Learning Associations That Run Counter to Biases in Learning: Overcoming Overshadowing and Learned Inattention

Andrew F. Heckler (heckler.6@osu.edu)
Department of Physics, Ohio State University
191 W. Woodruff Avenue, Columbus, OH 43210 USA

Jennifer A. Kaminski (kaminski.16@osu.edu)
Center for Cognitive Science, Ohio State University
210A Ohio Stadium East, 1961 Tuttle Park Place, Columbus, OH 43210, USA

Vladimir M. Sloutsky (sloutsky.1@osu.edu)
Center for Cognitive Science, Ohio State University
208C Ohio Stadium East, 1961 Tuttle Park Place, Columbus, OH 43210, USA

Abstract

Motivated by the possibility that some common scientific misconceptions are caused by learning biases that create undesired associations, we examine the effect of salience on associative learning tasks and test two methods to counter-train undesired associations learned during training. Experiment 1 tests the extent to which a cue can be learned in a novel task after it has been overshadowed or blocked in a previous learning task. We find an attenuation of learning of both the overshadowed and blocked cues, even though the overshadowed cue showed no evidence of being learned in the initial learning task. Thus the overshadowed cues are learned: they are learned to be ignored. Experiment 2 demonstrates that once a cue has been overshadowed, multiple kinds of positive examples are not effective in learning low salience cues, whereas negative examples attacking the validity of competing higher salience cues are effective in shifting attention to the low salience cues. The relevance of these results to scientific misconceptions is discussed.

Keywords: Overshadowing, blocking, associative learning, attention, scientific misconceptions.

Introduction

In the course of experiencing the world in which we live, not all information is learned equally well: it is often the case that people preferentially learn some features while ignoring others. For example, one might be more inclined to judge the weight of an object based only on its size, disregarding its composition. While a preference for a particular cue may be helpful in most every-day situations, it may also be the case that this preferred cue is incomplete or spurious and may hinder the learning of more relevant cues. In the example of determining weight, clearly both volume and composition of the object are needed for a proper estimate, thus attending to size alone is not sufficient for all cases. Nonetheless, people commonly assume that in general, larger things are necessarily heavier (see e.g., Pick and Pick, 1967).

Research reported here focuses on the case when biases in learning prevent the learner from acquiring specific target associations. This has special relevance to many scientific misconceptions, where students commonly learn undesired associations among variables. Often, these students fail to learn the relevant, fully predictive cues but rather preferentially learn an association between more salient, less predictive cues and a given outcome. In addition to the example above, another example is the common association between force and motion: both velocity and acceleration are associated with a net force, but velocity is a spurious cue—only acceleration is a perfect predictor. However, this is not how the association is usually learned: it is a common belief that a non-accelerating yet moving object is associated with a net force (e.g., Viennot, 1979, Clement, 1982; Halloun & Hestenes, 1985).

How then do we best address these misconceptions? If they are caused at least in part by common but undesired associations, then what is the best way to train learners on a target association that is not naturally aligned with associative learning biases, especially when the learner has already preferentially learned a different, undesired association?

In this study, we address this question by first training learners on specific, novel associations essentially to mimic the creation of an undesired association or "misconception". This allows us to examine more closely the nature of misconceptions and methods to address them through counter-training. In particular, in Experiment 1, we focus on learning biases due to differences in cue salience. The results provide evidence for the explanation that learned attention (or inattention) may be at least partially responsible for differential learning of competing cues with differential salience. We then use this as a framework for designing possible counter-training to shift attention to a given desired cue and target association. In Experiment 2 we again train learners to create an undesired association and implement two different counter-training methods to shift attention. The two kinds of counter-training examples we use are not only based on findings from previous associative learning research, but

they are also fundamental to science, namely *counter-example*, which reduces the validity of the incorrect association, and *induction*, in which the target cue is the only cue that is always paired with the target outcome. Standard error-reduction models of learning predict that both methods will shift attention to the desired cue, thus it is not a priori clear whether either kind will generally be more effective in addressing misconceptions.

Overshadowing, blocking and learned inattention

If, as proposed by several successful models of learning that attention directs learning (e.g., Mackintosh, 1975; Pierce and Hall, 1980), then to the extent that salience directs attention, salience may also direct learning. Certainly a difference in salience between competing cues has been shown to influence their learning—this is a special case of *overshadowing*, in which the learning of a cue is diminished if a second cue is presented simultaneously (Pavlov, 1927). For example, if two cues A and B are paired together, increasing the relative salience of cue A also increases its utilization, defined as the probability that the participant will choose the outcome when cue A is present (e.g., Edgell, 1996), and similarly, increasing the salience of a competing cue B will decrease the utilization of cue A (Mackintosh, 1976). Kruschke and Johansen (1999) have replicated these results with participants completing a series of probabilistic categorization tasks, and found that attentional shifting and learned attentional weights played an essential role in their successful connectionist model of the results.

Kruschke and Blair (2000) further used the classic associative learning phenomenon of *blocking* to test another important prediction of learned attention. In blocking, cues A and B are paired with an outcome, and the association of cue B with the same outcome is significantly weakened if it is also learned that A alone is associated with the outcome (Kamin, 1969). As proposed by Sutherland and Mackintosh (1971), low association of B with the outcome is due to a learned shift in attention away from B. It is not the case that B is not learned, rather B is learned to be ignored, thus it does not accrue significant associations. Improving upon an earlier experiment by Mackintosh and Turner (1971) with rats, Kruschke and Blair (2000) provided strong evidence for the diminution of the attentional weight of the blocked cue in humans by demonstrating that, compared to control cue, the learning of the blocked cue is attenuated in a subsequent novel learning task. That is, there is an attenuation of learning of a cue with a previously learned low attentional weight.

In Experiment 1 we wish to test whether an overshadowed cue also accrues a diminished attentional weight. Just like a blocked cue, is an *overshadowed* cue more difficult to learn in a subsequent novel learning task? This will be done using a method very similar to Kruschke and Blair's design. Attentional models predict that it will be more difficult, since by default the lower salience cue is initially not attended to, and after many iterations with feedback, its learned attentional weight should also decrease. Not only does this novel experiment further test attentional theory, but the context of always presenting two cues simultaneously—as opposed to the blocking design in which sometimes only one cue is presented—is more like real world learning in which most often both cues, such as the volume and the composition of an object, are always present. That is, overshadowing is likely more relevant than blocking for the study of misconceptions that may arise from associative learning.

Experiment 1

Method

Participants Ninety six undergraduate students from Ohio State University participated in the experiment and received partial credit for an introductory psychology course. Forty eight students each were assigned to the *Overshadowing* or *Blocking* condition (see Table 1).

Table 1. Design of Experiment 1

	Condition		
Phase	Blocking	Overshadowing	Number of trials
Training Phase 1	$A \rightarrow O_1$		10
	$H \rightarrow O_4$		10
	(pause)		
	$AB \rightarrow O_1$	$\mathbf{A}B \rightarrow O_1$	20
	$CD \rightarrow O_2$	$CD \rightarrow O_2$	20
	$\mathbf{EF} \rightarrow O_3$	$\mathbf{EF} \rightarrow O_3$	20
Test	B→?, D→?	B→?, D→?	
Training Phase 2	(X or Y)B→ O_5	(X or Y)B→ O_5	25
	(X or Y)D→ O_6	(X or Y)D→ O_6	25
	(X or Y)G→ O_7	(X or Y)G→ O_7	25

Note: Letters denote cues (computer chip components) and $O_1 - O_7$ denote outcomes (appliances). Bold type (e.g. **A**) denotes a high salience cue. In Phase 2, for every trial one of two novel cues X or Y is randomly paired with the indicated cue B, D, or G.

Design The design of Experiment 1 is shown in Table 1, and consists of two conditions, each counter balanced between subjects with four random cue combinations to control for any possible effects of specific cues. Both conditions consist of an initial training and testing phase and a final training and testing phase. Both conditions are identical except for the initial training phase. In the traditional blocking condition, the goal is to train the participants to block cue B via a traditional blocking design. Then the second training phase will associate B with a different outcome, in order to determine how easily B may be learned in a novel context after it has been

traditionally blocked. Note that in this second phase we are interested in finding differences in learning curves, therefore we designed the second learning phase task to be slightly more difficult by adding in random non-predictive cues X and Y. The blocking condition is meant to be a replication of Kruschke and Blair's results and used as a standard of comparison for the second condition, which is novel.

In the second condition, rather than being traditionally blocked, cue B is overshadowed by a more salient cue A, and both cues are always shown together. The overshadowed cue B will then be trained on a new outcome, in order to determine how easily it may be learned compared to control cues. In order to account for any possible effect of novelty, there are two kinds of control cues with which to compare the blocked or overshadowed cue during the final training phase. The first is a cue that is present in the initial training sessions, but is not blocked or overshadowed. The second is a completely novel cue introduced only in the final training phase when new outcomes are introduced.

Procedure All training and testing was presented to individual participants on a computer screen in a quiet room. They proceeded through training and testing at their own pace, and their responses were electronically recorded. The participants were given instructions that they were learning about a (clearly imaginary) appliance factory, and they were to learn about which computer chip components (cues) are installed in various kitchen appliances (outcomes). The computer chip components (cues) were simple geometric shapes superimposed on a simple diagram representing a circuit board, and a maximum of two components were placed randomly in one of four places on any given chip. The cues were classified as salient or non-salient. The salient cues were colored (e.g., red, blue, green) and somewhat larger than the non-salient cues which were all light gray.

In each trial the learners were given a multiple choice question in which a computer chip with a particular combination of components, say component A (green square) and component B (grey triangle), was presented with five pictures of familiar kitchen appliances displayed on the same screen. Participants were asked: "In which appliance is this computer chip used?" During the training phases, they were given immediate feedback, whether they were right or wrong, indicating which appliance was the correct one.

During the test phase, the participants are shown novel single cue and double cue combinations. For example, the learners are shown a chip with component B (cue B) only and are asked to chose in which appliance it would be installed. Likewise they are also asked about a chip with both components B and D.

The format of Phase 2 training is similar to Phase 1. After learning that a series of chip components are installed in specific kitchen appliances in Phase 1, the participants then learn in Phase 2 that some of these same components (B and D) are also installed in one of a selection of non-kitchen household appliances. For example in the first phase the participant learns that cue B (which is blocked or overshadowed, depending on the condition) is associated with a blender, then after this they learn that it is also associated with a radio. They learn this second association in way similar to the way in which the first association was learned: by presenting a pair of chip components and then asking which of a selection of appliances in which they are to be installed. Participants are then given immediate feedback as to the correct answer. Scores on each training and testing phase were electronically recorded.

Results and Discussion

The participants successfully learned during the initial training sessions, with an average score of 87% in the traditional blocking condition and 86% in the overshadowing condition (chance score is 16%). This excludes 6 participants who scored 2 standard deviations below average on at least three of the training cue types. In the initial testing phase, the participants clearly blocked cue B, with an average score of 40% correct on B and 75% on the control cue D [paired t-test, $t(43)=3.9$, $p<0.001$]. Likewise for the overshadowing condition, B was clearly overshadowed, with an average score of 21%, compared to 58% for the control cue D [paired t-test, $t(44)= 5.3$, $p<0.001$]. In fact, the participants in the overshadowing condition did not score significantly different from chance (16%) on cue B [$t(44)=0.8$, $p>0.4$], thus one might infer that nothing was learned about the overshadowed cue. However the subsequent learning task indicates this is not the case.

Traditional Blocking and attenuation of subsequent learning. As indicated by the learning curves in Figure 1, there was a significant attenuation of subsequent learning of the blocked cue B in the traditional blocking condition compared to subsequent learning of both the familiar control cue D and the novel control cue G. In particular, if we consider the total number of correct responses in the second training phase the learning as a measure of learning in the second phase, the score for cue B was 66%, which is reliably less than the average scores for the control cue D (73%) [paired t-test, $t(43) = 3.9$, $p< 0.001$]. While this is a somewhat small difference in scores, the effect size of the within-subject difference between the scores was 0.6. The effect was larger when comparing the score of cue B compared to the average score of novel cue G (80%) [paired t-test, $t(43) = 5.2$, $p< 0.001$], with an effect size for the within-subject difference of scores equal to 0.8. There was also a significant difference between the score for D and G [paired t-test, $t(43) = 3.5$, $p = 0.001$].

Overshadowing and attenuation of subsequent learning. The learning curves in Figure 1 indicate a key finding of this experiment, namely that the subsequent learning of the previously overshadowed cue B is significantly attenuated compared to the familiar control cue D and the novel control cue G. The total average score for the second training session for cue B was 55%, which is reliably less than the scores for the control cue D (61%) [paired *t*-test, $t(44) = 3.0$, $p = 0.004$]. The learning curve in Figure 1 indicates that the overshadowed cue always scored 5-10% below the control on all training blocks but the first. Much like the blocking condition, this difference is somewhat small, but the effect size of the within-subject difference in scores is 0.45. The score for cue B was also attenuated compared to the novel control cue G with a score of 63% [paired *t*-test, $t(44) = 3.4$, $p = 0.001$].

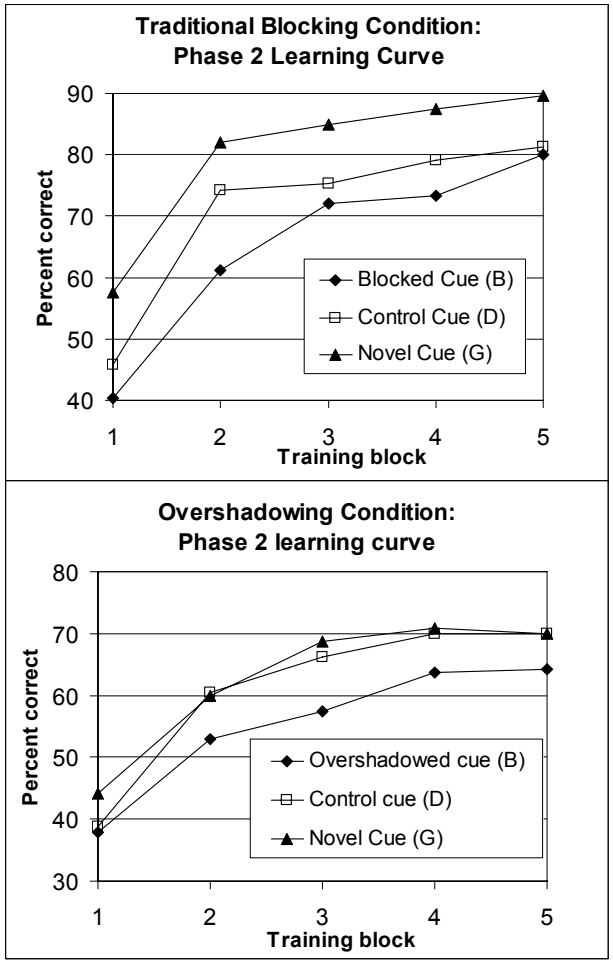

Figure 1. Experiment 1 learning curves for both conditions in Phase 2, in which cues are learned to be associated with a new set of outcomes. Cue B is the previously-blocked or overshadowed cue, depending in the condition. Cue D is a familiar, previously learned control cue and cue G is a novel control cue. Attenuation of learning of cue B occurs in both conditions.

Experiment 2

Given that an overshadowed, low salience cue may accrue a low attentional weight and becomes difficult to learn, how do we subsequently raise attention to this cue? We pursue two methods outlined in Table 2. The first method is suggested by the observation that when two cues compete, lowering the relative validity of one will increase the associative strength of the other (Wagner et al, 1968). This has been modeled in terms of learned attention: to decrease error, attention is rapidly shifted toward the cue(s) with more validity, and this cue in turn gains more associative strength (Kruschke and Johansen, 1999). This suggests that if we wish to raise attention to an overshadowed cue B, then one must lower the validity of the competing cue A (e.g. see Hall et al, 1977; Kruschke and Johansen, 1999). Our first counter-training method will reduce the validity of the overshadowing cue (A), by presenting direct counter-examples in which A is paired with a different outcome (i.e. not O1). Thus A is not 100% predictive of O1, and the learner will shift attention to the cue that is more predictive, namely the previously overshadowed cue B.

The second counter-training method, also outlined in Table 2, employs a method in which the overshadowed cue B is always paired with the desired target outcome O_1, whereas the previously overshadowing cue A and a novel cue X are only sometimes paired with the target outcome. This is similar to blocking in that there is a cue (B) which is always paired with a particular outcome, but cue A is not. Thus in some sense, one might expect A to be blocked during this counter-training. Models employing attentional learning suggest that in this case, attention should shift to cue B, resulting in higher association with O_1. On the other hand, this method is also somewhat similar to Phase 2 in Experiment 1 in that the method is attempting to train a previously overshadowed cue. Since Experiment 1 found inhibited learning of an overshadowed cue, one might argue that this method will not be effective. We label this method *induction* because the learner must infer from the examples that since B is the common cue, it must be the cue that is most strongly associated with the outcome.

Since both of these methods predict at least some shifting of attention to the overshadowed cue B, it is not clear which will be more effective.

Method

Participants Ninety undergraduate students from Ohio State University participated in the experiment and received partial credit for an introductory psychology course. Forty five students were assigned to each of the two conditions.

Materials and Design Table 2 shows the abstract design. The format of the materials, training trials, and general procedure were the same as in Experiment 1. The main difference is the sequence of training. The first phase of

training results in overshadowing of cues B and D. this phase is meant to simulate common learner experience with is responsible for the creation of a "misconception", namely that B is not associated with O_1. The second phase employs counter-training to cues A and B. This is meant to simulate instruction. The third phase simply repeats the training to the first phase, to simulate the common experience again, after instruction. The final testing phase tested on single cues. For example, the learners are shown a chip with single component B (cue B) only and are asked to chose in which appliance it would be installed.

Table 2. Design of Experiment 2

	Condition		
Phase	Counter-example	Induction	Trials
Training Phase 1	**A**B→O_1	**A**B→O_1	20
	CD→O_2	**C**D→O_2	20
	EF→O_3	**EF**→O_3	20
	GH→O_4	GH→O_4	20
Training Phase 2	**A**B→O_1	**A**B→O_1	20
	AX→O_5	**X**B→O_1	20
	CD→O_2	**C**D→O_2	10
	GH→O_4	GH→O_4	10
Training Phase 3	**A**B→O_1	**A**B→O_1	10
	CD→O_2	**C**D→O_2	10
	EF→O_3	**EF**→O_3	10
	GH→O_4	GH→O_4	10
Test	B→?, D→?	B→?, D→?	

Note: Letters A-H denote cues (computer chip components) and O_n denote outcomes (appliances). Bold type denotes a high salience cue.

Results and Discussion

The participants in both conditions successfully learned during the training sessions, with an average score of 84% correct in Phase 1, 95% in Phase 3 and 95% in Phase 2 on the control cues and 75% one the novel counter-training trials (chance was 16%). This excludes 10 participants who scored 2 standard deviations below average on the composite training scores or the any of the Phase 1 or 3 AB trials since these participants did not learn the critical cues. The scores on the training phases were independent of condition (ps>0.14) (excluding the novel counter-training trials in Phase 2, which differed by condition).

It is important to note that in the Phase 2 training, learners scored better on the novel trials: **X**B→O_1 in the induction condition (83% correct) compared to the learners on **A**X→O_5 novel trials in the counter-example condition (70% correct) [$t(77) = 2.3$, $p= 0.025$]. Since the scores in the trials near the end of the training phase are at virtually 100% for both conditions, this indicates that the learners in the counter-example condition underwent a greater amount of error correction during training.

The scores in the final testing phase indicate that there was a significant difference in final performance between training conditions. Figure 2 presents the scores for the overshadowed and trained cue compared to the overshadowed control cue for both conditions. Both between condition comparisons for trained cue B [$t(77) = 4.7$, $p< 0.001$], and within subject comparisons between the trained cue B and control cue D [paired t-test, $t(38) = 3.6$, $p= 0.001$] indicate that the counter-example training was significantly more effective, with effect sizes of 1.1 and 0.6 for between condition and within subject scores respectively.

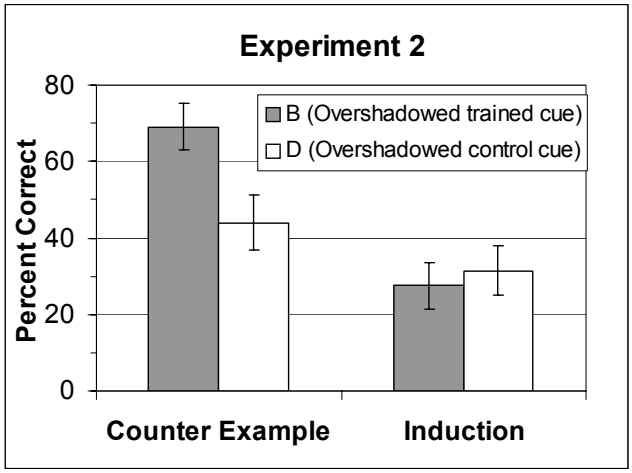

Figure 2. Experiment 2 final test scores (percent correct) on the overshadowed cues.

General Discussion

The first experiment demonstrates that when two cues of different salience are paired with an outcome over a number of trials, two important changes occur. The first is well known: the highly salient cue is strongly associated with the outcome, while the low salience cue is at most very weakly associated with the outcome. This is the classic overshadowing paradigm. A simple explanation of this result might be that the low salience cue is simply not appreciably noticed and in turn is not learned. This is supported by the first phase of Experiment 1 in which learners answer at chance when asked about the low salience cue. However, the second learning phase of the experiment uncovers a novel finding about overshadowing: it demonstrates that in fact something *was* learned about the low salience cue: it was learned to be ignored. Experiment 1 provides evidence that when a low salience cue is overshadowed, the subsequent learning of this cue is inhibited compared to a similar salience, non-overshadowed control cue. Similar to the corresponding effect in blocking, this result can be explained in terms of attentional learning models in which there is a learned decrease in the general attentional weight of the overshadowed cue.

This brings us to another result of Experiment 1. Since it has been demonstrated that a blocked cue also exhibits inhibited subsequent learning due to learned inattention, we replicated these previous finding (using a different learning task) in order to compare to the overshadowing case. While this effect on a blocked cue has been observed in humans by Kruschke and Blair (2000), we do not know of any subsequent replication.

The second experiment investigates the effectiveness of two different training methods to overcome the inhibited learning of an overshadowed, low salience cue. The results indicate that the "counter-example" method which lowers the relative validity of the competing overshadowing cue is more effective than the "induction" method in which the overshadowed cue is the only cue that is always paired with the target outcome.

We consider two possible explanations for the difference in training conditions. The first explanation is to consider a general rule that counter-examples are more effective than inductive examples in training overshadowed cues. In other words, in general multiple kinds of positive examples are not effective in learning low salience cues, whereas negative examples attacking the validity of competing higher salience cues are effective in shifting attention to the low salience cues.

The second explanation for the difference between training conditions is that, rather than the training method being inherently important, the factor determining the extent of learning and attentional shifting is the amount of error in training. This is supported the results in Experiment 2 in which learners in the counter-example condition made more initial errors in training. The extent to which initial errors in training are coupled to training method remains a question and a topic for future study.

These results may be very relevant to the learning of scientific misconceptions, which are notoriously resistant to training (e.g., Halloun & Hestenes, 1985; McDermott, 1991). In the case of velocity, acceleration and force, the association between velocity and force is the undesired association. Instead, the goal is to raise attention to the less salient cue of acceleration in order to associate it with force. However, it is continually overshadowed by velocity, and has a diminished attentional weight. Thus it may happen that if the lower salience cue is important, there are *two* strikes against it when trying to train the learner to recognize this: the first is that it has a low salience, so it may be difficult to notice, and the second is that the cue has a lowered attentional weight and is thus difficult to learn. One remedy appears to be to attack the validity of velocity as a predictor rather than only present positive examples where acceleration is predictive.

Acknowledgments

This research is supported by a grant from the Institute of Educational Sciences of the U.S. Department of Education (#R305H050125) to Andrew F. Heckler and Vladimir M. Sloutsky.

References

Clement, J. (1982). Students' preconceptions in introductory mechanics. *American Journal of Physics, 50*, 66-71.

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

Hall, G., Mackintosh, N. J., Goodall, G., & Dal Martello, M. (1977). Loss of control by a less valid or less salient stimulus compounded with a better predictor of reinforcement. *Learning and Motivation, 8,* 145-158.

Halloun, I. A., & Hestenes, D. (1985). The initial knowledge state of college physics students. *American Journal of Physics*, *53*, 1043-1055.

Kamin L. J. (1969). Predictability, surprise, attention, and conditioning. In B.A Campbell & R. M. Church (Eds.), *Punishment*. New York: Appleton-Century-Crofts.

Krushke J. K., & Johansen, M. K. (1999). A model of probabilistic category learning. *Journal of Experimental Psychology: Learning Memory and Cognition 25 (5),* 1083-1119.

Kruschke, J. K., & Blair, N. J. (2000). Blocking and backward blocking involve learned inattention. *Psychonomic Bulletin and Review, 7 (4),* 636-645.

Mackintosh NJ. 1975. A theory of attention:variations in the associability of stimuliwith reinforcement. *Psychol. Rev.* 82:276–98

Mackintosh, N. J., & Turner, C. (1971). Blocking as a function of novelty of CS and predictability of UCS. *Quarterly Journal of Experimental Psychology, 23,* 359-356.

McDermott, L. C. (1991). Millikan Lecture 1990: What we teach and what is learned—Closing the gap. *American Journal of Physics, 59*, 301-315.

Pavlov, I. P. (1927). *Conditioned Reflexes.* London: Oxford Univ. Press.

Pearce J.M.,& Hall G. (1980). A model for Pavlovian conditioning: variations in the effectiveness of conditioned but not unconditioned stimuli. *Psychol. Rev.* 87:332–52.

Pick, H. L., & Pick, A. D. (1967). A developmental and analytic study of the size-weight illusion. *Journal of Experimental Child Psychology, 5*, 362-371.

Sutherland, N. S., & Mackintosh, N. J., (1971). *Mechanisms of Animal Discrimination Learning.* New York: Academic Press.

Viennot, L. (1979). Spontaneous reasoning in elementary dynamics. *European Journal of Science Education, 1*, 205-221.

Wagner AR, Logan FA, Haberlandt K, Price T. (1968). Stimulus selection in animal discrimination learning. *J. Comp. Physiol. Psychol.* 76:171–80.