Psychonomic Bulletin & Review
2006, 13 (4), 662-667

Extinction of conditioned inhibition through nonreinforced presentation of the inhibitor

KLAUS G. MELCHERS
Universität Zürich, Zürich, Switzerland

and

SUSANN WOLFF and HARALD LACHNIT
Philipps-Universität Marburg, Marburg, Germany

In previous studies that have tried to extinguish conditioned inhibition through nonreinforced presentations of the inhibitor, researchers have repeatedly failed to find evidence for such extinction. The present study revealed that extinction can be achieved through nonreinforcement of the inhibitor, depending on properties of the reinforcer. In a human causal learning experiment, we found complete extinction in a scenario in which the reinforcer could take on negative values. Thereby, this scenario reflected the assumed symmetrical continuum on which associative strength can vary, according to the Rescorla–Wagner theory of associative learning. In contrast to this, the inhibitory cue retained its inhibitory potential in another condition, in which the scenario did not allow negative values of the reinforcer.

According to Rescorla and Wagner's (1972) theory of associative learning, conditioned excitation and conditioned inhibition are opposite ends of an underlying continuum on which a cue's associative strength V can vary. As a consequence, it should be possible to extinguish both through similar procedures. Strong evidence against this assumption, however, comes from studies in which a conditioned inhibitor failed to be extinguished through nonreinforcement, though this procedure extinguishes excitatory cues. The first of these studies came from Zimmer-Hart and Rescorla (1974); their study was followed by many more unsuccessful attempts in animal conditioning (see Williams, Overmier, & LoLordo, 1992, for a review) and in human causal learning (Yarlas, Cheng, & Holyoak, 1995). If an effect of the supposed extinction treatment was found at all, then it consisted of an increase in the inhibitory properties of the stimulus in question (e.g., DeVito & Fowler, 1987; Williams, Travis, & Overmier, 1986). As a consequence, many researchers have rejected the assumption that inhibition is the symmetrical opposite of excitation and have suggested various alternatives to account for conditioned inhibition (see Savastano, Cole, Barnet, & Miller, 1999, for a review). In contrast to this, the aim of the present study was to investigate a possible reason for these failures and to show that extinction can occur as a consequence of nonreinforced presentations of the inhibitor, depending on certain properties of the reinforcer.

The Rescorla–Wagner theory explains acquisition of conditioned excitation and of conditioned inhibition in a very similar manner. Learning in both cases is said to occur as a consequence of a discrepancy between the expected outcome and the actual outcome of a learning trial. Thus, the unexpected presence of a reinforcer should lead to an increase in the associative strength V and the unexpected absence of a reinforcer should lead to a decrease in V. As a consequence, V becomes positive (i.e., excitatory) when a cue is repeatedly paired with a reinforcer. Similarly, the omission of a reinforcer should make V negative (i.e., inhibitory) when the cue is repeatedly nonreinforced, either in an otherwise reinforced context or when it is presented together with another cue that is usually reinforced.

In this manner, the Rescorla–Wagner theory straightforwardly accounts for the acquisition of conditioned excitation and conditioned inhibition. Furthermore, it can also account for the extinction of conditioned excitation through nonreinforced presentations of the excitator. Extinction of excitation follows from the discrepancy arising when an organism expects the occurrence of a reinforcer on the basis of the given cue's positive associative strength, when this reinforcer is no longer paired with the cue in question.

The Rescorla–Wagner theory, however, fails to account for the results from experiments investigating extinction of conditioned inhibition. There, it predicts that nonreinforced presentations of an inhibitor should decrease its inhibitory potential due to the assumed discrepancy that stems from a negative value of V and the actual null outcome on the

Klaus G. Melchers has previously published under the surname "Lober." Susann Wolff is now working at the Max Planck Institute for Human Cognitive and Brain Sciences in Leipzig, Germany. We thank Holger Domsch for his help with preparing the computer program and Michael Brown and the reviewers for helpful suggestions on earlier versions of the manuscript. The research reported in this article was partially supported by a grant from the German Science Foundation (Deutsche Forschungsgemeinschaft) to H.L. (Grant La 564/12-3). Correspondence concerning this article should be addressed to K. G. Melchers, Psychologisches Institut, Universität Zürich, Binzmühlestrasse 14/12, CH-8050 Zürich, Switzerland (e-mail: k.melchers@psychologie.unizh.ch).

extinction trial. This discrepancy should, in turn, lead to an increase in V until it reaches a value of zero.

We propose that a likely reason for these failures concerns the nature of the reinforcers used. Specifically, these reinforcers did not reflect the symmetrical nature of the associative continuum assumed by the Rescorla–Wagner theory. According to this theory, an organism should expect the occurrence of less than no reinforcer as a consequence of a negative V. This expectation should be qualitatively different from the expectation of no reinforcer on the basis of a neutral cue with $V = 0$. However, the expectation of less than no reinforcer often has no real-world analogue, because the values of experimental reinforcers usually vary only unidirectionally. That means that reinforced trials are characterized by the presence of, for example, a food pellet or an electric shock, whereas nonreinforced trials are characterized by the absence of these events. However, it is impossible that these reinforcers take on values of less than zero. Therefore, the only reasonable consequence an organism can expect when a conditioned inhibitor is shown on its own is that no reinforcement will occur. Contrary to the assumptions of the Rescorla–Wagner theory, there is no discrepancy between the expected and the actual outcome on the alleged extinction trials. If anything at all, then the absence of a reinforcer confirms the expectation that no such reinforcer should occur.

We are not aware of any studies investigating the impact of the unidirectional nature of the reinforcers. Recent findings concerning the impact of cognitive factors on human causal learning and on Pavlovian conditioning in humans (see De Houwer, Beckers, & Vandorpe, 2005, for a review), however, indirectly support our suggestion. These findings show that the way learners conceive and think about a reinforcer may have considerable influence on the outcome of an experiment. In several experiments, for example, blocking was investigated (see, e.g., Mitchell & Lovibond, 2002). In a blocking procedure, a cue A is repeatedly paired with a reinforcer (A+). Then A is shown in compound with another cue B, and this compound is also reinforced (AB+). Blocking is said to occur when the response elicited by B in a later test is weaker compared with a condition in which A was not pretrained. The crucial finding was that the strength of blocking was related to whether participants assumed that reinforcers were additive—that is, that the presence of two valid predictors of reinforcement would lead to the application of two reinforcers (AB++). When participants could make such an assumption, blocking was stronger than in conditions in which participants could not make such an assumption but were encouraged to assume that reinforcers could not be larger than a certain ceiling value. Thus, during the AB+ trials, participants in the former condition could rule out that B had an effect on its own, whereas this was not possible in the latter condition due to the imposed ceiling.

When they experience a nonreinforced inhibitor, learning organisms are confronted with a similar problem as were the participants in the studies just mentioned. These trials are only informative if the reinforcer can potentially take on values below zero. In such a case, a real discrepancy between the organism's expectation and the actual outcome occurs, so that the inhibitor should lose its inhibitory potential. In contrast, it should remain inhibitory for conditions in which values below zero are not possible. In the present experiment, we therefore developed a causal learning scenario in which reinforcers could take on values larger and smaller than zero, and compared this with a condition in which the reinforcer could vary in only one direction.

METHOD

We used a medical prediction task similar to the task used in many human causal learning studies (e.g., Aitken, Larkin, & Dickinson, 2000; Melchers, Lachnit, & Shanks, 2004). In the present task, the participants had to learn which of several foods caused a change in the hormone level of a hypothetical patient. In Group Unidirectional, the hormone level could either increase or remain unchanged, which parallels the conditions from earlier causal learning studies investigating inhibition (e.g., Aitken et al., 2000; Chapman & Robbins, 1990). In contrast to this, the hormone level in Group Bidirectional could increase, remain constant, or decrease. Thus, potential expectations based on positive, neutral, or negative values of V had analogues in the values the reinforcer could take on.[1]

The experiment consisted of an acquisition and an extinction stage, each of which was followed by a test phase (Table 1). During acquisition, the participants were shown A+, AX^0 trials, where 0 depicts nonreinforcement of the AX compound. This should make X inhibitory. Furthermore, a cue to be used for a later summation test was also presented and reinforced (B+). Filler trials ensured that the participants also experienced nonreinforced presentations of a single cue as well as reinforced presentations of a compound.

An additional filler cue F was used to stress the unidirectional versus bidirectional nature of the reinforcer. After presentations of F, the hormone level remained constant in Group Unidirectional (F^0) but decreased in Group Bidirectional (F−). During the extinction stage, X^0 trials were shown instead of the AX^0 trials. Training of the other cues proceeded as before, with the exception that no more DE+ trials were shown.

Participants

The participants, 64 student volunteers, were tested individually and needed approximately 12 min to complete the experiment.

Procedure

Instructions and all necessary information were presented on a computer screen. The participants gave their answers by using the mouse. The following foods were used as cues for the experiment: bananas, broccoli, carrots, grapes, mushrooms, nuts, strawberries, and tomatoes. An incomplete Latin square was used for the alloca-

Table 1
Experimental Design

Type of Trials	Acquisition	Extinction	Test
Inhibition training	A+, AX^0	A+, X^0	A?, X?, N?
Transfer cue training	B+	B+	B?, BX?, BN?
Filler trials	C^0, DE+, F^0/F−	C^0, F^0/F−	C?, D?, E?, F?

Note—The letters depict different cues."+" means an increase of the hormone level, 0 means no change of the hormone level, and "−" means a decrease of the hormone level. For Group Unidirectional, F^0 filler trials were used, and for Group Bidirectional, F− filler trials were used. "?" means that participants were asked to rate the causal relationship between the different stimuli and the reinforcer.

tion of the various foods to the different cues, ensuring that each food was used equally often for each kind of cue.

The participants were told that for each day, they would be informed which foods the patient had eaten. On the basis of these foods, they had to make a prediction indicating whether or not they expected a change in the patient's hormone level. They were also told that they would get feedback for each day, allowing them to find out how the different foods influenced the hormone level.

Finally, the participants were informed that later in the experiment they would have to rate to which degree each food influenced the hormone level. Then, they were shown the rating scale (ranging from −10 to +10) used for the later tests. For both groups, a positive rating indicated that the food increased the hormone level, whereas a neutral rating (0) indicated that the food had no effect. The negative pole of the scale was labeled *prevents an increase* for Group Unidirectional (similar to studies in which the reinforcer could vary in only one direction, e.g., Aitken et al., 2000) and *decreases the level* for Group Bidirectional.

Both training stages consisted of six blocks. Each trial type was presented once per block, yielding 36 trials for the acquisition and 30 trials for the extinction stage. The order of presentation was determined randomly for each block. On each trial, the participants made their predictions by clicking a response button. For Group Unidirectional, two buttons labeled *increase* and *no change*, respectively, were shown. Group Bidirectional was also shown a third button, labeled *decrease*. After the participants had made their predictions, a feedback window appeared, showing the actual outcome of the respective trial.

After each learning stage, a test was conducted in which inhibition was assessed in two ways. For a *direct comparison*, the participants had to rate the putative inhibitor X on its own, as well as a neutral stimulus N not used during training. The comparison between X and N allowed assessment of the inhibitory properties of X on its own. Additionally, a *summation test* was conducted, for which the participants had to rate the compounds BX and BN (i.e., the excitatory transfer cue B was either combined with X or with N). Furthermore, the participants also had to rate the additional cues used in the learning stages. The order of presentation of the different stimuli was determined randomly for each participant.

RESULTS

Learning Stages

During the acquisition and extinction stages, the participants in both groups quickly learned about the experimental stimuli so that their predictions mirrored the actual contingencies as early as from Block 2 onward.

Causal Ratings After the Acquisition Stage

The mean ratings for the first test are displayed in the upper left and middle panels of Figure 1. In both groups, X was inhibitory after A+, AX^0 training, so that X was rated more negatively than N in both groups, and BX was rated lower than BN. This was confirmed by separate stimulus × group ANOVAs for the two manners used to assess inhibition (direct comparison: X vs. N, and summation test: BX vs. BN, respectively). The .05 level of significance was used for all analyses. In both cases, we found a significant main effect of stimulus [$F(1,62) = 20.17$, $MS_e = 35.10$, and $F(1,62) = 21.48$, $MS_e = 19.58$, respectively]. Furthermore, the main effect of group was also significant for the direct comparison [$F(1,62) = 9.79$, $MS_e = 32.90$], reflecting that X and N were both rated lower in Group Unidirectional than in Group Bidirectional. None of the remaining main effects or interactions was significant.

Causal Ratings After the Extinction Stage

The lower left and middle panels of Figure 1 show the results of the second test. It is evident that X remained inhibitory in Group Unidirectional but lost its inhibitory properties in Group Bidirectional. In Group Unidirectional, X was still rated more negatively than N in the direct comparison, and BX was still rated lower than BN in the summation test. In contrast to this, none of the comparisons showed evidence of X having retained inhibitory properties in Group Bidirectional. Accordingly, a significant stimulus × group interaction was detected for the direct comparison [$F(1,62) = 11.36$, $MS_e = 19.42$] and also for the summation test [$F(1,62) = 15.36$, $MS_e = 12.86$]. Furthermore, a significant main effect of stimulus was found for the direct comparison [$F(1,62) = 4.87$, $MS_e = 19.42$], as well as significant main effects of group for both comparisons [$F(1,62) = 8.87$, $MS_e = 20.34$, and $F(1,62) = 8.32$, $MS_e = 26.20$, respectively].

Due to the significant interactions, we conducted additional (Bonferroni-corrected) paired-samples *t* tests for both groups. For Group Unidirectional, we found significant differences between the ratings in the direct comparison and also in the summation test [$t(31) = 3.30$, and $t(31) = 3.28$, respectively]. In contrast to this, no significant differences were found for the respective comparisons in Group Bidirectional (both ts < 2.18).

Supplementary Analyses

Two aspects of the data shown in Figure 1 seem noteworthy. First, visual inspection of the ratings for X before and after extinction training suggests that although X retained considerable inhibitory potential, the conditioned inhibitor was weakened to some degree, even in Group Unidirectional. Two additional paired-samples *t* tests confirmed that X was rated significantly less negative after the extinction training than before, and that BX was rated significantly higher than before [$t(31) = 3.02$, and $t(31) = 2.41$, respectively].

Second, the absolute value of Group Bidirectional's ratings of X after the acquisition seems relatively low in comparison with previous studies (e.g., Aitken et al., 2000; Chapman & Robbins, 1990; Melchers et al., 2004). And even though the stimulus × group interactions did not reach significance for the ratings after the acquisition stage, this might suggest that the inhibitory effect of X (as compared with N) was somewhat stronger for Group Unidirectional than for Group Bidirectional. Closer inspection of the data revealed that 9 participants in Group Bidirectional failed to learn that X was an inhibitor and rated it positively after the acquisition stage, some of them even giving it the maximum rating of +10. In Group Unidirectional, on the other hand, only 2 such participants were found. We have no explanation for this finding, which is at variance with previous inhibition experiments from our laboratory using an allergy prediction

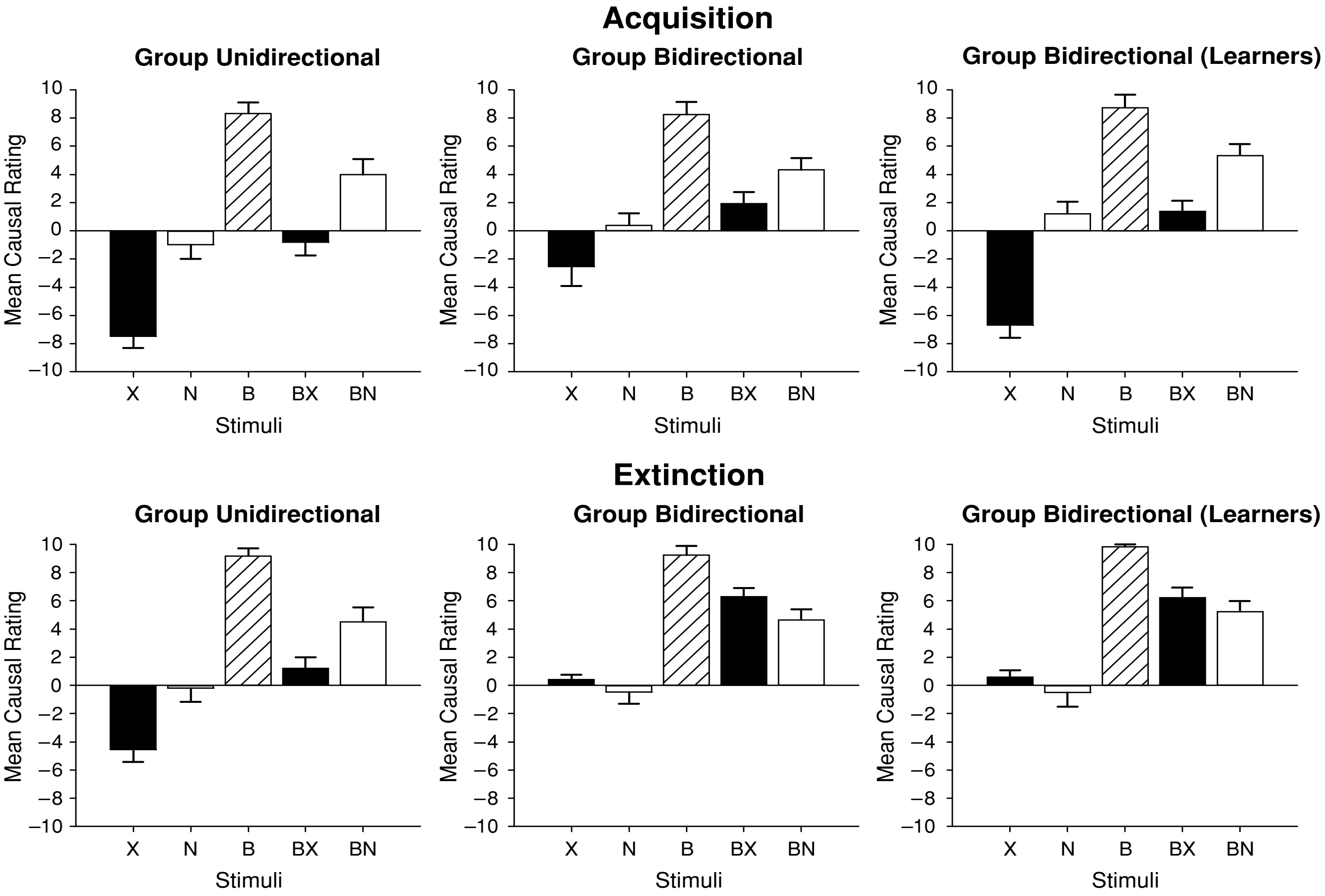

Figure 1. Mean causal ratings after the acquisition stage (top) and after the extinction stage (bottom).

task (e.g., Melchers et al., 2004, Experiment 2) and also with another, unpublished experiment that used the same scenario as the present study. To ensure that our finding of complete extinction of the inhibitory properties of X was not due to incomplete learning during the acquisition stage or to guessing behavior of some participants, we reanalyzed the data of only those participants who did not rate X positively after the acquisition stage. The data of the respective participants in Group Bidirectional (the learners) are displayed in the right section of Figure 1. The data for learners from Group Unidirectional were rather similar to the complete data set and are therefore not displayed in the figure.

As can be seen, the inhibition effect for the remaining participants in Group Bidirectional was much more pronounced after the acquisition stage but nevertheless was completely extinguished after X^0 training. ANOVAs of the learners' ratings after the acquisition stage generally showed the same pattern of results as for the complete data set; the only exception was that the inhibition tests revealed stronger effects.

For the ratings after the extinction training, the ANOVA for the direct comparison revealed a significant main effect of group [$F(1,51) = 8.08$, $MS_e = 20.84$] as well as the expected stimulus × group interaction [$F(1,51) = 9.16$, $MS_e = 21.21$], reflecting the different effects of the extinction training in the two groups. For the summation test, the results were rather similar, with a significant main effect of group [$F(1,51) = 8.57$, $MS_e = 26.44$] and also a significant stimulus × group interaction [$F(1,51) = 9.44$, $MS_e = 13.76$]. For both ANOVAs, the main effect of stimulus failed to reach significance (both Fs < 3.15).

Additional Bonferroni-corrected paired-samples t tests confirmed that X was still inhibitory for Group Unidirectional after the extinction training, as measured by both the direct comparison [$t(29) = 3.09$] and the summation test [$t(29) = 3.23$]. In contrast to this, X passed none of the inhibition tests in Group Bidirectional (both ts < 1.20).

DISCUSSION

The Rescorla–Wagner theory and its prediction of extinction of conditioned inhibition through nonreinforcement was the starting point for the present investigation. Contrary to various unsuccessful attempts, we were able to confirm this prediction. Our results suggest that the crucial factor for successful extinction was that the causal learning scenario used in Group Bidirectional made it possible for the participants to expect different outcomes after presentations of inhibitory cues, as opposed to neutral cues. Thus, when the underlying continuum of associative strength assumed by the Rescorla–Wagner theory was mirrored by a continuum on which the value of the reinforcer could vary, then (and only then) did the predicted extinction effect clearly occur. More in line with earlier unsuccessful attempts, the inhibitor retained most of its inhibitory potential after extinction training (albeit somewhat less pronounced than before) when the value of the reinforcer in Group Unidirectional could vary in only one direction.

It might be argued that the different effects of the extinction treatment in the two groups are hardly surprising, given that Group Bidirectional was taught a response to the inhibitor that was incompatible with its associative status (i.e., that it did not lead to the expected decrease in the hormone level), whereas this was not the case in Group Unidirectional. In our view, however, this is exactly the crucial point that was missed in earlier attempts to explain why inhibition could not be extinguished through nonreinforcement: Extinction trials are only informative when a reinforcer can vary in both directions so that a discrepancy between the participants' expectations and the actual outcome of a learning episode can occur. Nevertheless, future research should attempt to replicate our findings with procedures that are more parallel in both experimental groups than in the present study.

A potential limitation of our study is that we did not include a control condition (e.g., $G+$, GY^0) for the extinction treatment, to assess whether extinction was specific to the extinguished cue X or would also affect a nonextinguished inhibitor Y. Future studies should include such a control condition. Nonetheless, the omission of such a control in the present study does not diminish the finding that extinction did occur after nonreinforcement. To our knowledge, our study is the first to report such an effect.

Together with related findings concerning the role of cognitive factors in human causal learning or human conditioning, our results highlight the sensitivity of learners for aspects such as the nature of the reinforcer and the ability to take these aspects into account. Yet, in contrast to other researchers (e.g., De Houwer et al., 2005; Mitchell & Lovibond, 2002), we do not believe that this necessarily refutes associative theories as possible accounts of how humans learn about relationships between cues and their potential outcomes. Although some findings seem to be better explained by cognitive or inferential accounts (De Houwer et al., 2005), results from other studies question those accounts (see, e.g., Lober & Shanks, 2000; Melchers et al., 2004). Furthermore, with regard to nonextinction of conditioned inhibition, Rescorla (1973) has suggested a modification of the Rescorla–Wagner algorithm that allows that nonreinforced inhibitors retain their inhibitory potential. If the nature of the reinforcer determines the specific algorithm employed to calculate discrepancies between expectations and actual outcomes of learning episodes, then associative as well as inferential accounts can successfully describe the present findings for both groups. Thus, the present study does not help to decide whether the former or the latter account provides a more appropriate model of human causal learning.

In any case, however, we think that it is important to further investigate factors that influence the multifaceted ways in which organisms learn about relationships in their environment, especially in situations in which associative theories encounter difficulties.

REFERENCES

Aitken, M. R. F., Larkin, M. J. W., & Dickinson, A. (2000). Super-learning of causal judgements. *Quarterly Journal of Experimental Psychology*, **53B**, 59-81.

Chapman, G. B., & Robbins, S. J. (1990). Cue interaction in human contingency judgment. *Memory & Cognition*, **18**, 537-545.

De Houwer, J., Beckers, T., & Vandorpe, S. (2005). Evidence for the role of higher order reasoning processes in cue competition and other learning phenomena. *Learning & Behavior*, **33**, 239-249.

DeVito, P. L., & Fowler, H. (1987). Enhancement of conditioned inhibition via an extinction treatment. *Animal Learning & Behavior*, **15**, 448-454.

Lober, K., & Shanks, D. R. (2000). Is causal induction based on causal power? Critique of Cheng (1997). *Psychological Review*, **107**, 195-212.

Melchers, K. G., Lachnit, H., & Shanks, D. R. (2004). Within-compound associations in retrospective revaluation and in direct learning: A challenge for comparator theory. *Quarterly Journal of Experimental Psychology*, **57B**, 25-53.

Melchers, K. G., Üngör, M., & Lachnit, H. (2005). The experimental task influences cue competition in human causal learning. *Journal of Experimental Psychology: Animal Behavior Processes*, **31**, 477-483.

Mitchell, C. J., & Lovibond, P. F. (2002). Backward and forward blocking in human electrodermal conditioning: Blocking requires an assumption of outcome additivity. *Quarterly Journal of Experimental Psychology*, **55B**, 311-329.

Rescorla, R. A. (1973). A model of Pavlovian conditioning. In V. S. Rusinov, P. V. Simonov, & M. N. Rusalova (Eds.), *Mechanisms of the formation and inhibition of conditioned reflexes* (pp. 25-39). Moscow: Nauka Academy of Sciences of the U.S.S.R.

Rescorla, R. A., & 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 theory and research* (pp. 64-99). New York: Appleton-Century-Crofts.

Savastano, H. I., Cole, R. P., Barnet, R. C., & Miller, R. R. (1999). Reconsidering conditioned inhibition. *Learning & Motivation*, **30**, 101-127.

Williams, D. A., Overmier, J. B., & LoLordo, V. M. (1992). A reevaluation of Rescorla's early dictums about Pavlovian conditioned inhibition. *Psychological Bulletin*, **111**, 275-290.

Williams, D. A., Travis, G. M., & Overmier, J. B. (1986). Within-compound associations modulate the relative effectiveness of differential and Pavlovian conditioned inhibition procedures. *Journal of Experimental Psychology: Animal Behavior Processes*, **12**, 351-362.

Yarlas, A. S., Cheng, P. W., & Holyoak, K. J. (1995). Alternative approaches to causal induction: The probabilistic contrast versus the Rescorla–Wagner model. In J. F. Lehman & J. D. Moore (Eds.), *Proceedings of the Seventeenth Annual Conference of the Cognitive Science Society* (pp. 431-436). Hillsdale, NJ: Erlbaum.

Zimmer-Hart, C. L., & Rescorla, R. A. (1974). Extinction of Pavlovian conditioned inhibition. *Journal of Comparative & Physiological Psychology*, **86**, 837-845.

NOTE

[1] A task that could be modified straightforwardly, and that might seem more realistic, is the stock market task, in which different stocks that are traded represent the cues, and an increase in the overall value of the stock market represents the reinforcer (Chapman & Robbins, 1990). However, we recently found (Melchers, Üngör, & Lachnit, 2005) that cue-selection effects such as blocking (and potentially also conditioned inhibition) are less pronounced with this task than with medical prediction tasks.

(Manuscript received July 28, 2005;
revision accepted for publication January 10, 2006.)