Title
The Feature-Label-Order Effect In Symbolic Learning

Permalink
https://escholarship.org/uc/item/0g57k59z

Journal
Proceedings of the Annual Meeting of the Cognitive Science Society, 31(31)

ISSN
1069-7977

Authors
Dye, Melody
Kalchbrenner, Nal
Ramscar, Michael
et al.

Publication Date
2009

Peer reviewed

The Feature-Label-Order Effect In Symbolic Learning

Michael Ramscar, Daniel Yarlett, Melody Dye, & Nal Kalchbrenner
Department of Psychology, Stanford University,
Jordan Hall, Stanford, CA 94305.

Abstract

We present a formal analysis of symbolic learning that predicts significant differences in symbolic learning depending on the sequencing of semantic features and labels. A computational simulation confirms the Feature-Label-Ordering (FLO) effect in learning that our analysis predicts. Discrimination learning is facilitated when semantic features predict labels, but not when labels predict semantic features. A behavioral study confirms the predictions of the simulation. Our results and analysis suggest that the semantic categories people use to understand and communicate about the world might only be learnable when labels are predicted from objects.

Introduction

The ways in which symbolic knowledge is learned and represented in the mind are poorly understood. We present an analysis of symbolic learning—in particular, word learning—in terms of error-driven learning, and consider two possible ways in which symbols might be learned: learning to predict a label from the features of objects and events in the world; or learning to predict those features from a label. This analysis predicts significant differences in symbolic learning depending on the sequencing of objects and labels, confirmed in computational simulations and an empirical study. Discrimination learning is facilitated when semantic features predict labels, but not when labels predict semantic features. We call this the Feature-Label-Ordering (FLO) effect. Our results and analysis suggest that the semantic categories people use to understand and communicate about the world can only be learned if labels are predicted from objects.

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). This process is driven by discrepancies between what is expected and what is actually observed in experience (termed error-driven learning). The learned 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 cues are strengthened when relevant events are under-predicted, 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. 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 information about the relations between labels and their semantic features. Here we define labels as tokens of language, such as the word 'pan,' and semantic features as the properties of the objects and events communicated about in language. In turn, we can distinguish two possible forms that symbolic learning about labels and features can take:

(i) cues are labels and events are semantic features;

(ii) cues are semantic features and events labels.

In (i), which we call Label-to-Feature or **LF-Learning**, one learns to predict and expect certain features given a label. In (ii), which we call Feature-to-Label or **FL-Learning**, one learns to predict and expect labels given a feature or certain set of features. To explain the difference between what is learned in LF-learning versus FL-learning, it is important to note some 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 number of potentially discriminable cues that are simultaneously present.) Consider a situation in which say, a *pan* is encountered in the environment. A pan presents to a learner many discriminable features: shape, color, size, etc. However, because objects are not discrete (i.e., pans share many features with things that are not pans), some of these features will cue other labels as well.

By contrast, consider the label 'pan.' A native English speaker can parse this word into a sequence of phonemes [p^h *an*], but will be unable to discriminate many further features. 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. Other features of speech 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. Further, because phonemes occur in a sequence rather than simultaneously, there can be little to no direct competition between them as cues. Thus, labels such as 'pan' provides little competitive cue-structure: 'pan' essentially provides the learner with only a single cue, i.e. the label 'pan' itself.

The difference in cue-structure in turn affects the formal properties of the two forms of learning we described above. In LF-learning, because labels serve as cues and since individual labels have little cue-structure, learning involves predicting a set of features (the semantic features of objects and events) from a single cue (the label). Thus, LF-learning has a one-to-many form: one cue to many features.

By contrast, in FL-learning, when object or event serve as cues, learning involves predicting a single response (a label) from a large set of cues (the features of an event or object). Thus FL-learning, has a many-to-one form: from many semantic features to a label.

Cue-competition in learning

Where many cues are presented simultaneously, they can compete for relevance in the prediction of a particular event. If a cue successfully predicts an event over time (positive evidence), the associative strength between the cue and the event will increase. Conversely, when a cue unsuccessfully predicts a given event—i.e., the event does not follow the cue (negative evidence), the associative strength between the cue and the response will decrease.

In one-to-many LF-learning, a single cue will be predictive of each of the many 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 many-to-one FL-learning, because many cues are available to compete for relevance, learning will separate the highly reliable cues from the less reliable 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.

Cue-structure and symbolic 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 salient 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), a learner encounters an object with two poentially relevant features, shape-1 and red, and then hears the label 'wug'. The learner acquires information about two equally predictive relations, shape-⇒'wug' and red⇒'wug'. At (ii), the learner two new cues and a new label, and forms two new equally weighted predictive relations, shape-2⇒'niz' and blue⇒'niz'. Then at (iii), the learner encounters two previously seen cues, shape-1 and blue.

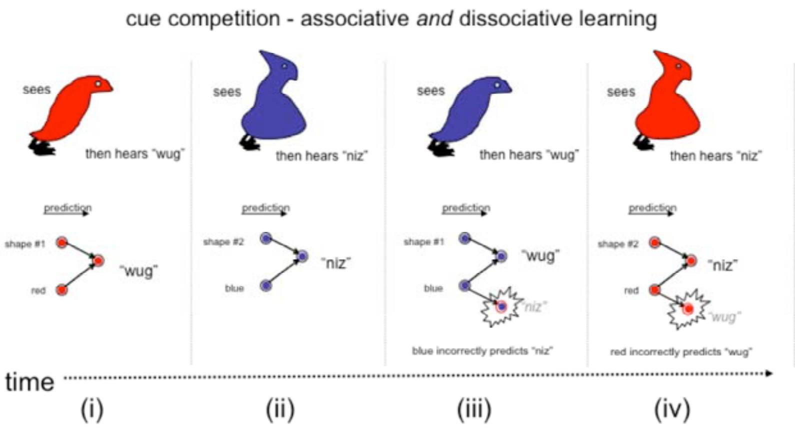

Figure 1. Cue competition in Feature-to-Label learning. The top panels depict the temporal sequence of events: an object is shown and then a word is heard over three trials. The lower panels depict the relationship between the various cues and labels.

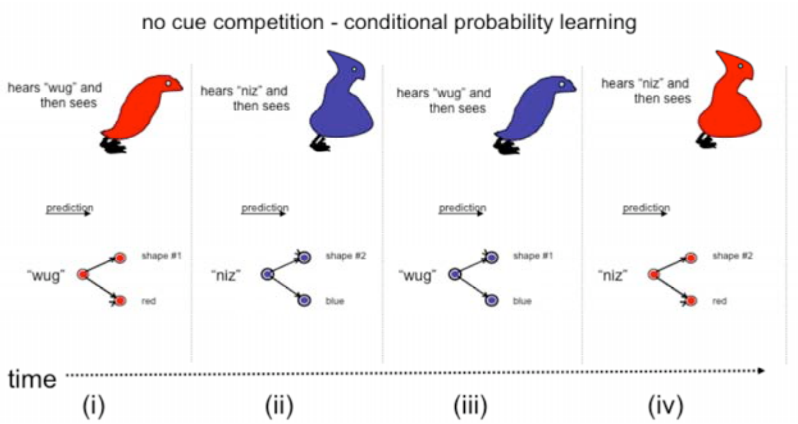

Figure 2. In Label-to-Feature Learning, the absence of cue competition results in a situation where the outcome of learning is simply a representation of the probability of the features given the label.

Given what the learner already knows—i.e., shape-1⇒'wug' and blue⇒'niz'—she expects 'wug' and 'niz,' but, only 'wug' actually occurs. As a result: (1) given positive evidence of the occurrence of 'wug', the associative strength for the relation shape-1⇒'wug' increases; and importantly (2) negative evidence about the non-occurrence of 'niz' causes blue⇒'niz' to lose associative strength. Crucially, as the associative strength of blue⇒'niz' decreases, its value *relative* to shape-2⇒ 'niz' changes as well. At (iv), a similar situation occurs. The learner encounters shape-2 and red and expects 'niz' and 'wug'. Only 'niz' is heard, so the associative strength of shape-2⇒'niz' increases, while red⇒'wug' loses associative strength.

FL-learning is thus *competitive*: if a cue loses associative strength, its value can change *relative* to other cues. Since one cue's loss can be another's gain, this may *shift* associative value from one cue to another.

Now consider LF-learning in a similar scenario (Figure 2). At (i), a learner encounters the label 'wug' and then an object with the two salient features, shape-1 and red. She thus learns about two equally valuable predictive relations 'wug' ⇒shape-1 and 'wug'⇒red. Similarly, at (ii), the learner acquires two further equally valued relations 'niz'⇒shape-2 and 'niz'⇒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 strength of 'wug'⇒shape-1; (2) 'wug'⇒blue becomes a new predictive relation; (3) negative evidence decreases the strength of 'wug'⇒red. However, since 'wug' is the only cue, this loss of associative strength is *not* relative to any other cues (likewise at iv). LF-learning is thus *non-competitive*, and simply results in the learning of the probabilities of events occurring given cues.

The Feature-Label-Order Hypothesis

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 hypothesis**.

We formally tested the FLO hypothesis in simulations using a prominent error-driven learning model (Rescorla &Wagner, 1972; see also; Allen and Siegel, 1996). We should note that the analysis of symbolic learning described here could be implemented in a number of other models (e.g., Pearce & Hall, 1980; Rumelhart, Hinton & McClelland, 1986; Barlow, 2001) and applied to learning other environmental regularities.

The Rescorla-Wagner model formally states how the associative values (V) of a set of cues i predicting an event j change as a result of learning in discrete training trials, where n indexes the current trial.

Equation (1) is a discrepancy function that describes the amount of learning that will occur on a given trial; i.e., the change in associative strength between a set of cues i and some event j:[1]

$$\Delta V_{ij}^{n} = \boldsymbol{\alpha}_i \boldsymbol{\beta}_j \, (\lambda_j - V_{TOTAL}) \qquad (1)$$

If there is a discrepancy between λ_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 α and the learning rate of the event β 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.

[1] V_{ij} is the change in associative strength on a learning trial n. α denotes the saliency of i, and β the learning rate for j.

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. Given an appropriate learning-rate, learning asymptotes at a level that minimizes the sum-of-squares prediction error for a set of observed cues to an event.

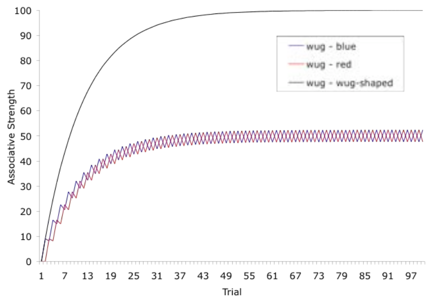

Figure 3. The development of cue values in a simulation of the LF-learning scenario depicted in **Figure 2**.

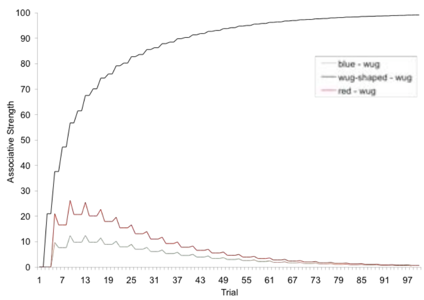

Figure 4. The development of cue values in a simulation of the FL-learning scenario depicted in **Figure 1.**

Discrimination and interference

Two computational simulations (in the Rescorla & Wagner, 1972 model, described below)[2] formally illustrate the differences in the representations of what gets 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'⇒red is about half of the associative strength of 'wug'⇒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 learned representations reflect the *value* of cues: the associative relationship 'wug'⇒wug-shaped is very reliable, and is highly valued relative to cues that generate prediction error. In this case the association 'wug'⇒red is effectively unlearned.

[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.

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. Imagine a world in which 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 "niz" about 98% of the time and "wug" 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 a 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 best discriminate one prediction from another. Thus, while both FL and LF-learning may produce successful response-discrimination in an ideal world, LF-learning will fail to discriminate events when their frequencies vary; and in the actual world, these frequencies inevitably will.

		Non discriminating features			Discriminating features					
		1	2	3	1	2	3	4	5	6
Category 1	75%	1	0	0	1	0	0	0	0	0
	25%	0	1	0	0	1	0	0	0	0
Category 2	75%	0	1	0	0	0	1	0	0	0
	25%	0	0	1	0	0	0	1	0	0
Category 3	75%	0	0	1	0	0	0	0	1	0
	25%	1	0	0	0	0	0	0	0	1

Figure 5: The abstract representations of the category structures used to train the Rescorla-Wagner models

Simulating interference

To illustrate the problem of response interference, we simulated category learning in the Rescorla-Wagner model using abstract representations of the category structures in Figure 5. The training set comprised 3 category labels and 9 exemplar features (3 non-discriminating features that were shared between exemplars belonging to different categories, and 6 discriminating features that were not shared with members of another category). The frequency of the sub-categories was manipulated so that each labeled category drew 75% of its exemplars from one sub-category and 25% of its exemplars from another subcategory. The two sub-categories that made up each labeled category did not share any features, such that learning to correctly classify one of the sub-categories paired with each label would provide no assistance with learning the other sub-category paired with that label. Finally, each low frequency sub-category shared its non-discriminating feature with the high frequency exemplars of a different labeled category. This manipulation was designed to create a bias towards the misclassification of the low-frequency exemplars. Learning to correctly classify low frequency exemplars necessarily required learning to weigh the discriminating feature more than the non-discriminating feature, despite its lower overall input frequency.

Two simulations were configured to created two networks of feature and label relationships. The first network learned associative weights from the 9 exemplar features (serving as cues) to the 3 labels (serving as events; “FL training”), while in the second case the network learned from the 3 labels (serving as cues) to the 9 features (serving as events; LF training). Each category had a high frequency exemplar, presented on 75% of the training trials for that category, and a low frequency exemplar (occurring 25% of the time). On each training trial a label and appropriate exemplar pattern were selected randomly to train each of the two networks. Training comprised 5000 trials, which allowed learning to reach asymptote. The model has several parameters that affect learning. For simplicity, the simulations assumed equally salient cues and events (α=0.01 for all i; β=0.01 for all j) and equal maximum associative strengths (= 1.0).

To test the FL-network, exemplar features were activated to determine the subsequent activation of the labels. Propagating these values across the weights learned by the network then determined the associative values that had been learned for each label given those features. Luce’s Choice Axiom (Luce, 1959) was used to derive choice probabilities for the 3 labels given these activations, revealing that the FL-trained network categorized and discriminated well (the probability of correct classification for the low and the high frequency exemplars was p=1).

LF-network testing involved activating the labels in order to determine subsequent activation of the features. In turn, each label was given an input value of 1, and this then produced activation levels in the features, which were determined by the associative values learned in training. In order to assess the network’s performance, the Euclidean distance between the predicted activations and the actual feature activations of the appropriate exemplar were calculated. For each label there were two sets of feature activations: those corresponding to the high and low frequency exemplars. To test learning of both exemplar types, a category and a frequency (either high or low) were selected, and the difference between the feature activations predicted by the network and the correct values for the category exemplars was computed. These differences were then converted to z-scores, and from these the probabilities of selecting the correct exemplar given the category label were calculated as follows:

$$P(x) = \exp(-z(dist(x,t)) \qquad (2)$$

where P(x) is the likelihood of the network selecting exemplar x, z(·) returns the z-score of its argument relative to its population, dist(·,·) is the Euclidean distance function, and t is the exemplar pattern generated by the network. The P(x) likelihoods were normalized using Luce’s Choice Axiom to yield normalized probability estimates. These revealed that the LF network performed poorly. At asymptote, it predicted the correct feature pattern with only p=.35 confidence for low frequency exemplars (chance), and p=.75 confidence for high frequency exemplars.

Testing the FLO Hypothesis

Consistent with our hypothesis, a notable Feature-Label-Order Effect was detectable in the simulations. The following experiment was designed to see whether human learning would show a similar effect.

Participants

32 Stanford Undergraduates participated for credit.

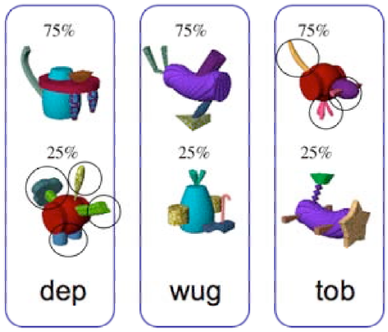

Figure 6. The category structures Experiment 1. (The stimuli are fribbles created by Michael Tarr’s lab at Brown University.) The features that need to be weighted to successfully distinguish the sub-categories are circled on the low-frequency “dep” and high-frequency “tob” exemplars.

Method and Materials

Three experimental categories of “fribbles” were constructed, each comprising two sub-categories clustered around a non-discriminating feature and a set of discriminating features. The two sub-categories that made up each labeled category did not share features, and so learning to correctly classify one of the sub-categories paired with each label provided no assistance with learning the other sub-category paired with that label. The sub-categories were again manipulated so that 75% of the exemplars of a category belonged to one sub-category, and 25% to another, and each non-discriminating feature was shared by high frequency and low frequency exemplars that belonged to different categories. Thus learning to correctly classify low frequency exemplars necessarily required learning to weigh the discriminating feature more than the non-discriminating feature. A control category served to check that there were no differences in learning between the two groups other than those we hypothesized: all its exemplars shared just one, highly salient feature (all were blue). Because learning this category involved a binary pairing blue⇒bim, there was no "predictive structure" to discover. In the absence of competing exemplars, learning was predicted to be identical for FL and LF training.

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 exemplar, 150 ms blank screen and 1000ms label ("that was a wug"). A 1000ms blank screen separated all trials (see Figure 10). A training block comprised 20 different exemplars of each experimental category – 15 high-frequency exemplars and 5 low-frequency exemplars – and 15 control category exemplars. Training comprised 2 identical blocks, with a short rest between the blocks.

Testing consisted of speeded 4 alternative forced-choice tasks. Half the participants matched an exemplar to the 4 category labels, and half matched a label to 4 previously exemplars drawn from each category. Participants were instructed to respond as quickly as they could (after 3500ms, a buzzer sounded and no response was recorded). Each sub-category (and the control) was tested 8 times, yielding 56 test trials.

Results and Discussion

The results of the experiment were remarkably consistent with our predictions; a 2 x 2 ANOVA revealed a significant interaction between exemplar-frequency and training ($F(1,94)=20.187$, $p<0.001$; Figure 6). The FL-trained participants classified high and low frequency items accurately (FL high p=.98; low p=.78), while the LF-trained participants only accurately classified high-frequency items (p=.86) and failed to classify the low frequency exemplars above chance levels (p=.36, $t(47)=0.536$, $p>0.5$). The control category was learned to ceiling in both conditions. Analyses of confusability (i.e., the rates at which exemplars were misclassified to the category with which they shared non-discriminating features) showed the same interaction between frequency and training ($F(1,94)=8.335$, $p<0.005$), with higher confusion rates after LF training (M=22.6%) than FL (M=6%; $t(16)=5.23$, $p<0.0001$). These differences were not due to a speed / accuracy trade-off; participants trained FL were faster as well as more accurate (LF M=2332ms, FL M=2181ms; $t(190)=1.677$, $p<0.1$).

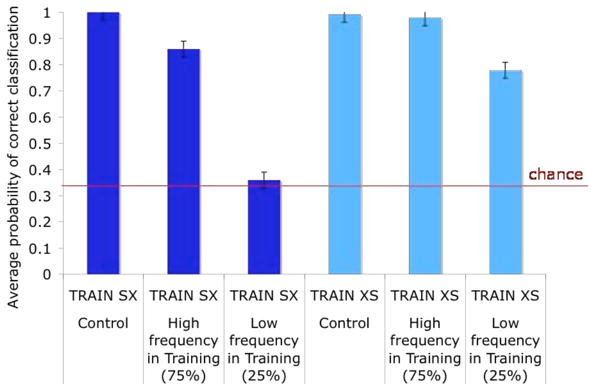

Figure 7: Performance of participants training and exemplar type. *Note here that SX corresponds to Label-to-Feature (LF) and XS to Feature-to-Label (FL).*

To the degree that learning is driven by prediction error (and there is considerable evidence that it is) the Feature-Label-Ordering effect may be an inevitable feature of learning. We believe it has many implications for our understanding of language and cognition.

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*, **12**, 241-253

Kamin L.J. (1969). Predictability, surprise, attention, and conditioning. In: Campbell B, Church R (eds). *Punishment and Aversive Behaviour*. Appleton-Century-Crofts: New York.

Luce, R. D. (1959). *Individual Choice Behavior: A Theoretical Analysis*. New York: Wiley

Pearce J.M. & Hall G. (1980) A model for Pavlovian learning. *Psychological Review*, 87:532-552

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*. New York: Appleton-Century-Crofts.

Rescorla R.A. (1988). Pavlovian Conditioning: It's Not What You Think It Is, *American Psychologist*, **43**(3), 151-160

Rumelhart, D. E., Hinton, G. E., & McClelland, J. L. A framework for Parallel Distributed Processing. in Rumelhart, McClelland, & the PDP research group. (1986). *Parallel distributed processing*. Volume I. Cambridge, MA: MIT Press

Siegel, S.G., & and Allan, L.G. (1996). The widespread influence of the Rescorla-Wagner model, *Psychonomic Bulletin and Review*, 3(3), 314-321