Reversal of inhibition by no-modulation training but not by extinction in human causal learning

Peter F. Lovibond

Julie Y.L. Chow

Cheryl Tobler

Jessica C. Lee

The University of New South Wales

This study was funded by an Australian Research Council Discovery Project Grant [DP190103738] to Peter Lovibond. Jessica Lee was supported by a Discovery Early Career Researcher Award from the Australian Research Council [DE210100292].

Please address correspondence to p.lovibond@unsw.edu.au

Abstract

One of the many strengths of the Rescorla-Wagner (1972) model is that it accounts for both excitatory and inhibitory learning using a single error-correction mechanism. However, it makes the counterintuitive prediction that non-reinforced presentations of an inhibitory stimulus will lead to extinction of its inhibitory properties. Zimmer-Hart and Rescorla (1974) provided the first of several animal conditioning studies that contradicted this prediction. However, the human data are more mixed. Accordingly, we set out to test whether extinction of an inhibitor occurs in human causal learning after simultaneous feature negative training with a conventional unidirectional outcome. In two experiments with substantial sample sizes we found no evidence of extinction after presentations of the inhibitory stimulus alone in either a summation test or causal ratings. By contrast, two "no-modulation" procedures that contradicted the original training contingencies successfully reversed inhibition. These results did not differ substantially as a function of participants' self-reported causal structures (configural/modulation/prevention). We hypothesise that inhibitory learning may be intrinsically modulatory, analogous to negative occasion-setting, even with simultaneous training. This hypothesis would explain why inhibition is reversed by manipulations that contradict modulation but not by simple extinction, as well as other properties of inhibitory learning such as imperfect transfer to another excitor.

Keywords: inhibition, extinction, prediction error, feature negative discrimination, modulation, causal structure

No extinction of inhibition in human causal learning with a unidirectional outcome

In the 1960s Robert Rescorla played a leading role in re-invigorating the study of inhibition in associative learning (Rescorla & LoLordo, 1965; Rescorla, 1967, 1969). In particular, he investigated Pavlov's (1927) conditioned inhibition procedure, in which a training excitor A is paired with the unconditioned stimulus (US) when presented alone but not when combined with an inhibitory stimulus B, and extended it to aversive conditioning and negative correlation designs. He showed that such a procedure (now often referred to as feature negative training) slowed subsequent excitatory learning to B (retardation test), and led to transfer of B's inhibitory properties to a separately trained excitor (summation test).

When Rescorla and Wagner formulated their famous error-correction theory (Rescorla & Wagner, 1972; Wagner & Rescorla, 1972), they maintained Pavlov's view of inhibition as the direct opposite of excitation. This opposition was implemented in their use of a single bi-directional scale for associative strength, with opposite signs for excitation and inhibition. The model accounted for the known properties of inhibitory learning, including the conditions for its establishment (negative prediction error) and the outcomes of summation and retardation tests. Furthermore, it did so with exactly the same formula as was used to account for excitatory learning and recently discovered stimulus competition effects such as blocking (Kamin, 1969), correlation learning (Rescorla, 1967) and relative validity (Wagner, 1969). The Rescorla-Wagner model has become the definitive model of associative learning and has also been highly influential in the human causal and predictive learning literature (e.g., Gershman, 2015; Shanks & Dickinson, 1987; Van Hamme & Wasserman, 1994).

However, the Rescorla-Wagner model makes the problematic prediction that non-reinforced presentations of an inhibitory stimulus alone will lead to extinction of its inhibitory properties. This prediction is counterintuitive because the outcome that occurs on such trials (no US) appears to be the same as that predicted by the inhibitor (no US), which should lead to

no change in associative strength. But because inhibition is represented with a negative sign, the model indicates that there will in fact be *positive* prediction error, arising from the difference between the actual outcome (zero) and the associative strength of the inhibitor (which is negative). Characteristically, Rescorla himself was the first to identify this potential shortcoming and to publish empirical findings that contradicted the model's prediction (Zimmer-Hart and Rescorla, 1974). Subsequent studies in animal conditioning have largely confirmed this result (e.g., Lysle & Fowler, 1985; Williams & Overmier, 1988).

In the present research we were concerned with extinction of inhibition in humans. This topic is of particular relevance to the human causal learning literature because inhibitory learning has been seen as a potential model for how people learn about preventive relationships (e.g., Shanks & Dickinson, 1987). However, the evidence regarding extinction of inhibition in human causal learning is mixed. Yarlas, Cheng and Holyoak (1995) provided a brief report of an experiment in which they used a hypothetical medical scenario to pair different combinations of cues (biochemical substances) with the presence or absence of an outcome (disease). After feature negative training, they presented the inhibitory feature I without the disease in one group (direct extinction condition) but not in a control group. There was no change in causal rating or outcome predictions for I in the Extinction group, nor any difference between the extinction and control groups. Thus, they found no evidence for extinction of inhibition after exposure to the inhibitor alone, consistent with the animal findings.

Melchers, Wolff and Lachnit (2006) revived the issue when they suggested that extinction of inhibition might occur if a bidirectional outcome were employed during training, since this would align better with the assumptions of the Rescorla-Wagner model regarding inhibition being opposite to excitation. The idea was that with such a procedure the inhibitor would elicit an expectancy that the outcome level would *decrease*, which would differ from the actual consequence given during the extinction phase (no change in outcome level), thus

generating prediction error. Melchers et al. (2006) tested this proposal by assessing extinction of inhibition as a function of the type of outcome employed, in a medical prediction task. They found that non-reinforced presentations of the inhibitor did in fact lead to a loss of its inhibitory strength in a group trained with a bidirectional outcome (hormone level with a non-zero baseline) but not in a group trained with a unidirectional outcome. This result was replicated, with additional controls, by Lotz and Lachnit (2009).

Although the results from the bidirectional conditions in Melchers et al. (2006) and Lotz and Lachnit (2009) make sense from an expectancy perspective, they do not really resolve the original concern with the Rescorla-Wagner model. In particular, the results from the unidirectional conditions are still problematic for the model, which does not postulate a bidirectional outcome dimension. Rather, it was designed from the outset to explain inhibitory learning with unidirectional outcomes such as food and shock which can vary in magnitude in a positive direction but cannot take on negative values. Furthermore, Baetu and Baker (2010) replicated the design used by Melchers et al. (2006) and Lotz and Lachnit (2009), modifying certain aspects such as the rating scale, and found evidence for extinction of inhibition in *both* the bidirectional and unidirectional groups. Their results thus favored the predictions of the Rescorla-Wagner (1972) model. Finally, Kutlu and Schmajuk (2012) observed some extinction of inhibition in a summation test after training with an outcome that was ostensibly unidirectional (high/low height of a bar, where the low outcome was near zero).

In the present research, we were specifically interested in the question of extinction of inhibition when a conventional unidirectional outcome is employed, because this is the condition relevant to evaluation of the Rescorla-Wagner (1972) model, and it is also the condition where the published results are in conflict. We investigated this question within a conventional causal judgment procedure developed by Van Hamme and Wasserman (1994), the allergist task. In this task, the predictive stimuli are foods eaten by a hypothetical patient in

a given meal, and the outcome is an allergic reaction which is either present or absent. Such an outcome is intrinsically unidirectional since it cannot take on negative values. We have conducted several studies on conditioned inhibition using this task, and have established a protocol that yields reliable inhibition in a summation test relative to a conservative control stimulus trained in a non-reinforced compound (Lee & Lovibond, 2021; Lovibond & Lee, 2021).

Using this procedure, we have identified a factor that may be critical in determining the impact of non-reinforced presentations of an inhibitor, namely the causal structure that participants infer when exposed to an inhibitory contingency. Specifically, we found that participants report having learned a variety of different causal structures after simultaneous feature negative training (A+/AB-). Some describe the inhibitory feature B as directly preventing the allergic outcome. We take this "prevention" structure to be analogous to the type of inhibitory learning inherent in the Rescorla-Wagner (1972) model, where inhibition directly opposes excitation. Others report that B prevents the training excitor from causing the outcome. We label this structure "modulation" and consider it to be analogous to Holland's view of negative occasion-setting, where the feature is thought to gate the association between the training excitor and the outcome (see Fraser & Holland, 2019, for a review). Finally, some participants report simply having learned what follows each combination of stimuli, with little consideration of the properties of individual stimuli. We have labelled this type of response as "configural" as it is similar to strategies that have previously been identified in causal learning (e.g., Williams et al., 1994) as well as in configural theories of learning (e.g., Pearce, 1987). In the present project, we tested whether these individual differences might be associated with different patterns of inhibitory extinction. In particular, we hypothesised that modulation participants would be unaffected by extinction of the inhibitory feature, because this type of learning is not directly contradicted by presentation of the inhibitor without its target excitor.

Conversely, prevention participants might show a loss of inhibition as predicted by the Rescorla-Wagner (1972) model.

Experiment 1

The aim of Experiment 1 was to test the effect of non-reinforced presentations of an inhibitor after training with a simultaneous conditioned inhibition (feature negative) design using a unidirectional outcome. We also included an assessment of self-reported causal structure in order to test whether it predicts the degree of extinction observed. The overall procedure followed that of our previous work (e.g., Lovibond & Lee, 2021). The experimental design is shown in Table 1. Letters represent foods eaten by a fictitious patient, and + represents the occurrence of an allergic reaction. In Phase 1, A+ and AB- trials were used to establish B as a putative inhibitor. C+ trials established C as an excitor for the summation test, and DE- trials provided a control cue D that had been trained in a similar way to B but not in the presence of an excitor. Causal structure was assessed by both open-ended and forced choice questions administered immediately after the inhibitory training. Phase 2 was the only phase that differed between groups. The Extinction group received B- trials to test the Rescorla-Wagner (1972) prediction of extinction of inhibition. The No-Mod group received A+ and AB+ trials intended to directly contradict B's ability to modulate A's relationship to the outcome. This "no-modulation" procedure was also employed by Zimmer-Hart and Rescorla (1974), who found that it effectively counteracted B's inhibitory properties. The Control group did not receive any trials involving B.

In the Test phase, CB trials served as a summation test for B with the separately trained excitor C, and CD trials provided a control. After the Test phase, participants rated the causal strength of individual stimuli on a cause-prevent scale and completed a second assessment of causal structure. Any reduction in the inhibitory properties of B would be seen in less transfer in the summation test and less inhibitory causal ratings. We expected the impact of the

extinction procedure to vary as a function of causal structure, as noted above. Conversely, we expected the no-modulation procedure to effectively reverse the inhibitory properties of B in all participants, because it contradicts all inhibitory structures. For example, the Rescorla-Wagner predicts that A+/AB+ training will increase B's associative strength to an asymptote of zero. Thus the No-Mod group served as a reference condition for comparison with the extinction manipulation.

Method

Participants

In this experiment 206 undergraduate students (129 female, *M* age = 19.3, *SD* age = 3.3) participated in exchange for course credit. This sample size was based on the rate of exclusions and the number of participants classified in the causal structure subgroups in our previous research, with the goal of achieving >80% power for detecting medium (d=0.5) effect sizes in between-group comparisons.

Apparatus and Stimuli

As in Lee and Lovibond (2021), the experimental stimuli (A-L) were selected randomly from a pool of 16 food pictures that included a verbal label (e.g., "chicken"). The allergic reaction outcome consisted of the text "Allergic Reaction!" accompanied by a sad face emoticon. No outcome consisted of the text "No allergic reaction". The experiment was programmed using the jspsych library (de Leeuw, 2015), hosted using JATOS (Lange, Kühn, & Filevich, 2015) and run on participants' web browsers.

Procedure

The project was approved by the University of New South Wales Human Research Ethics Advisory Panel C (approval number 3136), and participants provided online consent. Table 1 shows the sequence of phases. The design and procedure followed that used in our

previous research (Lee & Lovibond, 2021; Lovibond & Lee, 2021). In brief, participants were asked to play the role of an allergist trying to work out which foods cause allergic reactions in a patient "Mr X". Each trial represented a meal eaten by the patient. Participants were presented with the text "Mr X eats" followed by pictures of either one or two foods with their verbal labels. After 500ms, participants were asked to rate the likelihood that Mr X would show an allergic reaction on a visual analogue scale from "Definitely NO ALLERGIC REACTION" to "Definitely ALLERGIC REACTION". The predictive ratings were recorded on a numerical scale from 0 to 100. When participants had made their rating, the prediction scale and prompt were replaced by either the allergic outcome or no outcome. After 2s, the stimuli and feedback disappeared and the 1-s blank inter-trial-interval (ITI) period commenced.

Phase 1. As shown in Table 1, Phase 1 followed our previous training protocol (Lee & Lovibond, 2021; Lovibond & Lee, 2021) in which A+ and AB- trials were used to establish B as a conditioned inhibitor. C+ trials trained C as an excitatory (causal) stimulus for the later summation test. Stimulus D provided a control stimulus for B, having been similarly trained in a non-predictive compound but with no source of excitation. The remaining stimuli were fillers, included to help prevent participants from inferring general rules such as all single foods predict the allergic outcome. The order of trials was randomised, with the restriction that the same trial type could not be presented on successive trials. Each trial type was presented 6 times (twice within each of 3 blocks of 16 trials). The left-to-right presentation of the compound stimuli was counterbalanced within each block.

Phase 2. A similar procedure was followed in Phase 2, where each group received 5 trial types designed to implement the between-group manipulation. The Extinction group received B- trials to test for extinction of inhibition, while the No-Mod group received A+ and AB+ trials, intended to more strongly contradict the inhibitory training to B. Each group was also given novel trials (J+/JK+ or L-) to match the contingencies experienced by the other

group. The Control group received both sets of novel trials, and no trials involving B. All groups received C+ and F- trials to maintain some continuity with Phase 1.

Between Phases 1 and 2, participants completed an initial assessment of their inferred causal structure for stimulus B. This consisted of two parts. The first was an open-ended question asking them to describe what they had learned about stimulus B. The second was a 3-alternative forced choice (3AFC) question asking them to choose one of three descriptions regarding the causal status of food B. The three options (prevention, modulation, configural) were the same as in our previous studies (Lee & Lovibond, 2021; Lovibond & Lee, 2021). Specifically, the prevention option was "It prevented allergic reactions in general", the modulation option was "It prevented allergic reactions caused by specific foods", and the configural option was "It is hard to know the exact role of individual foods such as this one. I concentrated on remembering which combinations of foods caused an allergic reaction and worked from there". The order of the options was randomized for each participant.

Test phase. After training was completed, participants were told that they would now be presented with more meals and that they should continue to make predictive ratings, but they would no longer receive any feedback regarding the presence or absence of the allergic outcome. The trial types presented in the Test phase consisted of a) all the trial types from the Training phase except for the filler compound GH+; b) individual stimuli that had previously only been presented in a compound (D, E); c) a novel stimulus I; and d) two compounds for the summation test (CB and CD). Each trial type was tested twice, with the left-to-right presentation of compound stimuli counterbalanced.

After the Test phase, participants were presented with all of the individual stimuli except for the filler stimuli G and H. They were asked to rate "to what extent each food tended to prevent or cause an allergic reaction" on a visual analogue scale. The scale ranged from

"Strongly PREVENTED ALLERGIC REACTION" to "Strongly CAUSED ALLERGIC REACTION", with "No effect" labelled in the middle of the scale. The causal ratings were recorded on a numerical scale from -100 to +100. Finally, they completed the causal structure assessment for a second time, in order to test whether either of the reversal manipulations had altered participants' perception of the way in which the inhibitory stimulus B operated. Demographic questions were included at the start of the experiment, and an additional question at the end asked participants if they had written anything down during the experiment.

Table 1. *Design of Experiment 1.*

Group	Phase 1 ⇩	Phase 2	Test	Causal ratings ⇩
Extinction		**B-** J+, JK+ C+, F-		
No-Mod	A+, AB- C+, DE- , F-, GH+	**A+, AB+** L- C+, F-	A, AB, CB, CD, C, DE, D, E, F, I	A, B, C, D, E, F, I
Control		J+, JK+ L- C+, F-		

Note: Letters represent stimuli (foods); + represents the outcome (allergic reaction); – represents the absence of the outcome. The procedure for Phase 1, Test and Causal ratings were identical for the three groups. The critical Phase 2 manipulations are shown in bold type. Arrows indicate administration of the causal structure assessment.

Data Analysis

We analysed the data with planned contrasts using the PSY package (Bird, Hadzi-Pavlovic & Isaac, 2000). The critical analyses concerned the test data. Within-participant contrasts compared CB to CD in the outcome prediction ratings, and B to D in the causal ratings. Group contrasts tested all three pairwise comparisons between the Extinction, No-Mod and Control groups. Causal structure subgroup contrasts compared a) the modulation subgroup with the prevention subgroup, and b) the average of these two groups with the configural subgroup. We tested all contrasts as main effects (i.e., averaged over all other factors), as well as all possible interactions between contrasts. For each contrast of interest, we report the p value as well as the corresponding standardized 95% confidence interval (CI). The mean of the CI represents standardised effect size (Bird, 2004).

Results

Exclusion Criteria

The primary exclusion criteria were those used in Lee and Lovibond (2021) and Lovibond & Lee (2021). Specifically, participants were excluded if they reported having written anything down during the experiment (n=26) or if they failed the acquisition criterion which was an average rating > 75 for the stimuli that predicted the outcome (A+, C+, GH+) and an average rating < 25 for the stimuli that predicted no outcome (AB-, DE-, F-) in the last block of Phase 1 (n=26). In addition, we excluded participants who failed the instruction check more than twice (n=5). After applying all three criteria, 152 participants remained (47 in the Control group, 47 in the Extinction group, and 58 in the No-Mod group).

Causal Structure Assessment 1

Examination of the open-ended responses regarding causal structure at the first assessment revealed no novel causal structures not already covered by the forced choice

options. Therefore, in line with our previous research (e.g., Lee & Lovibond, 2021), we divided participants into subgroups based on their response to the first 3AFC question. These subgroups were used to analyse any effects of the initial causal structures that participants reported having formed during acquisition of the A+/AB- discrimination. There were 34 participants in the configural subgroup, 104 in the modulation subgroup, and 14 in the prevention subgroup. As shown in Table 2, there was no significant difference between the groups in their causal structure choice at the first assessment (X^2(df=2, N=152) = 3.50, p=.48).

Table 2. Causal structure classification by group at the first 3AFC assessment in Experiment 1

Group	Configural	Modulation	Prevention	Total
Control	11	34	2	47
Extinction	11	29	7	47
No-Mod	12	41	5	58
Total	34	104	14	152

Phase 1

Figure 1 shows the mean outcome prediction ratings across trials in Phase 1. Initial analyses confirmed very similar acquisition patterns across groups and causal structure subgroups, as would be expected given that all participants received the same training. There was a small but significant difference in mean ratings between the Control and Extinction groups (49.8 vs. 47.8 respectively; F(1,149) = 5.72, p=.02, 95% CI = 0.01, 0.14). However, none of the group or causal structure subgroup contrasts interacted significantly with the within-participant contrasts that assessed acquisition (largest F(1,149) = 2.61, p=.11, 95% CI = -0.04, 0.36); hence Figure 1 is collapsed across both group and subgroup.

Participants rapidly acquired differential responding to stimuli that predicted the outcome versus those that predicted no outcome. This resulted in a significant main effect for the contrast that compared predictive and non-predictive stimuli ($F(1,151) = 4082.5$, $p<.001$, 95% CI = 2.57, 2.73), as well as an interaction between this contrast and linear trend over trials ($F(1,151) = 2102.9$, $p<.001$, 95% CI = 1.32, 1.43). Ratings on the first AB- trial were higher than to the other stimuli, due to approximately half of the participants having previously received an A+ trial. This pattern led to a significant main effect for the contrast comparing AB- to the other two non-reinforced trial types (DE- and F-) averaged over trials ($F(1,151) = 64.3$, $p<.001$, 95% CI = -0.47, -0.29), as well as an interaction between this contrast and linear trend over trials ($F(1,151) = 72.3$, $p<.001$, 95% CI = 0.28, 0.45). Nonetheless, by the end of Phase 1 ratings were similarly low for the three non-reinforced trial types.

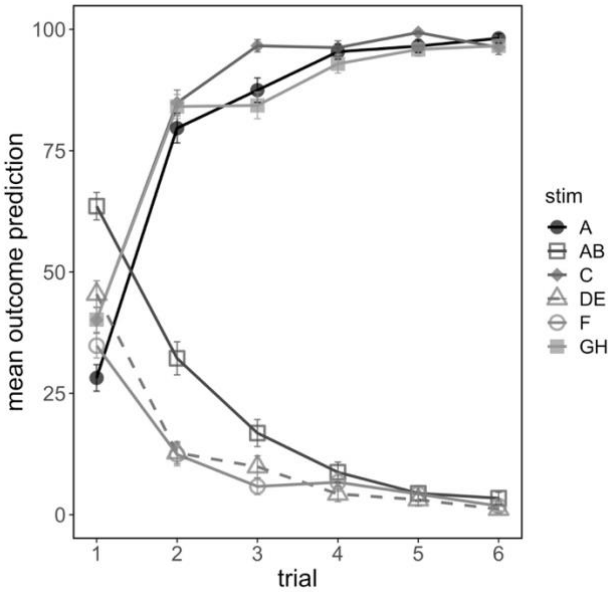

Figure 1. Mean outcome prediction ratings (+/– 1 SE) during Phase 1 in Experiment 1.

Phase 2

Figure 2 shows the mean outcome prediction ratings across trials in Phase 2, separated by group. There were only small differences between the causal structure subgroups in this phase so Figure 2 is collapsed across this factor. All groups maintained low ratings to F and

high ratings to GH throughout the phase. They also learned about the novel filler cues as expected. Statistical analysis focused on the critical trial types within the two experimental groups. In the Extinction group, prediction ratings to the inhibitory stimulus B were low from the first trial, despite the fact that these participants had never seen B outside of the AB compound previously. Nonetheless, the analysis revealed a small but significant decreasing linear trend across subsequent B- trials ($F(1,44) = 5.14$, $p=0.03$, 95% CI = -0.43, -0.03). This contrast also interacted with the comparison between the Configural and Prevention subgroups, reflecting a slightly higher mean rating for B on the first two trials in the Configural group ($F(1,44) = 5.16$, $p=0.03$, 95% CI = -1.19, -0.07). In the No-Mod group, initial ratings to the A and AB trial types were consistent with the end of Phase 1 training. However ratings to AB rapidly increased as participants learned that this compound now signalled the outcome. This pattern led to an interaction between the A vs AB comparison and linear trend over trials ($F(1,55) = 143.4$, $p<.001$, 95% CI = -1.48, -1.06). This pattern did not further interact with either of the causal structure subgroup contrasts, *F*s < 1.

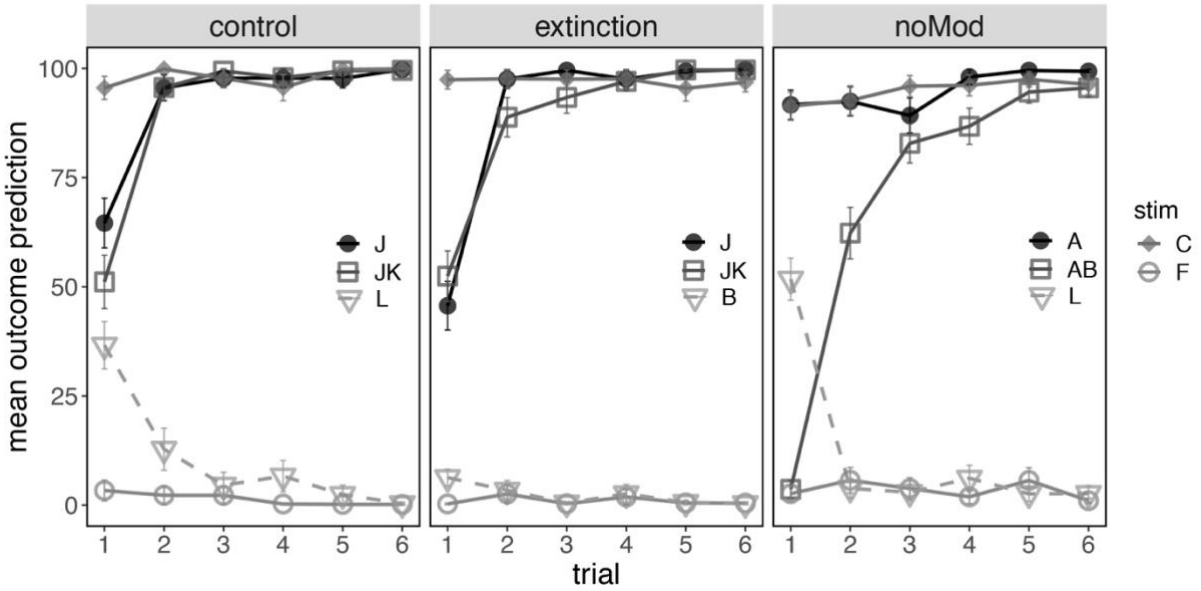

Figure 2. Mean outcome prediction ratings (+/– 1 SE) for each group during Phase 2 in Experiment 1.

Test

For analysis of the outcome prediction test data, each participant's prediction ratings were averaged over the two trials for each trial type. Ratings for the trained stimuli were very similar to the end of acquisition. Figure 2 shows mean ratings for the critical summation compounds for each group (see Supplemental Materials for full data). Looking first at the top panel, it can be seen that ratings to CB were clearly lower than to the control compound CD in the Control group, indicating successful transfer of the inhibitory properties of B to the novel excitor C. The data pattern for the Extinction group was very similar. By contrast, in the No-Mod group, ratings to CB were much higher and were close to CD. These differences were reflected in the statistical analysis. There was an overall main effect for the CB-CD difference, averaged over groups ($F(1,143) = 25.30$, $p<.001$, 95% CI = -1.40, -0.61). This contrast interacted with both the Control vs No-Mod comparison ($F(1,143) = 16.53$, $p<.001$, 95% CI = -3.19, -1.10) and the Extinction vs No-Mod comparison ($F(1,143) = 36.80$, $p<.001$, 95% CI = -3.34, -1.70). Importantly, however, the contrast comparing CB with CD did not interact with the Control vs Extinction comparison ($F<1$). Thus there was no evidence that the B- trials in Phase 2 in the Extinction group had any impact on B's ability to suppress responding to C in the summation test.

The bottom panel of Figure 3 shows the prediction test data further separated by self-reported causal structure as recorded after Phase 1. The three causal structure subgroups each followed a similar pattern to the sample as a whole. The ordering of ratings for the test compound CB was consistent with our previous work (strongest transfer in the Prevention subgroup and weakest transfer in the Configural subgroup; Lee & Lovibond, 2021; Lovibond & Lee, 2021). However, none of the interactions between the CB-CD comparison and the subgroup comparisons reached significance (largest $F(1,143) = 2.68$, $p=0.10$, 95% CI = -0.13, 1.40).

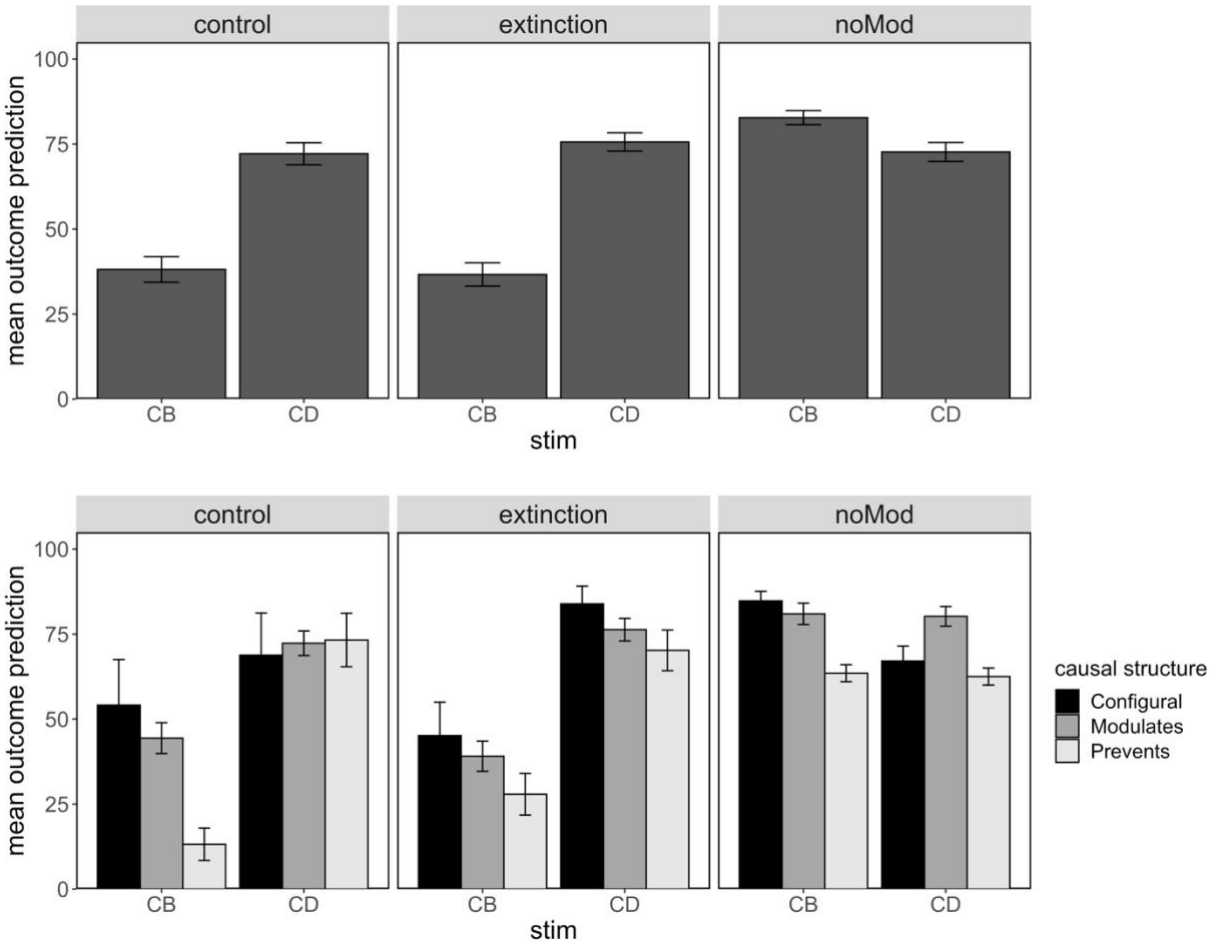

Figure 3. Top panel shows mean outcome prediction ratings (+/– 1 SE) for the summation (CB) and control (CD) compounds for each group in the Test phase of Experiment 1. Bottom panel is separated by causal structure subgroup.

Causal rating test

Causal ratings for the common training stimuli were as expected on the basis of their reinforcement history. The top panel of Figure 4 shows mean causal ratings for the critical test stimuli B and D (see Supplemental Materials for full data). This test differed from the outcome prediction test in that participants rated the causal strength of the individual stimuli on a cause-prevent scale, rather than making outcome predictions. Overall, participants gave significantly lower (more preventive) ratings for the inhibitory stimulus B than for the control stimulus D ($F(1,143=36.70$, $p<.001$, 95% CI = -1.55, -0.79). However, as in the prediction test data, this difference was not equivalent across groups. Interaction contrasts showed that the B-D

difference was significantly smaller in the No-Mod group than in either the Control group ($F(1,143=51.17$, $p<.001$, 95% CI = -2.96, -1.68) or the Extinction group ($F(1,143=79.12$, $p<.001$, 95% CI = -2.77, -1.77). However, the B-D difference did not differ between the Control and Extinction groups ($F<1$). We repeated these group comparisons for stimulus B alone and obtained the same pattern of results. Finally, we examined the causal structure subgroup factor (bottom panel of Figure 4). The pattern of group differences was consistent across subgroups. Although the Prevention participants in the No-Mod group appeared to show more negative ratings to B and D than the other participants, there were only 5 participants in this subgroup. There were no significant interactions involving the causal structure factor in either the B-D analysis (all $Fs<1$) or the analysis of B alone (largest $F(1,143)=1.40$, $p=0.24$, 95% CI = -0.26, 1.03).

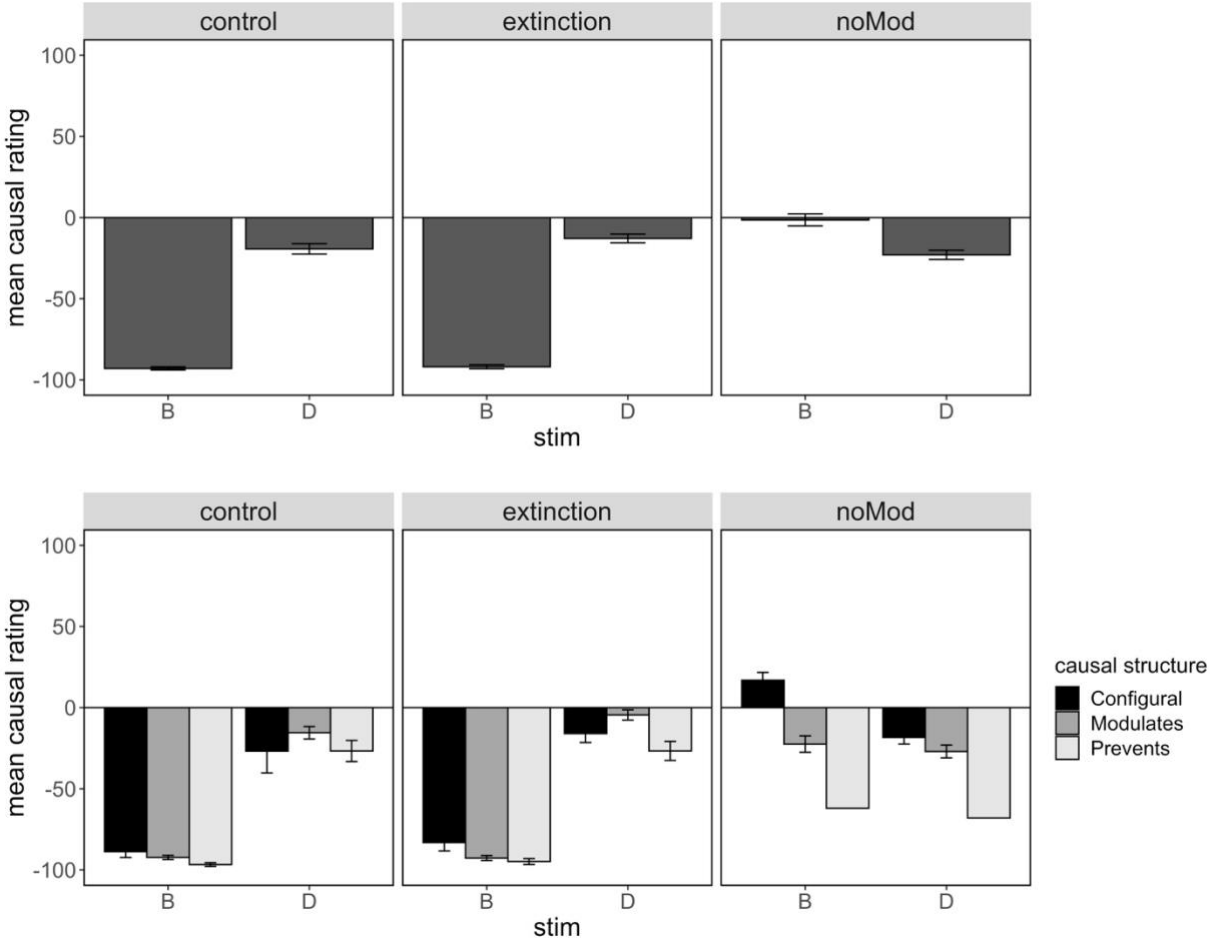

Figure 4. Top panel shows mean causal ratings (+/– 1 SE) for the critical test stimuli for each group in Experiment 1. Bottom panel is separated by causal structure subgroup. Positive ratings indicate causation, negative ratings indicate prevention, and zero indicates no effect.

Causal Structure Assessment 2

In the second causal structure assessment, 42 participants selected the configural option, 82 selected the modulation option, and 26 selected the prevention option. However, this pattern was not the same across groups (X^2(df=2, N=152) = 38.5, p<.001). As shown in Table 3, a greater proportion of participants in the No-Mod group chose the Configural option at the second assessment compared to the other two groups. This difference is presumably due to the contradictory experiences that the No-Mod group received in Phases 1 and 2.

Table 3. Causal structure classification by group at the second 3AFC assessment in Experiment 1

Group	Configural	Modulation	Prevention	Total
Control	5	31	11	47
Extinction	7	26	14	47
No-Mod	32	25	1	58
Total	44	82	26	152

Discussion

The results from Experiment 1 were very clear. Non-reinforced presentations of the inhibitor B in the Extinction group had no discernible influence on its inhibitory properties in either the summation test or causal ratings. The test data for this group were very similar to those from the Control group, which did not experience B- trials. By contrast to the Extinction group, the No-Mod group showed a complete reversal of B's inhibitory properties, on both test

measures. This condition is important because it demonstrates that it is possible to reverse inhibition, and that the tests we employed were sensitive enough to detect such a reversal.

In this experiment, the majority of participants reported inferring a configural or modulatory causal structure for B, as opposed to the prevention structure that would correspond to an inhibitory association that acted directly on the outcome representation (Wagner & Rescorla, 1972). The small number of participants who endorsed a prevention structure did not respond differently to the extinction procedure compared to the other subgroups. In fact, there were essentially no differences between participants in their response to the group manipulation as a function of their self-reported causal structure for B. This pattern will be considered further in the General Discussion.

Even though the results of Experiment 1 were clear cut, there are nonetheless two issues that impact on its interpretation. First, it is possible that interposing the causal structure assessment between Phase 1 and Phase 2 influenced the way in which participants represented the knowledge they had acquired in Phase 1, or the way in which they responded to the extinction manipulation in Phase 2. Second, the No-Mod procedure might have exerted its strong reversal effect by virtue of the fact that it directly contradicted the training contingencies. Specifically, the AB compound was non-reinforced in Phase 1 and reinforced in Phase 2. We addressed both of these issues in Experiment 2.

Experiment 2

In this experiment we included two design changes to extend and clarify the findings of Experiment 1. First, to eliminate any potential impact of the initial causal structure assessment on subsequent responding, we omitted this assessment in Experiment 2. Thus, there was no longer any break between Phases 1 and 2. Second, we added a fourth group to test a subtler version of the no-modulation procedure, as suggested by Baetu and Baker (2012). In this

group, labelled No-Mod Novel, participants were given intermixed M+ and MB+ trials, where M was a novel stimulus introduced in Phase 2. This design counteracts the modulatory role of B but does not directly contradict the training trials with stimulus A.

Method

Participants

Two hundred and forty four undergraduate students (163 female, *M* age = 20.0, *SD* age = 3.6) participated in Experiment 2 in exchange for course credit. The exclusion criteria were the same as for Experiment 1 with the addition that participants could not have participated in Experiment 1.

Apparatus and Stimuli

The apparatus and stimuli were the same as in Experiment 1, except that 13 rather than 12 food stimuli were randomly selected to serve as experimental stimuli A-M.

Procedure

The procedure for Experiment 2 followed that of Experiment 1, with two main exceptions. First, an additional group was included, No-Mod Novel. This group was treated in the same way as the No-Mod group, except that in Phase 2 the no-modulation training involved a novel stimulus M. Thus, this group received M+ and MB+ trials in Phase 2. The second change was that the first 3AFC causal structure assessment was omitted. Training proceeded from Phase 1 to Phase 2 with no demarcation. A final minor change was to include stimulus B alone in the Test phase. The full design table is shown in Table S2 in Supplemental Materials.

Data Analysis

We followed the same analysis strategy as in Experiment 1, except that we did not analyse causal structure subgroups since we only had a post-experimental assessment and Experiment 1 had demonstrated that the Phase 2 manipulations significantly altered

participants' responses from the first to the second assessment. In this experiment the planned contrasts for the group factor compared a) the Extinction with the Control group, 2) the No-Mod with the No-Mod Novel group, and c) the average of the Extinction and Control groups with the average of the No-Mod and No-Mod Novel groups.

Results

Exclusion Criteria

Eighteen participants failed the write check, 24 failed the instruction check, and 37 failed the acquisition criterion. After all exclusions had been applied, 175 participants remained (41 in the Extinction group, 48 in the No-Mod group, 45 in the No-Mod Novel group, and 41 in the Control group).

Phase 1

The training data were very similar to Experiment 1. There were no substantial differences between groups, so Figure 5 is collapsed across this factor. Participants again learned to discriminate cue combinations that predicted the outcome from those that did not. Acquisition was demonstrated by a significant difference between the reinforced and non-reinforced cues, averaged over trials ($F(1,163)=2846.0$, $p<.001$, 95% CI = 2.77, 2.98) as well as an interaction between this contrast and linear trend over trials ($F(1,163)=1663.1$, $p<.001$, 95% CI = 1.40, 1.55).

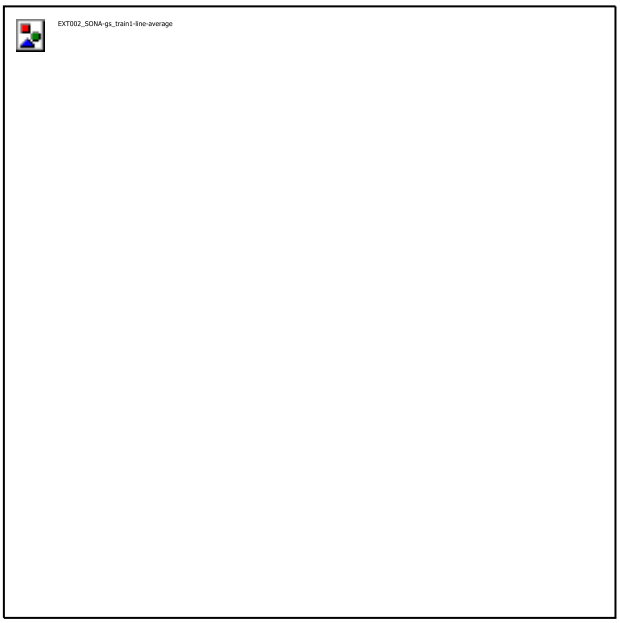

Figure 5. Mean outcome prediction ratings (+/– 1 SE) during Phase 1 in Experiment 2.

Phase 2

The Phase 2 predictive ratings are shown for each group in Figure 6. The data patterns for the Extinction, No-Mod and Control groups were very similar to those seen in Experiment 1. The Extinction group again gave low ratings to the inhibitory stimulus B on the first trial, and in this experiment showed no significant reduction over subsequent trials (linear trend: $F<1$). In the No-Mod Novel group, initial ratings to the novel cue M were near the middle of the scale but rapidly increased to asymptote. Ratings to the MB compound were somewhat lower, but also increased rapidly to asymptote. By contrast, in the No-Mod group, ratings to the AB compound increased more slowly, presumably due to the conflict with the Phase 1 contingency. These two groups were analysed together, to test the above pattern. There was a significant interaction between the No-Mod vs No-Mod Novel group comparison and the element (A or M) vs compound (AB or MB) comparison, ($F(1,87) = 37.00$, $p<.001$, 95% CI = 1.20, 2.37), confirming that averaged over trials, the A vs AB difference in group No-Mod was smaller than the M vs MB difference in group No-Mod Novel. This contrast further interacted with linear trend, indicating that the group difference in responding to the compound relative to

the element diminished across trials as participants learned the new contingencies, ($F(1,87) = 18.01$, $p<.001$, 95% CI = -1.60, -0.61).

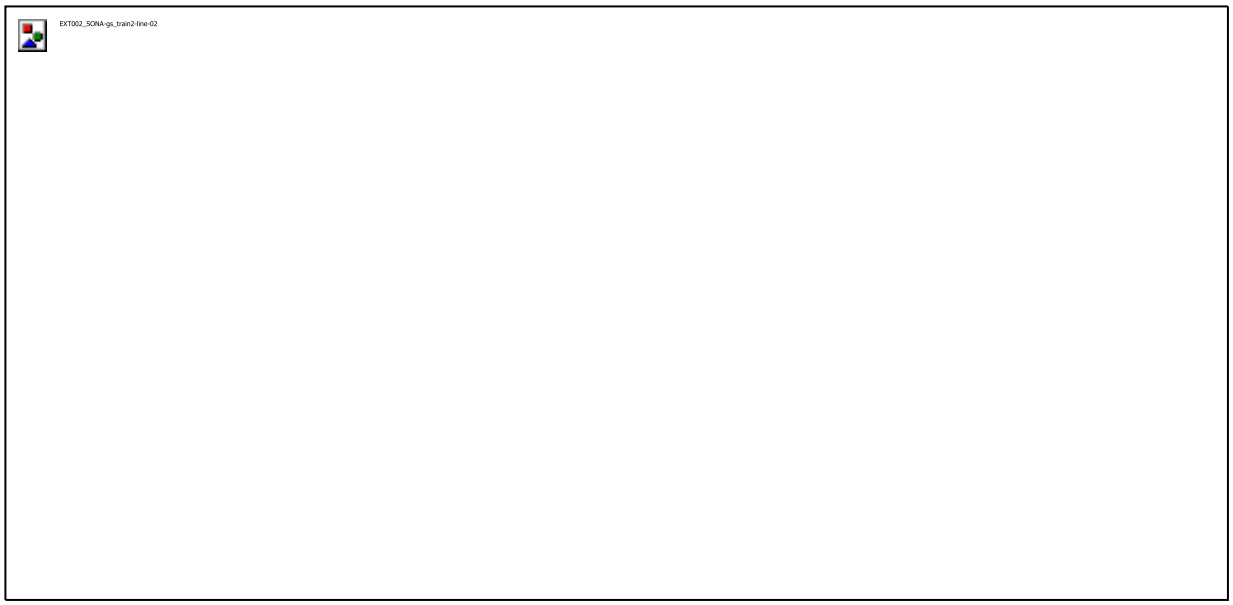

Figure 6. Mean outcome prediction ratings (+/– 1 SE) for each group during Phase 2 in Experiment 2.

Test

Figure 7 shows the mean predictive ratings to the critical test stimuli for each group (see Supplemental Materials for full data). In addition to the CB and CD summation test compounds, data for A and AB trials are included to allow comparison between the two no-modulation procedures on the original training discrimination. B trials are also included as a further test of the impact of the group manipulations. The results for the three groups in common with Experiment 1 were very similar to that experiment, except that the difference between the summation test compound CB and the control compound CD was somewhat smaller in this experiment. However, the pattern across groups replicated that seen in Experiment 1 and extended it to the No-Mod Novel condition which showed a very similar pattern to the No-Mod group. Averaged over groups, ratings to CB were lower than to CD ($F(1, 163)= 21.67$, $p<.001$, 95% CI = -0.64, -0.26), confirming the overall inhibitory properties of B. This contrast interacted with the comparison between the two no-modulation groups and

the other two groups (F(1, 163)= 22.81, p<.001, 95% CI = -1.29, -0.54), demonstrating that the no-modulation manipulations had reduced inhibitory transfer. The interaction between the CB vs CD comparison and the No-Mod vs No-Mod Novel comparison was non-significant (F<1), reflecting the very similar pattern seen in these two groups. Interestingly, the interaction between the CB vs CD comparison and the Extinction vs Control comparison just reached significance, in the direction of slightly *greater* inhibitory transfer in the Extinction group (F(1, 163)= 5.04, p=.026, 95% CI = 0.07, 1.11).

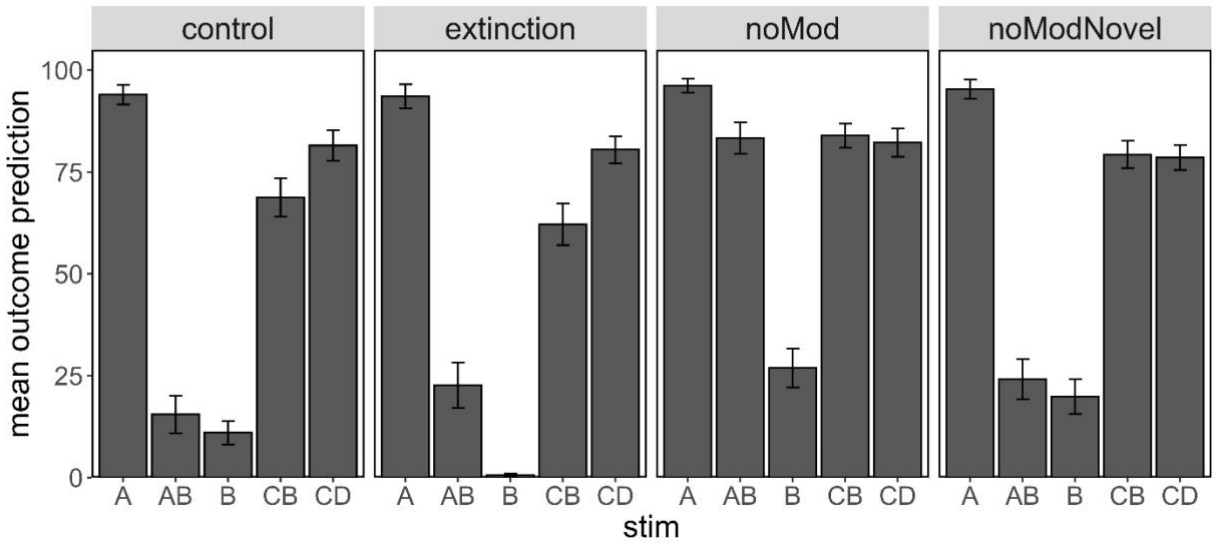

Figure 7. Mean outcome prediction ratings (+/– 1 SE) for critical test stimuli in each group in the Test phase of Experiment 2.

Although the two no-modulation groups showed a similar loss of transfer of B's inhibitory properties to the test excitor C, and both groups experienced B paired with the outcome in Phase 2, the No-Mod Novel group maintained low ratings to the AB compound whereas the No-Mod group showed much higher ratings. This pattern led to a significant interaction between the No-Mod vs No-Mod Novel comparison and the A vs AB comparison (F(1, 163)= 31.01, p<.001, 95% CI = -3.35, -1.60). We also compared the groups on their predictive ratings to stimulus B alone. Here, the principal finding was higher ratings to B in the two no-modulation groups compared to the Extinction and Control groups, consistent with the

summation test results (F(1, 163)= 8.92, p=.003, 95% CI = -1.07, -0.22). There was no difference between the two no-modulation groups (F<1) or between the Extinction and Control groups (F(1, 163)= 2.62, p=.11, 95% CI = -0.11, 1.05) in their ratings to B.

Causal Rating Test

Figure 8 shows mean causal ratings for the critical test stimuli B and D (see Supplemental Materials for full data). Averaged over groups, participants gave significantly lower (more preventive) ratings for the inhibitory stimulus B than for the control stimulus D (F(1,163)=33.64, p<.001, 95% CI = -1.21, -0.60). However, as in Experiment 1, the magnitude of this difference differed across groups. Specifically, the B-D difference was significantly smaller in the average of the two no-modulation groups than in the average of the Control and Extinction groups (F(1,163)=10.59, p=.001, 95% CI = -1.63, -0.40). However, the B-D difference did not differ between the No-Mod and No-Mod Novel groups, or between the Control and Extinction groups (Fs<1). We repeated these group comparisons for stimulus B alone and obtained the same pattern of results.

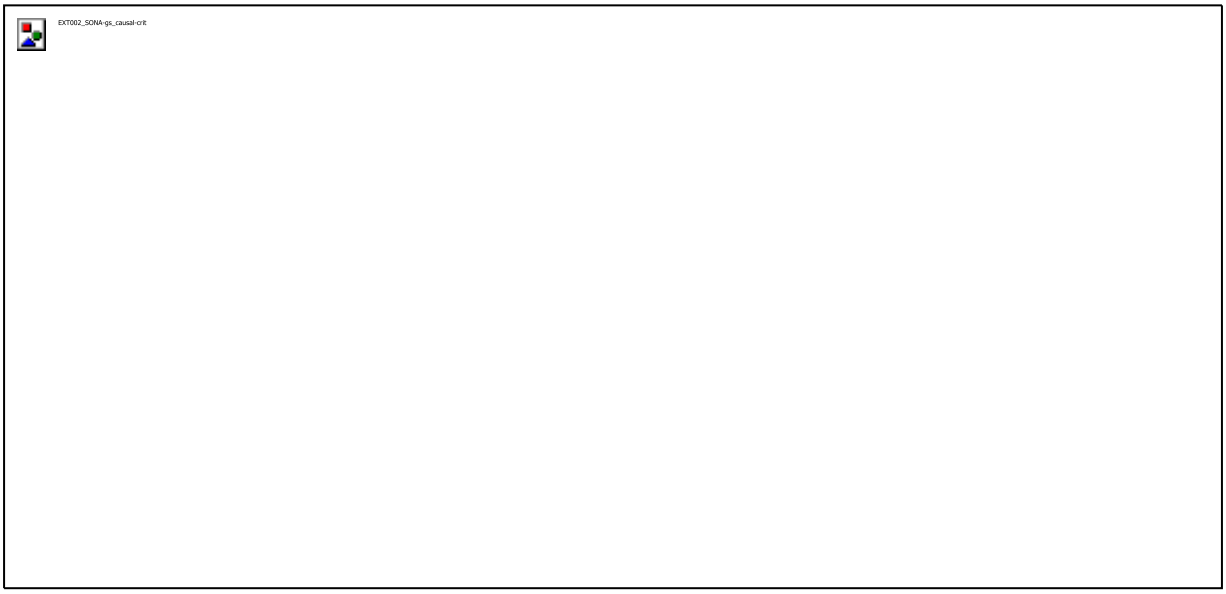

Figure 8. Mean causal ratings (+/– 1 SE of the mean) for test stimuli in each group in Experiment 2.

Discussion

This experiment replicated the main finding from Experiment 1, namely that there was no evidence of extinction of inhibition, in either the summation test or causal ratings. If anything, we saw a small effect in the opposite direction for the CB-CD comparison in the summation test. This outcome indicates that the absence of extinction in Experiment 1 was not due to the inclusion of a causal structure assessment prior to extinction. Nonetheless, the overall strength of inhibitory learning observed in both the summation test and causal ratings appeared to be slightly lower than in Experiment 1. This difference suggests that asking participants in Experiment 1 to reflect on causal structure, or exposing them to the options in the 3AFC question, enhanced their inhibitory performance and transfer.

Experiment 2 also showed that a no-modulation training procedure with a novel excitor (No-Mod Novel condition) was as effective in reversing inhibitory learning as the same procedure with the training excitor (No-Mod condition). This finding suggests that the critical feature needed for reversal of inhibitory learning is training that the inhibitor no longer has modulatory properties. Although no-modulation training with the training excitor might more successfully interfere with the original inhibitory learning, no-modulation training with a novel excitor might encourage greater transfer in a summation test with yet another excitor. Perhaps these two factors cancelled out in the present design. In any case, it is clear that both procedures were effective in reversing inhibition.

General Discussion

Both experiments showed that non-reinforced presentations of a conditioned inhibitor, established by feature negative training with a conventional unidirectional outcome, do not lead to any measurable reduction in its inhibitory properties. This outcome is consistent with the results reported by Yarlas et al. (1995) and the unidirectional outcome conditions in

Melchers et al. (2006) and Lotz and Lachnit (2009). In the present experiments, the lack of extinction of inhibition was observed in both a summation test with a novel excitor and in causal ratings, relative to a control group that did not receive presentations of the inhibitor alone. Both experiments had a substantial sample size, and the success of the no-modulation conditions in reversing inhibition demonstrated that the procedure was able to detect reversal if it occurred. To further assess the strength of the null findings for the Extinction-Control comparisons, we pooled the data from these two conditions from each experiment and carried out a Bayesian analysis. Bayes Factors showed that the evidence for the null hypothesis – no effect of the extinction manipulation relative to the control – was between 3 and 4 times higher than for the alternative hypothesis (BF01 = 3.81 for the CB-CD difference in predictive ratings, and BF01 = 3.77 for the B-D difference in causal ratings).

The results of the present experiments conflict with those of the Unidirectional group in Baetu and Baker (2010, Experiment 1), which showed evidence for extinction of inhibition. Comparison of the procedures suggests a critical difference that could account for the diverging outcomes. Specifically, the outcome used by Baetu and Baker (2010), change in hormone level from an unspecified baseline level, could have been interpreted by participants as bidirectional even in the Unidirectional group. The only procedural difference between the Unidirectional and Bidirectional groups was that the Unidirectional group did not experience any trials in which the hormone level decreased. Thus, the Unidirectional participants may still have inferred that the inhibitory cue would have produced such a decrease if presented alone. Direct support for this interpretation comes from their Test phase data, which were collected using a bidirectional magnitude scale that ranged from “decrease” through “not change” to “increase” in both groups. The Unidirectional group gave strong negative (decrease) ratings for both of the unextinguished inhibitory cues – in fact, stronger than for the Bidirectional group. Therefore it seems likely that the extinction of inhibition seen in the Unidirectional group in

Baetu and Baker (2010) occurred for the same reason as for the Bidirectional group, and in accordance with the account proposed by Melchers et al. (2006) – that is, presentations of the inhibitor with no change in outcome level contradicted the expectation that it would produce a decrease. By contrast, in the present research, we used a unidirectional outcome that could only vary between zero (no allergic reaction) and a positive value (allergic reaction).

A similar argument may apply to the results of Kutlu and Schmajuk (2012), who reported a loss of inhibitory strength after extinction of an inhibitory feature. These researchers used an outcome (line height) that was putatively unidirectional. However, they presented no causal scenario or instructions to participants to clarify whether cues led to a particular absolute line height or a change in line height from some starting level. Note that the outcome used by Melchers et al. (2009), Lotz & Lachnit (2009) and Baetu & Baker (2010), hormone level, is also intrinsically unidirectional – it is not possible to have negative values of a hormone. The outcome only became bidirectional in the Bidirectional conditions of those experiments by virtue of instructions stating that cues could cause a relative increase or a decrease from a non-zero starting value. In the Kutlu and Schmajuk (2012) design, the starting point for the line height was ambiguous, potentially allowing some participants to treat this outcome as bidirectional.

It therefore appears that when a conventional unidirectional outcome is employed, as in the present experiments and the majority of animal research, presentations of an inhibitor do not decrease its inhibitory properties. What does this mean for the way in which inhibitory learning is encoded? As noted earlier, it is problematic for the Rescorla-Wagner (1972) model, which incorrectly predicts a loss of inhibition. Rescorla (1979) pointed out that one way to accommodate this finding is in terms of the theory put forward by Konorski (1948). According to this theory, an inhibitor acts to increase the threshold required for an excitor to activate the US (outcome) representation. As such, non-reinforced presentations of an inhibitor would not

lead to extinction. However, this theory shares a limitation with the Rescorla-Wagner (1972) model, namely that it predicts strong transfer in a summation test, an outcome that is rare in animal research and even rarer in human research (e.g., Karazinov & Boakes, 2004; Lee & Livesey, 2012; Williams, 1995). Finally, unlike the Rescorla-Wagner (1972) model, Konorski's (1948) theory fails to capture the idea of expectancy violation that appears to be critical for inhibitory learning. It is also difficult to apply directly to causal learning in humans in terms of aligning with a distinct causal structure.

An alternative approach is suggested by the literature on negative occasion-setting (e.g., Rescorla, 1987; Fraser & Holland, 2019). This type of learning, which in animals is promoted by the use of a serial arrangement during feature negative training (A+ / B→A–), appears to endow the inhibitory stimulus with the ability to modulate responding to the target excitor. Importantly, this type of learning is typically somewhat specific to the target excitor and hence supports only partial transfer in a summation test. Furthermore, the occasion-setting property of a stimulus appears to be relatively unaffected by manipulations involving its direct association with the outcome. Both logically and empirically, presentation of an occasion-setter with or without the outcome does not affect its modulatory properties with respect to another stimulus (Fraser & Holland, 2019). Several aspects of the current results are consistent with the idea that inhibitory learning followed a modulatory associative structure. First, transfer to a novel excitor in the summation test was only modest. Second, non-reinforced presentations of the inhibitory stimulus did not affect its ability to modulate either the training excitor or a novel excitor. Third, the two manipulations that did contradict the modulatory properties of the inhibitor, No-Mod and No-Mod Novel, were highly effective in reversing inhibition. Finally, many of the participants reported having learned a modulatory causal structure in the self-report assessment in Experiment 1.

There are also aspects of the present results that do not align so well with the idea that participants were learning inhibition in a modulatory way. First, we used a simultaneous feature negative design, whereas in animals, modulatory learning (occasion-setting) is most commonly seen with serial designs. However, in our previous research using a similar procedure to the current experiments, we have found that a serial arrangement has little impact on self-reported causal structure (Lovibond & Lee, 2021). Perhaps humans are more predisposed to interpret inhibitory associations in a modulatory way, even when the stimuli are presented simultaneously. A second issue is that we only tested inhibitory learning with a summation test, not with a retardation test (Pavlov, 1927; Rescorla, 1969). An important topic for future research would be to test the prediction that follows from the occasion-setting literature, namely that an inhibitor with modulatory properties may not in fact show retarded excitatory acquisition, due to the apparent independence of modulatory and direct associations (Fraser & Holland, 2019).

A third complication with a modulatory account of our results is that not all participants reported a modulatory causal structure when we assessed this in Experiment 1 – a substantial number reported a configural or prevention structure. When we first investigated individual differences in inhibitory learning, we hypothesised that the different self-reported causal structures represented qualitatively different structures (Lee & Lovibond, 2021). Our subsequent work has led us to question this assumption, for several reasons. First, some participants might choose the configural option because it is the most conservative, rather than because it represents a discrete causal structure. In the present experiments, these participants gave substantial negative ratings to B in the causal rating test, suggesting they were in fact able to infer an inhibitory role for this element when it was presented outside the compound. Second, several manipulations we have tested, including serial vs. simultaneous training, and prior modulatory training of the test excitor, turned out to have a similar impact across all

causal structure subgroups (Lovibond & Lee, 2021). These results suggest an alternative possibility, that the differences in transfer we have consistently observed in a summation test (Lee & Lovibond, 2021; Lovibond & Lee, 2021) instead reflect participants' willingness to generalise the properties of the inhibitory stimulus to a novel excitor. Furthermore, this variation may be quantitative rather than qualitative, with preventers showing the greatest transfer in a summation test, configural participants the least, and modulators in between. Consistent with this perspective, we have recently argued that a more productive way to conceptualise the summation test is in terms of principles of generalisation, rather than the arithmetic summation of associative strengths (Chow, Lee & Lovibond, in press; see also Bonardi, Robinson & Jennings, 2017). Thus, participants who chose the prevention option may have had a greater tendency to generalize the inhibitory properties of the feature to a new excitor, presumably due to differences in their pre-experimental their reinforcement history. If participants vary primarily in their willingness to generalise, then it is less surprising that they respond equally to manipulations that are not related to this dimension, including the present manipulations of extinction and no-modulation.

Accordingly, it remains at least possible that there is a single causal structure that underlies inhibitory learning in humans, and that this structure is modulatory in nature. Such a view can explain the three key findings of the present experiments: no extinction of inhibition, successful reversal of inhibition by procedures that contradict modulation, and no substantial differences in these effects as a function of self-reported causal structure. It is possible that inhibitory learning in animals is also fundamentally modulatory in nature, even with simultaneous feature negative training. The apparent reversal of such inhibition by direct reinforcement (e.g., Holland, 1989) could be explained in terms of the superimposition of excitatory learning on the original modulatory learning. By this account, inhibition would be the symmetrical opposite not of excitation but of facilitation (positive occasion-setting), a view

put forward by Rescorla (1987) and extended to causal reasoning by Baetu and Baker (2012). Furthermore, it would not require abandonment of the central prediction error component of the Rescorla-Wagner (1972) model, only a modification of the way in which learning is represented: direct association in the case of excitatory learning, and modulation of an excitatory association in the case of inhibitory learning and facilitation. Whatever the merits of this particular position, it is clear that the insights provided by Robert Rescorla over his 50-year career will continue to play a central role in advancing our understanding of associative learning.

References

Baetu, I. & Baker, A.G. (2010). Extinction and blocking of conditioned inhibition in human causal learning. *Learning & Behavior*, *38*, 394-407.

Baetu, I. & Baker, A.G. (2012). Are preventive and generative causal reasoning symmetrical? Extinction and competition. *Quarterly Journal of Experimental Psychology*, *65*, 1675-1698.

Bird, K. D., Hadzi-Pavlovic, D. & Isaac, A. P. (2000). *PSY* [Computer software]. Available from http://www.psy.unsw.edu.au/research/psy.htm

Bird, K.D. (2004). *Analysis of variance via confidence intervals*. London: Sage Publications.

Bonardi. C., Robinson, J., & Jennings. D. (2017). Can existing associative principles explain occasion setting? Some old ideas and some new data. *Behavioural Processes*, *137*, 5-18.

Chow, J.Y-L., Lee, J.C. & Lovibond, P.F. (in press). Inhibitory summation as a form of generalisation. *Journal of Experimental Psychology: Animal Learning and Cognition*.

de Leeuw, J. R. (2015). jsPsych: A JavaScript library for creating behavioural experiments in a web browser. *Behaviour Research Methods*, *47*, 1-12.

Fraser, K. M., & Holland, P. C. (2019). Occasion Setting. *Behavioral Neuroscience, 133*, 145-175.

Gershman S.J. (2015). A unifying probabilistic view of associative learning. *PLoS Computional Biology*, *11*, e1004567.

Holland, P. C. (1989). Transfer of negative occasion setting and conditioned inhibition across conditioned and unconditioned stimuli. *Journal of Experimental Psychology: Animal Behavior Processes*, *15*, 311–328.

Kamin, L. J. (1969). Predictability, surprise, attention and conditioning. In B.A. Campbell & R.M. Church (Eds.), *Punishment and aversive behavior* (pp. 279-296). New York: Appleton-Century-Crofts.

Karazinov, D.M. & Boakes, R.A. (2004). Learning about cues that prevent an outcome: Conditioned inhibition and differential inhibition in human predictive learning. *Quarterly Journal of Experimental Psychology, 57,* 153-178.

Konorski, J. (1948). Conditioned reflexes and neuron organization. Cambridge: Cambridge University Press.

Kutlu, M.G. & Schmajuk, N.A. (2012). Deactivation and reactivation of the inhibitory power of a conditioned inhibitor: Testing the predictions of an attentional-associative model. *Learning & Behavior*, *40,* 83-97.

Lange, K., Kühn, S., & Filevich, E. (2015). "Just Another Tool for Online Studies" (JATOS): An easy solution for setup and management of web servers supporting online studies. *PLoS ONE, 10(6)*: e0130834.

Lee, J.C. & Livesey, E. J. (2012). Second-order conditioning and conditioned inhibition: Influences of speed versus accuracy on human causal learning. *PLoS ONE, 7(11),* e49899.

Lee, J.C. & Lovibond, P.F. (2021). Individual differences in causal structures inferred during feature negative learning. *Quarterly Journal of Experimental Psychology, 74(1),* 150-165.

Lotz, A & Lachnit, H. (2009). Extinction of conditioned inhibition: Effects of different outcome continua. *Learning & Behavior*, *37*, 85-94.

Rescorla, R.A. & LoLordo, V.M. (1965). Inhibition of avoidance behavior. *Journal of Comparative & Physiological Psychology*, *59,* 406–412.

Lovibond, P.F. & Lee, J.C. (2021). Inhibitory causal structures in serial and simultaneous feature negative learning. *Quarterly Journal of Experimental Psychology, 74*, 2165-2181.

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.

Melchers, K.G., Wolff, S. & Lachnit, H. (2006). Extinction of conditioned inhibition through nonreinforced presentation of the inhibitor. *Psychonomic Bulletin & Review, 13*, 662-667.

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

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

Rescorla, R.A. (1967). Pavlovian conditioning and its proper control procedures. *Psychological Review, 74*, 71–80.

Rescorla, R.A. (1969). Pavlovian conditioned inhibition. *Psychological Bulletin, 72*, 77–81.

Rescorla, R.A. (1979). Conditioned inhibition and extinction. In A. Dickinson and R.A. Boakes (Eds.), *Mechanisms of Learning and Motivation: A memorial volume to Jerzy Konorski* (pp. 83-110). Hillsdale, N.J.: Erlbaum.

Rescorla, R.A. (1987). Facilitation and inhibition. *Journal of Experimental Psychology: Animal Behavior Processes, 13*, 250-259.

Rescorla, R. A., & Wagner, A. R. (1972). A theory of Pavlovian conditioning: Variations in the effectiveness of reinforcement and nonreinforcement. *Classical conditioning II: Current research and theory, 2*, 64-99.

Shanks, D.R. & Dickinson, A.R. (1987). Associative accounts of human causality judgment. *The Psychology of Learning and Motivation, 21*, 229-261.

Van Hamme, L.J & Wasserman, E.A. (1994). Cue competition in causality judgments: The role of nonpresentation of compound stimulus elements. *Learning and Motivation, 25*, 127-151.

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: Dalhousie University Press.

Wagner, A.R. & Rescorla, R.A. (1972). Inhibition in Pavlovian conditioning: Application of a theory. In R.A. Boakes and H.S. Halliday (Eds.), *Inhibition and Learning*. London: Academic Press.

Williams, D.A. (1995). Forms of inhibition in animal and human learning. *Journal of Experimental Psychology: Animal Behavior Processes, 21*, 129-142.

Williams, D. A. & Overmier, J. B. (1988). Some types of conditioned inhibitors carry collateral excitatory associations. *Learning & Motivation, 19*, 345-368.

Williams, D.A., Sagness, K.E. & McPhee, J.E. (1994). Configural and elemental strategies in predictive learning. *Journal of Experimental Psychology: Learning, Memory and Cognition, 20*, 694-709.

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.

Supplemental Materials

Experiment 1

Table S1. Causal structure classification at the two time points in Experiment 1

	3AFC2			
3AFC1	Configural	Modulation	Prevention	Total
Configural	**14**	16	4	34
Modulates	28	**63**	13	104
Prevents	2	3	**9**	14
Total	44	82	26	152

Note: 3AFC1 refers to the first forced choice assessment of causal structure (after Phase 1) and 3AFC2 refers to the second assessment (at the end of the experiment). Bold font indicates participants whose choice was stable; that is, they selected the same option at each time point.

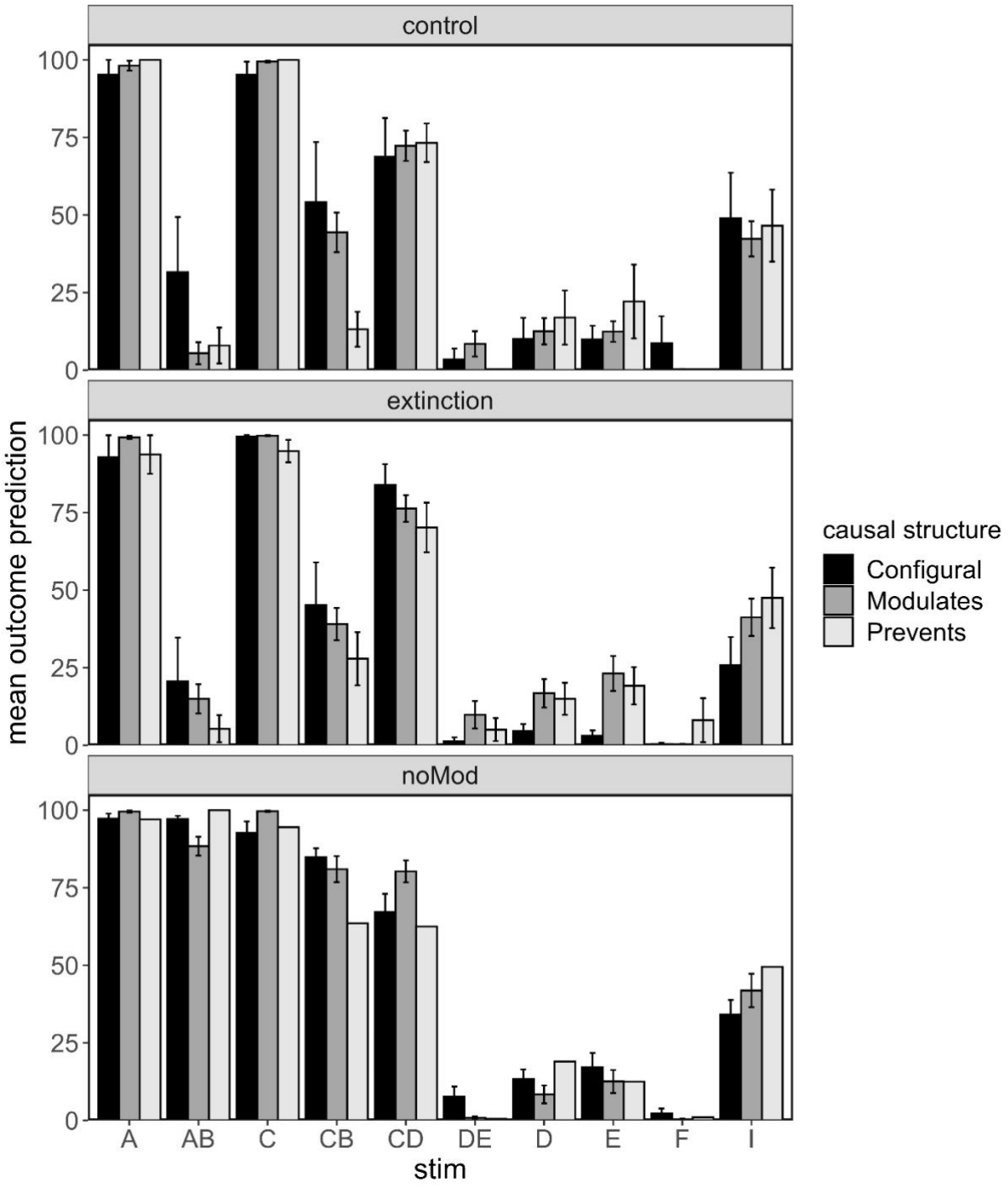

Figure S1. Mean outcome prediction ratings (+/– 1 SE) for all test stimuli in the Test phase of Experiment 1, separated by group and causal structure subgroup.

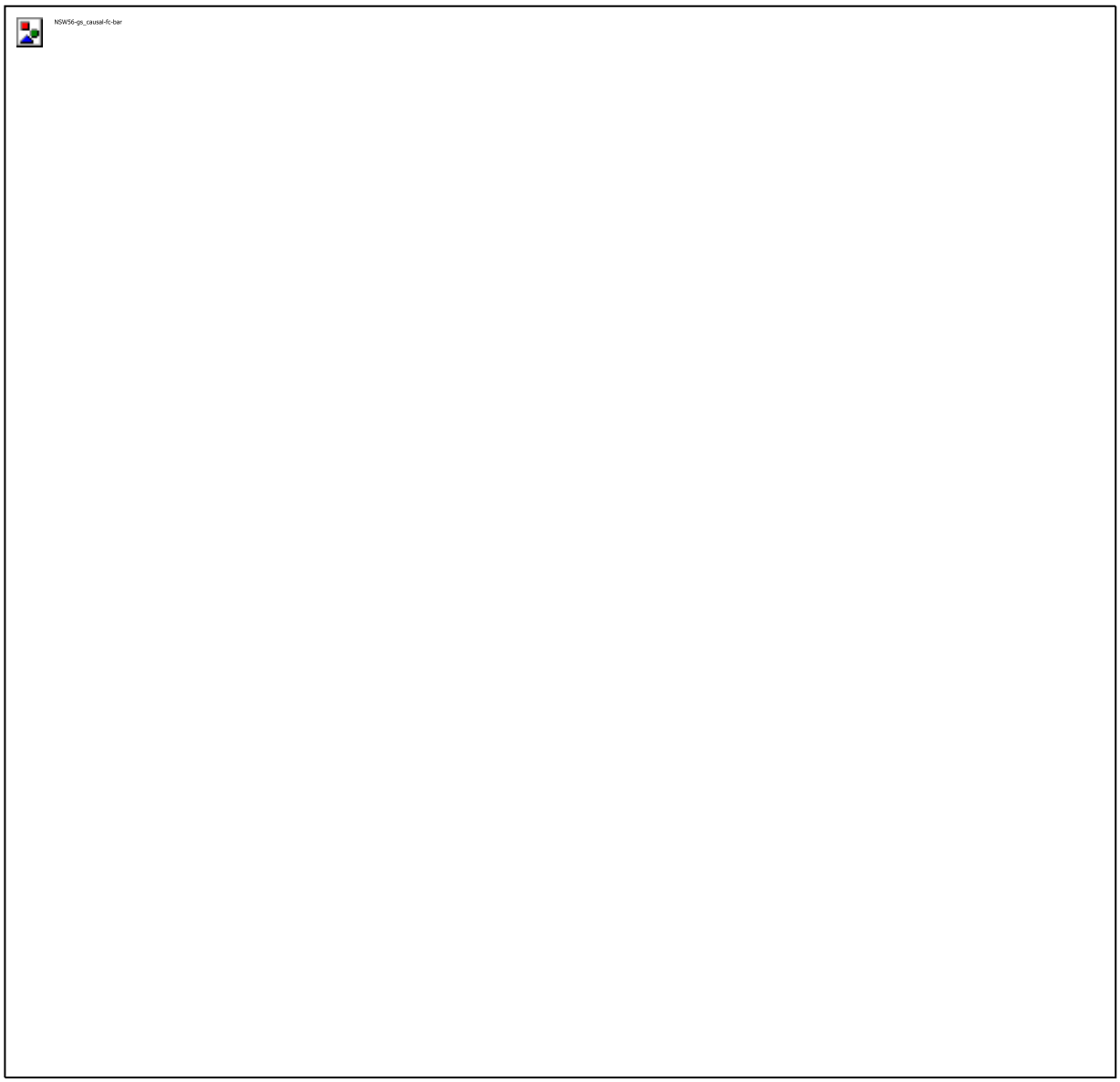

Figure S2. Mean causal ratings (+/– 1 SE) for all test stimuli in Experiment 1, separated by group and causal structure subgroup.

Table S2. Design of Experiment 2

Group	Phase 1	Phase 2	Test	Causal ratings ⇩
Extinction	A+, AB- C+, DE-, F-, GH+	**B-** J+, JK+ C+, F-	A, B, AB, CB, CD, DE, C, D, E, F, I	A, B, C, D, E, F, I
No-Mod		**A+, AB+** L- C+, F-		
No-Mod Novel		**M+, MB+** L- C+, F-		
Control		J+, JK+ L- C+, F-		

Notation as per Table 1 in the main manuscript.

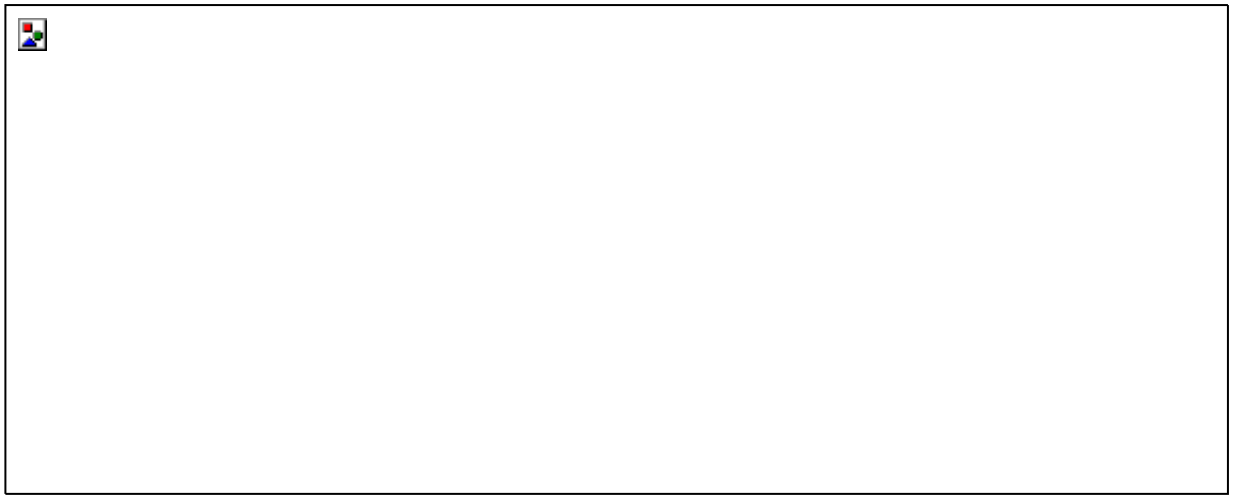

Figure S3. Mean outcome prediction ratings (+/– 1 SE) for all test stimuli in the Test phase of Experiment 2, separated by group.

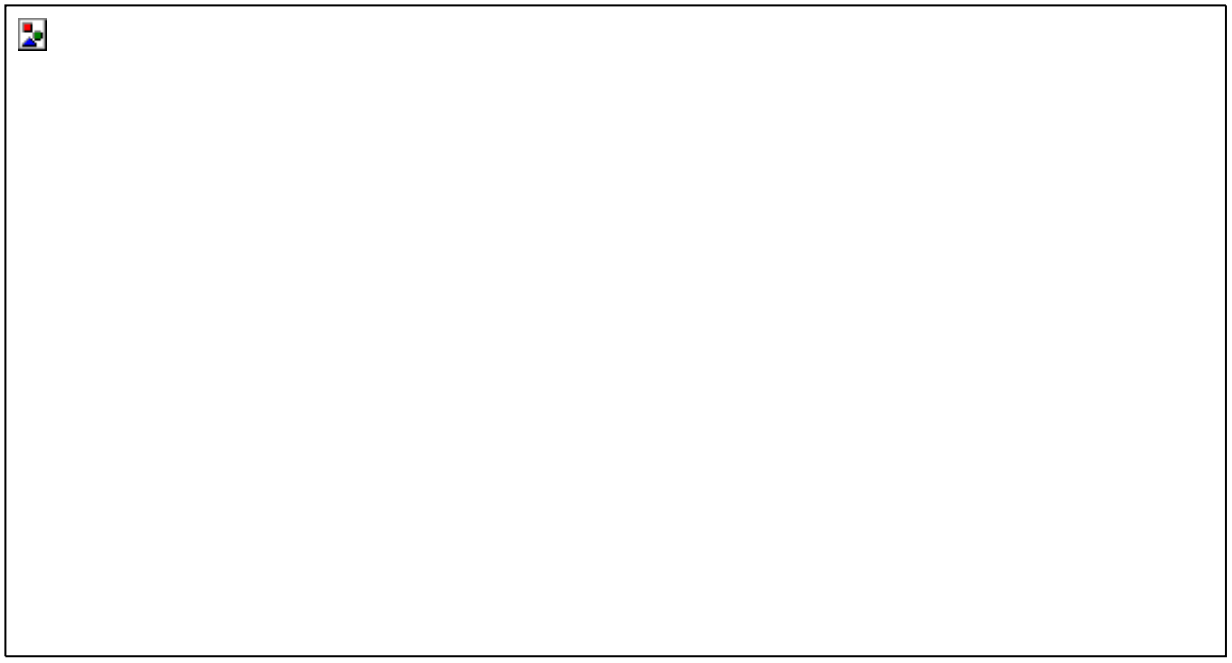

Figure S4. Mean causal ratings (+/– 1 SE) for all test stimuli in Experiment 2, separated by group.