Testing the deductive inferential account of blocking in causal learning.

Evan J. Livesey[1], Justine K. Greenaway[1], Samantha Schubert[1], & Anna Thorwart[2]

[1] University of Sydney

[2] Philipps-Universität Marburg

Word count:
Abstract: 215
Main Text: 8242

Running head: Blocking and the Additivity Pretraining Effect

Please address correspondence to:

Evan Livesey
School of Psychology
University of Sydney
Sydney NSW 2006 Australia
Email: evan.livesey@sydney.edu.au
Tel +61-2-9351 2845
Fax +61-2-9351 2603

Abstract

The sensitivity of the blocking effect to outcome additivity pretraining has been used to argue that the phenomenon is the result of deductive inference, and to draw general conclusions about the nature of human causal learning. In two experiments, we manipulated participants' assumptions about the additivity of the outcome using pretraining before a typical blocking procedure. Ratings measuring causal judgments, confidence, and expected severity of the outcome were used concurrently to investigate how pretraining affected assumptions of outcome additivity and blocking. In Experiment 1, additive pretraining led to lower causal ratings and higher confidence ratings of the blocked cue, relative to control cues, consistent with the notion that additive pretraining encourages deductive reasoning. However, Experiments 1 and 2 showed that removing additivity assumptions through nonadditive pretraining had no impact on a statistically reliable blocking effect observed in a blocking procedure with no pretraining. We found no evidence that the blocking effect in the absence of pretraining was related to the participants' assumptions about the additivity of the outcome. Although additive pretraining may enhance blocking by encouraging deductive reasoning about the blocked cue, the evidence suggests that blocking in causal learning is not reliant on this reasoning and that humans do not readily engage in deduction merely because they possess the assumptions that permit its use.

Keywords: Blocking, outcome additivity, deductive reasoning, causal learning, associative learning.

Testing the deductive inferential account of blocking in causal learning.

Learning theorists use the term *blocking* to refer to the tendency for people to either fail to learn about, or fail to attribute meaning to, a reliable signal of a relevant outcome simply because that signal is redundant. In a world where causally unrelated events frequently co-occur, this tendency is potentially very important. For instance, the well-learned knowledge that flicking the light switch causes the light to turn on prevents the learner from misattributing the light to other extraneous events that often occur around the same time such as the sun setting or someone entering the room. On the other hand, it may also prevent us identifying important causal relationships. For instance, if the stereotype about gender and math ability already provides a well-known "explanation" for a female student's performance, teachers and parents may not attribute problems in math to insufficient practice, thus missing an opportunity to effect positive change (Sanbonmatsu, Akimoto & Gibson, 1994). Since its initial report in animal conditioning, 50 years ago, the blocking effect (Kamin, 1968) has continued to stir theoretical controversy. Recently, the phenomenon has played an important role in debates about the mechanisms responsible for human causal learning (e.g. see Griffiths, Sobel, Tenenbaum & Gopnik, 2011; Mitchell, De Houwer & Lovibond, 2009; Shanks, 2007). In a blocking procedure, trials with cue A followed by an outcome (A+) are presented in an initial phase, followed by a second phase in which cues A and B occur together followed by the outcome (AB+). Participants are then asked to rate the extent to which various cues cause the outcome. Causal ratings for B are often lower than for control cues (C or D) that were followed by the same outcome but only ever appeared in compound (CD+).

Most contemporary models of associative learning, especially those based on a summed error-correction learning algorithm (e.g. Rescorla & Wagner, 1972), anticipate blocking by assuming that learning the association between B and the outcome is attenuated when the outcome is already well predicted (in this case, by A). Proponents of such accounts often assume that a basic associative memory mechanism serves as a source of information in many forms of cognition, including causal judgments, by operating as a domain-general psychological process responsible for contingency learning (e.g. see Shanks, 2007). According to this account, humans often base their assessment of causation on the strength of associative memories that link the putative causes with the effect. In other words, causation is determined by the extent to which each cue brings to mind the idea of the outcome.

Associationist theories are criticised for failing to address other factors that clearly influence judgments of causation, such as the participant's preconceived understanding of cause and effect within a given domain (Waldmann & Walker, 2005). Alternative accounts have argued that blocking is a product of the rational decision that the individual makes about the ability of the putative cause to generate the outcome in the absence of other causes (i.e. its *causal power,* see Cheng, 1997). These accounts focus on the formal computations necessary to arrive at an estimate of causal power when judging the blocked cue at test rather than characterising blocking as an attenuation of learning (e.g. Novick & Cheng, 2004). Bayesian approaches to causal learning have attempted to provide a framework by which one could arrive at a rational decision as to whether a cue is causally related to an outcome (Griffiths & Tenenbaum, 2005; Tenenbaum, Griffiths & Kemp, 2006). Theoretical treatments of the blocking effect are thus diverse and range from explicitly mechanistic, focusing on specific psychological processes, to explicitly computational, focusing on defining the formal problem and how a solution can be rationalised.

The purpose of this study was to examine the predictions of a specific inferential account proposed to explain blocking via a set of well-defined reasoning processes. This account proposes that the learning on which causal judgments are based is necessarily composed of explicit propositions and inferential reasoning, which the individual will engage in depending on the available evidence and the assumptions that they bring into each instance of learning (Mitchell et al., 2009). As part of this account, one explanation for blocking outlined by Lovibond, Been, Mitchell, Bouton & Frohardt (2003; see also De Houwer, Beckers & Glautier, 2002; Mitchell & Lovibond, 2002) is that participants deduce that B contributes nothing in addition to the already observed consequence of A and therefore cannot be causal. At test, the participant recalls that A led to an outcome and that AB led to an outcome of the same magnitude. The participant reasons that if B were causal then the outcome following AB should be more severe than the outcome following A alone. Since the AB outcome is *not* more severe, then B is *not* causal. Formally, this inference is equivalent to *modus tollens* whereby a logical proposition is falsified by a negative consequent. The participant rates B as being unlikely to cause the outcome with an elevated degree of confidence, relative to control cues that still have an ambiguous status as potential causes (see Vandorpe, De Houwer, & Beckers, 2005).

This deduction requires an “additivity” assumption, that the outcome should be larger in magnitude and/or probability in the presence of two causes. In contrast, if the participant assumes that the effects of two causes are nonadditive then the causal status of B is ambiguous and, according to Lovibond et al. (2003), should be regarded as having the same causal status as the control cues that are only trained in compound. Consequently, a participant might be expected to give the same causal rating for B, C, and D, which reflects this state of uncertainty. Importantly, although this is still a reasoned inference, it is very different to *modus tollens*. Instead of deducing the causal status of B, the participant must conclude that they lack the information that would be necessary to differentiate between the consequences of B, C, or D.

Lovibond et al. (2003) tested the effect of outcome additivity assumptions by giving one group of participants pretraining that explicitly demonstrated the additive nature of the outcome and another group of participants explicit pretraining that demonstrated the outcome was the same magnitude whether there were one or two causes present. The additive group subsequently displayed significantly greater blocking than the nonadditive group. Lovibond et al. (2003) found this result in both standard (forward) blocking, where the individual cue A is observed to cause the outcome prior to training with the AB compound, as well as backward blocking, where the individual training of cue A comes after the participant learns about the AB compound. This *additivity pretraining effect* has been replicated in several studies (e.g. Beckers, De Houwer, Pineno & Miller, 2005; Livesey & Boakes, 2004; Mitchell, Lovibond & Condoleon, 2005), supporting the deductive inferential account.

The significance of the additivity pretraining effect lies not in the fact that participants use deductive inference when strongly encouraged to do so under additive pretraining conditions but rather in what it suggests about the source of the blocking effect under more ambiguous circumstances. A corollary of the effect is that, in the *absence* of explicit pretraining, blocking must heavily depend on an assumption of outcome additivity held by at least some participants as only this would allow them to apply *modus tollens*. Some authors have made this conclusion explicitly (e.g. Beckers et al., 2005), whereas it is implicit in the arguments made by others. For instance, in line with this conclusion, many authors have interpreted the additivity pretraining result to mean that nonadditive pretraining reduces blocking by undermining the additivity assumption (e.g. Mitchell et al. 2009). Others have suggested that past failures to observe blocking are likely due to the use of conditions that encourage participants to assume that the outcome is nonadditive (Lovibond et al., 2003; Mitchell & Lovibond, 2002; Mitchell et al., 2005).

The deductive inferential hypothesis therefore has the potential to explain many of the blocking effects observed in human learning experiments, as well as failures to observe blocking where the additive outcome assumption is discouraged. Some experiments have shown blocking effects that are more accurately described as the result of a failure to encode the blocked cue-outcome relationship and are therefore not easily reconciled with this account of blocking as the inferences described require knowledge of the blocked cue-outcome relationship (e.g. Griffiths & Mitchell, 2008; Mitchell, Lovibond, Minard & Lavis, 2006). Nonetheless, the additivity pretraining effect still suggests that blocking is often a consequence of inferential (and specifically, *deductive*) reasoning.

The current study sought to further test this hypothesis and answer several outstanding questions concerning the additivity pretraining effect. In doing so, we hoped to better ascertain what the additivity pretraining effect tells us about the nature of blocking and human causal learning in general. To our knowledge, every published study of the additivity pretraining effect in humans has compared blocking after additive pretraining with blocking after nonadditive pretraining. Therefore, it remains to be seen whether additive pretraining enhances blocking, nonadditive pretraining reduces blocking, or both. Given the additivity pretraining effect has been used to draw broad conclusions about the nature of blocking, it makes sense to compare both types of pretraining condition against a standard blocking condition in which no pretraining is given. Both experiments therefore compared additive and/or nonadditive pretraining against a procedure with no pretraining.

In order to examine the relationship between the additivity assumption and blocking more directly, we developed a variant of the allergist task used by Lovibond et al. (2003) that requires several concurrent judgments during the test phase. Participants assumed the role of a doctor determining which foods were making their patient sick. In addition to the usual learning about food cues and a binary outcome of "no allergic reaction" or "allergic reaction", the magnitude of the outcome was indicated on a fictitious allergy severity scale, which provided the necessary information to assess the additive or nonadditive nature of the outcome. On test, participants then made three kinds of ratings, 1) a causal rating as in previous similar studies, 2) a confidence rating about their causal judgment, and 3) a rating of the expected severity of the outcome. The severity rating was used to test participants' additivity assumptions directly. Deductive inferential accounts of blocking expect three results on these measures to be closely related, namely reduction of causal ratings of B relative to C/D, should coincide with higher confidence in causal ratings for B, and the presence of an additivity assumption measured via severity ratings. If blocking in the absence of pretraining is due to inconsistent use of deductive inference across participants then 1) additive pretraining should strengthen *all three* of these effects, and 2) nonadditive pretraining should weaken all three of these effects.

Following the designs used by Lovibond et al. (2003), Experiment 1 compared forward and backward blocking effects with additive, nonadditive, or no pretraining. Experiment 2 replicated a key finding from Experiment 1 but controlled for effects of submaximality of the outcome on blocking and the additivity assumption.

Experiment 1

Experiment 1 had three aims. The first was to extend the findings of Lovibond et al. (2003) and Beckers et al. (2005), who found reduced blocking following nonadditive pretraining relative to additive pretraining, by comparing these conditions to a no pretraining control. According to the deductive inferential hypothesis, and assuming that pretraining is successful in manipulating additivity assumptions, blocking should be weaker after nonadditive pretraining than after no pretraining because participants who do not receive pretraining are more likely to entertain additivity assumptions by default. For similar reasons, blocking should be stronger after additive pretraining than after no pretraining because the additive pretraining participants will be further encouraged to entertain an additivity assumption and use this assumption to make deductions about the blocked cue. The second aim was to test the prediction that additive pretraining leads to increased confidence in the causal judgment for the blocked cue as the additivity assumption permits the use of deductive reasoning about the blocked cue.

The third aim was to test the extent to which additivity is assumed in the absence of pretraining, and whether changes in the strength of blocking reflect changes in this assumption when pretraining is applied. For example, according to the deductive inferential account, a strong reduction in the prevalence of the additivity assumption after nonadditive pretraining should be accompanied by a reduction in blocking. Lovibond et al. (2003) also suggested that a residual blocking effect in the nonadditive condition could be due to a small proportion of participants who continue to assume additivity despite all the contrary evidence. Thus there should also be a relationship between the additivity assumption and blocking at a participant level in the groups given no pretraining.

In Experiment 1, we tested both forward blocking and backward blocking. Lovibond et al (2003) found a reliable blocking effect after additive pretraining but no evidence of blocking after nonadditive pretraining, when the order of training was reversed (A+ trials were presented *after* AB+ training). Inferential accounts explain backward blocking with the same reasoning as forward blocking. Provided participants accurately remember the relevant blocking contingencies on which one needs to make inferences, inferential judgments about the blocked cue on test should not be affected by the order in which those contingencies were experienced. On the other hand, associative explanations typically require additional assumptions to account for any backward blocking effect (Aitken, Larkin, & Dickinson, 2001; Dickinson & Burke, 1996; Van Hamme & Wasserman, 1994). Therefore, observing a backward blocking effect that is the same magnitude and affected by the pre-training manipulations in the same way as forward blocking supports the hypothesis that both rely on inferential reasoning.

Crossing two between-subjects factors, pretraining (additive, nonadditive, no pretraining) and phase order (forward, backward) resulted in six experimental groups. As shown in Table 1, participants received two training phases including the blocking contingencies (A+5 and AB+5), relevant control trials (CD+5) and some additional cues (see below). For those in the backward blocking condition, training of the compounds AB and CD was followed by training of the single cue A. On test, blocking was taken to be evident if cue B receives significantly lower causal ratings than the average of the control cues C and D.

Table 1.

Design of Experiment 1

Pretraining (additive / nonadditive)	*Training Single*	*Training Compound*	*Test*
X +5	A +5	AB +5	A, B
Y +5		CD +5	C, D
XY +10 / +5	E +5	F +5	E, F, EF
W -	G -		EM, FM
Z -	GH -	GH -	
WZ -	IJ +5	KL -	K
	L-	L -	L

Note: Letters A-L and W-Z denote randomly allocated foods used as predictive cues. These cues were followed by either no allergic reaction (-) or an allergic reaction measuring 5 or 10 on a fictitious severity index (+5 or +10). Participants who were shown additive pretraining witnessed trials with XY +10 trials whereas participants shown nonadditive pretraining witnessed XY +5 trials.

Participants receiving pretraining were given trials on which cues X and Y each individually led to a medium severity (5 out of a possible 10) allergic reaction. The XY trials were followed by either a salient and severe allergic reaction (XY+10) for the additive groups or a mild allergic reaction (XY+5) for the nonadditive groups. Participants in the no-pretraining groups started with the training phases.

Additivity assumptions were examined by individually training two cues, E and F, each with an allergic reaction of severity 5. On test, participants rated the expected outcome severity after both single cues but also the novel compound EF. If outcome additivity is assumed, severity ratings for EF should be higher than the individual ratings of E and F.

The remaining trials in the lower part of Table 1 were included as filler trials, selected so that an allergic reaction occurred overall on half the trials of each phase and after both single and compound cues.

Method

Participants. One hundred and seventy-seven undergraduate psychology students from the University of Sydney participated in the experiment (134 females, mean age = 21.7 years). Forty-four participated for partial course credit and were tested individually in quiet testing cubicles, and the remaining 133 were tested in groups of up to 20 as part of a tutorial experiment. Participants under both testing conditions were allocated to one of six groups, nonadditive forward (n=30), no-pretraining forward (n=30), additive forward (n=29), nonadditive backward (n=30), no-pretraining backward (n=30), additive backward (n=28).

Apparatus and Stimuli. The experiment was programmed using the PsychToolbox for Matlab (Brainard, 1997; Pelli, 1997). The experiment was presented either on Apple Mac Mini computers or, for those who completed the experiment in tutorials, on iMac desktop computers, situated around three walls of a large teaching room. Experimental stimuli included images of a banana, apple, fish, lemon, cheese, milk, coffee, eggs, garlic, bread, pasta, peanuts, avocado, meat, mushrooms, strawberries, peas, rice and olive oil, accompanied by written labels. Single stimuli appeared in the center of the screen, whereas compounds were presented adjacent to one another with each stimulus presented on the left on half of the trials and right on the other half.

Procedure. All participants initially received instructions that they were to assume the role of a doctor whose hypothetical patient (Mr X) had presented with food allergies. Their task was to predict what foods were causing the reactions using trial and error. All participants were instructed to attend to reaction severity when combinations of foods were presented, as this would help in determining which foods were causing allergic reactions. Participants in the additive and nonadditive groups received additional instructions prior to pretraining, displayed in red text to increase their salience, which explicitly outlined the applicable rule (see Supplementary materials).

On each trial, one or two foods were presented and participants predicted what outcome ("no allergic reaction" or "ALLERGIC REACTION") occurred by clicking either option. After this choice, either "CORRECT" in green font or "INCORRECT" in red font appeared, followed by the correct answer, accompanied by a severity index. This entailed a continuous scale indicating minimum (severity index = 0), mild (severity index = 5) or maximum (severity index = 10) severity, with a red bar filling horizontally left-to-right to the magnitude of the outcome. If an allergic reaction occurred, a picture of a face accompanied the severity index. The face was either yellow and looking mildly ill (severity 5 reaction) or green and looking very ill (severity 10 reaction). The green face flashed intermittently for 1.5 seconds to emphasise its severity.

Pretraining consisted of four blocks, each containing two trials each of the six trial types (48 trials in total). In the two training phases, each of the four blocks again contained two trials of six trial types (48 trials in each training phase). Trial presentation order was randomized within blocks in all phases.

In the ratings test, participants were presented with seven individual cues from the training phase and three novel compounds (EF, EM and FM) and were asked to make three separate ratings for each. They were first asked "Does this food cause Mr X to have an allergic reaction?" and were required to make a rating on a linear analogue scale ranging from "definitely DOES NOT cause a reaction" to "definitely DOES cause a reaction". They were also asked to rate "How confident are you that your first rating is correct?" on a scale from "not at all confident" to "very confident". Once these two ratings were made, a third scale appeared and participants were asked to indicate "How severe will the reaction most likely be?" on a scale labeled from 0-10. Each response was made on a linear analogue scale, and causal and confidence ratings were transformed into a score of 0-100. No feedback was given. The order of presentation was randomized, with each stimulus presented only once.

After completing the ratings test, participants were given an additional trial recall test and manipulation check. For brevity, the methods and results of these tests can be found in supplementary material but they will not be discussed further here.

Results

Pretraining and Training Data. Predictions were very accurate and showed fast learning of all cue-outcome contingencies (see Figure S1 in supplementary materials). In the final blocks of pretraining (for those groups that received it), phase 1 training, and phase 2 training, accuracy of outcome predictions was 0.95 or higher for every contingency in all six groups. All participants performed well above chance.

Causal Ratings. Figure 1 shows causal ratings for B and for the mean of control cues C and D. The aim of Experiment 1 was to compare the effects of each form of pretraining to the no pretraining conditions, thereby gaining some indication of how pretraining affects the assumptions brought to bear over learning. To this end, separate analyses compared nonadditive pretraining groups to the no pretraining groups, and the additive pretraining groups to the no pretraining groups. First, comparing nonadditive pretraining to no pretraining, a (2) x 2 x 2 ANOVA with factors of trial type (B vs. C/D), phase order (forward vs. backward blocking) and pretraining condition (no pretraining vs. nonadditive pretraining) yielded significant main effects of trial type, $F(1,116) = 14.43$, $p <.001$, $\eta_p^2 = .111$ and phase order, $F(1,116) = 8.80$, $p = .004$, $\eta_p^2 = .071$, as well as a significant interaction between trial type and phase order, $F(1,116) = 4.60$, $p = .034$, $\eta_p^2 = .038$, indicating that the blocking effect was larger for forward than backward blocking conditions. However, neither the main effect of pretraining, nor any interaction with pretraining approached statistical significance, all *F*s < 1, revealing no evidence that the blocking effect and causal ratings in general were influenced by nonadditive pretraining.

To examine the evidence for blocking in each condition we conducted Bayes factor t-tests using the JASP software developed by Wagenmakers et al. (2018). Each test compared the null of no blocking effect against a directional alternative hypothesis reflecting evidence of blocking (i.e. B<C/D). All Bayesian analyses reported here used a default Cauchy prior scaled at r = 0.707. These analyses revealed anecdotal evidence in favor of the null (i.e. no backward blocking) in both the backward nonadditive pretraining group, $BF_{01} = 1.889$, and backward no pretraining group, $BF_{01} = 2.591$, as well as evidence for blocking in the forward nonadditive group, $BF_{10} = 17.601$, and forward no pretraining group, $BF_{10} = 7.613$. Furthermore, when we compared magnitude of forward blocking in the pretraining groups, evidence favored the null of no difference over the directional alternative that blocking should be more pronounced in the no pretraining group than in the nonadditive group, $BF_{01} = 4.275$.

In the second analysis comparing additive and no pretraining groups, significant main effects were found for trial type, $F(1,113) = 58.40$, $p < .001$, $\eta_p^2 = .341$, pretraining, $F(1,113) = 14.54$, $p = .001$, $\eta_p^2 = .114$, and phase order, $F(1,113) = 7.36$, $p = .008$, $\eta_p^2 = .061$, as well as for the interaction between trial type and pretraining, $F(1,113) = 19.20$, $p < .001$, $\eta_p^2 = .145$, indicating that blocking was larger in the additive pretraining conditions, and between trial type and phase order, $F(1,113) = 4.95$, $p = .028$, $\eta_p^2 = .042$, indicating that blocking was larger for forward than backward conditions. No other interactions were significant, Fs < 1.

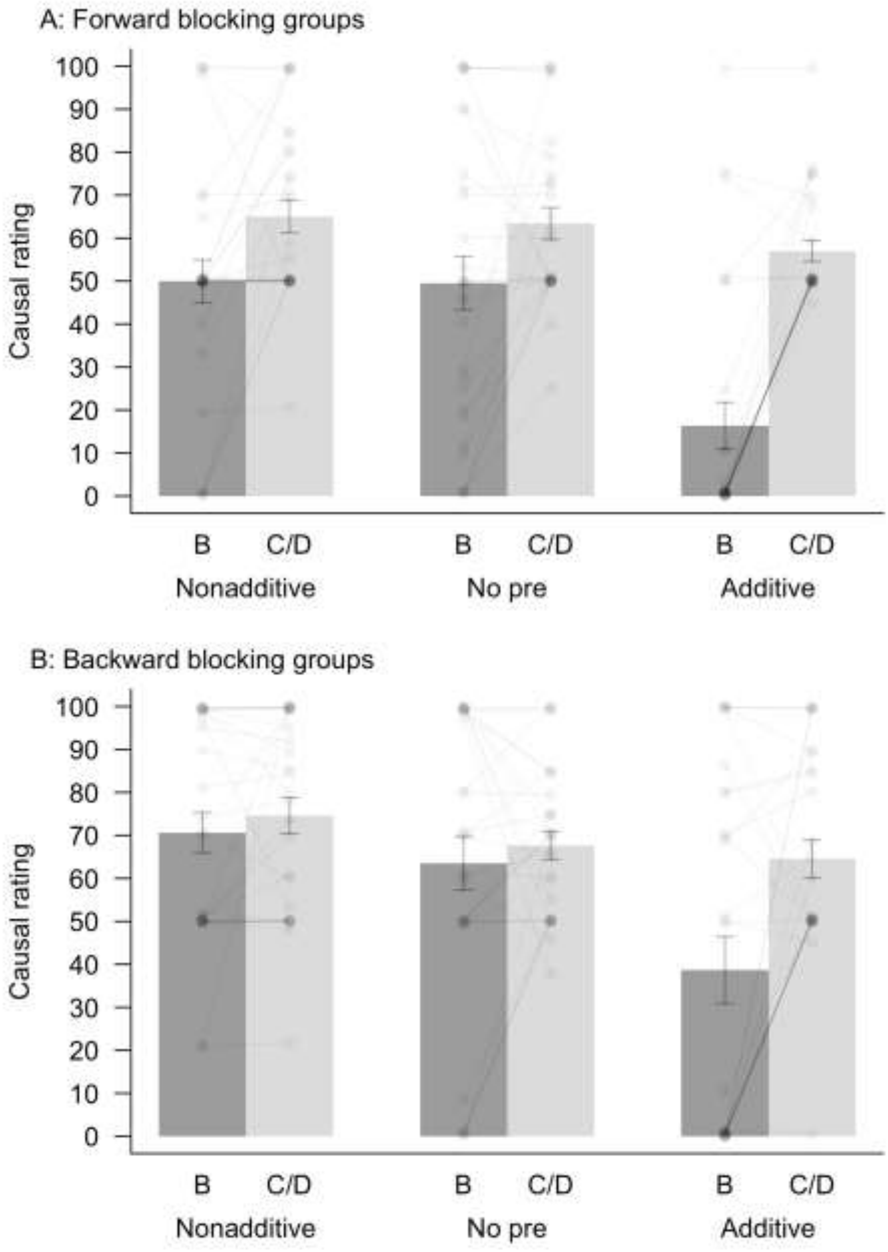

Figure 1. The blocking effect in Experiment 1. Dark grey bars show the causal rating of cue B; the light grey bars show the mean causal rating of cues C and D. Error bars indicate SE and connected data points represent data of individual participants.

Confidence Ratings. Figure 2 shows the confidence ratings for B and for the mean of cues C and D. Two (2) x 2 x 2 ANOVAs of the confidence ratings was conducted using trial type (B vs. C/D) as the within-subjects factor, and phase order (forward vs. backward blocking) and pretraining condition (no pretraining vs. nonadditive/additive) as between-subjects factors. Comparing the nonadditive and no pretraining groups, there were no significant main effects of trial type, pretraining, or phase order, largest $F(1,116) = 2.60$, $p = .110$, $\eta_p^2 = .022$ for main effect of phase order, and no significant interactions, largest $F(1,116) = 1.32$, $p = .253$, $\eta_p^2 = .011$, for the interaction between trial type and pretraining.

Comparing the additive and no-pretraining groups, there was a significant main effect of trial type, $F(1,113) = 22.17$, $p < .001$, $\eta_p^2 = .164$, but non significant main effects of pretraining and phase order, *F*s < 1. There was a significant interaction between trial type and pretraining, $F(1,113) = 12.73$, $p = .001$, $\eta_p^2 = .101$, indicating that the tendency to rate B with higher confidence than C/D was stronger after additive pretraining. There were no further significant interactions, (*Fs*<3.03, *ps*> .085)

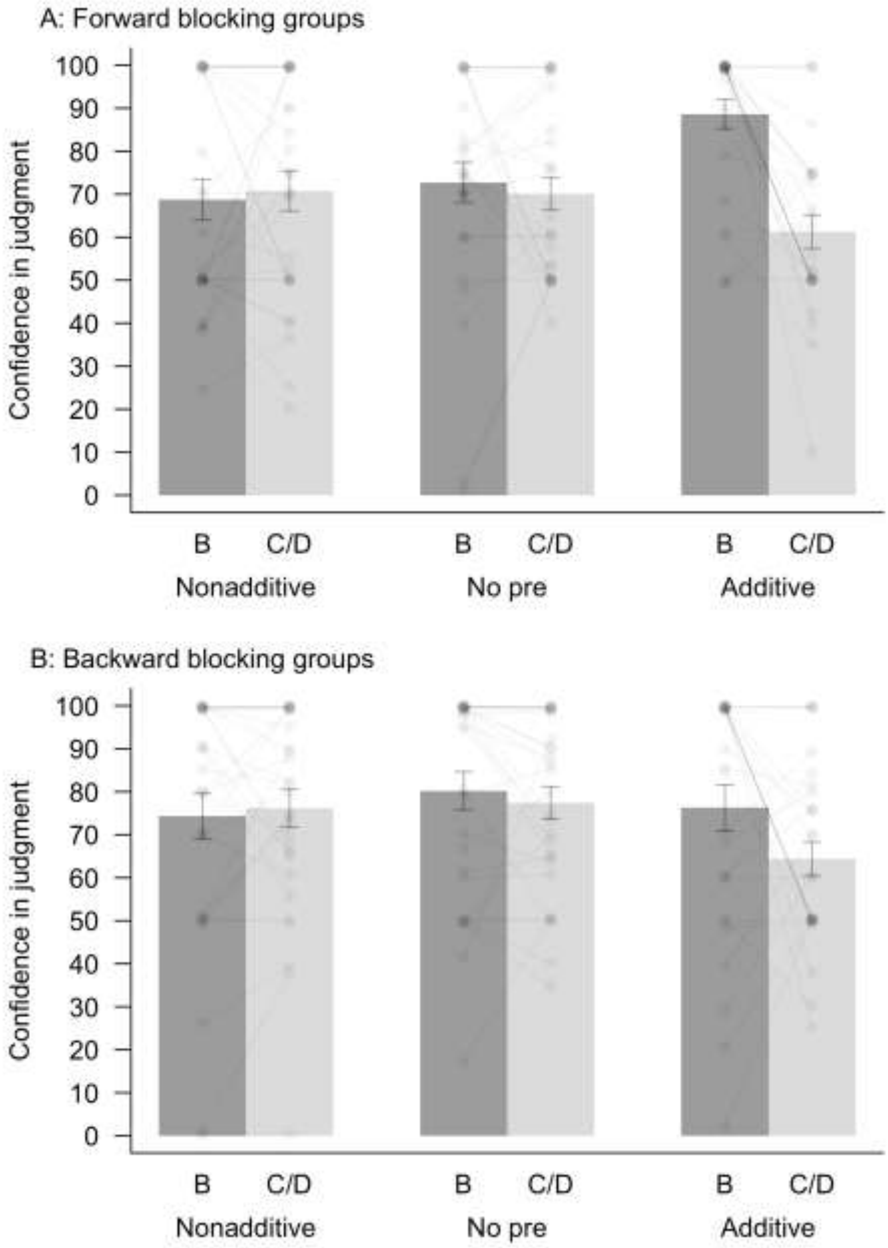

Figure 2. Confidence in causal rating of B vs C and D in Experiment 1. Dark grey bars show the confidence rating of cue B; the light grey bars show the mean confidence rating of cues C and D. Error bars indicate SE and connected data points represent data of individual participants.

Additivity test. Figure 3 shows the expected severity ratings for EF and for the mean of E and F individually. Two (2) x 2 x 2 ANOVAs were conducted using trial type (EF vs. E/F) as the within-subjects factor, and phase order (forward vs. backward blocking) and pretraining condition (no pretraining vs. nonadditive/additive) as between-subjects factors.

Comparing first the nonadditive and no-pretraining groups, there were significant main effects of trial type, $F(1,116) = 98.25$, $p < .001$, $\eta_p^2 = .459$, pretraining, $F(1,116) = 59.61$, $p < .001$, $\eta_p^2 = .339$, but not phase order, $F(1,116) = 2.49$, $p = .117$, $\eta_p^2 = .021$. The interaction between trial type and pretraining was significant, $F(1,116) = 91.61$, $p < .001$, $\eta_p^2 = .441$, indicating that the difference in expected severity between EF and E/F was much greater in the no pretraining groups than the nonadditive groups. There were no further significant interactions, largest $F(1,116) = 1.65$, $p = .202$, $\eta_p^2 = .014$.

Comparing the additive and no pretraining groups, there were significant main effects of trial type, $F(1,113) = 462.63$, $p < .001$, $\eta_p^2 = .804$, pretraining, $F(1,113) = 8.83$, $p = .004$, $\eta_p^2 = .072$, and phase order, $F(1,113) = 4.68$, $p = .033$, $\eta_p^2 = .040$. The interaction between trial type and pretraining was significant, $F(1,113) = 19.69$, $p < .001$, $\eta_p^2 = .148$, indicating that the difference in expected severity between EF and E/F was smaller in the no pretraining groups than the additive groups. All other interactions were non-significant, all *F*s < 1.

Bayes factor t-tests comparing the null (i.e. no additivity assumption) to a directional alternative (higher severity ratings for EF than E/F) were in favor of the null hypotheses for the backward nonadditive pretraining group, $BF_{01} = 5.684$, and to a lesser extent for the forward nonadditive pretraining group, $BF_{01} = 1.756$. All other groups showed strong evidence of assumed additivity in the severity ratings, smallest $BF_{10} = 2.969 \times 10^{15}$.

To further examine the relationship between assumed additivity and blocking in the absence of pretraining, we examined correlations between blocking scores (mean causal rating for C/D – rating for B) and additivity scores (predicted severity for EF – mean for E/F), using the two no pretraining groups. This correlation was weak and not significant for the forward group, $r(28) = .027$, $p = .887$, the backward group, $r(28) = .008$, $p = .968$, and combined, $r(58) = .028$, $p = .831$.

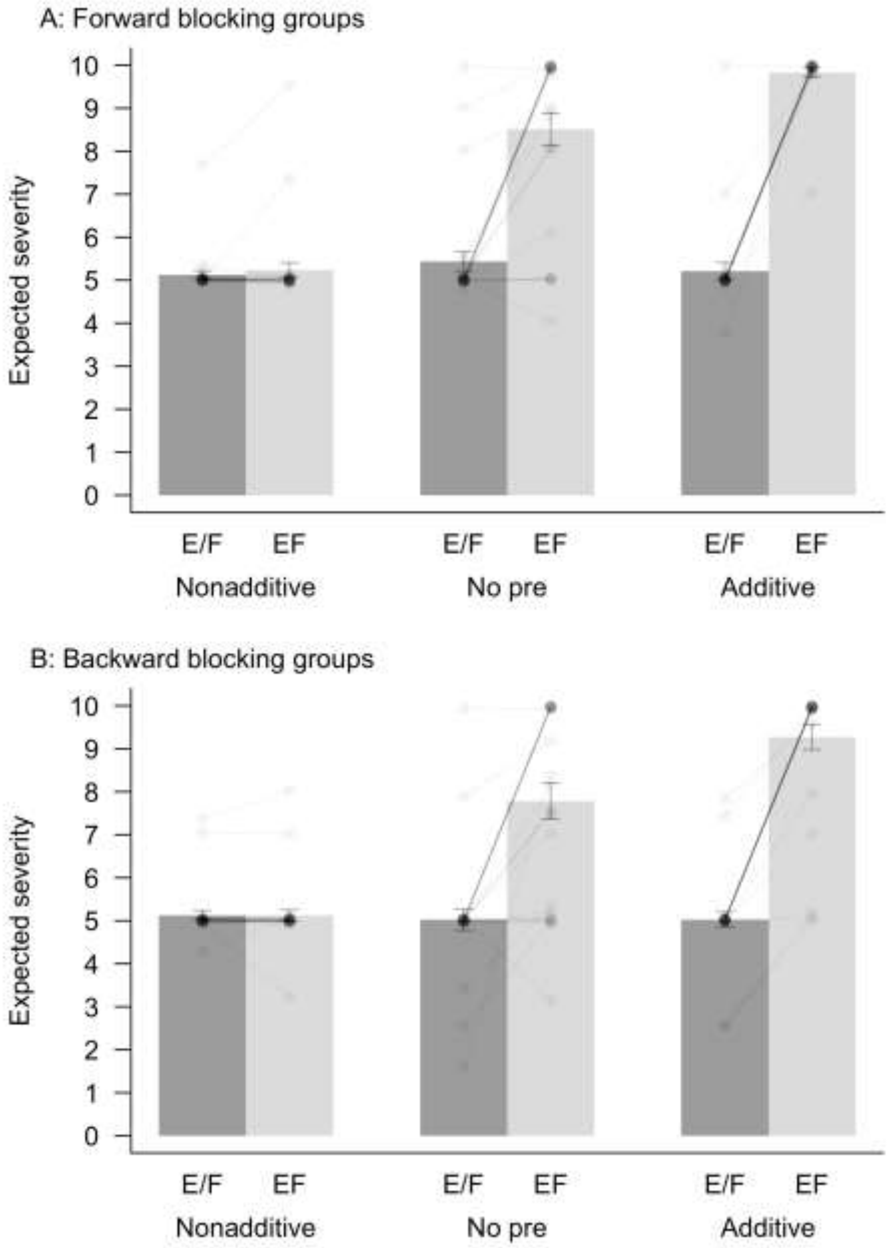

Figure 3. Additivity test in Experiment 1. Dark grey bars show the mean severity rating of training cues E and F; the light grey bars show the severity rating of the new test compound EF. Error bars indicate SE and connected data points represent data of individual participants.

Discussion

We replicated the additivity pretraining effect; consistent with the findings of Lovibond et al. (2003), blocking was considerably stronger in both additive groups. This pattern held even though we used a more conservative control group that received no pretraining. Additive pretraining significantly enhanced the blocking effect in causal ratings and, as predicted, significantly increased confidence for B relative to C/D. This is consistent with Vandorpe et al. (2005), who found that blocking enhanced by a strongly emphasised submaximal outcome was accompanied by greater confidence for the blocked cue. The result suggests that participants after additive pretraining deduce that B is not causal and thus can make a more confident and definitive judgment than for either C or D, for which the causal status remains uncertain. In line with this interpretation, additive pretraining successfully encouraged assumptions of outcome additivity expressed in severity ratings of the additivity test.

We also replicated an interesting aspect of Lovibond et al.'s (2003) data, that the nonadditive pretraining did not abolish the forward blocking effect. Indeed, it was of comparable magnitude to the no pretraining group. The deductive inferential hypothesis may account for this effect by assuming that some participants in these groups still entertain additivity assumptions. Indeed, the results of the additivity test suggest that outcome additivity is assumed by default in this task by a substantial proportion of participants, as suggested by Beckers et al. (2005). However, this assumption was successfully removed by nonadditive pretraining without having any effect on blocking. Inferential accounts would predict that the removal of assumed additivity among the nonadditive participants would eliminate blocking or, at the very least, reduce the magnitude of the effect. Further, our indices of blocking and the additive magnitude assumption were uncorrelated in the no-pretraining groups, suggesting that blocking under these conditions had little to do with the assumption of additivity. These results indicate that some of the observed blocking effect does not rely on inferential reasoning that requires additivity assumptions. This therefore contradicts suggestions that blocking observed in causal learning is typically based on such deductions (e.g. Beckers et al., 2005; Lovibond et al., 2003).

Backward blocking was substantially weaker than forward blocking across all pretraining conditions, including under additive pretraining. For each pretraining condition, the inferential reasoning processes operating at the time of test should be the same in the forward and backward conditions because the assumptions established during pretraining are identical and the order in which that ensuing information came to light is not relevant to the reasoning process. One explanation is that participants in the forward blocking conditions have greater opportunity to make appropriate inferential judgments during training. These participants are already aware of the causal status of A when they first make judgments about AB+ and thus may deduce the status of B when it is actually presented. This account would explain why forward blocking was larger than backward blocking after additive pretraining but still does not explain why forward blocking was larger than backward blocking after nonadditive pretraining (where the deductive inference account argues there should be no blocking).

The lack of any convincing evidence of backward blocking in the no pretraining and nonadditive pretraining groups is consistent with some previous failures to observe backward blocking in similar studies (e.g. Larkin, Aitken, & Dickinson, 1998). The sensitivity of blocking to trial order is anticipated by associative learning models in which a summed error term (Rescorla & Wagner, 1972) or selective associability (Le Pelley, 2004; Mackintosh, 1975) are assumed to play an important role in determining cue-outcome contingency learning. However, these models do not readily explain the effect of additive pretraining. One hypothesis to reconcile these two facets of the results is that participants often use the strength of the cue-outcome association to make causal judgments, but given the conditions necessary for *modus tollens*, some participants use deduction instead. Those that do use deduction will usually show forward and backward blocking, typified by lower and more confident causal ratings of B. Those who do not use deduction will tend to show a forward blocking effect much more readily than a backward blocking effect, for reasons that are well captured by associative learning models.

Experiment 2

Experiment 1 revealed for the first time that nonadditive pretraining reduces the assumption of outcome additivity but *not* the blocking effect. This result is of particular theoretical importance. However, participants in these conditions never witnessed a strong outcome, for instance one with the maximal severity of 10. It is possible that this fairly artificial and binary use of the severity scale contributed to there being little to separate the conditions in terms of blocking. The expected severity ratings made by participants in the no pretraining conditions suggest that a substantial proportion of participants consider the outcome to be both submaximal and additive (mean severity ratings for EF were substantially above the experienced severity 5). However, these ratings were made on test trials when the ratings scale itself made it obvious that the severity scale extends higher than the outcome severity that participants witnessed during training. Even though the severity index graphically implied that a more severe outcome was possible, one might still argue that the submaximality of the outcome was not sufficiently obvious during training to result in the correct inferential reasoning (see De Houwer et al., 2002; Beckers et al., 2005; Vandorpe et al., 2005).

Perhaps participants in the no pretraining group were not considering the potential submaximality of the outcome during training, and thus although they possessed the information necessary to deduce that the blocked cue B was *not* causal, they did not explicitly form the logical premises on which such a deduction would be made. In some respects, our argument assumes this to be the case since we have concluded thus far that participants are disinclined to use deductive reasoning unless explicitly encouraged to do so (e.g. via additivity pretraining).

The question then remains whether we underestimated the blocking effect in the no pretraining group and whether the blocking effect would still be equivalent in nonadditive and no pretraining conditions if continuous reminders of the submaximality of the outcome were given throughout training. In Experiment 2, an extra cue occurred on its own, followed either by a maximal outcome (M +10) or by the same submaximal outcome used for the other causal cues (M +5). Participants observing a maximal single cue should be reminded that the outcome following A, AB, and CD is submaximal. Thus, if deduction is readily used under assumptions of submaximality, participants with no pretraining and the maximal outcome (M +10) should have a greater propensity to deduce that B is not causal and thus show a stronger blocking effect. However, it is quite possible that participants in the no pretraining conditions already assume that the outcome is submaximal but simply do not use deductive reasoning. In this instance, we would expect to see no difference in the magnitude of blocking revealed in the presence of a maximal or submaximal single cue.

Method

Participants. Ninety-seven first year psychology students from the University of Sydney participated in return for partial course credit (75 females, mean age = 19.3 years). Participants were randomly allocated according to arrival time to the no pretraining, maximal single cue group (*n*=24), nonadditive pretraining, maximal single cue group (*n*=26), no pretraining, submaximal single cue group (*n*=23), or nonadditive pretraining, submaximal single cue group (*n*=24).

Apparatus and Stimuli. The apparatus and stimuli used were identical to Experiment 1 except that all testing was conducted in the laboratory.

Procedure. The same scenario and instructions from Experiment 1 were used, but with the following changes. As shown in Table 2, two trial types were added to Phase 1 and 2 of training. These trials involved the presentation of the causal cue M followed by an outcome of severity 10 in the maximal single cue groups and an outcome of severity 5 in the submaximal single cue conditions, and an additional non-causal cue N, added to keep the base rate probability of an allergic reaction the same as in previous experiments. Compounds EM and FM were removed from the test phase and replaced by cues M and N.

Table 2.

Design of Experiment 2

Pretraining	Training Phase 1	Training Phase 2	Test
X +5	A +5	AB +5	B
Y +5		CD +5	C, D
XY +5	E +5	F +5	E, F, EF
W -	G -		
Z -	GH -	GH -	H
WZ -	IJ +5	KL -	
	L -	L -	L
	M +5 / +10	M +5 / +10	M
	N -	N -	N

Note: Letters A-N and W-Z denote randomly allocated foods used as predictive cues. These cues were followed by either no allergic reaction (-) or an allergic reaction measuring 5 or 10 on a fictitious severity index (+5 or +10). Participants in the Max Single groups witnessed M +10 trials whereas participants in the Submax Single group witnessed M +5 trials.

Results

Training Data. Predictions during training were very accurate and showed rapid acquisition of the cue-outcome contingencies (see supplementary materials, Figure S2). By the final blocks of phase 1 and phase 2 training, both groups performed higher than 0.93 accuracy for every cue-outcome contingency. The nonadditive pretraining groups also performed higher than 0.95 for every contingency in the final block of pretraining. All participants performed well above chance.

Causal Ratings. Figure 4 shows causal ratings for B and for the mean of control cues C and D. A (2) x 2 x 2 ANOVA of the causal ratings was conducted using trial type (B vs.

C/D) as the within-subjects factor, and single cue (maximal vs. submaximal) and pretraining condition (nonadditive vs. no pretraining) as between-subjects factors. The main effect of trial type was significant, $F(1,93) = 24.39$, $p < .001$, $\eta_p^2 = .208$, indicating overall lower ratings for B than for C/D. Neither pretraining nor single cue factors interacted with trial type, *F*s < 0.1 and there was no 3 way interaction, $F < 1$, indicating that the magnitude of blocking was roughly equal in all groups. However the main effect of pretraining, $F(1,93) = 4.05$, $p = .047$, $\eta_p^2 = .042$, and main effect of single cue, $F(1,93) = 3.98$, $p = .049$, $\eta_p^2 = .041$, were both significant, indicating, respectively, that ratings were slightly higher overall after nonadditive pretraining and in the presence of the M+5. There was no interaction between pretraining and single cue, $F(1,93) = 1.45$, $p = .231$ $\eta_p^2 = .015$.

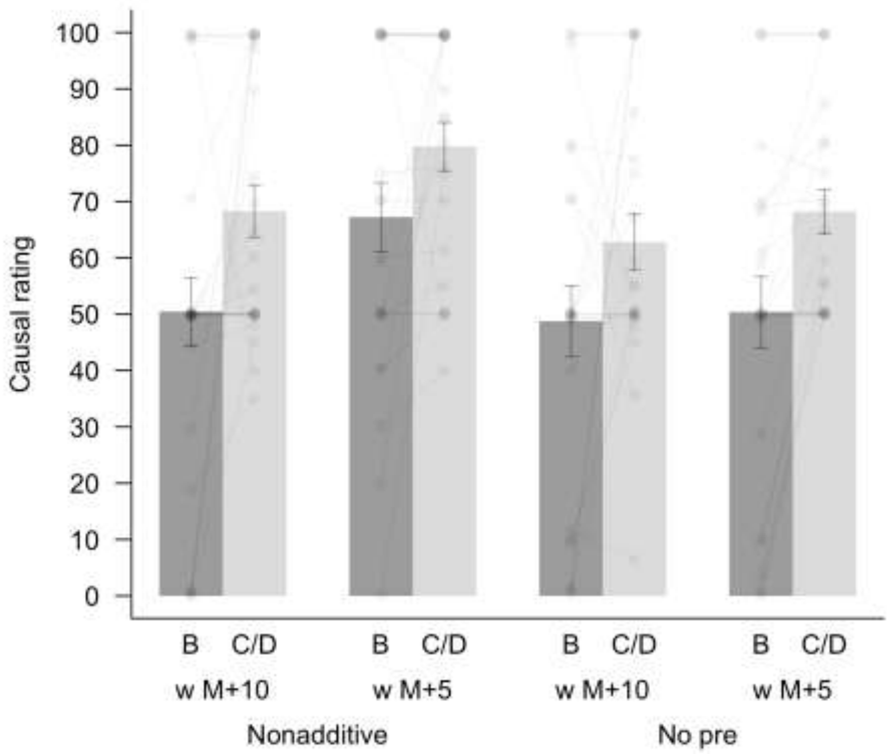

Figure 4. The blocking effect in Experiment 2. Dark grey bars show the causal rating of cue B; the light grey bars show the mean causal rating of cues C and D. "w M+10" refers to groups who completed training with an additional cue that resulted in a maximum severity (10) allergic reaction; "w M+5" refer to groups who received the additional cue paired with submaximal severity (5) reaction. Error bars indicate SE and connected data points represent data of individual participants.

To examine the relative magnitude of the observed blocking effects, we conducted Bayes factor t-tests on the blocking scores comparing nonadditive and no pretraining groups in each of the single-cue conditions (maximal and submaximal). We again compared the null against a directional hypothesis that blocking should be more pronounced in the no pretraining than in the nonadditive groups. We found evidence that the blocking effects did not differ according to nonadditive vs no pretraining for the maximal single cue conditions, $BF_{01} = 4.521$. Note that this is the condition in which M+10 serves as a frequent reminder that the outcome is *submaximal* for the blocking compound, and thus we would expect these conditions to provide the best opportunity to observe a difference in blocking between nonadditive and no pretraining. The evidence that the blocking effects were equivalent in the submaximal single cue conditions, $BF_{01} = 1.864$, was more equivocal and cannot be said to favor either hypothesis when considered in isolation, but note that this result still effectively replicates Experiment 1, where the evidence for the null was somewhat stronger.

Confidence Ratings. A (2) x 2 x 2 ANOVA of the confidence ratings was conducted using trial type (B vs. C/D) as the within-subjects factor, and single cue (maximal vs. submaximal) and pretraining condition (nonadditive vs. no pretraining) as between-subjects factors. The main effect of trial type was significant, $F(1,93) = 6.55$, $p = .012$, $\eta_p^2 = .066$. This indicates that, overall, confidence was marginally lower for B (M = 68.9) than for C/D (M = 75.6). There were no significant interactions with trial type, largest $F(1,93) = 1.20$, $p = .277$, $\eta_p^2 = .013$, and no further main effects or interactions, largest $F(1,93) = 2.03$, $p = .157$, $\eta_p^2 = .021$. Means for individual conditions and groups are reported in supplementary materials (see Table S3 and Figure S3).

Additivity Test

Figure 5 shows the expected severity for EF versus the mean of E and F individually. A (2) x 2 x 2 ANOVA of the severity ratings was conducted using trial type (EF vs. E/F) as the within-subjects factor, and single cue (maximal vs. submaximal) and pretraining condition (nonadditive vs. no pretraining) as between-subjects factors. The main effect of trial type, $F(1,93) = 102.21$, $p < .001$, $\eta_p^2 = .524$, and pretraining condition, $F(1,93) = 28.44$, $p < .001$, $\eta_p^2 = .234$, were significant, as was their interaction, $F(1,93) = 48.31$, $p < .001$, $\eta_p^2 = .342$. This interaction indicates that the difference in severity ratings for EF vs E/F was much larger for the no pretraining groups than for the nonadditive pretraining groups. The main effect of single cue approached conventional levels of significance, $F(1,93) = 3.14$, $p = .080$, $\eta_p^2 = .033$, but this factor did not interact significantly with any other, largest $F(1,93) = 1.43$, $p = .234$, $\eta_p^2 = .015$.

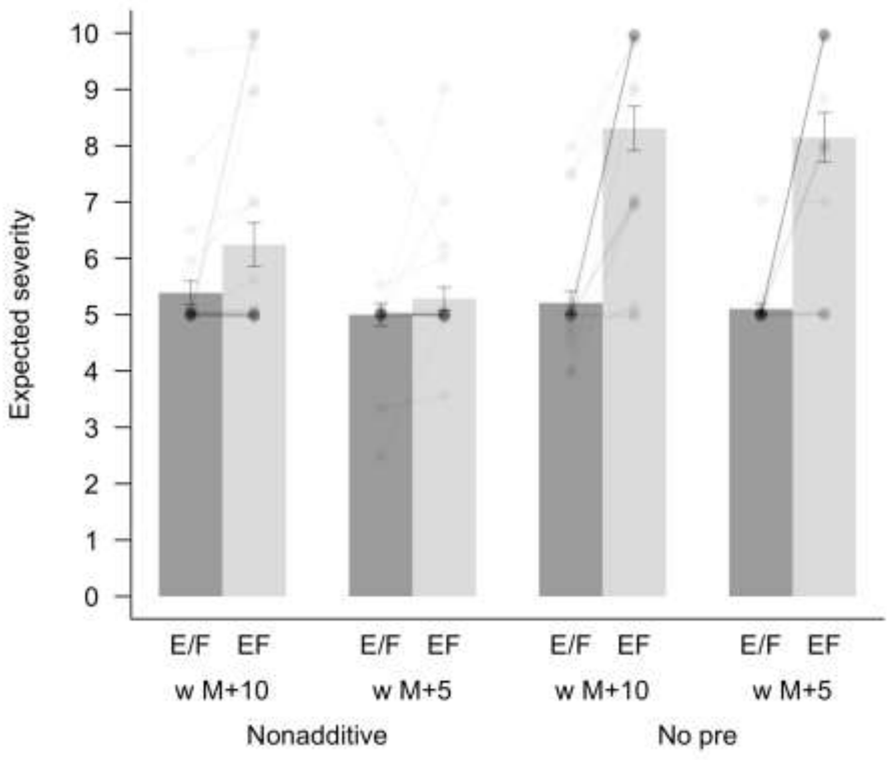

Figure 5. Additivity test in Experiment 2. Dark grey bars show the mean severity rating of training cues E and F; the light grey bars show the severity rating of the new test compound EF. Error bars indicate SE and connected data points represent data of individual participants.

Bayes factor t-tests comparing the null of no difference in severity ratings to an alternative that severity ratings should be higher for the compound EF than for the individual cues E and F, revealed strong evidence for assumed additivity in the no pretraining submaximal and no pretraining maximal groups, BF_{10} = 32963 and BF_{10} = 652621 respectively. The nonadditive maximal group also displayed evidence of assumed additivity, albeit weaker than the evidence in the no pretraining groups, BF_{10} = 5.903. Conversely, the nonadditive submaximal group displayed anecdotal evidence in favor of the null, BF_{01} = 1.383.

As in Experiment 1, we examined correlations between blocking scores (mean causal rating for C/D – rating for B) and additivity scores (predicted severity for EF – mean for E/F), in the two no pretraining groups. Again, the correlation was weak and not significant for the maximal single cue group, $r(22) = -.070$, $p = .744$, the submaximal single cue group, $r(21) = .120$, $p = .585$, and combined, $r(45) = .004$, $p = .978$.

Discussion

In Experiment 2 we added a reminder cue with maximal outcome severity to the standard blocking procedure. This addition could have potentially served as a reminder about the submaximality of the outcome and thus may have increased the propensity for deductive reasoning, particularly in the no pretraining condition. However, the addition of this maximal cue did not increase blocking and did not increase the confidence of participants' ratings about cue B, and thus gave no indication that there was any deductive reasoning in this group or indeed any shift towards the results observed after explicit additive pretraining. The addition of the single cue with maximal outcome had a negligible effect on ratings of expected severity in the no pretraining groups, though it may have had a small effect on severity ratings in the nonadditive groups. This suggests that the presence of the maximal cue may have slightly undermined the nonadditive pretraining, resulting in a small difference between the ratings for EF versus E/F. On the basis of these results it seems reasonable to assume that participants in the no pretraining group already held an assumption of outcome submaximality even without the added reminder cue.

Many causal learning experiments that have demonstrated a robust blocking effect have used outcomes with no magnitude information at all. In contrast, even without the M+10 reminder, our procedure involves a graphical depiction of a continuous variable that extends beyond the magnitude shown in training. Therefore, in the absence of pretraining, our procedure is probably slightly biased towards encouraging assumptions that the outcome is submaximal and, by extension, additive in nature. The proportion of participants holding these assumptions may therefore be relatively high in these experiments compared to many others. It might well be the case that the current blocking results, even those observed with nonadditive pretraining, were in part dependent on the submaximality of the outcome that we used. However, even if this is true, it is not plausible in this case that submaximality produced blocking by encouraging an assumption of outcome additivity and thereby encouraging engagement in deductive reasoning based on this assumption. Removal of this assumption by nonadditive pretraining made no difference to the average magnitude of the blocking effect.

General Discussion

The most common explanation for why additive pretraining enhances blocking is that it encourages the use of deduction (e.g. Lovibond et al., 2003; Beckers et al., 2005). Our results are consistent with this interpretation; under additive pretraining, participants were more confident about their causal rating for B and showed relatively strong forward and backward blocking effects. On the basis of this and other examples of the additivity pretraining effect, one could hypothesise that since additive pretraining enhances blocking, most examples of blocking in human learning may be due to the use of deductive reasoning by a subset of participants who assume outcome additivity even in the absence of explicit pretraining or instruction. Our results, however, suggest that this hypothesis is incorrect.

In both experiments, nonadditive pretraining dramatically reduced expectation that the outcome would be more severe when two causal cues were combined, and yet produced levels of blocking that were almost identical to groups given no pretraining. Furthermore, possessing an additive assumption about outcome magnitude, expressed in expected severity ratings, was not associated with a larger blocking effect in the groups who received no pretraining. The blocking effect typically observed in causal learning scenarios is robust in the absence of any additive assumptions and is not based on the same form of deduction as blocking after additive pretraining. Thus, even though our results corroborate that blocking after additive pretraining probably involves deduction, we argue that this form of inferential reasoning is not typical of the decisions made by most participants in causal learning experiments. While a minority of participants might still use deduction without explicit encouragement, the blocking effect is not restricted to the judgments of these individuals.

Lovibond et al., 2003 (see also Beckers et al., 2005; Mitchell & Lovibond, 2002) argue that, when the outcome is explicitly nonadditive, the natural inference made about the blocked cue is that its causal status is ambiguous and that it should be given the same causal rating to control cues that have been observed in compound only. Whether or not this inference is logically correct[1], the form of reasoning from which it is derived is very different to the deduction encouraged by additivity pretraining. Since the nonadditive and no

[1] According to an analysis of the problem in terms of classical probability theory, this conclusion would actually be incorrect. The blocked cue still has a lower probability of being a cause of the outcome than either of the overshadowing control cues (see Livesey, Lee & Shone, 2013 and McCormack, Butterfill, Hoerl, & Burns, 2009).

pretraining groups displayed equivalent and statistically significant blocking effects, it seems that this form of reasoning is not strongly encouraged by nonadditive pretraining. Participants may find deductive inference easier to apply provided they are given instructions and pretraining that explicitly encourage its use. Without this explicit encouragement, our participants did not appear to have a natural proclivity to engage in deductive reasoning just because they held assumptions that the outcome magnitude was additive.

These experiments tested a hypothesis derived from specific inferential reasoning processes and found a pattern of results that is inconsistent with that hypothesis. We argue that the two reasoning processes outlined in previous literature on additivity pretraining (one conducive to blocking, the other not) are actually uncommon in human blocking and that, in the absence of additivity pretraining, deductive reasoning is not the main cause of the blocking effect. This leaves an obvious question; what accounts *are* consistent with these results? Theories that view causal reasoning as a process of induction do not necessarily rely on the same set of assumptions as the deductive hypothesis tested here. Waldmann (2007), for instance, proposed that individuals could assume that the causal effects of two cues either sum or average depending on the causal model that they apply based on their prior knowledge, and this will affect the extent to which they show blocking. This account thus explains most of the additivity pretraining effects reported in previous studies without appealing to explicit deductive reasoning as an explanatory mechanism. The challenge, however, for this and other accounts that focus on the assumed or learned functional form of the causal relationships is that it still fails to address the lack of association between assumed additivity and blocking observed across our experiments. The current results are, however, consistent with a hypothesis that participants will use deductive reasoning when strongly encouraged to do so but usually (for instance, in the absence of explicit instructions or pretraining) base their causal judgments on a psychologically more intuitive decision such as the strength to which a cue brings to mind an outcome (e.g. see Le Pelley, Griffiths & Beesley, 2017; López, Cobos & Caño, 2005; Thorwart & Livesey, 2017). This hybrid approach calls upon associative memory to inform inductive inferential reasoning but leaves open the possibility that judgments about cause and effect will be made on the basis of other decision processes in circumstances that encourage the use of alternative modes of reasoning.

Explanations of causal learning in terms of associative retrieval and simple inferential reasoning (including the deductive reasoning hypothesis investigated here) share in common a desire to specify a psychological mechanism responsible for the explicit judgments made about cause and effect. In this respect, they differ from theories whose main aim is to formalize the problem space involved in making such decisions (e.g. Tenenbaum, Griffiths and Kemp, 2006). Functional and mechanistic theories of causal learning are not mutually exclusive and they may in fact be complementary, provided their aims are clear.

Like other formal approaches to causal learning (e.g. Griffiths & Tenenbaum, 2009, Waldmann, 2007), the deductive inferential hypothesis provides a natural explanation for why blocking is less reliable in some causal learning scenarios than others. We have deliberately chosen to use a task that lends itself to producing blocking and acknowledge that other scenarios and tasks do not show blocking so readily. This might be because some tasks encourage a causal model or a form of reasoning that does not lend itself to cue competition (Waldmann & Holyoak, 1992). Alternatively, it might be for reasons to do with the learning of the association between the cues and outcomes, and in particular the encoding of the cues (Haselgrove, 2010; Livesey & Boakes, 2004). Maes et al. (2016) have recently highlighted that even in traditional animal models of conditioning, the boundary constraints on the blocking effect are not clearly defined, leading to the suggestion that the effect may be more elusive than once thought. In this respect, it is noteworthy that we find the blocking effect in human learning to be highly consistent when using the task parameters reported in this study. It remains a challenge for future research to identify the conditions that permit a replicable blocking effect, but our research suggests that it will not simply be a matter of whether outcome additivity assumptions are encouraged by the task. Consistent with this, blocking has been observed in tasks where the additive properties of the outcome are arguably irrelevant, for instance tasks that have no causal scenario (e.g. Luque et al., 2018) and tasks that imply that the cue is *caused by* the outcome (Don et al., 2018).

In conclusion, our results highlight that although additive pretraining enhances blocking, probably by encouraging deductive reasoning about the blocked cue, this does not mean that individuals readily engage in deductive reasoning when learning about cause and effect in other situations. Removing additivity assumptions through nonadditive pretraining had no influence on a statistically reliable blocking effect and we found no evidence that blocking in the absence of pretraining was related to the participants' assumptions about the additivity of magnitude of the outcome. This suggests that the presence of the additivity assumption is not a major factor in producing the blocking effect unless deduction is strongly encouraged.

References

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

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*, 238–49.

Brainard, D. H. (1997). The Psychophysics Toolbox. *Spatial Version, 10,* 433-436.

Cheng, P. W. (1997). From Covariation to Causation: A Causal Power Theory. *Psychological Review*, *104*, 367-405.

De Houwer, J., Beckers, T., & Glautier, S. (2002). Outcome and cue properties modulate blocking. *Quarterly Journal of Experimental Psychology, 55A*, 965-985.

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

Don, H. J., & Livesey, E. J. (2018). Is the blocking effect sensitive to causal model? It depends how you ask. In T.T. Rogers, M. Rau, X. Zhu, & C. W. Kalish (Eds.), *Proceedings of the 40th Annual Conference of the Cognitive Science Society* (pp. 1630-1635). Austin, TX: Cognitive Science Society.

Griffiths, O. & Mitchell, C. J. (2008) Selective Attention in Human Associative Learning and Recognition Memory. *Journal of Experimental Psychology: General, 137*, 626-648.

Griffiths, T. L., Sobel, D., Tenenbaum, J. B., & Gopnik, A. (2011). Bayes and Blickets: Effects of knowledge on causal induction in children and adults. *Cognitive Science, 35(8)*, 1407-1455.

Griffiths, T. L., & Tenenbaum, J. B. (2009). Theory-Based Causal Induction. *Psychological Review*, *116*, 661-716.

Griffiths, T. L., & Tenenbaum, J. B. (2005). Structure and strength in causal induction. *Cognitive Psychology, 51,* 334-384.

Haselgrove, M. (2010). Reasoning Rats or Associative Animals? A Common-Element Analysis of the Effects of Additive and Subadditive Pretraining on Blocking. *Journal of Experimental Psychology: Animal Behavior Processes, 36*, 296–306.

Kamin, L. J. (1968). “Attention-like” processes in classical conditioning. In M. R. Jones (Ed.), *Miami symposium on the prediction of behavior: Aversive stimulation* (pp. 9–31). Miami, FL: University of Miami Press.

Larkin, M. J., Aitken, M. R. F., & Dickinson, A. (1998). Retrospective revaluation of causal judgments under positive and negative contingencies. *Journal of Experimental Psychology: Learning, Memory, & Cognition, 24,* 1331-1352.

Le Pelley, M. E. (2004). The role of associative history in models of associative learning: A selective review and a hybrid model. *Quarterly Journal of Experimental Psychology, 57B*, 193–243.

Le Pelley, M. E., Griffiths, O., & Beesley, T. (2017). Associative accounts of causal cognition. In M. R. Waldmann (Ed.), The Oxford Handbook of Causal Reasoning (pp.13-28). Oxford, UK: Oxford University Press.

Livesey E. J., & Boakes R. A. 2004. Outcome additivity, elemental processing and blocking in human causality judgements. *Quarterly Journal Experimental Psychology, 57B*, 361-79.

Livesey, E. J., Lee, J. C., & Shone, L. (2013). The relationship between blocking and inference in causal learning. In M. Knauff, M. Pauen, N. Sebanz, & I. Wachsmuth (Eds.), Proceedings of the 35th Annual Conference of the Cognitive Science Society. Austin, TX: Cognitive Science Society. (2920-2925).

López, F. J., Cobos, P. L., & Caño, A. (2005). Associative and causal reasoning accounts of causal induction: Symmetries and asymmetries in predictive and diagnostic inferences. *Memory & Cognition, 33*, 1388-1398.

Lovibond P. F., Been S.-L., Mitchell C. J., Bouton M. E., & Frohardt R. (2003). Forward and backward blocking of causal judgment is enhanced by additivity of effect magnitude. *Memory & Cognition 31*, 133–42.

Luque, D., Vadillo, M. A., Gutiérrez-Cobo, M. J., & Le Pelley, M. E. (2016). The blocking effect in associative learning involves learned biases in rapid attentional capture. *The Quarterly Journal of Experimental Psychology, 71*, 522-544

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

Maes, E., Boddez, Y., Alfei, J. M., Krypotos, A. M., D'Hooge, R., De Houwer, J., & Beckers, T. (2016). The elusive nature of the blocking effect: 15 failures to replicate. *Journal of Experimental Psychology: General*, *145*, e49-e71.

McCormack, T., Butterfill, S., Hoerl, C., & Burns, P. (2009). Cue Competition Effects and Young Children's Causal and Counterfactual Inferences. *Developmental Psychology, 45,* 1563-1575.

Mitchell C. J., De Houwer J., & Lovibond P. F. (2009). The propositional nature of human associative learning. *Behavioral and Brain Science*, *32*, 183-246.

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

Mitchell C.J., Lovibond P. F., & Condoleon, M. (2005). Evidence for deductive reasoning in blocking of causal judgments. *Learning and Motivation. 36*, 77-87.

Mitchell C.J., Lovibond P. F., Minard E, & Lavis Y. (2006). Forward blocking in human learning sometimes reflects the failure to encode a cue-outcome relationship. *Quarterly Journal of Experimental Psychology 59*, 830-844.

Novick, L. R., & Cheng, P. W. (2004). Assessing interactive causal influence. *Psychological Review*, *111(2)*, 455-485.

Pelli, D. G. (1997). The Video Toolbox software for visual psychophysics: Transforming numbers into movies. *Spatial Vision, 10*, 437-442.

Rescorla, R. A., & Wagner, A. R. (1972). A theory of Pavlovian conditioning: Variations in the effectiveness of reinforcement and nonreinforcement. In A. H. Black & W. F. Prokasy (Eds.), *Classical conditioning II: Current research and theory* (pp. 64-99). New York: Appleton-Century-Crofts.

Shanks, D. R. (2007). Associationism and cognition: Human contingency learning at 25. *Quarterly Journal of Experimental Psychology, 60,* 291–309.

Sanbonmatsu, D. M., Akimoto, S. A. & Gibson, B. D. (1994). Stereotype-Based Blocking in Social Explanation. *Personality and Social Psychology Bulletin*, 20, 71-81. doi:10.1177/0146167294201007.

Tenenbaum, J. B., Griffiths T. L., & Kemp, C. (2006). Theory-based Bayesian models of inductive learning and reasoning. *Trends in Cognitive Sciences, 10(7),* 309-318.

Thorwart, A. & Livesey, E. L. (2017). Three Ways That Nonassociative Knowledge May Affect Associative Learning Processes. *Frontiers in Psychology*, *7,* 1-15.

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

Vandorpe, S., De Houwer, J., & Beckers, T. (2005). Further evidence for the role of inferential reasoning in forward blocking. *Memory & Cognition*, *33*, 1047-1056.

Wagenmakers, E. J., Love, J., Marsman, M., Jamil, T., Ly, A., Verhagen, J., … Morey, R.D. (2018). Bayesian inference for psychology. Part II: Example applications with JASP. *Psychonomic Bulletin & Review, 25,* 58-76.

Waldmann, M. R. (2007). Combining Versus Analyzing Multiple Causes: How Domain Assumptions and Task Context Affect Integration Rules. *Cognitive Science*, *31*, 233-256.

Waldmann, M. R. & Holyoak, K. J. (1992). Predictive and Diagnostic Learning Within Causal Models: Asymmetries in Cue Competition. *Journal of Experimental Psychology: General*, *121*, 222-236.

Waldmann, M. R. & Walker, J. M. (2005). Competence and performance in causal learning. *Learning & Behavior*, *33*, 211-229.

Open Practices Statements

None of the experiments in this study were formally pre-registered. The data for these experiments will be made available on the Open Science Framework at https://osf.io/sk2n9/.

Author note

This research was supported by a postdoctoral research fellowship of the German Academic Exchange Service (DAAD) to Anna Thorwart and a University of Sydney Bridging Support Grant to Evan Livesey.

Supplementary material for:

Testing the deductive inferential account of blocking in causal learning.

Evan J. Livesey, Justine K. Greenaway, Samantha Schubert, & Anna Thorwart

Pretraining Instructions administered to the additive and non-additive pretraining conditions in Experiments 1 & 2. These instructions were accompanied by extensive trial-by-trial pretraining of relevant contingencies, as outlined in the main article.

Additivity instructions: "Throughout this experiment, it is important that you remember the following: If two foods that produce a mild allergic reaction when eaten alone are eaten together, they WILL produce a more severe allergic reaction. That is, the foods will cause a stronger allergic reaction when eaten together as they each did when eaten separately".

Nonadditivity instructions**:** "Throughout this experiment, it is important that you remember the following: If two foods that produce a mild allergic reaction when eaten alone are eaten together, they will NOT produce a more severe allergic reaction. That is, the foods will cause the SAME mild allergic reaction when eaten together as they each did when eaten separately".

Experiment 1

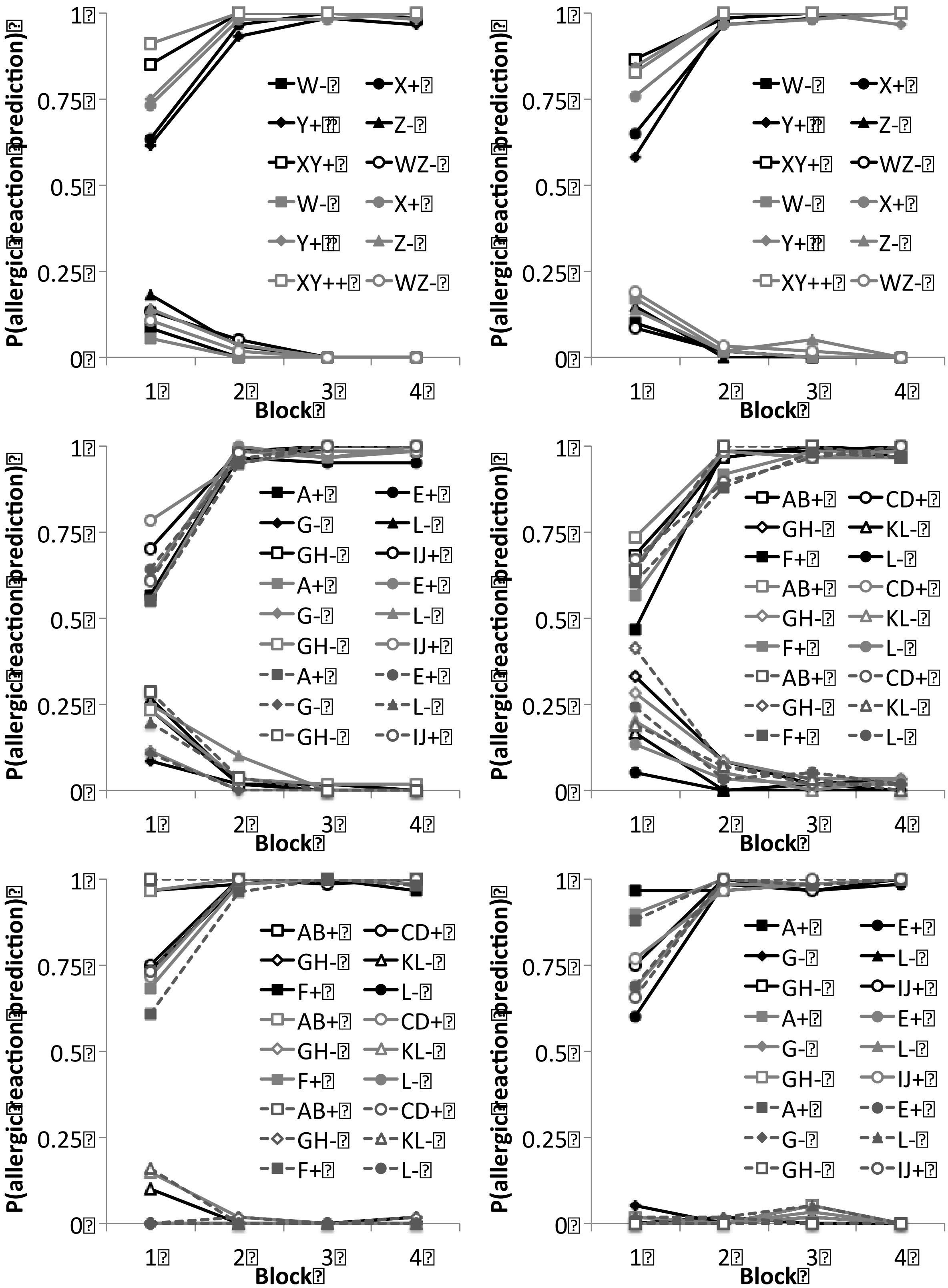

Figure S1. Predictions made during the three stages of training in Experiment 3, pretraining (top panel), phase 1 (middle panel), and phase 2 (bottom panel). Left panels show data for the forward blocking groups, right panels show data for the backward blocking groups. Black symbols show data for the Nonadditive groups, light grey symbols show data for the Additive groups, and grey symbols with dashed lines show data for the No pretraining groups.

Table S1.
Mean ratings and SE for the test phase of Experiment 1.

			B	C	D	E	F	EF	EM	FM	H	L
Causal Ratings	Forward	Nonadd	50.0	68.8	61.3	96.8	99.4	94.8	92.8	93.5	1.5	1.1
			5.0	*4.3*	*4.3*	*1.8*	*0.1*	*3.5*	*2.6*	*2.7*	*0.7*	*0.3*
		No Pre	49.5	62.1	64.7	99.0	99.2	97.7	90.6	90.4	1.9	4.5
			6.2	*4.2*	*4.3*	*0.4*	*0.3*	*1.7*	*3.1*	*3.5*	*0.9*	*3.3*
		Add	16.4	56.7	57.2	98.9	99.3	99.2	91.9	90.1	1.9	0.8
			5.4	*3.9*	*4.5*	*0.4*	*0.2*	*0.2*	*4.0*	*3.3*	*0.9*	*0.1*
	Backward	Nonadd	70.7	77.6	71.7	97.2	97.5	96.3	88.3	89.8	7.8	0.8
			4.7	*4.5*	*5.3*	*1.5*	*1.2*	*2.2*	*3.0*	*3.8*	*4.6*	*0.2*
		No Pre	63.5	72.6	62.6	98.7	95.6	99.1	85.9	82.7	10.7	1.0
			6.2	*4.9*	*5.2*	*0.5*	*3.3*	*0.4*	*3.9*	*5.1*	*4.8*	*0.2*
		Add	38.7	65.5	63.5	99.0	92.9	94.8	87.0	80.3	0.5	0.7
			7.8	*4.9*	*4.4*	*0.4*	*4.3*	*3.6*	*4.4*	*5.9*	*0.1*	*0.1*
Confidence Ratings	Forward	Nonadd	68.7	73.1	68.4	95.4	98.9	95.8	90.9	91.3	91.7	95.8
			4.7	*5.3*	*4.6*	*2.3*	*0.3*	*2.4*	*2.9*	*3.2*	*4.6*	*3.3*
		No Pre	72.7	68.5	71.7	98.1	99.0	94.9	83.7	84.3	85.3	91.6
			4.7	*4.0*	*3.9*	*0.7*	*0.3*	*3.1*	*4.7*	*4.4*	*5.6*	*4.6*
		Add	88.7	61.5	61.0	97.9	98.6	98.6	91.7	91.8	95.5	91.1
			3.4	*4.1*	*4.6*	*0.8*	*0.4*	*0.7*	*3.2*	*2.8*	*2.2*	*4.9*
	Backward	Nonadd	74.3	77.6	74.8	95.0	96.0	95.5	87.2	87.7	91.1	95.9
			5.3	*5.2*	*4.7*	*2.1*	*1.6*	*2.2*	*3.8*	*3.3*	*4.6*	*3.3*
		No Pre	80.2	79.0	75.9	98.1	97.9	95.9	83.4	87.6	89.2	98.9
			4.4	*4.2*	*4.1*	*0.7*	*0.9*	*1.8*	*3.8*	*3.2*	*4.8*	*0.3*
		Add	76.3	64.1	64.6	97.2	91.8	97.4	85.5	82.8	94.9	95.7
			5.3	*4.1*	*4.2*	*1.4*	*3.8*	*1.7*	*3.8*	*4.6*	*3.1*	*3.4*
Expected Severity	Forward	Nonadd	3.8	4.7	4.2	5.1	5.1	5.2	4.9	4.9	0.2	0.1
			0.3	*0.2*	*0.3*	*0.1*	*0.1*	*0.2*	*0.1*	*0.1*	*0.1*	*0.0*
		No Pre	3.2	4.5	4.7	5.5	5.4	8.5	5.6	5.5	0.4	0.1
			0.5	*0.3*	*0.3*	*0.2*	*0.2*	*0.4*	*0.3*	*0.3*	*0.2*	*0.0*
		Add	1.3	3.7	4.6	5.3	5.1	9.8	5.1	5.2	0.6	0.3
			0.4	*0.3*	*0.3*	*0.2*	*0.2*	*0.1*	*0.1*	*0.2*	*0.3*	*0.2*
	Backward	Nonadd	4.6	4.8	4.5	5.1	5.1	5.1	5.1	5.0	0.8	0.6
			0.3	*0.3*	*0.3*	*0.1*	*0.1*	*0.1*	*0.0*	*0.2*	*0.4*	*0.3*
		No Pre	3.7	4.3	3.4	5.1	5.0	7.8	4.7	4.4	0.5	0.3
			0.4	*0.4*	*0.4*	*0.3*	*0.3*	*0.4*	*0.2*	*0.3*	*0.2*	*0.2*
		Add	2.4	4.8	4.7	5.3	4.8	9.3	5.4	5.1	0.5	0.3
			0.5	*0.2*	*0.2*	*0.2*	*0.3*	*0.3*	*0.2*	*0.3*	*0.3*	*0.2*

Note: "No pre" refers to the No Pretraining groups, "Nonadd" refers to the Nonadditive Pretraining groups and "Add" refers to the Additive Pretraining groups. Letters in top row refer to test cues. The values represent mean rating and SE (second number, in italics) for each cue in each group.

Trial recall test

The motivation for including this test was to determine whether reversing the order of learning phases in the case of backward blocking might result in a poorer capacity of the participant to remember the relevant contingencies at the time of making causal judgments, potentially explaining why forward and backward blocking differ in magnitude.

Method: After completing the ratings test, participants were given an additional *trial recall* test. On each trial, participants were presented with one food cue and used the mouse to select the food cue with which it was paired and then the outcome that followed this pair of foods. They then pressed space bar to continue. Participants were tested on all food cues that had been consistently paired with one other cue during training, namely A, B, C, D, G, H, I, J, K, and L. The instructions for this test were as follows:

"In the next phase, we are going to test how well you remember which foods went together during the earlier phases of the experiment. On each trial, you will be shown one food that was eaten by Mr X. In the earlier phases of the experiment, Mr X ate this food in combination with one (and only one) other food. Your task is to complete the combination by choosing the correct food (i.e. the food that was eaten in combination with the one that is shown on screen). You will also be asked to choose which outcome (e.g. allergic reaction or no reaction) resulted from eating this combination of foods."

Results. Full results of the trial recall test can be found in Table S2. In order to make inferences about the blocked cue B on test, it is necessary to accurately recall trial information about B but also about A. Comparing across the cues and pretraining conditions, the probability of correctly choosing the paired cue was usually slightly higher in the backward blocking groups than in the forward blocking groups, when presented with cue B, (χ^2 (1, N = 177) = 5.00, p = .025) but not when presented with cue A (χ^2 (1, N = 177) = 3.78, p = .052). Chi-square tests comparing pretraining groups (collapsing across phase order) were not significant, (B: χ^2 (2, N = 177) = 2.05, p = .358, A: χ^2 (2, N = 177) = 4.08, p = .130. Probability of correctly choosing the paired outcome did not differ for either A or B depending on either the phase order (B: χ^2 (1, N = 177) = 3.29, p = .070, A: χ^2 < 1) or pretraining group (B: χ^2 < 1, A: χ^2 (2, N = 177) = 2.26, p = .323).

Table S2.
Proportion of participants who correctly recalled the cue (paired cue recall) and outcome (outcome recall) in each of the six groups of Experiment 1.

			A	B	C	D	G	H	I	J	K	L
Paired Cue Recall	Forward	Nonadd	0.70	0.70	0.80	0.80	0.90	0.83	0.77	0.80	0.63	0.67
		No Pre	0.70	0.70	0.83	0.77	0.73	0.83	0.80	0.80	0.70	0.70
		Add	0.86	0.75	0.89	0.89	0.93	0.96	0.89	0.93	0.75	0.82
	Backward	Nonadd	0.83	0.87	0.83	0.87	0.93	0.87	0.90	0.93	0.83	0.77
		No Pre	0.83	0.77	0.77	0.80	1.00	0.93	0.90	0.87	0.83	0.83
		Add	0.93	0.93	1.00	0.93	0.97	0.93	0.93	0.97	0.79	0.83
Outcome Recall	Forward	Nonadd	0.87	0.80	0.73	0.87	1.00	1.00	0.77	0.73	0.90	0.93
		No Pre	0.80	0.70	0.77	0.90	0.90	0.90	0.87	0.83	0.87	0.80
		Add	0.93	0.82	0.93	0.93	0.96	1.00	0.89	0.86	0.93	0.86
	Backward	Nonadd	0.83	0.83	0.93	0.87	1.00	0.93	0.83	0.90	0.93	1.00
		No Pre	0.90	0.90	0.80	0.83	1.00	0.97	0.80	0.93	0.93	0.90
		Add	0.93	0.90	0.97	0.86	1.00	0.97	0.90	0.76	0.97	0.93

Note: “No pre” refers to the No Pretraining groups, “Nonadd” refers to the Nonadditive Pretraining groups and “Add” refers to the Additive Pretraining groups. Letters in top row refer to the presented cue.

For completeness, in Table S3 we also report an analysis of the relationship between blocking and the participants’ ability to report which cue went with the pretrained cue A, the blocked cue B, and the control cues C and D. Note that the paired cue recall test was not ideally suited to this analysis because the results suffer from restriction of range (i.e. many participants score an accuracy of 1) and the test may not be sensitive enough for this purpose. The only significant correlation was in the forward nonadditive pretraining group, between the blocking score and the participants’ ability to remember cue A as the cue paired with cue B, r = -.444, p = .014, suggesting that participants who answered this test item incorrectly displayed more blocking. Given the lack of evidence of any other correlations, we are cautious about interpreting this one significant correlation. Perhaps more notably, blocking in the backward groups did not correlate with any of the indices of paired cue recall.

Table S3.
Correlations between blocking score (mean causal rating for control cues C and D minus causal rating for cue B) and accuracy on six relevant measures of paired cue recall.

		n	A	B	A&B	C&D	ABCD
Forward	Nonadd	30	-.243	**-.444***	-.358	-.033	-.245
	No Pre	30	-.207	.021	-.101	.027	-.048
	Add	28	.163	.212	.206	.119	.183
Backward	Nonadd	30	-.020	.051	.015	-.079	-.041
	No Pre	30	.115	-.027	.044	.076	.083
	Add	29	-.100	-.063	-.095	.197	.016

Note: "No pre" refers to the No Pretraining groups, "Nonadd" refers to the Nonadditive Pretraining groups and "Add" refers to the Additive Pretraining groups. Letters in top row refer to the presented cue, A&B represents the sum of A and B test trials, C&D represents the sum of C and D test trials, ABCD represents the sum of A, B, C, D test trials. * indicates $p < .05$.

Discussion. Cued recall of the critical blocking contingencies was better or at least as good in the backward blocking conditions compared to the forward blocking conditions and did not differ according to pretraining group. Therefore, the differences in the magnitude of forward and backward blocking are unlikely to be a result of an enhanced or diminished capacity to retrieve the information necessary to reason about B on test.

Post-Experiment Manipulation Check

The post experiment questionnaire tested outcome additivity assumptions using a single 3-alternative forced choice question. All participants were presented with the same hypothetical scenario in which Mr X ate two novel foods (orange and sultanas) each of which produced a mild allergic reaction individually. Participants predicted what they thought would happen if he ate the two foods together. Three buttons appeared which read Mr X suffered no allergic reaction, Mr X suffered a mild allergic reaction or Mr X suffered a severe allergic reaction. Participants selected an option and then rated their confidence in their response. Two participants in the forward no pretraining group and one participant in each of the three backward blocking groups indicated that the reaction should be less severe following the compound than following each of the single cues. Removing these 5 participants did not alter any of the critical results in relation to blocking. The remaining 59 nonadditive pretraining participants all indicated that the compound should produce the same magnitude reaction as the single cues and 53 out of the remaining 56 additive pretraining participants indicated that the compound should produce a stronger reaction. The remaining 3 participants were all in the backward additive group, but still produced a strong blocking score in causal ratings (M = 28.7). Thirteen participants in the forward no pretraining group and 14 in the backward no pretraining group indicated that the compound should result in a more severe reaction than the single cues. Fifteen participants from each no pretraining group indicated that the severity for the compound and single cues should be the same.

Experiment 2

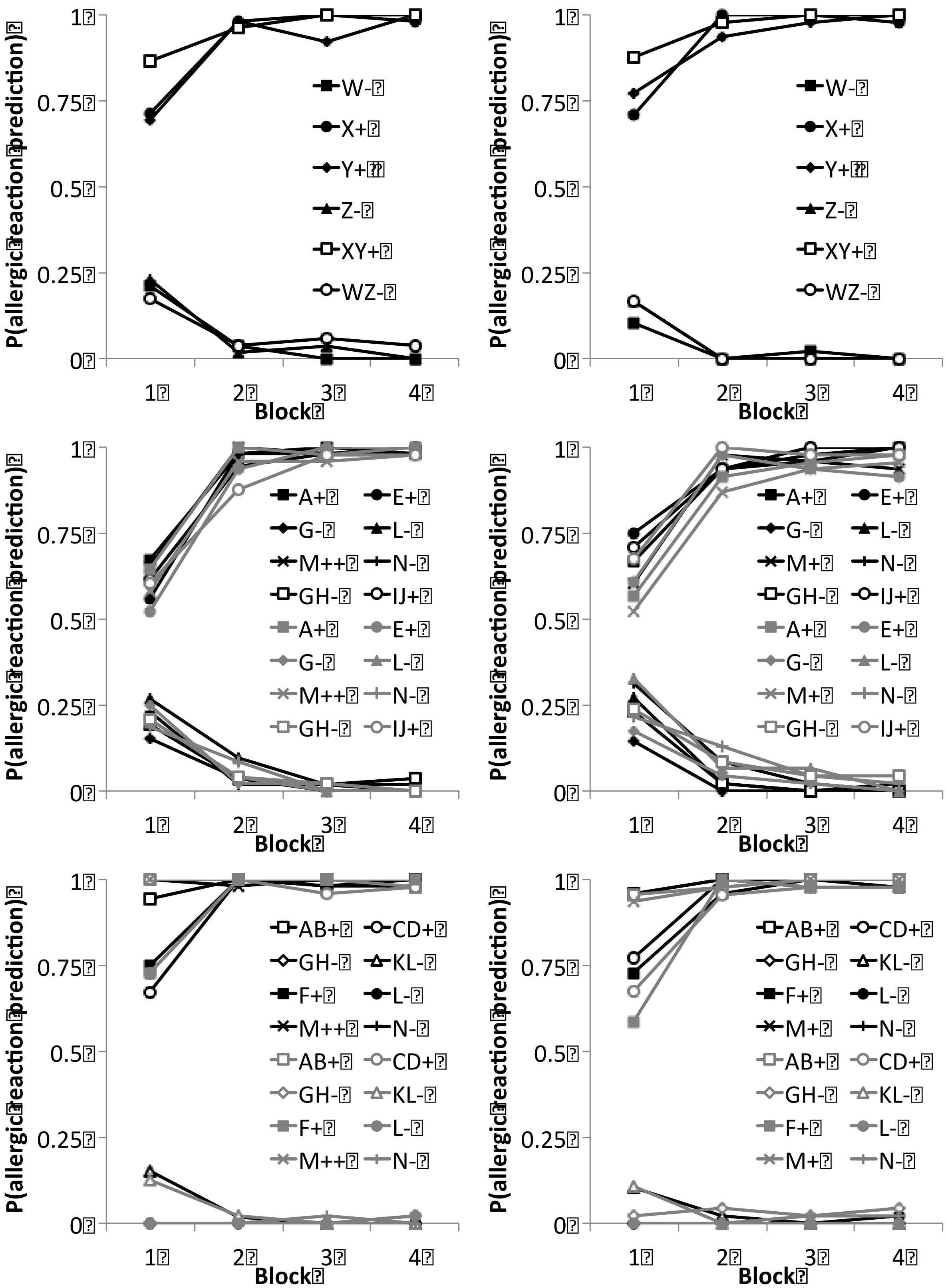

Figure S2. Predictions made during the three stages of training in Experiment 2, pretraining (top panel), phase 1 (middle panel), and phase 2 (bottom panel). Left panels show data for the maximal single cue (M+10) groups, right panels show data for submaximal single cue (M+5) groups. Black symbols show data for the Nonadditive groups, grey symbols show data for the No pretraining groups.

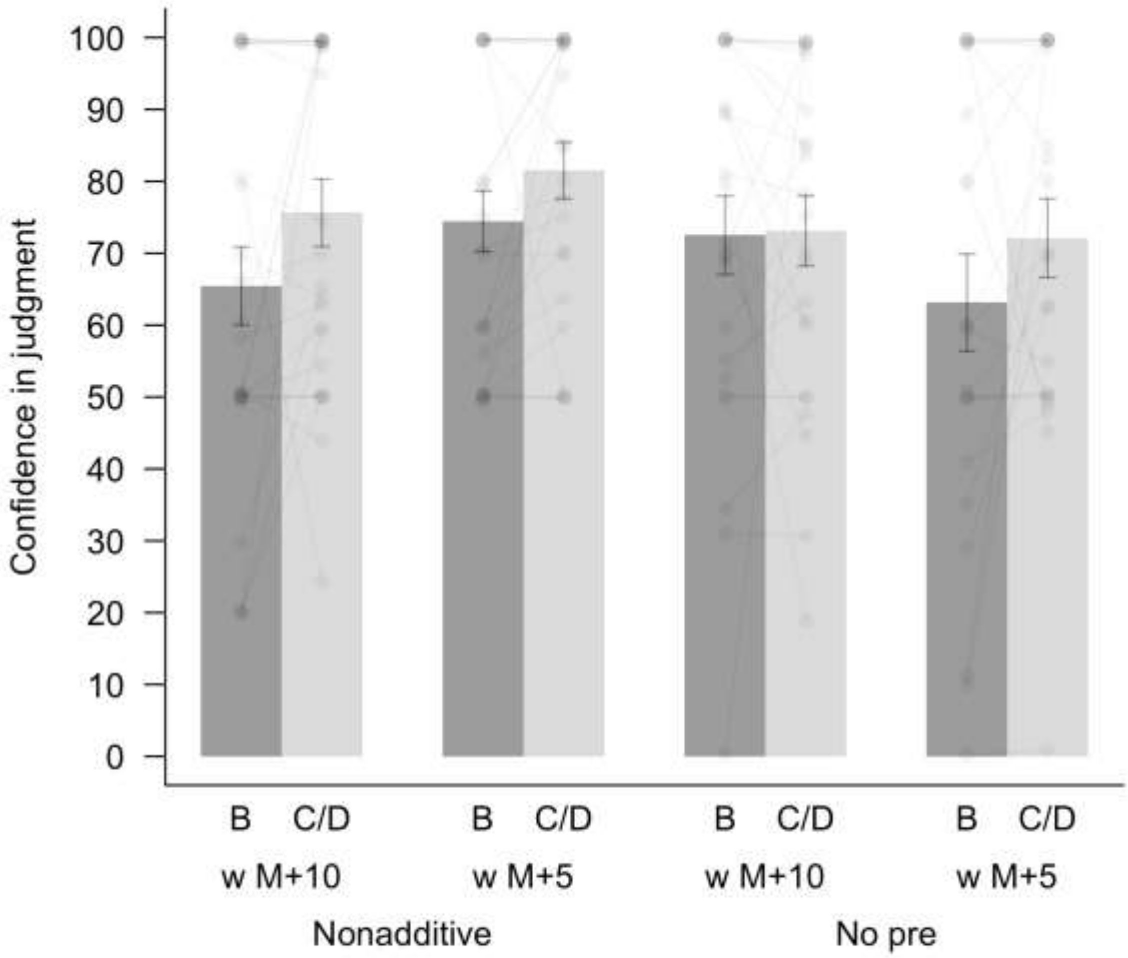

Figure S3. Confidence in causal rating of B vs C and D in Experiment 2. Dark grey bars show the confidence rating of cue B; the light grey bars show the mean confidence rating of cues C and D. Error bars indicate SE and connected data points represent data of individual participants.

Table S4.
Mean ratings and SE for the test phase of Experiment 2.

			B	C	D	E	F	EF	H	L	M	N
Causal Ratings	maximal single cue M+10	Nonadd	50.4	70.7	65.8	96.2	96.0	99.1	2.8	0.8	98.4	1.0
			6.1	*4.7*	*5.4*	*2.1*	*2.2*	*0.5*	*1.9*	*0.2*	*0.8*	*0.5*
		No Pre	48.7	65.3	60.4	91.7	90.9	98.1	3.9	1.7	99.3	0.5
			6.3	*5.7*	*6.0*	*4.6*	*4.9*	*1.2*	*3.3*	*1.1*	*0.1*	*0.1*
	submax single cue M+5	Nonadd	67.2	78.3	81.2	93.9	99.4	99.7	0.7	0.4	97.5	0.5
			6.1	*4.6*	*4.7*	*4.2*	*0.2*	*0.0*	*0.2*	*0.1*	*2.0*	*0.1*
		No Pre	50.4	62.1	74.2	96.1	98.0	98.4	8.0	1.8	96.8	0.7
			6.3	*5.2*	*4.4*	*2.4*	*1.2*	*0.6*	*4.4*	*0.9*	*1.6*	*0.4*
Confidence Ratings	maximal single cue M+10	Nonadd	65.4	77.8	73.5	97.4	93.3	98.0	86.3	94.7	98.6	94.8
			5.4	*4.6*	*5.4*	*1.0*	*3.9*	*1.0*	*5.8*	*3.8*	*0.6*	*3.6*
		No Pre	72.5	73.2	73.0	94.7	95.1	96.3	90.8	92.0	98.3	94.0
			5.5	*5.4*	*5.5*	*2.4*	*2.3*	*1.8*	*5.6*	*4.6*	*1.0*	*4.1*
	submax single cue M+5	Nonadd	74.5	82.5	80.5	97.1	99.2	99.5	88.5	95.4	97.6	95.3
			4.2	*3.8*	*4.8*	*1.4*	*0.3*	*0.1*	*5.6*	*4.1*	*1.9*	*4.1*
		No Pre	63.1	67.4	76.8	94.8	98.9	95.5	94.9	92.5	97.2	98.3
			6.8	*6.2*	*5.6*	*3.4*	*0.5*	*2.7*	*2.0*	*4.7*	*1.3*	*0.9*
Expected Severity	maximal single cue M+10	Nonadd	3.9	5.2	4.7	5.4	5.4	6.2	0.2	0.1	9.3	0.4
			0.5	*0.3*	*0.3*	*0.2*	*0.2*	*0.4*	*0.1*	*0.0*	*0.3*	*0.3*
		No Pre	3.1	4.3	3.8	5.1	5.3	8.3	0.3	0.3	9.8	0.2
			0.4	*0.3*	*0.4*	*0.2*	*0.3*	*0.4*	*0.3*	*0.2*	*0.1*	*0.1*
	submax single cue M+5	Nonadd	4.0	4.8	5.0	4.9	5.1	5.3	0.2	0.4	4.9	0.6
			0.4	*0.3*	*0.2*	*0.2*	*0.2*	*0.2*	*0.1*	*0.3*	*0.2*	*0.3*
		No Pre	3.3	4.1	4.5	5.1	5.1	8.2	0.1	0.4	5.1	0.3
			0.4	*0.3*	*0.2*	*0.1*	*0.1*	*0.4*	*0.0*	*0.2*	*0.1*	*0.2*

Note: "No pre" refers to the No Pretraining groups, "Nonadd" refers to the Nonadditive Pretraining groups and "Add" refers to the Additive Pretraining groups. Letters in top row refer to test cues. The values represent mean rating and SE (second number, in italics) for each cue in each group.