration.

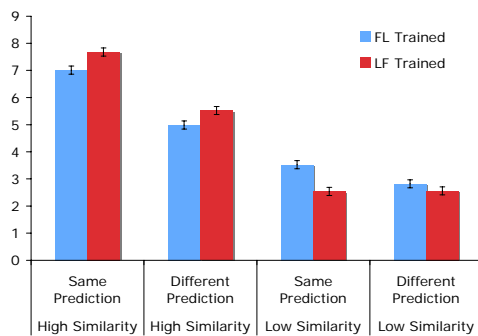
Participants passively viewed exemplars of the six “species of aliens” in Figure 7. The species were put into pairs, which predicted, or were predicted by, sounds (an explosion, a

spring, and breaking glass). The two families that predicted each sound shared only the few features described above, and each shared body type with a family that predicted a different sound. To enforce LF and FL training, and to minimize our participants ability to consciously track variance in the categories, 75% of the exemplars that predicted each sound were from one family (25% were from the other), and each non-discriminating feature was shared by high frequency exemplars predicting one sound and low frequency exemplars predicting another. Presentation utilized the speeded paradigm described above, except that instead of seeing a label, participants heard a 500ms sound file.

Because our predictions are specifically about the effect learning predictive structure has on similarity, we took two measures of the effects of this training: similarity, and inferential soundness (see also Gentner, Ratterman & Forbus, 1993). Participants were first asked to rate the similarity between pairs of *fribbles*, and then asked to rate the likelihood that if one *fribble* had a given property, another *fribble* shared that property. There were 2 blocks of tests, each comprising 12 similarity judgments and 12 inferential soundness judgments.

**Results and Discussion**

A 2 (LF vs FL training type) x 2 (high versus low initial similarity) x 2 (same or different prediction) ANOVA of our participants' ratings revealed the interaction between initial similarity and learning predicted by our analysis ( $F(1,1582)=11.594, p<0.001$ ).



**Figure 8:** Average similarity and inferential soundness ratings in experiment 2 (error bars are SEM)

Post hoc tests showed the FL-training decreased both the similarity (*Sim*) and inferential soundness (*IS*) ratings of *fribbles* that shared body type but which predicted different sounds (*Sim*  $t(262)=2.193, p<0.05$ ; *IS*,  $t(262)=2.43, p<0.05$ ), while increasing the similarity between dissimilar *fribbles* that did predict the same sounds (*Sim*  $t(130)=2.265, p<0.05$ ; *IS*  $t(130)=2.43, p<0.001$ ).

While participants' ratings of dissimilar object that predicted different sounds were the same after both LF and FL-training, ratings of *fribbles* in the same families (which predicted the same sounds) were reliably *lower* after FL-training (*Sim*  $t(196)=3.723, p<0.001$ ; *IS*  $t(196)=2.709, p<0.01$ ). In the context of learning a conceptual system, the same training that increases participants' ability to classify and discriminate correct and incorrect classifications of items from memory, *reduced* their perceived similarity in inferential soundness!

**General Discussion**

We have shown how conceiving of symbols and symbolic learning in the context of the predictive structure of the environment can provide new insights into the nature and learning of cognitive representations. While our view of learning is associative, it is compatible with approaches that emphasize the role of structural alignment in similarity, analogy and categorization (Gentner, 1983; Medin, Goldstone & Gentner, 1993; Markman, 1996). It is, however, incompatible with the idea that similarity *drives* categorization (see also Goodman, 1972). In our view, similarity and categorization are *products* of learning. We believe the framework in which we present these ideas has much to offer our understanding of cognition.<sup>5</sup>

**REFERENCES**

Elman, J. L. (1991) Finding structure in time. *Cognitive Science*, 14, 179-211  
 Fodor, J. (1998). *Concepts*. New York: OUP  
 Gentner, D. (1983). Structure-mapping: A theoretical framework for analogy. *Cognitive Science*. 7 (2), 155-170.

<sup>5</sup> This material is based on work supported by NSF Grant Nos. 0547775 and 0624345 to Michael Ramscar.

- Gentner, D., Ratterman, M., & Forbus, K. (1993). The role of similarity in transfer: Separating retrievability from inferential soundness. *Cognitive Psychology*, 25, 524-575
- Goodman, N. (1972) *Problems and Projects*. Indianapolis: Bobbs-Merril
- Love, B.C., Medin, D.L., & Gureckis, T.M. (2004). SUSTAIN: A Network Model of Category Learning. *Psychological Review*, 111, 309-338
- Medin, D.L., Goldstone, R.L., & Gentner, D. (1993). Respects for similarity. *Psychological Review*, 100, 254-278
- Markman, A.B. (1996). Structural alignment in similarity and difference judgments. *Psychonomic Bulletin and Review*, 3(2), 227-230.
- Markman, A. B. & Ross, B. H. (2003) Category use and category learning. *Psychological Bulletin*, 129, 592-613
- Murphy, G. L. (2002). *The big book of concepts*. Cambridge: MIT Press
- Penn D.C., Holyoak K.J., Povinelli D.J. (2008) Darwin's mistake: Explaining the discontinuity between human and nonhuman minds. *Behavioral & Brain Science*, 31:109-178
- Quine, W.V. (1960) *Word and Object*, Cambridge, MA: MIT Press
- Ramscar, M., Yarlett, D., Dye, M., Denny, K., & Thorpe, K. (in submission) *The Feature-Label-Order Effect And Its Implications For Symbolic Learning*.
- Rescorla R.A. and Wagner A.R. (1972). A Theory of Pavlovian Conditioning: Variations in the Effectiveness of Reinforcement and Nonreinforcement. In Black & Prokasy (Eds.), *Classical Conditioning II: Current Research and Theory* (pp. 64-99). New York: Appleton-Century-Crofts
- Wittgenstein, L (1953). *Philosophical Investigations*. Oxford: Blackwell