First published in Journal of Experimental Psychology:
Animal Behavior Processes **31(1)**, 107–111 (2005).

Retrospective revaluation as simple associative learning

Stefano Ghirlanda
University of Bologna and University of Stockholm

Backward blocking, unovershadowing and backward conditioned inhibition are examples of "retrospective revaluation" phenomena, that have been suggested to involve more than simple associative learning. Models of these phenomena have thus employed additional concepts, e.g. appealing to attentional effects or more elaborate learning mechanisms. I show that a suitable representation of stimuli, paired with a careful analysis of the discriminations faced by animals, leads to an account of these and other phenomena in terms of a simple "elemental" model of associative learning, with essentially the same learning mechanism as the Rescorla and Wagner (1972) model. I conclude with a discussion of some implications for theories of learning.

Phenomena of "retrospective revaluation" (where responding established to a stimulus is modified by later experience with other stimuli) have been considered "perhaps the biggest challenge to traditional theories of associative learning" (Le Pelley, Cutler, & McLaren, 2000) and "the most daunting empirical challenge to the Rescorla and Wagner (1972) model" (Wasserman & Berglan, 1998). Models of retrospective revaluation have thus appealed to additional constructs, e.g. invoking "attentional effects", often implemented as changes in stimulus associability (Le Pelley & McLaren, 2003; Mackintosh, 1975; Kruschke & Blair, 2000), or postulating more elaborate learning rules and stimulus representations (Dickinson & Burke, 1996; Le Pelley et al., 2000). The perceived shortcomings of simple "elemental" models of associative learning, however, are not supported by a general proof of impossibility. Rather, they stem from analyses of particular cases such as the classical Rescorla and Wagner (1972) model (from now on: the RW model). Here I show that a different scheme of stimulus representation, together with a careful analysis of what observed behavior implies, yields an elemental model in which the phenomena of backward blocking, unovershadowing and backward conditioned inhibition emerge naturally. The model accounts as well for external inhibition, a simpler phenomenon also thought to lie outside the scope of elemental models (Pearce, 1987).

Model

I consider an "elemental" model in which stimuli are represented as graded patterns of activity in an array of N "stimulus elements". I write S_i the activity induced in element i by stimulus S ($i = 1, \dots, N$). To each element, a weight W_i is attached, and the strength or likelihood of responding to S is assumed to increase with the weighted sum r_S:

$$r_S = \sum_i W_i S_i \qquad (1)$$

To apply the model to data from learning experiments, we need further assumptions in two main areas. First, to determine the set of S_i values representing a given stimulus S we need a scheme of stimulus representation. Second, to determine the weights corresponding to a given learning experiment we need a learning algorithm.

Learning is modeled here with a refinement of the RW learning rule, first derived by Widrow and Hoff (1960) and introduced to animal learning theory by Blough (1975). In this algorithm weight W_i changes according to

$$\Delta W_i = \alpha(\lambda - r_S) S_i \qquad (2)$$

where λ is the maximum value that responding to S can attain given the applied reinforcer, and α mainly regulates the speed of learning (Widrow & Stearns, 1985). Equation (2) is capable, through repeated applications, of establishing a different response to each of many stimuli, provided the stimulus representations satisfy certain technical requirements ("linear separability", Hsiung & Mao, 1998; Haykin, 1999). I refer the reader to the literature for further details (Blough, 1975; Haykin, 1999; McClelland & Rumelhart, 1986; Rumelhart & McClelland, 1986; Widrow & Stearns, 1985). Weights are assumed to have a value of zero at the

Stefano Ghirlanda, Department of Psychology, University of Bologna, Italy, and Group for Interdisciplinary Cultural Research, University of Stockholm, Stockholm, Sweden.

This research has been partially funded by Riksbankens Jubileumsfund, Marcus och Marianne Wallenberg Stiftelse, the Italian Ministry of Education, University and Research (through grants to E. A. Jannini, Department of Experimental Medicine, Universtiy of L'Aquila and V. Di Lazzaro, Institute of Neurology, Catholic University of Rome), and my family. I thank Ian P. L. McLaren for drawing my attention to unovershadowing, and for encouragement. Many conversations with Magnus Enquist have contributed to the ideas presented here. I also thank John M. Pearce and Allan R. Wagner for fruitful discussion, and an anonymous reviewer for insightful comments.

Correspondence should be addressed to Stefano Ghirlanda, Department of Psychology, Viale Carlo Berti-Pichat 5, 40127 Bologna, Italy. Electronic mail may be sent to stefano.ghirlanda@unibo.it.

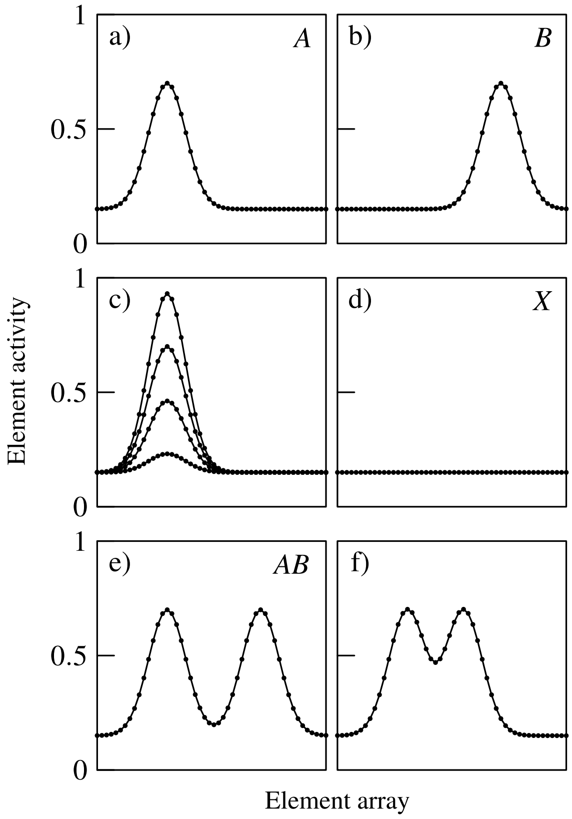

Figure 1. A simple scheme of stimulus representation whereby each stimulus elicits a graded pattern of activity in an array of 50 "stimulus elements". Each dot is the activity of one element. Letters at top right of panels identify stimuli used in this paper. The panels depict: a) pattern of activity elicited by a stimulus A (in context X, see below); b) pattern of activity elicited by a stimulus B which is physically rather different from A; c) patterns of activity elicited by varying the physical intensity of A (a lower curve corresponds to lower intensity); d) a "background" or "contextual" stimulus X (e.g. a dimly illuminated Skinner box) that elicits low activity in all elements; e) pattern of activity elicited by the compound stimulus AB obtained by presenting A and B simultaneously; f) pattern of activity elicited by A compounded with a more similar stimulus.

start of training (drawing weights at random from a distribution symmetrical around zero would lead to the same conclusions).

How to represent stimuli in an elemental model is the subject of enduring debate (see the Discussion). The results below depend only on four rather general assumptions: 1) the activity of each element is a positive number; 2) a stimulus elicits a graded pattern of activity in the elements, activating different elements to different extents; 3) physically more similar stimuli activate the elements in more similar patterns (conversely, very different stimuli elicit very different activation patterns); 4) higher intensity of stimulation corresponds to higher element activity. This representation scheme is summarized in Figure 1.

Results

I consider phenomena that arise in experiments using two stimuli A and B as well as their compound AB.[1] In actual experiments, A and B are typically very different, e.g. a light and a tone. Hence I assume that they elicit different patterns of activity in the elements. It will be crucial to consider also a fourth stimulus X, corresponding to the experimental setting in the absence of A and B. Given that element activity is assumed proportional to stimulus intensity, it is often appropriate to assume that X activates the elements to a rather low level, corresponding to e.g. a dimly illuminated Skinner box with low ambient noise (this assumption, however, is not crucial). Figure 1 shows the stimulus representations used in the following.

The results below are derived from simple and general arguments requiring little formal mathematics, and are illustrated by computer simulations of learning experiments (see the legend to Figure 2 for details). The results can also be derived formally by standard methods of linear algebra and the theory of linear differential equations (Ghirlanda, 2002; Hsiung & Mao, 1998; Simmons, 1974).

External inhibition

The experimental design of external inhibition is simply A^+ (reinforced presentations of A) followed by testing with A and AB, where B is a novel stimulus. Responding to AB is typically less than responding to A, $r_{AB} < r_A$. External inhibition is arguably a simpler phenomenon than retrospective revaluation, and has been known since much longer (Pavlov, 1927). Yet it is not exhibited by the RW model (Pearce, 1987).

To derive a prediction from the present model, we need to understand what weights are produced by training. One way is to infer properties of the weights from the established stimulus-response relationships. For instance, that the animal responds to A means that the weights must satisfy the equation

$$\sum_i W_i A_i = r_A \gg 0 \tag{3}$$

where $\gg 0$ indicates "substantially greater than zero" (the exact value is immaterial). It is also crucial to note that the animal does not respond when A is not present, e.g. when a response key is dark or a sound CS is not played. Labeling X such a stimulus situation we can write

$$\sum_i W_i X_i = r_X \simeq 0 \tag{4}$$

where $\simeq 0$ indicates "close to zero" (the exact value is immaterial). Equation (4) implies that, since $X_i > 0$ by hypothesis, some weights must be positive and some negative (equation (3) implies that the weights cannot all be zero). In

[1] Actual experiments are usually more complex (see the Discussion). Part of this complexity, however, arises from the need to infer internal processes from behavioral observations. Analyzing models is simpler because we have direct access to the model internals, i.e. the weights.

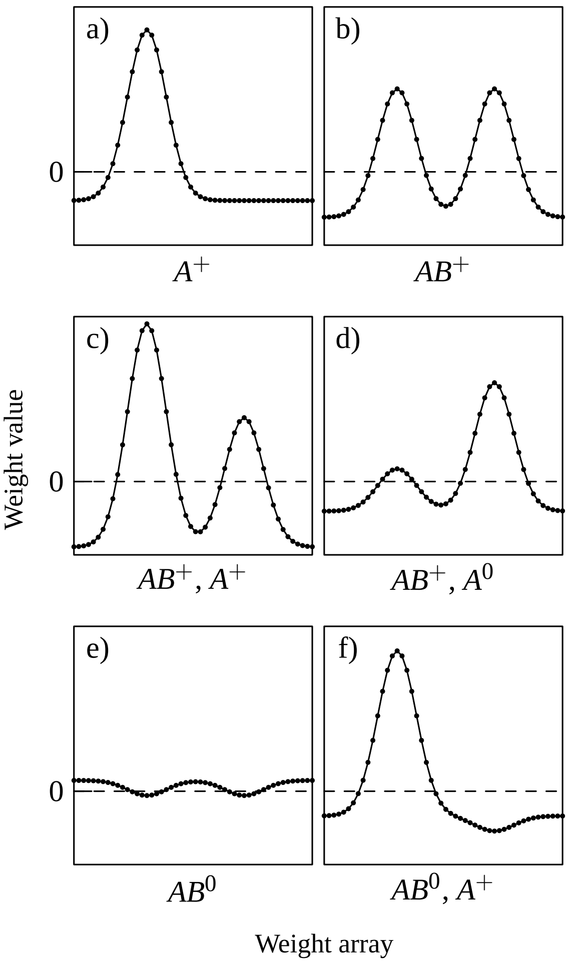

Figure 2. Weights produced by different experimental designs, as noted under each panel. All designs also include unreinforced presentations of X (X^0). Each dot is the value of one weight. Dashed lines are drawn at 0 to highlight which weights are positive and which negative. The scale is the same on all panels, allowing direct comparison of weight values. Each weight array is the result of a computer simulations of the relevant design, using the learning equation (2) and the stimuli in Figure 1. Reinforced trials corresponded to $\lambda = 0.75$, unreinforced trials to $\lambda = 0.05$. Training continued until asymptote.

other words, if the animal does not react to X we infer that inhibitory and excitatory tendencies caused by X balance. Equation (3) further implies that the elements that carry negative weight cannot be those most excited by A, otherwise r_A could not be large. The only way to get the required responses to A and X (i.e. to satisfy equations (3) and (4) simultaneously) is that the elements most excited by A carry positive weight, and the others the negative weight necessary to produce $r_X \simeq 0$. As an example, Figure 2a shows the weights that are obtained in a computer simulation of learning, when A and X are the stimuli in Figure 1. Note the negative weights relative to elements not strongly excited by A.

What happens now when B is added to A? To the extent that B is different from A, by hypothesis it activates different elements, hence elements carrying negative weight. It follows that AB will stimulate elements carrying negative weights more than A, yielding $r_{AB} < r_A$, or external inhibition. Simulation results were $r_A = 0.75$ and $r_{AB} = 0.59$.

Backward blocking

A "backward blocking" experiment consists of an AB^+ X^0 stage followed by an A^+ X^0 stage (where X^0 means unreinforced presentations of the background X). It is found that responding to B decreases between the first and second stage (Dickinson & Burke, 1996; Shanks, 1985; Wasserman & Berglan, 1998). With the same reasoning as above, we infer that discrimination training between AB and X produces a mixture of positive and negative weights such that $r_{AB} \gg 0$ and $r_X \simeq 0$. The negative weights pertain to elements not strongly activated by AB, as shown in Figure 2b. What happens now when we withdraw B and continue to reinforce responding to A? At first a drop in responding will be observed, because some elements that carry positive weight are less activated by A than by AB. Learning will thus proceed to increase responding to A, by increasing the appropriate weights. However, the constraint $r_X \simeq 0$ requires that, if some weights increase, others must decrease. The latter include weights relative to elements activated by B, leading to a drop in responding to B. This can be seen comparing the weights in Figure 2b and Figure 2c. Simulation results were $r_B = 0.40$ after the first stage and $r_B = 0.25$ after the second stage.

Unovershadowing

The unovershadowing design consists of AB^+ X^0 training followed by A^0 X^0 training. It is found that responding to B increases between the two stages (Dickinson & Burke, 1996; Wasserman & Berglan, 1998). The explanation in the present model is similar to that for backward blocking. During A^0 training, weights attached to elements strongly excited by A are lowered. Due to the constraint $r_X \simeq 0$, other weights are bound to increase, including those linked to elements activated by B. Hence A^0 training is expected to increase responding to B. The weights at the end of the AB^+ and A^0 stages are shown in Figure 2b and Figure 2d, respectively. Simulation results were $r_B = 0.40$ after the first stage and $r_B = 0.51$ after the second stage.

Backward conditioned inhibition

The backward conditioned inhibition design consists of AB^0 X^0 training followed by A^+ X^0 training, whereby B is shown to acquire inhibitory properties (Chapman, 1991). The account of this phenomenon in the present model is, perhaps surprisingly, similar to the account of external inhibition. Indeed, the initial stage AB^0 X^0 of backward conditioned inhibition produces a rather uniform weight array (Figure 2e). Thus the subsequent A^+ X^0 stage has very similar effects as the A^+ X^0 stage in external inhibition, resulting in negative weight being attached to the elements most strongly activated by B. This is seen in the similar weight arrays in Figure 2a and Figure 2f. Simulation results where $r_B = 0.05$ after the first stage and $r_B = -0.21$ after the second stage.

Discussion

The theory of "elemental" models has been recently developed in several directions (Ghirlanda & Enquist, 1999; McLaren & Mackintosh, 2000, 2002; Wagner & Brandon, 2001; Wagner, 2003). My aim was not to evaluate these proposals, but to show that even simple elemental models are not yet fully understood. On one hand, it is obvious that a weighted sum of element activities cannot model all learning phenomena. For instance, animals are certainly capable of more complex processing of stimuli (including e.g. perceptual learning and within-compound associations), which has been suggested to play a role in retrospective revaluation (e.g. Dickinson & Burke, 1996; Wasserman & Berglan, 1998; Aitken, Larkin, & Dickinson, 2001). On the other hand, without a thorough analysis of simple models it is difficult to know when more complex models and concepts are really necessary. The results above appear to question the status of some retrospective revaluation phenomena as revealing different or more sophisticated processes than simple associative learning.

The designs of actual experiments on retrospective revaluation are often more complex than the ones analyzed here (e.g. Wasserman & Berglan, 1998; Aitken et al., 2001). For instance, initial training on AB^+ CD^+ X^0 is followed by A^+ C^0 X^0. It is typically found that the second stage of training has caused responding to B and D to change in opposite directions. I have chosen to consider simpler designs for several reasons (see also footnote 1). First, enough interesting results seem to follow. Second, my main points are not specific to any experimental design or theoretical model. For instance, the implications of lack of responding to background stimuli may be considered even if we add element-element connections to the model (e.g. to represent within-compound associations). It is not unlikely that a fuller understanding of retrospective revaluation may arise from applying the reasoning above to more powerful models than the one analyzed here.

Comparison with the RW model

The inability of the RW model to account for retrospective revaluation has been typically attributed to inadequate assumptions on learning, but the above results show that it can as well be traced back to stimulus representations, at least partly. The RW model is equivalent to letting each "simple" stimulus S (heuristically identified by the researcher) activate a single element. Responding to S is then controlled by a single weight W_S ("associative strength"), that can be modified only by experience with S. In the present model, a stimulus S activates the whole element array (though some elements may be little affected, see Figure 1). Hence responding to S depends on all weights: whatever changes the weights may affect responding to S, be it experience with S or other stimuli. An instructive example of these differences is what lack of responding to background stimuli X implies. In the RW model, $r_X \simeq 0$ is obtained simply by $W_X \simeq 0$, with no consequences for responding to other stimuli. If stimuli activate overlapping sets of elements, however, $r_X \simeq 0$ has important implications for the whole weight array (see above).

What is "retrospective revaluation"?

The fact that experience with a stimulus may change responding to other stimuli may be surprising from the standpoint of some learning models, but it is a basic finding of stimulus control. For instance, a neutral stimulus may acquire the ability to elicit a response following experience of a physically different S^+ (Mackintosh, 1974; Ghirlanda & Enquist, 2003). Whether we call this "generalization" or "retrospective revaluation" may stem more from our incomplete understanding of underlying processes than from fundamental differences between the processes themselves. It is true that learning experiments often use very different stimuli, whereas stimulus control studies are typically concerned with continuous variation in similarity, but this may not be of any deep significance. Both kinds of stimuli may be represented as patterns of element activity, with the only difference that such patterns would overlap more for similar stimuli than for different ones. The advantage is that stimulus control and learning phenomena can be approached within the same model, by asking how stimuli activate the elements and what weights are produced by training. Previous work (Blough, 1975; Ghirlanda & Enquist, 1999; Rescorla & Wagner, 1972) and the results above show that this approach can account for many stimulus control and learning phenomena, even within a simple weighted-sum model.

The representation of stimuli

The elemental approach to stimuli does not come without pitfalls. It has especially been criticized for failing to justify the rules that link physical stimuli with patterns of element activity (Blough, 1975; Pearce, 1987). This criticism is justified, but requires qualification. First, while details of stimulus representations are often important (e.g. to compute similarity between stimuli), there may be phenomena that are largely insensitive to such details. The phenomena considered here appear to be of this kind. In Figure 1, and in the simulations leading to Figure 2, I have use bell-shaped activity profiles for illustrative purposes, but this assumption is not used in the mathematical arguments that lead to the results.

Second, an elemental representation of stimuli has in fact empirical justification. The evidence is overwhelming that stimuli are represented in nervous systems as graded patterns of activity in ensembles of many neurons (reviewed in Kandel, Schwartz, & Jessell, 2000; Toates, 2001). This holds at the level of sense organs (where the "elements" may be interpreted as model receptors, Ghirlanda & Enquist, 1999), at the level of retinal and cochlear ganglion cells (Warren, 1999; Kandel et al., 2000), at the level of cerebral and cerebellar cortex (Eichenbaum, 1993; Kandel et al., 2000; Knudsen, DuLac, & Esterly, 1987; Maunsell & Newsome, 1987; Thompson, 1965). Many details of such representations are unknown, but some general properties are established beyond doubt. The simple representation scheme used here captures some such properties, e.g. the representation of similarity as overlap between activity patterns.

Third, it is difficult to give a treatment of stimuli that does not rely on an elemental analysis. Even theories that seemingly refrain from such an analysis (e.g. by considering stimulus "configurations") actually rely on breaking stimuli into elements to calculate relationships (similarity) between them (Pearce, 1987, p. 65). Such theories may complement elemental analyses in interesting ways (Ghirlanda, 2002; Pearce, 1987, 1994, 2002), but are not independent of them. This, and the results above, suggests that further development of elemental schemes of stimulus representation is today a pressing need for psychological theory.

References

Aitken, M. R. F., Larkin, M. J. W., & Dickinson, A. (2001). Reexamination of the role of within-compound associations in the retrospective revaluation of causal judgements. *Quarterly Journal of Experimental Psychology*, *54B*(1), 27-51.

Blough, D. S. (1975). Steady state data and a quantitative model of operant generalization and discrimination. *Journal of Experimental Psychology: Animal Behavior Processes*, *104*(1), 3–21.

Chapman, G. B. (1991). Trial order affects cue interaction in contingency judgements. *Journal of Experimental Psychology: Learning, Memory, and Cognition*, *17*, 837-854.

Dickinson, A., & Burke, J. (1996). Within-compound associations mediate retrospective revaluation of causality judgements. *Quarterly Journal of Experimental Psychology*, *49B*, 60-80.

Eichenbaum, H. (1993). Thinking about brain cell assemblies. *Science*, *261*, 993-994.

Ghirlanda, S. (2002). Intensity generalization: physiology and modelling of a neglected topic. *Journal of Theoretical Biology*, *214*(2), 389-404.

Ghirlanda, S., & Enquist, M. (1999). The geometry of stimulus control. *Animal Behaviour*, *58*, 695–706.

Ghirlanda, S., & Enquist, M. (2003). A century of generalization. *Animal Behaviour*, *66*, 15-36.

Haykin, S. (1999). *Neural Networks: A Comprehensive Foundation* (2 ed.). New York: Macmillan.

Hsiung, C. Y., & Mao, G. Y. (1998). *Linear algebra.* Singapore: World Scientific.

Kandel, E., Schwartz, J., & Jessell, T. (2000). *Principles of neural science* (4 ed.). London: Prentice-Hall.

Knudsen, E., DuLac, S., & Esterly, S. D. (1987). Computational maps in the brain. *Annual Review of Neuroscience*, *10*, 41-65.

Kruschke, J. K., & Blair, N. J. (2000). Blocking and backward blocking involve learned inattention. *Psychonomic Bulletin & Review*, *7*, 636-645.

Le Pelley, M. E., Cutler, D. L., & McLaren, I. P. L. (2000). *Retrospective effects in human causality judgment.* Available at http://www.cis.upenn.edu/~ircs/cogsci2000/PRCDNGS/SPRCDNGS/posters/lepcumc.pdf. (Poster presented at the 22nd annual meeting of the Cognitive Science Society)

Le Pelley, M. E., & McLaren, I. P. L. (2003). Learned associability and associative change in human causal learning. *Quarterly Journal of Experimental Psychology*, *56B*(1), 68-79.

Mackintosh, N. J. (1974). *The psychology of animal learning.* London: Academic Press.

Mackintosh, N. J. (1975). A theory of attention: Variations in the associability of stimuli with reinforcement. *Psychological Review*, *82*, 276-298.

Maunsell, J. H. R., & Newsome, W. T. (1987). Visual processing in monkey extrastriate cortex. *Annual Review of Neuroscience*, *10*, 363.

McClelland, J. L., & Rumelhart, D. E. (Eds.). (1986). *Parallel Distributed Processing: Explorations in the Microstructure of Cognition* (Vol. 2). Cambridge, MA: MIT Press.

McLaren, I. P. L., & Mackintosh, N. J. (2000). An elemental model of associative learning: I. Latent inhibition and perceptual learning. *Animal Learning & Behavior*, *28*(3), 211–246.

McLaren, I. P. L., & Mackintosh, N. J. (2002). Associative learning and elemental representation: II. Generalization and discrimination. *Animal Learning & Behavior*, *30*, 177-200.

Pavlov, I. P. (1927). *Conditioned reflexes.* Oxford: Oxford University Press.

Pearce, J. M. (1987). A model for stimulus generalization in Pavlovian conditioning. *Psychological Review*, *94*(1), 61–73.

Pearce, J. M. (1994). Similarity and discrimination: a selective review and a connectionist model. *Psychological Review*, *101*, 587–607.

Pearce, J. M. (2002). Evaluation and development of a connectionist theory of configural learning. *Animal Learning and Behavior*, *30*(2), 73-95.

Rescorla, R. A., & Wagner, A. R. (1972). A theory of Pavlovian conditioning: Variations in the effectiveness of reinforcement and nonreinforcement [incollection]. In *Classical conditioning: current research and theory.* Appleton-Century-Crofts.

Rumelhart, D. E., & McClelland, J. L. (Eds.). (1986). *Parallel Distributed Processing: Explorations in the Microstructure of Cognition* (Vol. 1). Cambridge, MA: MIT Press.

Shanks, D. R. (1985). Forward and backward blocking in human contingency judgement. *Quarterly Journal of Experimental Psychology*, *37B*(1), 1-21.

Simmons, G. F. (1974). *Differential Equations.* New Delhi: Tata McGraw-Hill.

Thompson, R. F. (1965). The neural basis of stimulus generalization. In D. I. Mostofsky (Ed.), *Stimulus generalization.* Stanford, CA: Stanford University Press.

Toates, F. M. (2001). *Biological psychology: An integrative approach.* Prentice Hall.

Wagner, A. R. (2003). Context-sensitive elemental theory. *Quarterly Journal of Experimental Psychology*, *56B*, 7-29.

Wagner, A. R., & Brandon, S. E. (2001). A componential theory of associative learning. In R. R. Mower & S. B. Klein (Eds.), *Contemprorary learning: theory and applications.* Hillsdale, NJ: Erlbaum.

Warren, R. M. (1999). *Auditory perception: A new analysis and synthesis* (2 ed.). UK: Cambridge University Press.

Wasserman, E. A., & Berglan, L. R. (1998). Backward blocking and recovery from overshadowing in human causal judgement: The role of within-compound associations. *Quarterly Journal of Experimental Psychology*, *51B*(2), 121-138.

Widrow, B., & Hoff, M. E. J. (1960). Adaptive switching circuits [inproceedings]. In *Ire wescon convention record* (Vol. 4, pp. 96–104). New York: IRE.

Widrow, B., & Stearns, S. D. (1985). *Adaptive signal processing.* Englewood Cliffs, NJ: Prentice-Hall.