Extinction and blocking of conditioned inhibition in human causal learning

IRINA BAETU AND A. G. BAKER
McGill University, Montréal, Québec, Canada

Two experiments investigated extinction and blocking of a conditioned inhibitor in a human contingency-learning task. Lotz and Lachnit (2009) and Melchers, Wolff, and Lachnit (2006) reported extinction of inhibition only when participants experienced outcome levels lower than those used in training. In Experiment 1, which used more neutral instructions than the previously mentioned studies, we found that extinction of inhibition occurred, whether or not participants experienced lower outcome levels. In Experiment 2, we applied this outcome manipulation to blocking of a conditioned inhibitor. We found blocking of inhibition both when participants had experience with lower outcomes and when they did not. The results of our two experiments are consistent with Rescorla and Wagner's (1972) associative model, and inconsistent with an inferential account of causal learning (De Houwer, Beckers, & Vandorpe, 2005).

Learning about cues that predict when outcomes occur is useful for people and animals, for obvious reasons. However, learning when these outcomes will not occur can sometimes be equally important. In this article, we examine possible descriptions of how the absence of an event is learned and how different theories represent this learning. In a causal scenario, we will investigate whether the values that an outcome can take, and how they are mapped onto a rating scale, have an impact on people's judgments of such cues.

We report two experiments investigating learning about cues that predict the absence of an outcome. Although several procedures have been used to produce such cues, we will focus on the *Pavlovian conditioned inhibition procedure* (Pavlov, 1927). In animal conditioning, this procedure consists of pairing a cue that we will refer to as A (e.g., a tone) with an outcome (such as food or electric shock). Cue A is also presented on other trials with another cue, B, and the outcome is omitted. We denote the presence of the outcome by $+$ and its absence by 0 (A^+ , AB^0). B is an inhibitor because it predicts the absence of the outcome that would normally have occurred with A. The *summation test* is one way of assessing conditioned inhibition (Rescorla, 1969). This test consists of pairing a separately trained exciter that predicts the occurrence of the outcome with the inhibitor. If the inhibitor successfully reduces the conditioned response to the new exciter, this indicates that it might possess some inhibitory properties independent of the compound in which it was originally trained. In human contingency-learning experiments, participants can also be directly asked whether a cue has preventive or inhibitory powers (e.g., Baker, Mercier, Vallée-Tourangeau, Frank, & Pan, 1993; Dickinson, Shanks, & Evenden, 1984).

We were interested in procedures that might reduce the strength of conditioned inhibitors. One such procedure that we investigated in Experiment 1 is extinction of inhibition. An extinction procedure consists of repeatedly presenting a trained cue without the outcome. Extinction of inhibition thus consists of presenting Cue B in the absence of the outcome (B^0) after the A^+ , AB^0 training.

Rescorla and Wagner's (1972) associative model predicts that the extinction procedure should diminish the inhibitory properties of Cue B. The model was inspired in part by Kamin's (1969) work, showing that a cue will acquire a conditioned response only if it is followed by an unexpected unconditioned stimulus; this suggests that only unexpected or surprising outcomes support new learning. According to the model, learning consists of the development of associations between the cues and the outcomes. On every trial, the association between a cue presented on that trial and the outcome changes as a function of how discrepant the associative strength of the presented cue is from the strength of the outcome on that trial. Thus, the Rescorla–Wagner model represents a formalization or theory of the concept of surprise, because the discrepancy between the associative strength of the cues and the strength of the actual outcome mirror the extent to which the actual outcome was unexpected. It is a true theory of “surprise,” because surprise is not a primitive of the model; rather, the predictions from the model that mirror some aspects of “surprise” are an emergent property of the workings of the model.

Training with A^+ , AB^0 trials results in an excitatory association between A and the outcome, and also in an inhibitory association between B and the outcome. The negative association develops because, on AB^0 trials, the

I. Baetu, irina.baetu@mail.mcgill.ca

outcome is omitted but A creates an expectation of it, so there is a discrepancy. Inhibition is represented by a negative value that cancels out excitation when it is paired with an excitatory cue. Extinction of an exciter occurs because there is a discrepancy between the associative strength of the exciter and the absence of reinforcement. Thus, the model adjusts the strength of the exciter to zero. Likewise, when a negative cue is presented by itself in the absence of a reinforcer, there is a discrepancy between its negative strength and zero (the absent outcome), so it is adjusted up to zero. Thus, extinction of inhibition is predicted. Unlike excitatory extinction, this prediction of inhibitory extinction is counterintuitive because the inhibitor is followed by the same "event," nonreinforcement, on both training (AB^0) and extinction (B^0) trials.

The prediction that inhibition should extinguish is considered to be a weakness of the model because it is difficult to find extinction (see Miller, Barnet, & Grahame, 1995, for a review of the model's successes and failures). Indeed, although some researchers found extinction of inhibition in animals (Detke, 1991; Holland, 1985; Holland & Gory, 1986; Robbins, 1990), most have failed (Detke, 1991; DeVito & Fowler, 1986, 1987; Hallam, Grahame, Harris, & Miller, 1992; Lysle & Fowler, 1985; Miller & Schachtman, 1985; Owren & Kaplan, 1981; Pearce, Nicholas, & Dickinson, 1982; Rescorla, 1982; Williams, 1986; Williams & Overmier, 1988; Williams, Travis, & Overmier, 1986; Witcher & Ayres, 1984; Zimmer-Hart & Rescorla, 1974). Yaras, Cheng, and Holyoak (1995) found no extinction in a human contingency-learning task. Thus, a number of researchers have abandoned the idea that excitation and inhibition reflect underlying symmetrical associative processes and have developed different views of inhibition (e.g., Konorski, 1948; Rescorla, 1973; Zimmer-Hart & Rescorla, 1974).

Melchers, Wolff, and Lachnit (2006) suggested that the nature of the outcome can determine whether inhibition extinguishes or not. They proposed that the Rescorla-Wagner model is correct and that inhibition will extinguish when the outcome can take on negative values. However, it incorrectly predicts extinction of inhibition when, as in conditioning procedures, the outcome cannot take on negative values. For example, an unconditioned stimulus, such as food, is either present or absent, so this outcome cannot reflect the negative expectation represented by the inhibitor's negative associative strength. At most, an animal might expect the outcome to be absent, which is exactly what happens. If the outcome cannot confirm a negative expectation, the Rescorla-Wagner model fails. However, if the outcome became negative or, at least, lower than the level occurring on the extinction trials, it could confirm negative expectations, and the Rescorla-Wagner model's prediction concerning extinction of inhibition would be expected. Thus, the absence of the outcome on B^0 trials would be inconsistent with the model's negative prediction, which could have been reflected in negative or lower outcomes. Thus, according to Melchers et al., an organism should be sensitive to whether a negative outcome expectation could be con-

firmed in the real world by an actual negative outcome. They used this notion to modify the original Rescorla-Wagner model. Only when the negative expectation might be confirmed should an organism behave in a way consistent with the inhibitory postulates of the Rescorla-Wagner model.

Melchers et al. (2006) tested this hypothesis in a human contingency-learning experiment by manipulating the type of outcome. One group (Group Unidirectional) only observed outcomes that were either present or absent (positive or zero); the other group (Group Bidirectional) observed outcomes that could also take on negative values (i.e., positive, zero, and negative). According to their hypothesis, extinction of inhibition should occur only in Group Bidirectional, which experienced the possibility of a negative outcome. In these experiments, a hypothetical patient's hormone level could vary following the consumption of certain foods. In Group Unidirectional, the hormone level increased or remain unchanged (the positive and zero outcomes). In Group Bidirectional, the hormone level could also decrease, so only Group Bidirectional had the opportunity to observe negative outcomes. All participants received A^+ , AB^0 training, followed by B^0 trials. A and B were the foods; + denotes the hormone level increase; and 0 , the unchanged hormone level. Following the extinction phase, B was no longer inhibitory in the group that experienced negative outcomes—that is, Group Bidirectional. However, B did not lose its inhibitory properties in the group that did not experience negative outcomes (Group Unidirectional), although it was somewhat less inhibitory than before the extinction phase. Thus, experience with negative outcomes seemed to have resulted in stronger extinction of inhibition. This result was replicated by Lotz and Lachnit (2009).

Melchers et al. (2006) pointed out that their results were consistent with inferential theories of learning (De Houwer, Beckers, & Vandorpe, 2005). These theories might posit that participants in Group Unidirectional perceived a floor in the outcome level and reasoned that it was not possible to observe this effect because the lowest outcome level was zero, although on B^0 trials B might have had a negative effect on the outcome. Hence, they would not reevaluate B as noninhibitory during the extinction phase. Group Bidirectional, on the other hand, was exposed to negative outcomes and, hence, had reason to believe that if B was inhibitory, it should have reduced the hormone level. Extinction (B^0) trials were inconsistent with this inference, so participants in this group reevaluated B as neutral.

Melchers et al. (2006) and Lotz and Lachnit (2009), however, used different instructions and rating scales for the two groups. This might have biased their results. Both groups were told that a positive rating meant that the food increased the hormone level, and a zero rating meant that the food did not influence the hormone level. Group Unidirectional, which experienced only positive or zero outcomes, was further informed that a negative rating on their rating scale indicated that the food *prevented* an increase in the hormone level. Group Bidi-

rectional, which did experience negative outcomes, was told that a negative rating meant that the food *decreased* the hormone level. The different instructions suggest different mechanisms of inhibition. Suggesting that an inhibitor prevents the hormone level increase caused by another cue is similar to Zimmer-Hart and Rescorla's (1974) conception of inhibition as a threshold, so the inhibitor should have an effect only in the presence of an exciter. On the other hand, suggesting that an inhibitor directly reduces the hormone level is consistent with the Rescorla–Wagner model's assumption that exciters and inhibitors have symmetrical but opposite effects on the outcome. The description of inhibition as a threshold implies that an inhibitor should have no effect on its own and, therefore, B⁰ trials are not informative, whereas the representation in which inhibitors directly reduce the outcome suggests the opposite. These differences might encourage participants to infer extinction only in Group Bidirectional.

EXPERIMENT 1

In Experiment 1, we tested Melchers et al.'s (2006) hypothesis that extinction of inhibition should occur more readily for participants who experienced negative outcomes. We used the same general scenario and procedure, but a different design. All participants received the same instructions and used the same rating scale. The participants observed one or two foods on every trial and guessed what would happen to the hormone level. They were then given feedback. To make a prediction, they could click on one of two buttons labeled "Increase" and "No change" in Group Unidirectional, or on one of three buttons labeled "Increase," "No change," and "Decrease" in Group Bidirectional.

The test phase also involved presenting one or two foods and predicting the resulting hormone level. However, instead of having a choice of two or three buttons to make their predictions, participants could predict the change in hormone level on a continuous scale. The left end of the scale was labeled "The level will decrease," the right end was labeled "The level will increase," and the middle, "The level will not change." In the test phase, participants were not given feedback following their predictions. We chose to ask participants to simply predict the hormone level during the test phase, because making a prediction is less abstract than a causal question such as "What is the effect of food X on the hormone level?" Also, asking a causal question forces one to label the ends of the rating scale with the possible effects of the cue, and we wanted to avoid this. Labeling the positive end of a causal rating scale is straightforward, since a positive cue in this scenario increases the hormone level, but labeling the negative end means choosing to describe one of two possible inhibitory mechanisms: either the inhibitory cue prevents an increase, or it directly decreases the hormone level. By asking participants to predict the hormone level, we avoided having to describe the possible mechanisms for the effects of an inhibitory cue.

Furthermore, to minimize the influence of the instructions associated with the rating scale, our participants evaluated the various cues only at the end of the experiment. In the studies by Melchers et al. (2006) and Lotz and Lachnit (2009), participants were asked to evaluate the effect of the cues both after the initial training with the A⁺ and AB⁰ trials, and, after the extinction, B⁰ trials, so that the ratings of B before and after the extinction phase could be compared. The effect of their different instructions might have been particularly strong, given that their participants were exposed to the rating scale immediately before the extinction phase. We, on the other hand, asked participants to evaluate the effect of the cues only at the end of the experiment, and compared the extinguished inhibitor, B, with a separately trained inhibitor that did not undergo extinction. Our design should diminish the effect of rating instructions during learning. In addition, our design has the advantage that B's reduced inhibition, compared with the nonextinguished inhibitor, might not be attributed to interference or decaying memories, since both inhibitors were trained at the same time and, presumably, were equally influenced by the passage of time or the interference caused by the trials that followed their training.

The experimental design involved two training phases and a testing phase at the end of the experiment. The design of Experiment 1 is summarized in Table 1. During the first training phase, participants observed A⁺ and AB⁰ trials, and B was subsequently extinguished (B⁰) during the second phase. Cue A received more pairings with the hormone level increase during Phase 2, as in Melchers et al. (2006). To demonstrate that a decline in outcome could occur following certain foods, in Group Bidirectional one single food (K⁻) and one compound of two foods (LM⁻) were paired with a decline in hormone level. In Group Unidirectional, these foods were paired with no change in the hormone level (K⁰, LM⁰).

To provide cues for summation tests, and to provide various control conditions, participants observed a number of other cue–outcome relationships. They observed a control cue, D, that received training trials similar to Cue B but that wasn't a conditioned inhibitor, because it was never paired with an exciter. Participants observed C⁰, CD⁰ trials in Phase 1 and C⁰, D⁰ trials in Phase 2. Cue D should have remained relatively neutral following this training;

Table 1
Design of Experiment 1

Phase 1	Phase 2	Test
A ⁺ , AB ⁰	A ⁺ , B ⁰	A, B, AB, NB
C ⁰ , CD ⁰	C ⁰ , D ⁰	C, D, ND
E ⁺ , EF ⁰	E ⁺	E, F, NF
G ⁺	G ⁺ , GH ⁰	G, H, NH
IJ ⁺	IJ ⁺	I, J
K ^{0/-} , LM ^{0/-}	K ^{0/-} , LM ^{0/-}	K, L, M, NK, NL, NM
N ⁺	N ⁺	N

Note—⁺, hormone level increase; ⁰, no change in hormone level; ⁻, hormone level decrease. In Group Unidirectional, Cues K and LM were followed by no hormone level change (⁰), and in Group Bidirectional, these cues were followed by a hormone level decrease (⁻).

we used this cue as a control to determine whether Cue B retained some inhibition. We also trained a second conditioned inhibitor, F, in Phase 1 by giving participants E^+ and EF^0 trials. F did not appear during the second phase and should have retained its inhibitory potential acquired in the first phase. Not only should F be more inhibitory than D, but it should also be more inhibitory than B, if B underwent extinction during Phase 2.

We trained a third conditioned inhibitor, H, to determine whether there was some loss of the inhibitory power of the conditioned inhibitors trained in Phase 1. During Phase 2, participants observed G^+ and GH^0 trials. Since H was trained more recently, loss of inhibition should be detected if, at test, H is rated as more inhibitory than F, the inhibitor trained in Phase 1. As had Melchers et al. (2006), we included a compound of two cues followed by a hormone level increase (IJ^+). This was done to prevent participants from learning that a compound was never followed by a hormone level increase. Finally, food N was followed by a hormone level increase (N^+), and this exciter was used in the summation tests. The IJ , K , LM , and N trials were presented during both training phases.

Method

Participants. Fourteen McGill University students were assigned to Group Unidirectional (11 women and 3 men), and another 14 were assigned to Group Bidirectional (10 women and 4 men). Fourteen participants were used in each group, so we could assign one to each cell of the incomplete Latin-square design for the cues' assignment described below. They participated in exchange for credit in a psychology course or a payment of \$CAD 8 each. We used a learning criterion according to which participants had to correctly solve the A^+ , AB^0 and the E^+ , EF^0 discriminations on the last block of trials in the first phase in order to be included in the analyses. We used this criterion to ensure that all participants had learned the discrimination involving the two most important cues, B and F. One female participant in each group failed to reach this criterion. They were replaced by 2 female participants.

Apparatus and Stimuli. The experiment was programmed in RealBasic and was run on Apple Mac computers with color monitors. The following 14 foods were used for the 14 cues: corn, peppers, pears, basil, peaches, bran, beef, bananas, potatoes, blueberries, carrots, cabbage, broccoli, and chicken. Each food name was accompanied by a picture of that food. Fourteen different food-cue assignments were generated using an incomplete Latin-square design, and each food-cue assignment was used for 1 participant in each group.

Procedure. The instructions on the computer screen (found in the Appendix) informed participants that their task was to discover whether different foods eaten by a hypothetical client influence a hormone level. Participants were further informed that they would observe several records showing one or two foods that the client had eaten, that they would have to predict what would happen to the hormone level, and that they would be given feedback. They were also informed that, after they had seen all the records, they would make additional predictions without feedback. They subsequently observed the trials in Phases 1 and 2. There was no obvious transition between the phases. There were six blocks of trials in each phase. Each block consisted of one of each of the trial types for the respective phase shown in Table 1. The trials in each block were randomly mixed. Thus, in each phase, participants observed six trials of each type.

On every trial, the picture(s) of one or two foods appeared at the top of the screen. Immediately below, the following question ap-

peared: "What do you think will happen to the hormone level? (click on one of the buttons below)." In Group Unidirectional, two buttons labeled "No Change" and "Increase" were displayed. In Group Bidirectional, a third button labeled "Decrease" was displayed.

Following the two training phases, participants were asked to rate how much the hormone level would increase or decrease if the patient ate each food individually or ate one of eight combinations of two foods (refer to Table 1 for the test trials). The compound trials represented the summation tests for inhibition for B and the control stimuli. The test trials were randomly intermixed. Participants in both groups were asked the following question on every test trial: "What do you think will happen to the hormone level? Use the scale below to indicate how much the hormone level will change." They made their predictions on a scale on which the middle point was labeled "The level will not change," the left extremity was labeled "The level will decrease," and the right extremity was labeled "The level will increase." The position of the cursor on the rating scale was recorded as a number between -100 and 100.

Results

Figure 1 shows the average ratings given during the test phase in the two groups for the critical cues involved in assessing whether or not conditioned inhibition extinguished in Groups Bidirectional and Unidirectional. A complete list of all the ratings given in Experiment 1 is shown in Table 2. In both groups, participants reported that the nonextinguished conditioned inhibitor, F, would decrease the hormone level, but that the extinguished inhibitor, B, would have a much smaller effect. In fact, the ratings given to B are very similar to the ones given to D, the control cue, suggesting that the B^0 trials in Phase 2 completely abolished B's inhibitory properties in both groups. Therefore, it would appear that conditioned inhibition extinguished regardless of whether a negative outcome was observed during training.

We performed a 3 (cue: B, D, F) $\times$ 2 (group: unidirectional, bidirectional outcome) ANOVA on the ratings of these three cues and found that the three cues were rated differently [$F(2,52) = 11.1$, $MS_e = 718$, $p < .001$, $\eta_p^2 = .299$]. Moreover, the ratings of the three cues were more negative, on average, in Group Unidirectional than in Group Bidirectional [$F(1,26) = 6.48$, $MS_e = 329$, $p = .017$, $\eta_p^2 = .200$]. There was no interaction between the two factors ($F < 1$), indicating that the pattern of results was similar in the two groups. Post hoc contrasts revealed that the ratings given to Cue F were reliably more negative than the averaged ratings given to Cues B and D [$F(1,26) = 17.9$, $MS_e = 1,312$, $p < .001$, $\eta_p^2 = .408$], and that the ratings given to Cues B and D did not reliably differ ($F < 1$). As in the main analysis, neither of these contrasts reliably interacted with the group factor (maximum $F < 1$). The individual ratings given to Cues B, D, and F suggest that, contrary to the findings of Melchers et al. (2006), there was a similar extinction of inhibition in the two groups.

We observe a similar pattern in the summation tests that appear on the right-hand side of Figure 1. Although the compound NF received lower ratings than did ND, demonstrating a successful summation test for the unextinguished inhibitor F, the difference between NB and ND is much smaller, suggesting that B had lost much of its con-

Figure 1. Average ratings given to Cues B, D, F, and H, and their respective summation tests in Experiment 1. The average ratings given in Group Unidirectional are shown in the upper panel; those given in Group Bidirectional are shown in the lower panel. Error bars represent the standard errors of the means.

ditioned inhibition after it received extinction (B^0) trials in Phase 2. A 3 (compound: NB, ND, NF) $\times$ 2 (group: unidirectional, bidirectional outcome) ANOVA showed that the three compounds were rated differently [$F(2,54) = 7.17$, $MS_e = 652$, $p = .002$, $\eta_p^2 = .216$]. The main effect of group was not reliable, nor was the interaction (maximum $F < 1$). As was the case with the individual ratings, F was more inhibitory than B and D, since the ratings for NF were reliably lower than the averaged ratings for NB and ND [$F(1,26) = 10.6$, $MS_e = 1,217$, $p = .003$, $\eta_p^2 = .289$]. Cue B seemed to be completely extinguished, since the ratings attributed to NB and ND were not significantly different [$F(1,26) = 1.61$, $MS_e = 986$, $p = .215$, $\eta_p^2 =$

.058]. The group factor did not interact with either of these two contrasts [maximum $F(1,26) = 2.32$, $MS_e = 986$, $p = .140$, $\eta_p^2 = .082$]. Thus, we found evidence for the extinction of inhibition in both the individual cue ratings as well as in the summation tests, so this effect was similar in the two groups.

Given that an extinction design always involves at least two phases (an acquisition and an extinction phase), we wondered whether some of the conditioned inhibition acquired in Phase 1 would be lost during Phase 2. This might have occurred due to the passage of time, or to the interference caused by the trials observed during Phase 2, or perhaps to some generalized extinction, since Cue B,

Table 2
Results of Experiment 1

Tested Cues	Group Unidirectional		Group Bidirectional	
	<i>M</i>	<i>SE</i>	<i>M</i>	<i>SE</i>
A	78.14	6.82	80.57	6.48
B	-10.14	7.21	0.00	0.00
C	-1.29	1.29	6.00	6.00
D	-17.29	9.08	-0.93	0.93
E	72.86	6.95	87.29	5.51
F	-49.00	9.32	-23.14	11.13
G	81.07	6.65	83.93	4.96
H	-48.43	12.44	-43.21	11.79
I	26.21	15.91	56.50	6.88
J	30.57	16.47	48.86	7.88
K	-7.14	7.14	-88.79	4.57
L	2.14	8.55	-59.57	9.23
M	-3.57	6.12	-58.86	9.71
N	82.64	6.63	90.86	3.83
AB	26.71	10.30	26.43	9.43
NB	27.07	9.58	46.00	10.78
ND	43.64	11.52	44.50	8.55
NF	12.79	7.57	25.00	7.97
NH	-5.14	5.77	3.93	3.15
NK	34.29	11.14	-1.93	1.71
NL	44.57	10.62	12.71	9.78
NM	31.14	10.51	-6.21	7.70

the extinguished inhibitor, had received training similar to Cue F in Phase 1 and was extinguished in Phase 2. Nevertheless, any evidence showing loss of inhibition does not undermine our finding that extinction of inhibition occurred, since Cues B and F were both trained in Phase 1 and the passage of time or the interference caused by the trials seen in Phase 2 presumably had an equal effect on the two cues. To test whether some loss of inhibition occurred, we compared Cues F and H, the two conditioned inhibitors trained in Phase 1 and Phase 2, respectively.

A 2 (cue: F, H) $\times$ 2 (group: unidirectional, bidirectional outcome) ANOVA revealed no reliable main effect of cue or group and no interaction [maximum $F(1,26) = 2.45$, $MS_e = 608$, $p = .129$, $\eta_p^2 = .086$]. Thus, the difference observed in the ratings of F and H in Group Bidirectional was not large enough to be reliable. However, we did find strong loss of inhibition in the summation tests. A 2 (compound: NF, NH) $\times$ 2 (group: unidirectional, bidirectional outcome) ANOVA showed that the NF ratings were reliably more positive than the NH ratings [$F(1,26) = 10.6$, $MS_e = 501$, $p = .003$, $\eta_p^2 = .290$]. There was no main effect of group and no interaction [maximum $F(1,26) = 2.45$, $MS_e = 647$, $p = .130$, $\eta_p^2 = .086$]. Hence, although the individual ratings showed no significant loss of inhibition, the summation tests did. This is an interesting result, since it suggests that the summation tests might be more sensitive to a decrease in the inhibitory properties of a cue, perhaps because they involve a transfer of the acquired inhibitory potential to a new exciter.

The ratings of the other cues (shown in Table 2) followed an expected pattern. All the cues that were individually paired with a hormone level increase (A⁺, E⁺, G⁺, and N⁺) were given strong positive ratings, whereas Cues I and J (IJ⁺) were given weaker positive ratings, demon-

strating overshadowing (i.e., lower responding to each of two cues when they predict the presence of an outcome together, compared with responding to a cue that has been singly paired with the outcome). Indeed, a comparison of Cue N and the average of Cues I and J revealed a significant overshadowing effect [$F(1,26) = 43.8$, $MS_e = 1,365$, $p < .001$, $\eta_p^2 = .627$].

Cues K, L, and M were given ratings close to zero in Group Unidirectional, in which they were followed by no hormone level change (K⁰, LM⁰). These cues were given strong negative ratings in Group Bidirectional, in which they were paired with a hormone level decrease (K⁻, LM⁻). This was confirmed by a reliable difference between the ratings that the two groups attributed to Cues K, L, and M [$F(1,26) = 115$, $MS_e = 268$, $p < .001$, $\eta_p^2 = .815$]. This is also visible in the summation tests. In Group Unidirectional, the NK, NL, and NM test trials were similar to the control summation test ND [$F(1,13) = 1.14$, $MS_e = 600$, $p = .306$, $\eta_p^2 = .080$], suggesting that cues K, L, and M were not more inhibitory than the neutral control cue, D. In Group Bidirectional, on the other hand, Cues K, L, and M passed the summation test for inhibition, compared with ND [$F(1,13) = 19.9$, $MS_e = 1,300$, $p = .001$, $\eta_p^2 = .605$]. In Group Bidirectional, Cues L and M overshadowed each other, since they predicted the hormone level decrease together. Their ratings were less negative than those of K, a cue that predicted the hormone level decrease by itself [$F(1,13) = 26.3$, $MS_e = 465$, $p < .001$, $\eta_p^2 = .669$]. However, this overshadowing effect appears only in the individual ratings and not in the summation tests, since the summation tests NK, NL and NM did not reliably differ [$F(2,46) = 1.71$, $MS_e = 807$, $p = .201$, $\eta_p^2 = .116$].

In conclusion, the results of Experiment 1 demonstrate extinction of conditioned inhibition, as measured by a direct and a summation test. These results are, however, inconsistent with those of Melchers et al. (2006), since we found reliable extinction of inhibition, whether our participants observed negative outcomes or not. We will discuss the possible reasons for this discrepancy in the General Discussion. Nevertheless, our results support the Rescorla–Wagner model and are inconsistent with Melchers et al.'s hypothesis that extinction of inhibition should be unlikely to occur unless the participants observe negative outcomes. The results also do not support the inferential account of learning.

In Experiment 2, we extended this analysis using a blocking of inhibition paradigm. As we will explain shortly, applying the same manipulation as in Experiment 1, in which half of the participants observed negative outcomes, changes the predictions of the inferential account, but not those of Melchers et al.'s (2006) hypothesis or those of the original Rescorla–Wagner model; so this is another way of contrasting the predictions of these accounts.

EXPERIMENT 2

Blocking, or cue competition, has been studied most frequently in excitatory conditioning (e.g., Kamin, 1969),

but has been investigated in conditioned inhibition as well (e.g., Suiter & LoLordo, 1971). In a conventional excitatory blocking experiment, a single cue (A) is reinforced alone (A^+) but is also reinforced in compound with the target cue (B—i.e., AB^+). Compared with a group that only has had experience with the compound conditioning trials, the extra training with A blocks or reduces conditioning to B, presumably because A predicts all of the outcomes, making B redundant. A^+ trials are typically presented before AB^+ trials, although the two types of trials might be intermixed (Wagner, 1969). We used an analogous procedure to study blocking of conditioned inhibition, in which the target cue (here, C) signaled only some of the trials on which the outcome was omitted, but another cue (B) predicted all of them. We used an A^+ , AB^0 , ABC^0 design, in which Cue A signaled a hormone level increase when presented alone, but not in the presence of B, or B and C. The Rescorla–Wagner model predicts that B should acquire inhibitory associative strength, but that C should not, because B already predicts the absence of the outcome. Thus, the more valid inhibitor B should block C. Again, the Rescorla–Wagner model predicts this, whether or not participants have specific experience with negative outcomes. Melchers et al.'s (2006) hypothesis, which focuses on whether negative expectancies can be confirmed, also makes this prediction, regardless of the possibility of negative outcomes. This is true because, as was not the case with Experiment 1, the outcome expectation created by the cues presented on a given trial never takes on negative values, because the inhibitors are never presented in the absence of an exciter, as they are in extinction experiments. Hence, in this design, the Rescorla–Wagner model never generates a negative expectation of the outcome, and exposing participants to negative outcomes should not affect the magnitude of blocking.

The inferential account, on the other hand, predicts that experience with negative outcomes should increase the amount of blocking. According to this account, participants should judge B as an inhibitor, given that it cancels the effect of A on AB^0 trials. On ABC^0 trials, however, if there is no evidence that the outcome could be lower than zero, it is uncertain whether C has an inhibitory effect that is masked by the floor in the outcome level, or whether C truly has no effect on the outcome. If, on the other hand, negative outcomes were possible, it would be certain that, on ABC^0 trials, C would have no effect beyond that of B on AB^0 trials. Thus, participants should be more likely to judge C as neutral in the latter case than in the former. This kind of reasoning parallels the one proposed for blocking of excitatory cues when there is a ceiling in the outcome level (Beckers, De Houwer, Pineño, & Miller, 2005; De Houwer, Beckers, & Glautier, 2002; Vandorpe, De Houwer, & Beckers, 2005). In an A^+ , AB^+ blocking design, learning about Cue B should be blocked more if it were known that stronger outcomes could be possible, but B did not increase the outcome level beyond the increase caused by Cue A on A^+ trials.

Table 3 summarizes the trial types in Experiment 2. In addition to the A^+ , AB^0 , and ABC^0 trials, we included

Table 3
Design of Experiment 2

Training	Test
A^+ , AB^0 , ABC^0 , ADE^0	A, B, C, D, E, NB, NC, ND, NE
AF^+ , AFG^+	F, G
H^0 , HI^0 , HIJ^0	H, I, J, NH, NI, NJ
$K^{0/-}$, $LM^{0/-}$	K, L, M, NK, NL, NM
N^+	N
OP^+ , OPQ^+	O, P, Q

Note— $+$, hormone level increase; 0 , no change in hormone level; $-$, hormone level decrease. In Group Unidirectional, Cues K and LM were followed by no hormone level change (0), and in Group Bidirectional, these cues were followed by a hormone level decrease ($-$).

ADE^0 trials to demonstrate overshadowing in conditioned inhibition. D and E cancel the hormone level increase caused by A together, so the Rescorla–Wagner model predicts that each should become inhibitory, but less so than B. This is a form of *overshadowing* and has often been reported for excitatory cues (as in our Experiment 1). Overshadowing is the proper control procedure for excitatory blocking (Kamin, 1969), so inhibitory overshadowing is an important comparison treatment for inhibitory blocking. We would expect D and E to be less inhibitory than B through overshadowing, and C to be less inhibitory than D and E through blocking.

We used the same excitatory cue (A) to train blocked Cue C and overshadowed Cues D and E, for several reasons. The simplest reason is that it reduces the total number of cues and trials. Moreover, because all the critical cues are paired with the same exciter, this design has the advantage that the target inhibitors C, D, and E are paired with a cue that must support the same amount of inhibition. Had we paired different exciters with C and with D and E, a difference between C and D and E could be attributed to potential differences between the strengths of the exciters with which they were paired. Using a single excitatory cue also has the advantage that it shows that different cues can act differently on the same cue in future summation tests. Moreover, we included AF^+ and AFG^+ trials in order to prevent participants from learning that Cue A was always followed by no hormone level change when presented in a compound.

A further cue, J, had training similar to our critical blocked cue (C), except that it should have remained more neutral because it was never paired with an exciter (i.e., H^0 , HI^0 , HIJ^0). We used J as a control cue to test for conditioned inhibition. J might become slightly inhibitory, because it is paired with a null outcome and positive outcomes sometimes occur in its absence (*differential inhibition*). If so, comparing the target inhibitors with J is a more conservative test than comparing the target inhibitors with a neutral cue, but this would not change the conclusions we might draw from our tests concerning blocking and overshadowing of inhibition. As in Experiment 1, Cues K, L, and M were followed by no hormone level change in Group Unidirectional (K^0 , LM^0) and by a hormone level decrease in Group Bidirectional (K^- , LM^-). Cue N was followed by a hormone level increase (N^+) and was used in the summation tests. OP^+ and OPQ^+ trials

were included in order to reduce the high percentage of compounds followed by no hormone level change. There was only one training phase in this experiment, and all trial types were intermixed. Lotz, Vervliet, and Lachnit (2009) have recently reported finding blocking of inhibition using a bidirectional design similar to ours. They did not directly compare bidirectional and unidirectional procedures, so little could be concluded about the different accounts of blocking. Moreover, they did not compare the blocked cue with a control cue; rather, they compared it with a value of zero. That is, they simply checked whether the rating was negative. As can be seen from the results of Experiment 1, the presence and absence of negative outcomes can bias the participants' use of rating scales, so comparisons against a fixed value are not ideal.

Method

Participants. Seventeen McGill University students were assigned to Group Unidirectional (12 women and 5 men), and another 17 were assigned to Group Bidirectional (8 women and 9 men). As in Experiment 1, cue assignment determined the number of participants. They participated in exchange for credit in a psychology course. We used a learning criterion, according to which participants had to correctly solve the discrimination involving A (A^+ , AB^0 , ABC^0 , ADE^0 , AF^+ , AFG^+) in the last block of trials of the training phase in order to be included in the analyses. Five participants in Group Unidirectional (3 women and 2 men) and 2 in Group Bidirectional (1 woman and 1 man) failed to reach this criterion and were replaced by 7 female participants.

Apparatus and Stimuli. The experiment was programmed in RealBasic and was run on Apple Mac computers with color monitors. The following 17 foods were used for the 17 cues: corn, peppers, pears, basil, bran, beef, bananas, potatoes, blueberries, carrots, cabbage, broccoli, chicken, grapes, tomatoes, green beans, and oranges. Each food name was accompanied by a picture of that food. Seventeen different food-cue assignments were generated, and these were used in both groups.

Procedure. The procedure was the same as the one used in Experiment 1, except that the trial types shown in Table 3 were presented. The training phase consisted of eight blocks of trials. Each block consisted of one of each trial type, shown in the table randomly intermixed. Thus, participants observed eight trials of each type.

Results

Figure 2 shows the average ratings given during the testing phase for the critical cues. The mean ratings for all the cues in Experiment 2 can be found in Table 4. The ratings of the blocking inhibitor (B), the target blocked inhibitor (C), and the control overshadowed inhibitors (D and E) follow the ordinal predictions generated by the Rescorla-Wagner model in both groups, although the pattern is less pronounced in Group Unidirectional. Nonetheless, the ratings of B are more negative, on average, than the ratings of D and E, demonstrating overshadowing. Furthermore, the ratings of C are less negative than those of D and E, demonstrating blocking. Hence, C, the inhibitor that was trained along with the established inhibitor B (A^+ , AB^0 , ABC^0), did not acquire as much inhibitory strength as D and E, the two inhibitors that were only trained together (ADE^0). Indeed, C was rated similarly to the control cue, J, suggesting that complete blocking occurred (i.e., Cue C did not acquire any conditioned inhibition).

The ratings of the two overshadowed cues, D and E, did not reliably differ either in Group Bidirectional ($F < 1$) or in Group Unidirectional [$F(1,16) = 3.00$, $MS_e = 831$, $p = .103$, $\eta_p^2 = .158$], although the difference between the ratings in Group Unidirectional was relatively large (-26.82 and -43.94 for D and E, respectively). This nonreliable difference between D and E in Group Unidirectional, however, was probably spurious, because the summation tests with D and E barely differed (23.94 and 18.00 for ND and NE, respectively; $F < 1$). Therefore, we calculated the average score of D and E for every participant and compared it with the ratings of the blocked (C) and blocking (B) cues to determine whether the blocking and overshadowing effects were significant. These averages are shown in Figure 2. We analyzed the ratings of the blocked cue C, the blocking cue B, the average of the overshadowed cues D and E, and the control cue J in a 4 (cue) $\times$ 2 (group: unidirectional, bidirectional outcome) ANOVA. The main effect of cue was significant [$F(3,96) = 11.9$, $MS_e = 954$, $p < .001$, $\eta_p^2 = .271$], but neither the main effect of group nor the interaction reached statistical significance [maximum $F(1,32) = 2.14$, $MS_e = 411$, $p = .153$, $\eta_p^2 = .063$]. Ratings of Cue C were significantly less negative than those of Cues D and E [$F(1,32) = 7.50$, $MS_e = 1,108$, $p = .010$, $\eta_p^2 = .190$], demonstrating a reliable blocking effect. As we mentioned above, Figure 2 shows that the ratings of Cue C are very similar to those of Cue J, the control cue, which suggests that Cue C was completely blocked. The ratings of the two cues did not differ reliably ($F < 1$). The overshadowing effect was also reliable, since B was given significantly more negative ratings than was the overshadowed cues, D and E [$F(1,32) = 7.33$, $MS_e = 2,119$, $p = .011$, $\eta_p^2 = .186$]. Neither of these contrasts interacted reliably with the group factor [maximum $F(1,32) = 1.54$, $MS_e = 2,119$, $p = .223$, $\eta_p^2 = .046$], indicating that the effects were similar in the two groups.

The summation tests, in which the various inhibitory cues are paired with the novel exciter, N, support these conclusions as well. The means of these compounds are shown in the right-hand four bars of Figure 2. B is the most inhibitory cue, since ratings for the NB compound are lowest, followed by ND and NE, confirming overshadowing. C is the weakest inhibitor. Moreover, as was the case for the individual ratings, the summation test for C was very similar to that for the control cue, J, indicating nearly complete blocking. As we did for the individual ratings, we analyzed the average of the summation tests for the overshadowed cues (ND and NE) since they did not reliably differ in either group (both F s < 1). We analyzed the ratings of NJ (the control cue summation test), NB, NC, and the average of ND and NE in a 4 (compound) $\times$ 2 (group: unidirectional, bidirectional outcome) ANOVA. The compounds were rated differently [$F(3,96) = 13.8$, $MS_e = 624$, $p < .001$, $\eta_p^2 = .301$]. The main effect of group and the interaction were not reliable [maximum $F(1,32) = 2.58$, $MS_e = 382$, $p = .118$, $\eta_p^2 = .075$]. There was a reliable blocking effect, since the ratings of the compound NC were more positive than the ratings of ND and NE [$F(1,32) = 10.9$, $MS_e = 871$, $p = .002$, $\eta_p^2 = .254$].

Figure 2. Average ratings given to cues B, C, D, E, and J, and their respective summation tests in Experiment 2. The figure shows the average score of the two overshadowed cues, D and E. The same is true for their summation tests, ND and NE. The average ratings given in Group Unidirectional are shown in the upper panel; those given in Group Bidirectional are shown in the lower panel. Error bars represent the standard errors of the means.

As can be seen in Figure 2, the summation tests show that Cue C was completely blocked, because the ratings for the compound NC did not reliably differ from those of NJ; J was the control cue [$F(1,32) = 2.52$, $MS_e = 1,672$, $p = .122$, $\eta_p^2 = .073$]. Neither of these comparisons interacted reliably with the group factor (maximum $F < 1$). Complete blocking of Cue C in the summation tests in both groups parallels our findings in the individual ratings of the cues. A significant overshadowing effect was demonstrated by lower ratings for the compound NB than for the overshadowed cues [$F(1,32) = 12.4$, $MS_e = 972$, $p = .001$, $\eta_p^2 = .279$]. Although the size of the overshadowing

effect is somewhat larger in Group Bidirectional than in Group Unidirectional, this interaction did not quite reach significance [$F(1,32) = 3.04$, $MS_e = 972$, $p = .091$, $\eta_p^2 = .087$]. Nevertheless, because this interaction almost reached the significance level, we compared the average of ND and NE with NB in each group. The overshadowing effect was significant in Group Bidirectional [$F(1,16) = 11.9$, $MS_e = 1,128$, $p = .003$, $\eta_p^2 = .427$], but it was not in Group Unidirectional [$F(1,16) = 1.88$, $MS_e = 816$, $p = .189$, $\eta_p^2 = .105$].

One might argue, however, that the blocking effect (i.e., lower ratings for D and E vs. C) might have been

Table 4
Results of Experiment 2

Tested Cues	Group Unidirectional		Group Bidirectional	
	<i>M</i>	<i>SE</i>	<i>M</i>	<i>SE</i>
A	90.24	3.68	95.82	2.55
B	-46.94	10.34	-53.35	10.97
C	-20.12	8.54	-6.18	6.90
D	-26.82	8.41	-23.24	8.05
E	-43.94	9.95	-21.12	7.32
F	66.59	7.98	58.35	10.38
G	25.76	9.31	32.41	12.24
H	-20.47	9.53	0.00	0.00
I	-8.76	6.21	-2.53	2.53
J	-20.29	8.63	-0.29	4.32
K	-17.76	9.06	-96.18	1.86
L	-5.35	9.77	-69.82	8.25
M	-24.18	9.00	-75.41	5.60
N	87.00	4.48	96.88	1.90
O	54.18	9.72	79.47	5.51
P	71.53	8.31	71.47	7.88
Q	15.12	11.22	49.59	9.80
NB	11.47	4.73	7.94	4.31
NC	35.47	8.20	55.00	8.88
ND	23.94	8.15	32.41	7.97
NE	18.00	10.06	39.76	7.80
NH	43.76	10.75	51.29	8.31
NI	36.47	9.62	54.65	7.41
NJ	34.82	7.65	46.76	8.81
NK	36.12	8.62	0.47	0.47
NL	43.59	8.99	0.12	0.08
NM	30.41	7.21	14.47	6.43

affected by the outcome manipulation, despite the lack of an interaction with the group factor, the interaction not being sensitive enough to pick up subtle differences between the two groups. In that case, it would be appropriate to test for blocking in each group. In Group Unidirectional, the blocking effect was reliable in the single-cue tests [$F(1,16) = 4.48$, $MS_e = 442$, $p = .050$, $\eta_p^2 = .219$] but did not quite reach the significance level in the summation tests [$F(1,16) = 3.12$, $MS_e = 573$, $p = .096$, $\eta_p^2 = .163$]. Conversely, in Group Bidirectional, blocking was reliable in the summation tests [$F(1,16) = 10.2$, $MS_e = 298$, $p = .006$, $\eta_p^2 = .389$], but not quite in the single-cue tests [$F(1,16) = 3.27$, $MS_e = 665$, $p = .089$, $\eta_p^2 = .170$]. Considering the single-cue tests and the summation tests together, however, the blocking effect was reliable in both groups [$F(1,16) = 9.84$, $MS_e = 383$, $p = .006$, $\eta_p^2 = .381$, and $F(1,16) = 6.74$, $MS_e = 769$, $p = .019$, $\eta_p^2 = .296$, for Groups Unidirectional and Bidirectional, respectively]. This blocking effect did not reliably interact with the test factor in either group (both $F_s < 1$). It should also be noted that the effect size was 29% larger in Group Unidirectional than in Group Bidirectional in the combined tests for blocking. Thus, over the two tests, there is little statistical support for the claim that blocking is weaker in Group Unidirectional. The fact that the size of this effect was similar in the two groups supports the Rescorla–Wagner model over the inferential account.

The ratings of the other cues, reported in Table 4, followed expected patterns. The ratings of cues F, G, O,

P, and Q demonstrate blocking and overshadowing of excitatory cues, as would be expected from a number of accounts, including the Rescorla–Wagner model. Experience with A⁺, AF⁺ and AFG⁺ trials resulted in strong ratings for A, weaker ratings for F, which was blocked by A [$F(1,32) = 18.6$, $MS_e = 1,703$, $p < .001$, $\eta_p^2 = .368$], and even weaker ratings for G, which was blocked by both A and F [$F(1,32) = 19.5$, $MS_e = 1,944$, $p < .001$, $\eta_p^2 = .379$]. Similarly, training with OP⁺ and OPQ⁺ trials resulted in overshadowing of O and P compared with N, a cue that predicted a hormone level increase on its own (N⁺) [$F(1,32) = 213$, $MS_e = 2,255$, $p < .001$, $\eta_p^2 = .870$]. This overshadowing effect, however, is much stronger in Group Unidirectional than in Group Bidirectional [$F(1,32) = 15.4$, $MS_e = 2,255$, $p < .001$, $\eta_p^2 = .324$]. The two overshadowed cues, O and P, successfully blocked Q [$F(1,32) = 34.8$, $MS_e = 1,323$, $p < .001$, $\eta_p^2 = .521$], since O and P were sufficient to predict the hormone level increase, and Q was a redundant predictor.

Finally, Cues K, L, and M were given significantly more negative ratings in Group Bidirectional, in which they predicted a hormone level decrease, than in Group Unidirectional, in which they predicted no change in the hormone level [$F(1,32) = 52.4$, $MS_e = 680$, $p < .001$, $\eta_p^2 = .621$]. Similarly, in the summation tests, participants judged Cues K, L, and M to be inhibitory, compared with the control cue, J, in Group Bidirectional [$F(1,32) = 21.7$, $MS_e = 1,364$, $p < .001$, $\eta_p^2 = .576$], but not in Group Unidirectional ($F < 1$). Once again, the single-cue causal ratings in Group Bidirectional revealed an overshadowing effect on the cues that predicted a hormone level decrease. Cues L and M predicted a hormone level decrease together (LM⁻) and thus overshadowed each other, compared with K, which predicted a hormone level decrease on its own (K⁻) [$F(1,32) = 13.1$, $MS_e = 720$, $p = .002$, $\eta_p^2 = .450$].

It seems that participants in the two groups mapped their judgments onto the rating scale differently, depending on whether they observed negative outcomes. Those in Group Bidirectional used both extreme ends of the scale, since they attributed strong positive ratings to cues such as A and N and strong negative ratings to Cues K, L, and M. Accordingly, their judgments of the cues that did not receive conditioned inhibition training—H, I, and J—are centered around zero. Participants in Group Unidirectional, on the other hand, used the positive end of the scale much more than the negative end, because none of the cues in this group was paired with a hormone level decrease during training. Nonetheless, they stretched their ratings toward the unused negative extremity of the scale, such that even H, I, and J were given slightly negative ratings. Indeed, Cues H, I, and J were given reliably more negative ratings in Group Unidirectional than in Group Bidirectional [$F(1,32) = 5.83$, $MS_e = 353$, $p = .022$, $\eta_p^2 = .154$]. A similar, but smaller, effect had occurred in Experiment 1. Indeed, in Experiment 1, Cues C and D received more negative ratings in Group Unidirectional than in Group Bidirectional [$F(1,26) = 4.72$, $MS_e = 414$, $p = .039$, $\eta_p^2 = .154$].

GENERAL DISCUSSION

Experiment 1 demonstrated extinction of inhibition, regardless of whether or not participants had observed negative outcomes, and this supports the Rescorla–Wagner model, which does not differentiate between outcome expectations that could be confirmed in reality and those that could not. Because of this, it predicts extinction of inhibition with bidirectional and unidirectional outcomes. Our participants behaved in a similar way. This finding is inconsistent both with Melchers et al.'s (2006) hypothesis (that experience with negative outcomes is necessary for extinction of inhibition to occur) and with the inferential account of learning, both of which predicted stronger extinction of inhibition when negative outcomes were observed. One possible reason for the discrepancy between our results and those of Lotz and Lachnit (2009) and Melchers et al. is the different instructions given to the participants. Perhaps the reduced extinction of inhibition observed by Lotz and Lachnit and by Melchers et al. in Group Unidirectional depended on the participants' perceiving conditioned inhibitors as cues that prevented the hormone level increase caused by another cue and, thus, had little effect on their own. In that case, following the inhibitor by no change in outcome level during extinction should not affect it, since it would need to be paired with an exciter in order for its inhibitory effects to be visible. Participants in Group Bidirectional in Lotz and Lachnit's and in Melchers et al.'s experiments, on the other hand, were encouraged to perceive an inhibitor as a cue that would directly decrease the hormone level. Extinction trials would be inconsistent with this view, since the inhibitor should be followed by an outcome level decrease when presented by itself, and not merely by no change in outcome level. This might have encouraged participants in this group to devalue the inhibitory properties of the extinguished cue. However, their claim was that the effect was due to the differential outcome levels, and not to the inferences encouraged by their instructions. In this experiment, we did not confirm that differential outcome levels influence extinction of inhibition.

An alternative explanation lies in the complexity of the two designs. Melchers et al. (2006) used a total of 7 cues and 7 trial types, and Lotz and Lachnit (2009) used 9 cues and 10 trial types, whereas we used 14 cues and 14 trial types in Experiment 1. Perhaps participants are able to make causal inferences in simpler designs that involve fewer cues, and more complex designs limit participants to using associative-like mechanisms to learn these more complex structures (LePelley, Oakeshott, & McLaren, 2005). This might explain why we found strong evidence for the Rescorla–Wagner model with our more complex design. Our participants were faced with a more complex learning problem and were not given specific instructions describing the possible effects of the cues, and they seemed to use learning processes consistent with associative mechanisms. This suggests that, if there is an inferential reasoning process, it might occur only with designs

involving relatively few cues or a simple structure and with the appropriate instructions. Less simple learning problems, on the other hand, seem to recruit associative-like mechanisms.

Melchers et al. (2006) were the first to test the predictions of associative and inferential accounts using an extinction of inhibition design. In Experiment 2, we used a blocking of conditioned inhibition design to test the predictions of the Rescorla–Wagner model and those of the inferential account. Here again, our results favored the Rescorla–Wagner model. This experiment is interesting because, to our knowledge, the associative and inferential accounts of blocking of inhibition have not been tested before. There are reports of blocking of inhibition, however, in studies that did not manipulate outcome values. Reduction of inhibition has been found by Lotz et al. (2009) and in a relative validity design in human learning with probabilistic cues (e.g., Baker et al., 1993). Furthermore, blocking of inhibition was found in at least one experiment involving animal conditioning (Suiter & LoLordo, 1971), but others failed to find the effect (Moore & Stickney, 1985; Schachtman, Matzel, & Miller, 1988).

Taken together, our two experiments showed little influence of exposure to negative outcomes on extinction and blocking of inhibition. Nevertheless, one might argue that our statistical analyses were not sensitive enough to differences in the size of the effects in the two groups. Indeed, although the relevant interactions did not reach the significance level, the size of the effects slightly differed in the two groups. These differences, nonetheless, were not always in the same direction in the single-cue and summation tests, and it is hard to interpret these small differences, given the judgments' variability. Moreover, the rating scale might have been used differently in the two groups, which makes direct comparisons of the size of the effects unsafe. However, the fact that both the extinguished inhibitor in Experiment 1 and the blocked inhibitor in Experiment 2 were judged to be similar to their respective control cues in both groups is evidence that the extinction and blocking effects were near-asymptote in both groups. These comparisons with control cues are important, given the fact that participants might have used the scale differently in the two groups. Because participants in Group Unidirectional did not use the extreme negative end of the scale, as did the ones in Group Bidirectional for Cues K, L, and M (these cues were paired with a hormone level decrease; K^- , LM^-), they might have stretched some of their ratings toward the negative end. There is some evidence for this in both experiments, because the control cues and some of the other cues were given ratings that were more negative in Group Unidirectional than in Group Bidirectional. If participants used the rating scale differently in the two groups, presumably, they did so for all cues. Hence, the comparisons with control cues that showed complete extinction and blocking of inhibition in both groups are the best evidence that the size of the effects was similar in the two groups.

Although the Rescorla–Wagner model predicts our results fairly well, there is at least one aspect of the data that is inconsistent with the model. In both experiments, we observed a drastic reduction in the ratings when the exciter, N, was paired in the summation tests with the control cues that did not receive conditioned inhibition training. For example, the mean rating of the exciter, N, in Group Unidirectional of Experiment 2 was 87, but the compound of it and Cue J was rated as 35. These cues might have acquired some inhibition simply by virtue of being paired with no hormone level increase, when the mean effect of all other cues was a slight increase in hormone level. However, the fact that participants estimated that these cues had little effect on the hormone level when presented alone is inconsistent with the strong effect they had on the exciter, N. There are at least two reasons why a truly neutral cue might decrease the response associated with an exciter. The first is generalization decrement, whereby presenting an exciter in compound with another cue is perceived differently from presenting the exciter alone (Pearce, 1987). To the extent that the two presentations are similar, the response associated with the exciter will generalize to the compound. The more dissimilar the two presentations are perceived as being, the weaker will be the generalized response and, hence, the greater the generalization decrement. The second possible mechanism is Pavlov's (1927) external inhibition. According to this explanation, even a neutral cue can decrease an exciter's conditioned response, if the novelty of the compound is surprising, and thus disrupt the response generated by the exciter.

Although we have focused on the Rescorla–Wagner model, it is worth noting that other associative models predict extinction and blocking of inhibition. Other models that implement a learning rule similar to the Rescorla–Wagner model (generally called the *delta rule*), and that, like the original Rescorla–Wagner model, do not formalize a mechanism for the effect of perceived possible outcomes on learning, make similar predictions. One such model is the *auto-associator* (Baetu & Baker, 2009), a connectionist network that allows all stimuli presented to become associated. This class of model was described by McClelland and Rumelhart (1988); we used a version of it to simulate learning phenomena beyond the scope of the Rescorla–Wagner model (Baetu & Baker, 2009). Pearce's (1987) configural model also predicts extinction and blocking of inhibition. This model has the advantage of also predicting generalization decrement on the summation tests, as mentioned above. Although the modified SOP model (Aitken & Dickinson, 2005) does not directly use the delta learning rule, it does compute a contrast between the actual and expected outcomes on every trial. Because of this, cues that are redundant predictors of the outcome are learned about less. Thus, the model can account for the results of Experiment 2, since it predicts blocking and overshadowing of inhibition. It does not, however, predict extinction of inhibition, and this is inconsistent with the results of Experiment 1, but consistent with numerous animal experiments that did not

find extinction of inhibition and that were mentioned at the beginning of the article.

According to our simulations, Stout and Miller's (2007) formalization of the extended comparator hypothesis (SOCR) predicts blocking of inhibition with the design we used in Experiment 2 (A^+ , AB^0 , ABC^0 , ADE^0), although its predictions are unstable if other, seemingly equivalent, designs are used (such as training the blocked and the overshadowed inhibitors separately) or if contextual cues are included in the simulations. The model predicts extinction of inhibition, which is consistent with the results of Experiment 1. It is worth noting that, in our extinction design, the exciter that was used to train the target inhibitor was further presented with the outcome during the extinction phase (i.e., participants observed A^+ and AB^0 trials in Phase 1 and A^+ and B^0 trials in Phase 2). A similar design was used by Melchers et al. (2006) and Lotz and Lachnit (2009). A^+ trials in Phase 2 may influence extinction of B. Our simulations indicate, however, that SOCR predicts complete extinction of B, whether or not A^+ trials are administered during the extinction phase.

The discussion of the two different descriptions of the inhibitory action of a cue used by Lotz and Lachnit (2009) and Melchers et al. (2006) raises some important theoretical issues. Although an excitatory cue is typically assumed to act directly on the outcome, an inhibitory cue might act in at least three ways to prevent the outcome from happening. The first is by exerting an inhibitory force directly on the outcome, as the Rescorla–Wagner model assumes. The other two ways involve modulating the action of an exciter instead of directly influencing the outcome: by raising the threshold of activation of the outcome, thereby making it more difficult for any exciter to activate the outcome (Zimmer-Hart & Rescorla, 1974); and by inhibiting the association between an exciter and the outcome. This type of hierarchical association that acts on another association has been proposed to account for occasion setting (Ross & Holland, 1981). These three conceptualizations of conditioned inhibition entail different potential consequences of presenting the inhibitor alone, or in the presence of exciters other than the one it was trained with. Given that humans and animals can behave as if they had each of these three representations of inhibitory cues in different situations, the question is not which representation is more valid but, rather, under what circumstances they are more likely to be used.

To conclude, our two experiments have shown that, contrary to inferential accounts, exposure to negative outcomes is not sufficient to enhance extinction and blocking of inhibition. We have also shown that people can selectively learn about inhibitory cues, just as they selectively learn about excitatory cues. Numerous studies have demonstrated that animals and humans tend to learn only about excitatory cues that convey unique information about the occurrence of an outcome, and that redundant cues are ignored (Dickinson et al., 1984; Kamin, 1969). In Experiment 2, our participants learned less about a redundant inhibitor, and this blocking of inhibition parallels the well-known blocking of excitation effect in humans.

AUTHOR NOTE

This research was supported by a Discovery Grant awarded to A.G.B. and a Postgraduate Fellowship awarded to I.B. by the Natural Sciences and Engineering Research Council of Canada. Correspondence concerning this article should be addressed to I. Baetu, Department of Psychology, McGill University, 1205 Dr. Penfield Avenue, Montréal, QC, H3A 1B1 Canada (e-mail: irina.baetu@mail.mcgill.ca).

REFERENCES

- AITKEN, M. R., & DICKINSON, A. (2005). Simulations of a modified SOP model applied to retrospective revaluation of human causal learning. *Learning & Behavior*, **33**, 147-159.
- BAETU, I., & BAKER, A. G. (2009). Human judgments of positive and negative causal chains. *Journal of Experimental Psychology: Animal Behavior Processes*, **35**, 153-168.
- BAKER, A. G., MERCIER, P., VALLÉE-TOURANGEAU, F., FRANK, R., & PAN, M. (1993). Selective associations and causality judgments: Presence of a strong causal factor may reduce judgments of a weaker one. *Journal of Experimental Psychology: Learning, Memory, & Cognition*, **19**, 414-432.
- BECKERS, T., DE HOUWER, J., PINEÑO, O., & MILLER, R. R. (2005). Outcome additivity and outcome maximality influence cue competition in human causal learning. *Journal of Experimental Psychology: Learning, Memory, & Cognition*, **31**, 238-249.
- DE HOUWER, J., BECKERS, T., & GLAUTIER, S. (2002). Outcome and cue properties modulate blocking. *Quarterly Journal of Experimental Psychology*, **55A**, 965-985.
- 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.
- DETKE, M. J. (1991). Extinction of sequential conditioned inhibition. *Animal Learning & Behavior*, **19**, 345-354.
- DE VITO, P. L., & FOWLER, H. (1986). Effects of contingency violations on the extinction of a conditioned fear inhibitor and a conditioned fear excitator. *Journal of Experimental Psychology: Animal Behavior Processes*, **12**, 99-115.
- DE VITO, P. L., & FOWLER, H. (1987). Enhancement of conditioned inhibition via an extinction treatment. *Animal Learning & Behavior*, **15**, 448-454.
- DICKINSON, A., SHANKS, D., & EVENDEN, J. (1984). Judgement of act-outcome contingency: The role of selective attribution. *Quarterly Journal of Experimental Psychology*, **36A**, 29-50.
- HALLAM, S. C., GRAHAME, N. J., HARRIS, K., & MILLER, R. R. (1992). Enhanced negative summation following operational extinction of a Pavlovian conditioned inhibitor. *Learning & Motivation*, **23**, 43-62.
- HOLLAND, P. C. (1985). The nature of conditioned inhibition in serial and simultaneous feature negative discriminations. In R. R. Miller & N. E. Spear (Eds.), *Information processing in animals: Conditioned inhibition* (pp. 267-297). Hillsdale, NJ: Erlbaum.
- HOLLAND, P. C., & GORY, J. (1986). Extinction of inhibition after serial and simultaneous feature negative discrimination training. *Quarterly Journal of Experimental Psychology*, **38B**, 245-265.
- KAMIN, L. J. (1969). Selective association and conditioning. In W. K. Honig & N. J. Mackintosh (Eds.), *Fundamental issues in associative learning* (pp. 42-64). Halifax, NS: Dalhousie University Press.
- KONORSKI, J. (1948). *Conditioned reflexes and neuron organization*. New York: Cambridge University Press.
- LE PELLEY, M. E., OAKESHOTT, S. M., & McLAREN, I. P. L. (2005). Blocking and unblocking in human causal learning. *Journal of Experimental Psychology: Animal Behavior Processes*, **31**, 56-70.
- LOTZ, A., & LACHNIT, H. (2009). Extinction of conditioned inhibition: Effects of different outcome continua. *Learning & Behavior*, **37**, 85-94.
- LOTZ, A., VERVLIT, B., & LACHNIT, H. (2009). Blocking of conditioned inhibition in human causal learning: No learning about the absence of outcomes. *Experimental Psychology*, **56**, 381-385.
- LYSLE, D. T., & FOWLER, H. (1985). Inhibition as a "slave" process: Deactivation of conditioned inhibition through extinction of conditioned excitation. *Journal of Experimental Psychology: Animal Behavior Processes*, **11**, 71-94.
- MCCLELLAND, J. L., & RUMELHART, D. E. (1988). *Explorations in parallel distributed processing: A handbook of models, programs, and exercises*. Cambridge, MA: MIT Press.
- MELCHERS, K. G., WOLFF, S., & LACHNIT, H. (2006). Extinction of conditioned inhibition through nonreinforced presentation of the inhibitor. *Psychonomic Bulletin & Review*, **13**, 662-667.
- MILLER, R. R., BARNET, R. C., & GRAHAME, N. J. (1995). Assessment of the Rescorla-Wagner model. *Psychological Bulletin*, **117**, 363-386.
- MILLER, R. R., & SCHACHTMAN, T. R. (1985). Conditioning context as an associative baseline: Implications for response generation and the nature of conditioned inhibition. In R. R. Miller & N. E. Spear (Eds.), *Information processing in animals: Conditioned inhibition* (pp. 51-88). Hillsdale, NJ: Erlbaum.
- MOORE, J. W., & STICKNEY, K. J. (1985). Antiassociations: Conditioned inhibition in attentional-associative networks. In R. R. Miller & N. E. Spear (Eds.), *Information processing in animals: Conditioned inhibition* (pp. 209-232). Hillsdale, NJ: Erlbaum.
- OWREN, M. J., & KAPLAN, P. S. (1981). *On the failure to extinguish Pavlovian conditioned inhibition: A test of a reinstatement hypothesis*. Paper presented at the meeting of the Midwestern Psychological Association, Detroit, MI.
- PAVLOV, I. P. (1927). *Conditioned reflexes* (G. V. Anrep, Trans.). London: Oxford University Press.
- PEARCE, J. M. (1987). A model for stimulus generalization in Pavlovian conditioning. *Psychological Review*, **94**, 61-73.
- PEARCE, J. M., NICHOLAS, D. J., & DICKINSON, A. (1982). Loss of associability by a conditioned inhibitor. *Quarterly Journal of Experimental Psychology*, **33B**, 149-162.
- RESCORLA, R. A. (1969). Pavlovian conditioned inhibition. *Psychological Bulletin*, **72**, 77-81.
- RESCORLA, R. A. (1973). A model of Pavlovian conditioning. In V. S. Rusinov (Ed.), *Mechanisms of formation and inhibition of conditional reflex*. Moscow: Academy of Sciences of the U.S.S.R.
- RESCORLA, R. A. (1982). Some consequences of associations between the excitator and the inhibitor in a conditioned inhibition paradigm. *Journal of Experimental Psychology: Animal Behavior Processes*, **8**, 288-298.
- RESCORLA, R. A., & WAGNER, A. R. (1972). A theory of Pavlovian conditioning: Variations in the effectiveness of reinforcement and non-reinforcement. In A. H. Black & W. F. Prokasy (Eds.), *Classical conditioning II: Current theory and research* (pp. 64-99). New York: Appleton-Century-Crofts.
- ROBBINS, S. J. (1990). Mechanisms underlying spontaneous recovery in autoshaping. *Journal of Experimental Psychology: Animal Behavior Processes*, **16**, 235-249.
- ROSS, R. T., & HOLLAND, P. C. (1981). Conditioning of simultaneous and serial feature-positive discriminations. *Animal Learning & Behavior*, **9**, 293-303.
- SCHACHTMAN, T. R., MATZEL, L. D., & MILLER, R. R. (1988). Retardation of conditioned excitation following operational inhibitory blocking. *Animal Learning & Behavior*, **16**, 100-104.
- STOUT, S. C., & MILLER, R. R. (2007). Sometimes competing retrieval (SOCR): A formalization of the comparator hypothesis. *Psychological Review*, **114**, 759-783.
- SUITER, R. D., & LoLORDO, V. M. (1971). Blocking of inhibitory Pavlovian conditioning in the conditioned emotional response procedure. *Journal of Comparative & Physiological Psychology*, **76**, 137-144.
- VANDORPE, S., DE HOUWER, J., & BECKERS, T. (2005). Further evidence for the role of inferential reasoning in forward blocking. *Memory & Cognition*, **33**, 1047-1056.
- WAGNER, A. R. (1969). Stimulus validity and stimulus selection in associative learning. In N. J. Mackintosh & W. K. Honig (Eds.), *Fundamental issues in associative learning* (pp. 90-122). Halifax, NS: Dalhousie University Press.
- WILLIAMS, D. A. (1986). On extinction of inhibition: Do explicitly unpaired conditioned inhibitors extinguish? *American Journal of Psychology*, **99**, 515-525.
- WILLIAMS, D. A., & OVERMIER, J. B. (1988). Some types of conditioned inhibitors carry collateral excitatory associations. *Learning & Motivation*, **19**, 345-368.
- WILLIAMS, D. A., TRAVIS, G. M., & OVERMIER, J. B. (1986). Within-compound associations modulate the relative effectiveness of dif-

- ferential and Pavlovian conditioned inhibition procedures. *Journal of Experimental Psychology: Animal Behavior Processes*, **12**, 351-362.
- WITCHER, E. S., & AYRES, J. J. B. (1984). A test of two methods for extinguishing Pavlovian conditioned inhibition. *Animal Learning & Behavior*, **12**, 149-156.
- 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. D. Moore & J. F. Lehman (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.

APPENDIX

Instructions

Imagine you are a nutritionist trying to discover whether a specific hormone level will change when your client eats certain foods.

Each day a record that displays whether the hormone level changed and some of the foods the client has eaten is made (any other foods in that meal are already known to have no effect on the hormone level). On each record, you will first observe either one or two foods that your client ate. Then you will predict whether the hormone level will change or not. Following the prediction, you will be given feedback showing what actually happened. At first you will need to guess because you won't know which food(s) change the hormone level and which don't. However, after seeing a few records you will probably give more accurate predictions.

After the final record you will predict whether the hormone level will change if the client eats various foods. In this last phase you will not be given feedback indicating whether your prediction was correct.

(Manuscript received January 20, 2010;
revision accepted for publication April 21, 2010.)