

No Representation without Taxation: The Costs and Benefits of Learning to Conceptualize the Environment

Melody Dye and Michael Ramscar

Department of Psychology, Stanford University,
Jordan Hall, Stanford, CA 94305.

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 based on their highly-discriminating features makes people less likely to notice or remember the same objects' less-discriminating features. Gains in response-discrimination between categories thus come at a cost to within category discrimination. We suggest that the mechanisms of human learning obey a simple principle: there can be no representation without taxation.

Introduction

While we perceive the world through our senses, we do not experience the world in terms of raw sense data; rather 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. 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. In particular, in language-learning tasks, improved response-discrimination—i.e., improved accuracy in dividing up perceptual input into conceptual categories—comes at a cost to the representation of the original input. 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 appear to home in on the particular 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 what follows, we lay out these ideas in detail and present empirical evidence in support of them. We argue that the basic principle of no representation without taxation amounts to a fundamental law of learning.

Learning

Formally, learning can be conceived of as a process by which probabilistic information is acquired about the relationships between important regularities in the environment (such as objects or events) and the cues that allow those regularities to be predicted (Rescorla & Wagner, 1972). The learning process is driven by discrepancies between what is expected and what is actually observed in experience (termed error-driven learning). The learned predictive value of a given cue produces 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 are under-predicted by that cue, and weakened when they are over-predicted (Kamin, 1969; 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. Learning is thus the product of both positive and negative evidence in the environment. That is, if one takes a cue *C* to predict an Event *E* and *E* occurs, then the associative strength between *C* and *E* increases to the degree that *E* was underpredicted (positive evidence). However, if one

takes C to predict E, but E does not occur, then the associative strength between C and E decreases to the degree that E was overpredicted (negative evidence). This process produces patterns of learning that are very different from what would be expected if learning were shaped by positive evidence alone (a common portrayal of Pavlovian conditioning, Rescorla, 1988).

Symbolic learning

Language learning involves acquiring probabilistic information about the predictive relations between two aspects of the environment: labels and their semantic features. By a label we mean a token of language, such as the word ‘pan,’ and by semantic feature, we mean the features of objects, events or any ‘thing’ that is communicated about in symbolic language.

Since the predictive relations relevant to language learning are relations between labels and features, we can distinguish two possible sequential forms that symbolic learning might take: (i) *when cues are labels and events are semantic features*; (ii) *when cues are semantic features and events are labels*.

In case (i), which we will call **LF-learning**, one learns to predict and expect a certain feature from a given label. In case (ii), which we will call **FL-learning**, one learns to predict and expect a certain label from a given feature or set of features. To understand the difference between what is actually learned in LF-learning as opposed to FL-learning, it is important to first note some important differences between labels, as they are employed in language, and the aspects of the environment they typically describe.

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 discrete, and possess much denser cue-structure. (By cue-structure we mean the amount of salient and discriminable cues that are simultaneously present in the thing—label or object—in question.)

Consider a situation in which an object—say, a *pan*—is encountered in the environment. Even if one focuses on the pan and ignores other features of the background, one still encounters many discriminable features at once: shape, color, size, etc. Thus, when a pan predicts its label, it simultaneously provides a learner with many potentially discriminable cues to that label. Further, because objects are not discrete (pans share many features with things that are not pans), when the features of pans serve as cues to the label ‘pan,’ some will cue other labels as well. In learning, all of these features will compete for relevance as better or worse predictors of ‘pan.’

By contrast, consider the label ‘pan.’ A native English speaker can rapidly parse this word into a

sequence of phonemes [*p^h an*], but will otherwise be largely unable to discriminate many further features within these sounds. This is not to say that there are no other discriminable features within speech (such as emphasis, volume, or pitch contour), but rather to say that the dominant semantic feature is at the level of the phoneme. Ordinarily, other features of speech, such as pitch contour, do not compete with phonemes in predicting meaning in the same way that *color* might vie for relevance with *shape* in predicting an object label.¹ And because phonemes occur in a sequence rather than simultaneously (see McClelland & Elman, 1986), there can be little to no direct competition between them as cues. Thus, when the label ‘pan’ serves as a cue, the label comes with little competitive cue-structure: ‘pan’ essentially provides the learner with only one potential cue, i.e. the label ‘pan’ itself.

The difference in cue-structure between labels and objects allows us to make a distinction between the two forms of learning. In LF-learning, since only one label at a time serves as a cue and since individual labels have little cue-structure, learning involves predicting a complex response from a single cue. LF-learning thus has a **one-to-many** form: one cue (the label) to many responses (the features).

On the other hand, in FL-learning, when an object serves as the cue set, learning involves predicting a single response from a dense set of cues. FL-learning thus has a **many-to-one** form: many cues (the features) to one response (the label).

The impact of Cue-Competition on Learning

To see how these factors affect symbolic learning, consider a simplified environment in which there are two kinds of objects: wugs and nizes. These objects have two discriminating features: their shape and their color. Wugs are wug-shaped and can be either blue or red. Likewise, nizes are niz-shaped and can be either blue or red. Suppose now that one is learning what wugs and nizes are under FL-learning conditions. Figure 1 represents FL-learning in this simplified environment.

At (i), the learner encounters an object with two salient features, shape-1 and red, and then hears the label ‘wug.’ The learner acquires equal information about two predictive relations, shape⇒‘wug’ and red⇒‘wug.’ At (ii), the learner encounters two new cues and a new label, and forms two new equally weighted predictive relations, shape-2⇒‘niz’ and

¹ In speech, it seems that there are complementary cue streams rather than cues in competition. For example, the ordinary way that words are stressed is complementary to their phonemic structure and the ways in which they are used. In English, phonemes and stresses do not *compete* for relevance as cues.

blue \Rightarrow 'niz'. Then at (iii), the learner encounters two previously seen cues, shape-1 and blue. Given what the learner already knows—i.e., shape-1 \Rightarrow 'wug' and blue \Rightarrow 'niz'—she expects to encounter both 'wug' and 'niz,' but only 'wug' actually occurs. As a result: (1) the associative strength for the relation shape-one \rightarrow 'wug' increases, given the positive evidence of the occurrence of 'wug'; and importantly (2) the negative evidence for the non-occurrence of 'niz' causes a loss of associative strength of blue \rightarrow 'niz.' Crucially, as the associative strength of blue \Rightarrow 'niz' decreases, it's value *relative* to shape-2 \Rightarrow 'niz' decreases as well. At (iv), a

similar situation occurs. The learner encounters shape-2 and red and expects 'niz' and 'wug' to occur. When only 'niz' is heard, the associative strength of shape-2 \Rightarrow 'niz' increases, while red \Rightarrow 'wug' loses associative strength.

FL-learning is thus **competitive**: when cues lose associative strength, this changes their values *relative* to other cues. Importantly, associative value can *shift* from one cue to another; as one cue loses value, that value can be subsumed—i.e., gained—by another.

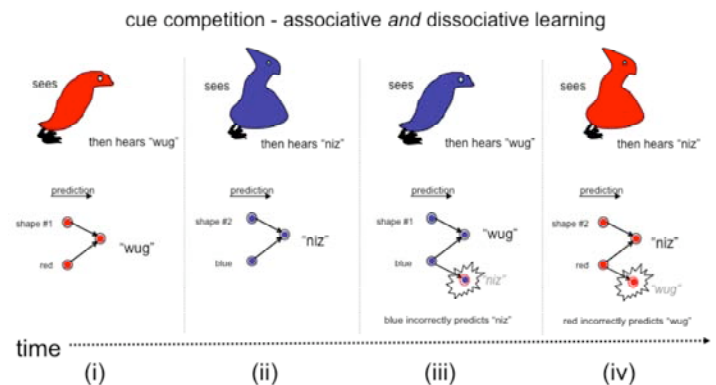


Figure 1. Cue competition in learning. The top panels depict the temporal sequence of events: an object is shown and then a word is heard over three trials. The lower panels depict the relationship between the various cues and labels in word learning.

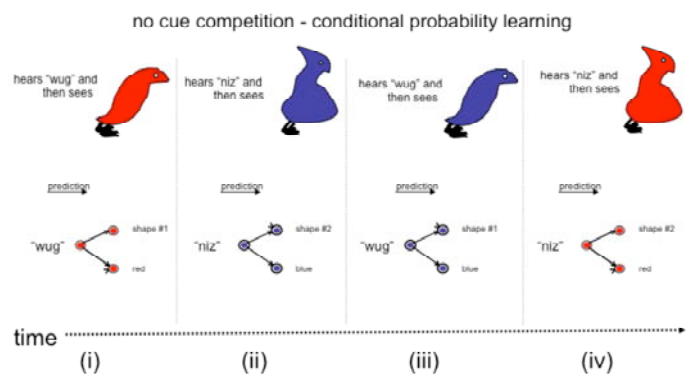


Figure 2. When **labels predict features**, the absence of cue competition results a situation where the outcome of learning is simply be a representation of the probability of the features given the label.

Now consider LF-learning in a similar scenario (Figure 2). At (i), the learner encounters the label 'wug' and then an object with the two salient features, shape-1 and red. She thus learns about two equal 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 the responses red and shape-1. However, shape-1 occurs and blue occurs. This has three consequences: (1) positive evidence causes an increase in the associative strength of

'wug' \Rightarrow shape-1; (2) 'wug' \Rightarrow blue becomes a new predictive relation; (3) negative evidence decreases the associative strength of 'wug' \Rightarrow red. However, since 'wug' is the only cue, and there is no cue competition, this loss of associative strength does not occur relative to any other cues. Likewise at (iv), we have an increase in 'niz' \Rightarrow shape-2, a new relation 'niz' \Rightarrow red and a decrease in 'niz' \Rightarrow blue. But again, these losses and gains in associative strength do not occur relative to other cues, since 'niz' is the sole cue.

LF-learning is thus termed **non-competitive** and results in learning the probabilities of events occurring given a particular cue. We call this “conditional probability learning.”

The Feature-Label-Order Hypothesis

Both FL and LF-learning capture probabilistic information about the predictive relations between cues and responses in the environment; in each case, the relations are affected both by positive and negative evidence. However, there are fundamental distinctions between the two forms of learning. In FL-learning, since the cue-structure tends to be dense, cue-competition tends to be strong; thus, FL-learning is competitive in addition to being non-competitive. On the other hand, LF-learning fails to satisfy dense cue-structure and remains solely non-competitive.

The following two computational simulations (in the Rescorla & Wagner, 1972 model)² formally illustrate the differences in the representations of what one might would expect to get learned in LF and FL-learning. As Figure 3 shows, LF-learning simply 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 strength of ‘wug’ \Rightarrow wug-shaped, because ‘wug’ predicts red successfully only 50% of the time while wug-shaped successfully 100% of the time. In FL-learning (Figure 4), the learned representations reflect the *value* of cues: e.g., the associative relationship wug-shaped \Rightarrow ‘wug’ is very reliable, and is highly valued relative to cues that generate prediction error. In this case the association red \Rightarrow ‘wug’ is effectively unlearned, since red is a poor predictor of ‘wug.’

It appears, therefore, that the sequencing of labels and features has a significant effect learning. *We call this the Feature-Label-Order hypothesis.*

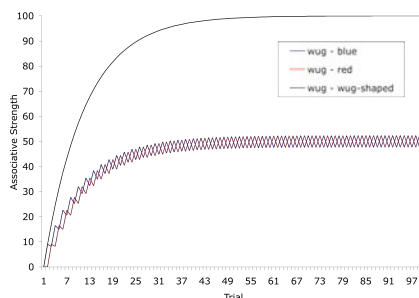


Figure 3. The development of cue values in a simulation of the LF-learning scenario depicted in **Figure 2**.

² The simulations assume either a *niz* or a *wug* is encountered in each trial, that each species and color is equally frequent in the environment, and that color and shape are equally salient.

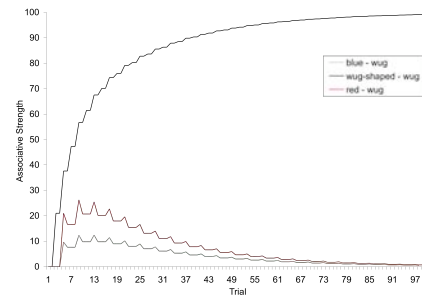


Figure 4. The development of cue values in a simulation of the FL-learning scenario depicted in **Figure 1**.

Feature-Label-Order And Representation

Our analysis predicts that the lack of cue structure in labels will inhibit category-learning when words serve as cues (LF-learning) as compared to when they are predicted by objects (FL-learning). Further, as the simulations reveal, in FL-learning a sacrifice is made in terms of the representation of items at the cue-level in order to gain discriminatory accuracy at the response-level. This suggests a complementary prediction: if LF-learning produces less distortion in the representation of cues, we should expect that items learned about as cues in LF-training ought to be represented more accurately in memory.

To examine both sides of the *no representation without taxation hypothesis*—i.e., that information gains at one level of representation come at a cost to another level—we conducted an experiment to see whether the increases in category discrimination brought about by FL-learning would be accompanied by decreases in the completeness of people’s representations of the items they had encountered in FL-training as compared to items they had encountered in LF-training.

24 Stanford Undergraduates learned the names of six “species of aliens”. The six families were divided into three family pairs, with each pair sharing the same body type (Figure 5 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* categories³ 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 predictive relationships in training, we minimized participants' opportunities to strategize. All six categories were trained simultaneously, with exemplars interspersed in a semi-randomized order so that exemplars of each category were presented in a non-predictable sequence. Exemplars were presented for only 175ms to inhibit participants' ability to consciously search for features (Woodman & Luck, 2003). LF-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. FL-training trials comprised 175 ms exemplar, 150 ms blank screen and 1000ms label ("that was a wug"). A 1000ms blank screen separated trials.

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. The second test yielded 96 test trials.

Results

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 feature-label pairings when trained FL than LF ($t(21) = 2.09$, $p < 0.05$). However, 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 hypothesized, improved learning about the categorization task appears to have come at the expense of accurate memory for the items that this information was learned from. Participants' gains at one level of representation appear to have come at a cost to another.

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.

Discussion

In this paper, we have sought to take seriously the task of accounting for both learning and representation. We have taken the view that in order for something to be represented it must be learned. Representations must therefore be subject to the constraints that are imposed by learning mechanisms.

The principle we have discussed here—that learning distorts inputs—is implicitly enshrined in the mechanisms of neural network models (Rosenblatt, 1959; Rescorla & Wagner, 1972; Rumelhart, Hinton & McClelland, 1986). To the extent that these models capture some aspects of the way learning works (Miller, Barnet & Grahame, 1995; Siegel & Allan, 1996), there seems to be reason to believe that human learning may be governed by the same principle (Hollerman & Schultz, 1998; Barlow, 2001), i.e., that learning about things can only come at some cost to the completeness of the representations of the things learned from. The results of our experiment support this analysis. We believe that this principle of no representation without taxation governs the processes by which we conceptualize the world.

What might this contribute to the study of perceptual and cognitive representation? Briefly: it follows from our account that in order for the contents of a perceptual (phenomenal) representation to be instantiated or realized, the elements in the representation must be individually cashed out. In order to make those elements discrete, they must be discriminated from one another, and this will necessarily involve a loss of information at the level of the perceptual input. By this line of reasoning, a purely perceptual representation doesn't represent any thing per se; rather it is unanalyzed, or holistic. What we call cognitive

³ Created by Michael Tarr's lab at Brown University.

representations, on the other hand, are the result of discrimination (and discrimination-learning); that is, they are arrived at by taking a perceptual representation and slicing it up, discarding information for a purpose.

This approach offers a new perspective from which to consider a wide range of cognitive phenomena. For example, in Sperling's (1960) partial report experiments, observers are required to identify a subset of the characters within the visual display. At various intervals after the removal of the visual display a tone is sounded to indicate to observers which particular set of characters within the display they are to report (e.g., the top, middle or bottom row). Participants are able to recollect four to five characters, irrespective of how many other characters were present within the display, and frequently report the phenomenal impression of many more elements immediately after display presentation. This is often taken as evidence that observers have a representation of all of the elements in the array, and that non-reported elements are lost to decay in memory.

No representation without taxation suggest a more nuanced perspective of this phenomenon. We suggest that it is a mistake to think that there is, say, a "7" or "k" in a phenomenal representation; the point is, there isn't. The information to *discriminate* a "7" or a "k" is there, but moving from a phenomenal representation to a cognitive representation necessarily involves a loss of information that allows the "7" and "k" to be discriminated cognitively from the rest of the phenomenal representation.

We do not wish to claim here that this phenomena reduces simply along these lines. The ways in which multiple memory systems integrate their representations to produce human experience and behavior is likely to be far more complex than this simple sketch. Our hope is simply that by better illuminating the nature of cognitive representations and the constraints on their development, we may help further our understanding of how it is that we come to represent the world.

Acknowledgments

This material is based upon work supported by NSF Grant Nos. 0547775 and 0624345 to Michael Ramscar

References

- Barlow H. (2001). Redundancy reduction revisited, *Network: Computation in Neural Systems*, 12, 241-253.
- Hollerman J.R., Schultz W. (1998) Dopamine neurons report an error in the temporal prediction of reward during learning. *Nature Neuroscience*, 1: 304-309.
- Kamin L.J. (1969). Predictability, surprise, attention, and conditioning. In: Campbell B, Church R (eds).
- Punishment and Aversive Behaviour. Appleton-Century-Crofts: New York.
- Miller, R. R., Barnett, R. C., & Grahame, N. J. (1995). Assessment of the Rescorla-Wagner model. *Psychological Bulletin*, 117, 363-386
- McClelland, J.L., & Elman, J.L. (1986). The TRACE model of speech perception. *Cognitive Psychology*, 18, 1-86.
- Ramscar, M. & Yarlett, D. (2007) Linguistic self-correction in the absence of feedback: A new approach to the logical problem of language acquisition *Cognitive Science*. Vol. 31, pp 927-960
- Ramscar, M., Yarlett, D., Dye, M., Denny, K., & Thorpe, K. *in submission* (in submission) The Feature-Label-Order Effect In Symbolic Learning
- Rescorla R.A. and Wagner A.R. (1972). A Theory of Pavlovian Conditioning: Variations in the Effectiveness of Reinforcement and Nonreinforcement. In A.H. Black & W.F. Prokasy (Eds.), *Classical Conditioning II: Current Research and Theory* (pp. 64-99). New York: Appleton-Century-Crofts.
- Rosenblatt, F (1959). *Principles of Neurodynamics*. Spartan Books. New York. 1959
- Rumelhart, D. E., Hinton, G. E., & McClelland, J. L. (1986). A general framework for parallel distributed processing. In Rumelhart, D. E., & McClelland, J. L. (Eds.) *Parallel Distributed Processing: Explorations in the Microarchitecture of Cognition*, Volume I. Cambridge, MA: MIT Press.
- Rosenblatt, F. (1959) "Two Theorems of Statistical Separability in the Perceptron". In *Mechanisation of Thought Processes, Vol. I*. London: H.M. Stationery Office, pp.419-72.
- Siegel and Allan (1996). The widespread influence of the Rescorla-Wagner model, *Psychonomic Bulletin and Review*, 3(3), 314-321
- Sperling, G. (1960). The information available in brief visual presentations. *Psychological Monographs: General and Applied*, 74(11), 1-30.
- Woodman, G. F., & Luck, S. J. (2003). Electrophysiological measurement of rapid shifts of attention during visual search. *Journal of Experimental Psychology: Human Perception and Performance*, 29, 121-138.