2022-02

Theory protection: do humans protect existing associative links?

Spicer, S

10.1037/xan0000314
Journal of Experimental Psychology: Animal Learning and Cognition
American Psychological Association

Theory protection: do humans protect existing associative links?

Stuart G. Spicer, Chris J. Mitchell, Andy J. Wills, Katie L. Blake, and Peter M. Jones

University of Plymouth

Running head: Theory protection in learning

Author note:

The experiments reported here were conducted as part of Stuart Spicer's PhD. Correspondence concerning this article should be addressed to Dr. Stuart G. Spicer, School of Psychology, University of Plymouth, Plymouth, PL4 8AA, United Kingdom. Email: stuart.spicer@plymouth.ac.uk

Author contribution

Stuart G. Spicer (lead author): Co-contributer to the rationale, theoretical basis and design of the experiments. Programmed the experiments, collected and analysed the data as part of PhD project. Wrote up the experiments.

Peter M. Jones: Co-contributer to the rationale, theoretical basis and design of the experiments. Contributed to analysis and write up as PhD supervisor.

Chris J. Mitchell: Co-contributer to the rationale, theoretical basis and design of the experiments. Contributed to analysis and write up as co-author.

Katie Blake: Contributed to Experiment 4 as part of final year undergraduate project.

Andy J. Wills: Contributed to analysis, interpretation, and write up as PhD supervisor.

Abstract

Theories of associative learning often propose that learning is proportional to prediction error, or the difference between expected events and those that occur. Spicer et al. (2020) suggested an alternative, that humans might instead selectively attribute surprising outcomes to cues that they are not confident about, in order to maintain cue-outcome associations about which they are more confident. Spicer et al. reported three predictive learning experiments, the results of which were consistent with their proposal ("theory protection") rather than a prediction error account (Rescorla, 2001). The four experiments reported here further test theory protection against a prediction error account. Experiments 3 and 4 also test the proposals of Holmes et al. (2019), who suggested a function mapping learning to performance that can explain Spicer et al.'s results using a prediction-error framework. In contrast to the previous study, these experiments were based on inhibition rather than excitation. Participants were trained with a set of cues (represented by letters), each of which was followed by the presence or absence of an outcome (represented by + or -). Following this, a cue that previously caused the outcome (A+) was placed in compound with another cue (B) with an ambiguous causal status (e.g. a novel cue in Experiment 1). This compound (AB-) did not cause the outcome. Participants always learned more about B in the second training phase, despite A always having the greater prediction error. In Experiments 3 and 4, a cue with no apparent prediction error was learned about more than a cue with a large prediction error. Experiment 4 tested participants' relative confidence about the causal status of cues A and B prior to the AB-stage, producing findings that are consistent with theory protection and inconsistent with the predictions of Rescorla, and Holmes et al. (2019).

Keywords: associative learning, prediction error, theory protection, uncertainty, confidence

One of the most widespread ideas in the field of associative learning is that learning is determined, at least in part, by prediction error. This is the discrepancy between an organism's expectation about what might happen, and the events that occur. If expectations match events then there is no prediction error and little learning occurs, but if there is a mismatch then the organism learns, in order to better predict future events. Furthermore, learning is thought to be proportional to the size of the prediction error. All other things being equal, a larger prediction error should result in greater learning. This commonplace idea is embedded in numerous influential models of learning (e.g. Bush & Mosteller, 1951; Mackintosh, 1975; Pearce and Hall, 1980; Rescorla & Wagner, 1972).

Rescorla (2001) conducted a series of experiments to examine one way in which prediction error might influence learning. These experiments were designed to measure the relative amounts of learning for two cues that were presented together on the same learning trials. In other words, these experiments were not concerned with the overall amount of learning resulting from an unexpected event, but with investigating whether prediction error determined which of the two cues would be learned about most. In one experiment, rats were initially trained with two cues that were followed by the delivery of food, A+ and C+, and two cues that were not, B- and D- (where + and – indicate the presence and absence of the food outcome). After completion of this initial training, the rats were then trained with a compound that was not followed by food (AB-). This compound contained one cue that had been paired with food in the first stage (A), and one that had not (B). If prediction error determined which of these two cues would be learned about most then more learning would have accrued to A than B during AB- trials. To test this idea, rats received a final test in which compounds AD and BC were presented. Each of these compounds contained one cue that had been paired with food in the first stage of the experiment (A or C) and one that had not (B or D). Hence, if the AB- training had not occurred, the conditioned response to the AD and BC compounds would have been the same. Any difference between the two must therefore be attributable to differences in

learning for A and B that took place on the AB- trials[1]. Rescorla observed less conditioned responding on AD trials than BC trials, and consequently inferred that the rats had learned more about A than B during AB- training. Since the outcome that occurred (no food) was consistent with the prior training with B, but not A, Rescorla's results suggest that the relative amounts of learning for these cues were determined by prediction error.

Analogous results to those of Rescorla (2001) have also been observed in humans in a causal learning paradigm. Haselgrove and Evans (2010) asked participants to play the role of a doctor and to determine which foods caused an allergic reaction in a fictional patient. On each trial, participants were told that the patient had consumed one or more foods, and asked to make a prediction about whether or not the reaction would occur. After making each prediction, they received feedback. The abstract design of the experiment was the same as that reported by Rescorla, with pairings of two foods with the allergic reaction (A+ and C+) and two foods with the absence of the reaction (B- and D-). Participants subsequently received AB- trials, followed by the same kind of test (AD vs. BC) that Rescorla used. This test suggested that more learning had taken place for A than for B during AB- trials. Taken together, the results of these experiments suggest that both humans and non-humans sometimes learn most about the cues that have the largest prediction error.

However, recent evidence suggests that this principle might not always apply. Spicer et al. (2020) proposed that a different process might affect learning in humans, based on 'theory protection'. This idea is quite different from the notion of prediction error. According to theory protection, learning about each cue in a compound is influenced by the extent to which the outcome is consistent with existing knowledge about each cue. Spicer et al. proposed that humans maintain existing associations as far as possible, interpreting new information in ways that are consistent with what is already known. One way of doing this is to attribute unexpected outcomes to cues about which

[1] This seemingly-complex compound testing procedure (see also Rescorla, 2000) provides a way of comparing the amount of learning for two cues that are trained from different starting associative strengths. It is necessary because the relationship between associative strength and responding may not be linear (e.g. Gluck & Bower, 1988).

existing knowledge is known to be incomplete, i.e. cues whose relationship with the outcome is uncertain. This idea is analogous to processes in other psychological fields, such as new information being incorporated into existing schemata (Bartlett, 1932), and existing beliefs being updated in a minimal fashion in the face of novel information (e.g. Harman, 1986).

Spicer et al. (2020) presented findings consistent with theory protection in a series of three experiments, all of which used a similar compound conditioning procedure to Rescorla (2001). In one experiment, initial training included A+, AX+, BY+, and CY- trials. Following initial training, X and C were presented together in a compound conditioning phase. These two cues were chosen because prior evidence (Jones et al., 2019) demonstrated that cues trained in this way differed not just in the extent to which participants thought they predicted the outcome, but also how certain participants were in these judgements. During initial training, X was only encountered in compound with a better predictor of the outcome, A. This means that participants lacked evidence of the relationship between X and the allergic reaction. In contrast, training with C allowed participants to infer that C did not cause the allergic reaction. In subsequent test trials with individual cues, Jones et al.'s participants rated the likelihood of the outcome as greater for X than for C, but had more confidence in their ratings for C than for X. Based on these findings, Spicer et al. presented participants with XC+ compound conditioning trials. If learning on these trials was determined by prediction error then more learning should have occurred for C than for X, since C had the larger prediction error. However, a test phase analogous to that used by Rescorla showed that more learning had taken place for X. Spicer et al. suggested that this was because participants had been less certain about X than C prior to XC+ trials, and had therefore been able to interpret the new information without changing their existing beliefs by attributing the allergic reaction to X.

In some respects, Spicer et al.'s (2020) results are at odds with those reported by Haselgrove and Evans (2010). Both experiments used causal learning paradigms with human participants, and yet in

one case learning was greater for the cue with the smaller prediction error, and in the other case it was greater for the cue with the larger prediction error. There were, of course, many differences between the two procedures. The purpose of the current series of experiments is to evaluate two possible explanations for these differing results, to clarify the circumstances in which results resembling prediction error and theory protection might be obtained.

One notable difference between the two procedures is that the outcome was present during the compound conditioning phase in Spicer et al.'s (2020) experiment (XC+), but absent during the equivalent phase of Haselgrove and Evans' (2010) experiment (AB-). Perhaps the distribution of excitatory learning between cues is governed by theory protection in humans, while inhibitory learning is governed by prediction error. Evidence to either support or reject this (somewhat unlikely) possibility is currently inconclusive. Le Pelley and McLaren (2004) reported an experiment using a similar compound conditioning procedure. They used a more complex experimental design, in which a single group of participants experienced both excitatory and non-reinforced compound training. Two cues were trained in a compound that predicted an outcome (AB+) alongside two cues that predicted the absence of that outcome (CD-). Two of these cues (A and C) were subsequently paired in a compound that either predicted the outcome (AC+) or the absence of the outcome (AC-). Subsequent testing revealed more learning about the cue with the smaller prediction error for excitatory (AC+) compound training, but a null result for neutral (AC-) compound training. The current experiments provide a further test of this idea, using a simpler design.

Alternatively, the discrepancy between Spicer at al.'s (2020) results and Haselgrove and Evans' (2010) results might be attributable to differences in certainty about the causal status of the cues. Spicer et al. suggested that participants could learn about the relationship between X and the outcome on XC+ trials because, prior to those trials, they had not been certain about whether X

caused the outcome. For the same principle to apply to Haselgrove and Evans' AB- trials, participants would need to be uncertain about one of the cues. However, in their experiment participants had been trained with each cue alone (A+ and B-), leaving less room for uncertainty. Furthermore, the food allergist framework used by Haselgrove and Evans might have further reduced uncertainty. Consider cue B in their experiment, which was initially trained in the absence of the outcome (B-). In some situations, participants might be uncertain about whether B prevents the occurrence of the outcome, or whether it does nothing. If this were so, and if theory protection were to determine learning, we might expect more learning about B than A during subsequent AB- trials; participants would protect their theory that A causes the outcome by learning that B prevents it. However, this is unlikely in the food allergist paradigm because foods do not often prevent allergic reactions in this way in the real world.

Zaksaite and Jones (2019) similarly argued that the food allergist paradigm does not easily permit inhibitory learning. They demonstrated conditioned inhibition more readily in an alternative scenario in which the cues were chemicals that were equally likely to cause or prevent the outcome (changes in hormone levels). Perhaps if a non-food based scenario like that used by Zaksaite and Jones were used to present Haselgrove and Evans' design, a result different to theirs would be obtained. That is, increasing uncertainty about B – specifically, allowing B to be either non-causal or inhibitory – might permit theory protection, rather than prediction error, to dominate. Note that Le Pelley and McLaren (2004) also used the food allergist paradigm, when they failed to see more learning about a neutral cue than an excitor trained together in the absence of any outcome. It is possible that a result resembling theory protection would have been observed in that study if a scenario that more readily allowed inhibitory learning had been used; in order to protect the theory that A is causal following A+ then AB- training, it is necessary that B has the capacity to prevent the outcome that would otherwise have been caused by A. Here we present four experiments that test this idea.

The experiments reported here used a design similar to Haselgrove and Evans (2010), but with two modifications designed to increase uncertainty about cue B. Firstly, in Experiment 1 the initial training trials with B were omitted. B was therefore novel at the start of the compound AB- training, and participants should have been maximally uncertain about the relationship between B and the outcome. Experiment 2 employed a variant of this procedure in which B was presented during initial training, but with information about the presence or absence of the outcome obscured from participants. Secondly, all experiments here used a scenario similar to that used by Zaksaite and Jones (2019), in which cues could both cause and prevent the outcome. In Experiments 3 and 4, B was presented in the initial training phase and followed by the absence of the outcome (B-), as in Haselgrove and Evans' experiment – and Rescorla's (2001) experiment. If this training leaves participants uncertain about whether B is neutral or preventative with respect to the outcome, then more learning might accrue to B than to A during the subsequent AB- trials.

These experiments also provide a test of a more recent account of Rescorla's (2001) findings, proposed by Holmes et al. (2019). Rescorla's interpretation of his results was that a larger change in responding for A than for B implied more learning for A during AB- trials, but this inference is only valid if we assume that the relationship between learning and responding is linear when cues are combined in compound. Holmes et al. suggested an alternative explanation in which *learning* about A and B may have been equal (as predicted by Rescorla & Wagner, 1972) even though there was a greater change in *responding* to A, because the function relating associative strength to responding is non-linear. This means that different changes in responding to cues with different initial associative strengths can occur despite equal learning, depending on the slope of the response function at the point where each cue is located. Holmes et al. suggested that the shape of this function is a double sigmoid (see Figure 1), with the flattest region around zero. Cues with starting associative strengths close to zero (e.g. B at the start of AB- training) therefore undergo minimal

changes in responding during learning because they sit at the flattest part of the learning-performance curve, whereas learning translates into a change in responding more readily for cues that have associative strengths further from zero, where the curve is steeper. Holmes et al. proposed this an as explanation of the Rescorla (2001) experiment (i.e. A was located on a steeper part of the curve after initial training; the top right quadrant in Figure 1). Chan et al. (2021) suggested that this response function can also account for the Spicer et al. (2020) results, because C would have an associative strength close to zero at the start of XC+ training. Jones et al. (2021) accepted the logic of this potential explanation, but suggested a scenario in which it might be tested using human participants: the Rescorla (2001) design. As outlined above, according to a theory protection account, if participants are provided with a scenario in which they are uncertain about the status of B during initial training, more learning will accrue to B during the subsequent AB- trials, and so ratings of B will fall further than those of A. This is inconsistent with the prediction of Holmes et al.'s account, which predicts a greater change in ratings for A than B (because A is further from zero associative strength than B at the beginning of AB- training). Even if A has asymptotic associative strength (i.e. located at a similarly flat part of the learning-performance curve to B), this should still only result in equal changes in ratings for both cues, rather than a greater change in responding for B. Experiments 3 and 4 tested this prediction.

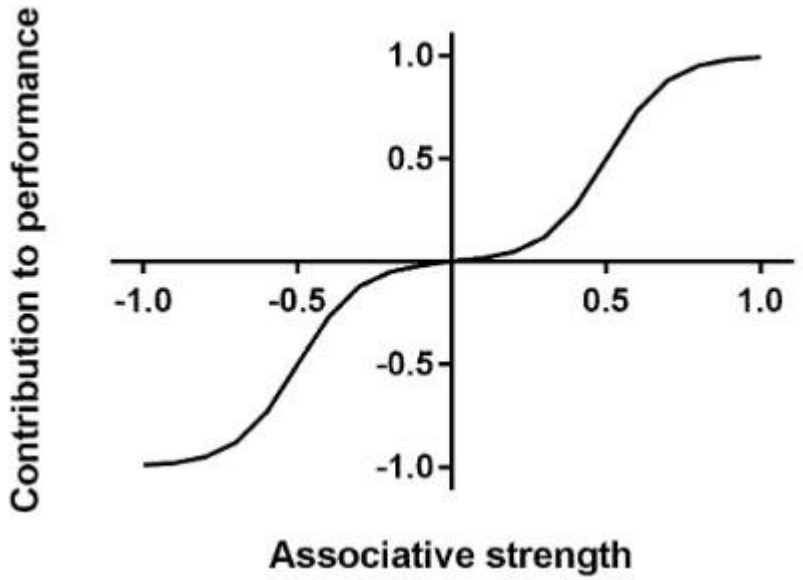

Figure 1. Holmes et al. (2019) learning-performance curve, where excitatory cues are located in top right quadrant, neutral cues are located at intersection of the two axes, and inhibitory cues are located in bottom left quadrant.

Experiment 1

The experiments presented here used an allergist task in which participants had to learn whether an allergic reaction would occur, on the basis of different chemicals being ingested. Participants were presented with a fictional scenario in which they were working in a drug research setting, trying to work out which chemicals cause the side effect of a stomach ache in a test patient. Participants were told that chemicals could be causal, neutral, or preventative with respect to the stomach ache outcome. On each training trial, participants were presented with one or more chemicals and were asked to predict whether or not the patient would experience a stomach ache after consuming them. Once participants had made their prediction, they were provided with feedback as to whether or not a stomach ache occurred. Following training, participants were tested by being asked to make ratings indicating how likely they thought a stomach ache would be after the patient ingested specific chemicals singly or in pairs. The design of the experiment is shown in Table 1. Participants were initially trained with two causal cues (A+ and C+), and two non-causal cues (E- and F-). The non-causal cues were added as fillers, so that participants experienced both the presence and the

absence of the stomach ache during Stage 1. Next, participants were trained with a non-causal compound AB- that consisted of a previously causal cue (A) and a novel cue (B). The novel cue was chosen to minimize the extent to which participants might have any causal theory about B at the start of AB- training. If learning about each cue is proportional to its prediction error then there should be more learning about A, since it was consistently paired with the outcome during Stage 1. However, theory protection predicts more learning about B, since participants should protect their theory that A is causal, and instead attribute the absence of the outcome to the novel B.

Table 1. The design of experiments 1-3

Experiment	Stage 1	Stage 2	Test
1	A+ C+ E- F-	AB-	AD BC A B C D E F
2	A+ C+ E- F- B? D?	AB-	AD BC A B C D E F
3	A+ C+ B- D-	AB-	AD BC A C B D

Key:
Letters A-F = different cues
+ = stomach ache
- = no stomach ache
? = outcome concealed from participants

Learning about A and B was compared using a final test discrimination equivalent to Rescorla (2001) and Haselgrove and Evans (2010). In addition to being asked for the likelihood of the outcome for each individual cue, participants were asked to give likelihood ratings for two compounds, AD and BC. Each of these compounds contained one cue that had been trained as causal in Stage 1, and one that had not been encountered in Stage 1. In the absence of Stage 2 training, these two compounds should have been assigned the same likelihood ratings at test. Consequently, any difference between these compounds must have been the result of the AB- training in Stage 2, and would indicate differing amounts of learning about A and B during that

stage. If learning is governed by prediction error, participants should have rated the likelihood of the outcome as being lower for AD than BC, as a result of A decreasing in associative strength during the AB- trials. However, if learning is determined by theory protection, then BC should have been assigned lower ratings than AD at test, indicating that participants learned during the AB- trials that B is preventative of the outcome, in order to protect their existing theory that A is causal.

Method

Participants

Forty psychology students from the University of Plymouth participated in this experiment, in return for course credits (28 female, 11 male, 1 non-binary; mean age = 23.05 , SD = 7.03). This sample size has adequate power to detect medium-sized within-subjects effects (87% power at d = 0.5). This sample size of all experiments was decided a priori. Comparable experiments by Spicer et al. (2020), using more complex designs, resulted in small to medium-sized effects, so medium-sized effects were regarded as a plausible assumption for these experiments. Forty participants were chosen as the optimum stopping rule for data collection, although a range of no more than five participants either side of this number was deemed acceptable when factoring in practical issues, such as recruitment, resource and cancellations. People who had previously taken part in similar experiments were excluded from this study, to ensure participants were naive to the purpose of the experiment.

Materials

Participants were tested in the same lab at the University of Plymouth. The experiment used 22-inch LED displays, with participants at a typical distance (of approximately 40-80 cm) from the screen. The experiment was designed and executed in Psychopy (Peirce, 2007). Participants made their responses by pressing keys on a standard UK computer keyboard during the training stages, and by using mouse clicks during the test stage. The six individual cues were represented on screen as

different coloured images of shapes: blue oval, green square, grey triangle, pink diamond, purple star, yellow circle. All the coloured shapes were presented within a white square. The dimensions of each cue (including the white square) were 300 x 300 pixels on a 1920 x 1080 pixel screen. For each participant, the coloured shapes were randomly assigned to serve as A, B, C, D, E, and F. The two outcomes, 'stomach ache' and 'no stomach ache', were represented by text on screen and a photograph of a man clutching his stomach, or a man giving a 'thumbs up', respectively. The outcome images were presented within a white rectangle. The dimensions of the outcome images (including the white rectangle) were 291x332 pixels. All experimental text, including instructions, was white. A black background was used throughout the experiment.

Design

The experiment used a within-subjects design, as outlined in Table 1. During Stage 1, participants were presented with twelve blocks of training. The four trial types (A+, C+, E-, F-) each appeared once in a random order within each block. Stage 2 consisted of six AB- trials. During the Test stage, participants were presented with two blocks of test cues. The eight trial types (A, B, C, D, E, F, AD, BC) each appeared once in a random order within each block.

Procedure

Participants were required to read an information sheet and sign a consent form prior to participating in the experiment. The experimental instructions were presented on the screen at the start of the experiment. They were adapted from Spicer et al. (2020) and are included in the Appendices.

For each trial during the training stages, the cues were presented on either the left-hand side of the screen or the right-hand side of the screen. When only one image was presented, the opposite side of the screen contained a blank space. The cues were randomly assigned to either the left or right

position on each trial. Text at the top of the screen stated that ‘The patient ingests the following:’, with the cues presented below this. Underneath the cues, further text stated ‘Which outcome do you expect? Please use your keyboard to respond’. Participants were instructed to respond by pressing the appropriate key on their keyboard; Z for ‘No Stomach Ache’ and M for ‘Stomach Ache’. After participants made their response, the feedback for that trial was shown. The feedback screen consisted of the appropriate outcome image along with its accompanying text, indicating either ‘Stomach Ache’ or ‘No Stomach Ache’. The feedback was shown on screen for two seconds, after which the next trial began.

After the completion of Stage 1, Stage 2 started with no trial break, so that from the perspective of participants this was a seamless continuation of the training. Stage 2 consisted of a previously unseen compound AB- presented six times in a row. As in Stage 1, the cues were randomised on each trial to appear on either the left- or right-hand side of the screen. The on-screen text and responding via the keyboard was the same as in Stage 1. The process for displaying the trial feedback was also the same, except that all six trials resulted in ‘No Stomach Ache’ as the outcome.

After the completion of Stage 2, a further instruction screen was shown before commencement of the Test stage (see Appendices).

For each trial during the Test stage, the cues were visually presented on either the left- or right-hand side of the screen. When only one image was presented, the opposite side of the screen again contained a blank space. The cues were randomly assigned to either the left or right position on each trial. As before, text at the top of the screen stated that ‘The patient ingests the following:’, with the cues presented below this. Underneath the cues, further text stated ‘How likely are they to suffer a stomach ache? (0 = Very Unlikely; 10 = Very Likely)’. Participants were instructed to respond by clicking on an 11-point rating scale using their mouse pointer, to indicate how likely

they thought the occurrence of a stomach ache would be. The rating scale was located in the lower part of the screen, with the 11-point scale running from left to right, in ascending numerical order. After participants made their response, a blank screen appeared for 0.4 secs, after which the next test trial was presented. Following the completion of the experiment, participants were provided with a debrief form.

Analysis

The data were processed and analysed using R (R Core Team, 2018). The difference between the AD and BC test compounds was assessed using paired-samples t-tests. Some additional analyses were also conducted on key single test stimuli, to test for specific predicted differences between cue ratings. The alpha level was set to $p < .05$ for all tests. As these tests were done on the basis of specific prior predictions, the values are reported without correcting for multiple comparisons. However, all differences reported here as significant would have survived a Bonferroni correction. Bayesian t-tests were also conducted, using the procedure recommended by Dienes (2011) and implemented as R code by Baguley and Kaye (2010). A uniform distribution was specified as the prior for each test, with a lower limit of -10 and an upper limit of 10 (in terms of the mean difference between ratings), because these are the largest mean differences in either direction permitted by an 11-point test rating scale. In keeping with accepted conventions (e.g. Jeffreys, 1961), a Bayes factor of over three was considered to provide substantial evidence for a difference, while a Bayes factor of less than one third was considered to provide substantial evidence for the absence of a difference. Values between a third and three were considered to provide inconclusive evidence.

Results and Discussion

The trial-level raw data and analysis script for this experiment will be available, upon publication of this manuscript, at https://osf.io/bzerh/. The descriptive statistics for the Experiment 1 training

stages are shown in Figure 2. The data from Stage 1 indicate that participants learned sufficiently about the four different cues by the time Stage 1 was complete. Similarly, the data from Stage 2 indicate that participants learned that the AB- compound was non-causal by the end of Stage 2.

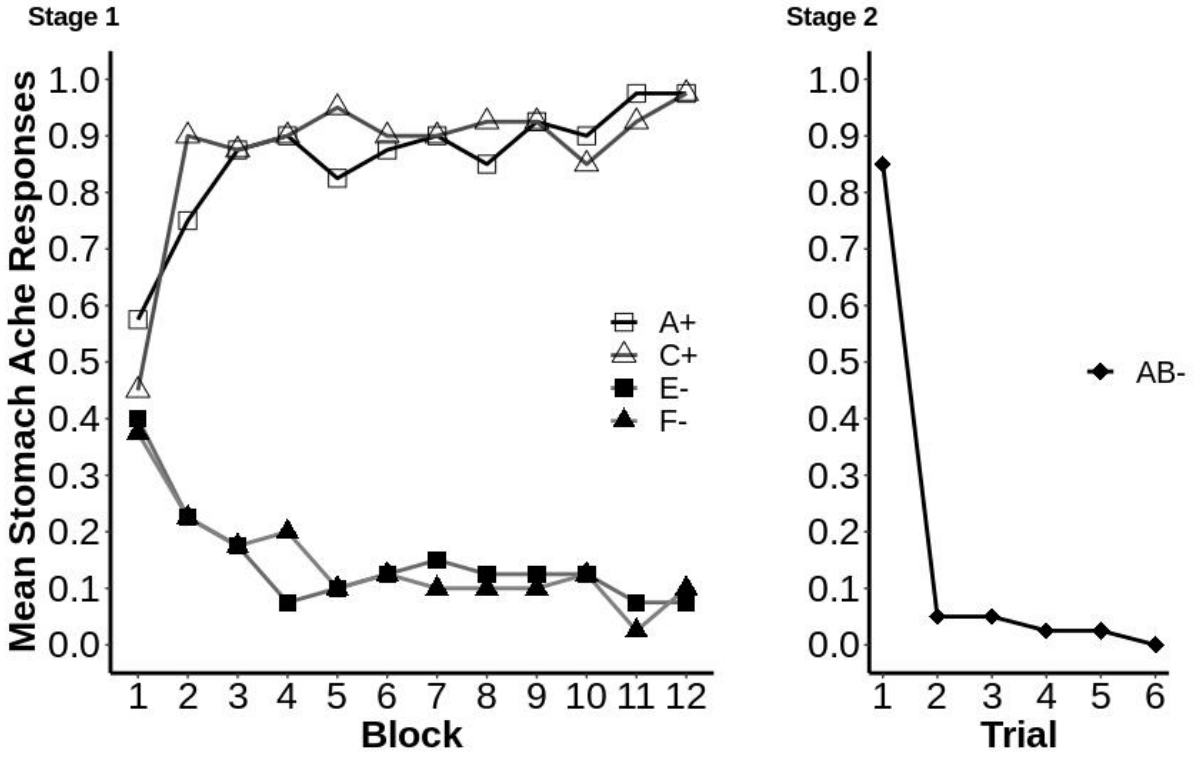

Figure 2. Experiment 1 training Stage 1 and Stage 2 data.

Descriptive statistics for the Experiment 1 Test stage are shown in Figure 3. Ratings for BC were significantly lower than for AD; $t(39) = 5.61$, $p < .001$, $BF = 4.08 \times 10^5$, $SE = .49$, $d = .89$. Further testing revealed lower ratings for the single cue B compared to D; $t(39) = 8.48$, $p < .001$, $BF = 2.23 \times 10^{14}$, $SE = .43$, $d = 1.34$. Conversely, there was evidence of no difference between the ratings assigned to A and C; $t(39) = .04$, $p = .919$, $BF = .05$, $SE = .37$, $d = .02$, indicating that participants did not learn about A during Stage 2. Taken as a whole, these findings show that the differences between the compounds were specifically driven by learning about B during AB- trials, with participants maintaining their association between A and the occurrence of stomach ache. Figure 3 panel B shows inter-subject variability on the key compound test difference. As would be expected from the mean test ratings, most individual participants rated BC lower than AD, although the

variability did extend to some participants assigning a lower rating to AD. It is also worth noting that the intermediate mean rating for D during the Test stage is consistent with the idea that participants do not know the causal status of novel cues, and are consequently unlikely to assign high or low ratings. These data appear to support theory protection (Spicer at al., 2020), as opposed to a prediction error account (Rescorla, 2001), suggesting that theory protection is seen when the outcome is absent (as well as present). However, one possible limitation of this experiment is that novel cues are often regarded as being more salient than familiar cues (e.g. Lubow & Moore, 1959). Consequently, B may have been more salient than A during Stage 2, and it may have been this that led to more learning about B, rather than theory protection. Experiment 2 addressed this possibility by replacing the two novel cues (B and D) with familiar cues that had an unknown causal status.

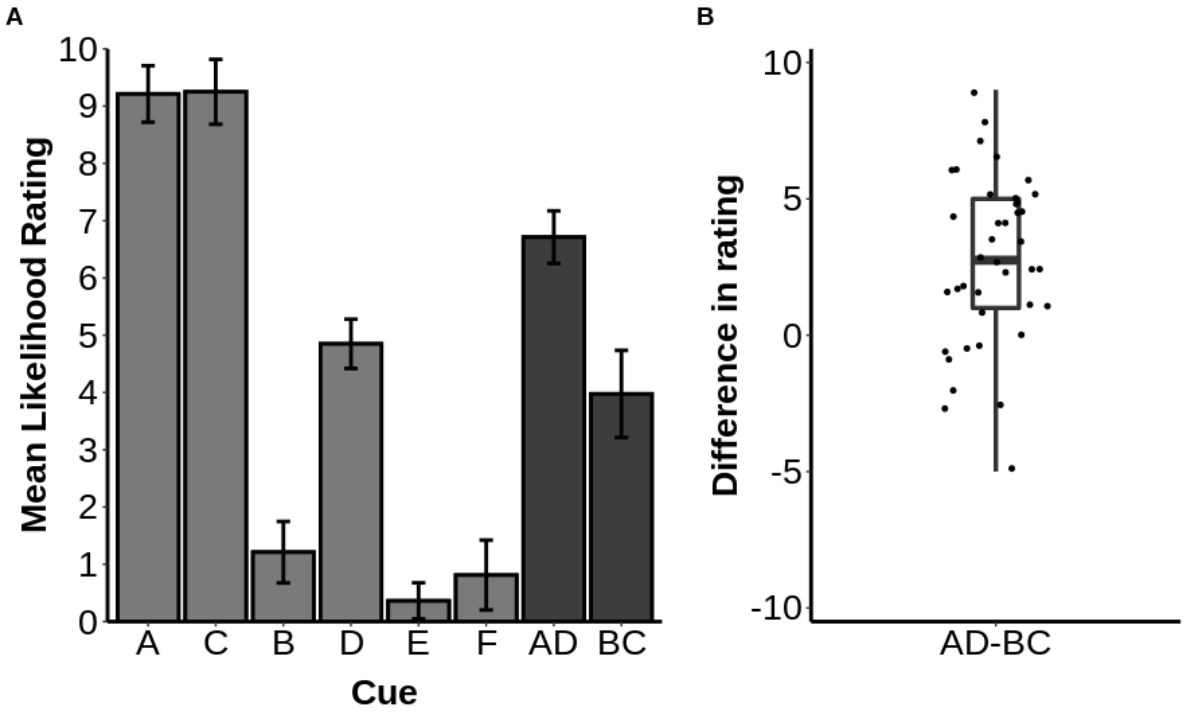

Figure 3. Panel A shows Experiment 1 Test stage ratings for single cues and compounds, with the error bars representing within-subjects confidence intervals. Panel B shows inter-subject variability on the key AD-BC difference. Each dot is one participant, with jitter applied for readability. The boxplot shows the median and interquartile range.

Experiment 2

Experiment 2 was a variant of Experiment 1 in which cues B and D were causally ambiguous but not novel. This was achieved by presenting cues B and D during Stage 1, but concealing the presence or absence of the outcome on those trials. In keeping with the chemical allergy paradigm, participants were informed that the patient information was missing on these trials, rather than being told whether or not a stomach ache occurred. This allowed B and D to be familiar at the end of Stage 1 of the experiment, but for participants to still hold no information about their causal status. It also permitted an examination of the generality of Experiment 1 by testing whether analogous results would be obtained when a different type of causally-ambiguous cue was employed. The design of Experiment 2 is shown in Table 1. All of the other experimental details were the same as Experiment 1. The experimental predictions were also the same, in that cue A would have the greater prediction error at the start of Stage 2, and participants would be less likely to have a theory about B. As before, we expected participants to maintain their belief that A is a cause during Stage 2, instead learning about B, and giving lower ratings for the BC compound than the AD compound at test.

Method

Participants

Thirty-six psychology students from the University of Plymouth participated in this experiment, in return for course credit (26 female, 10 male; mean age = 24.67, SD = 6.81). This sample size has adequate power to detect medium-sized within-subjects effects (83% power at $d = 0.5$). The slightly smaller sample size in this experiment, relative to Experiment 1, was a consequence of recruitment practicalities, but still fell within the threshold of our stopping rule. People who had previously taken part in similar experiments were excluded from this study, to ensure participants were naive to the purpose of the experiment.

Materials

The materials used for Experiment 2 were the same as those used for Experiment 1, with the exception of an additional image and text used for the concealed-outcome trials. The 'information missing' trial feedback was represented by text on screen and an image of a black question mark. As with the other trial feedback, the image was presented within a white rectangle. The dimensions of the image (including the white rectangle) were 291 x 332 pixels.

Design

The experiment used a within-subjects design, as outlined in Table 1. During Stage 1, participants were presented with twelve blocks of training. The six trial types (A+, B?, C+, D?, E-, F-) appeared in a random order within each block. Each trial type was only presented once within each block. Stage 2 and the Test stage were identical to Experiment 1.

Procedure and analysis

Apart from the changes described above, the procedure and analysis for Experiment 2, including Bayesian priors, were the same as for Experiment 1.

Results and Discussion

The trial-level raw data and analysis script for this experiment will be available, upon publication of this manuscript, at https://osf.io/amubk/. The descriptive statistics for the Experiment 2 training stages are shown in Figure 4. The data from Stage 1 indicate that participants learned sufficiently about the six different trial types by the time training was complete. The intermediate responses for B and D are consistent with participants being unsure of the causal status of these cues. Similarly, the data from Stage 2 indicate that participants learned that the AB- compound was non-causal by the time training was complete.

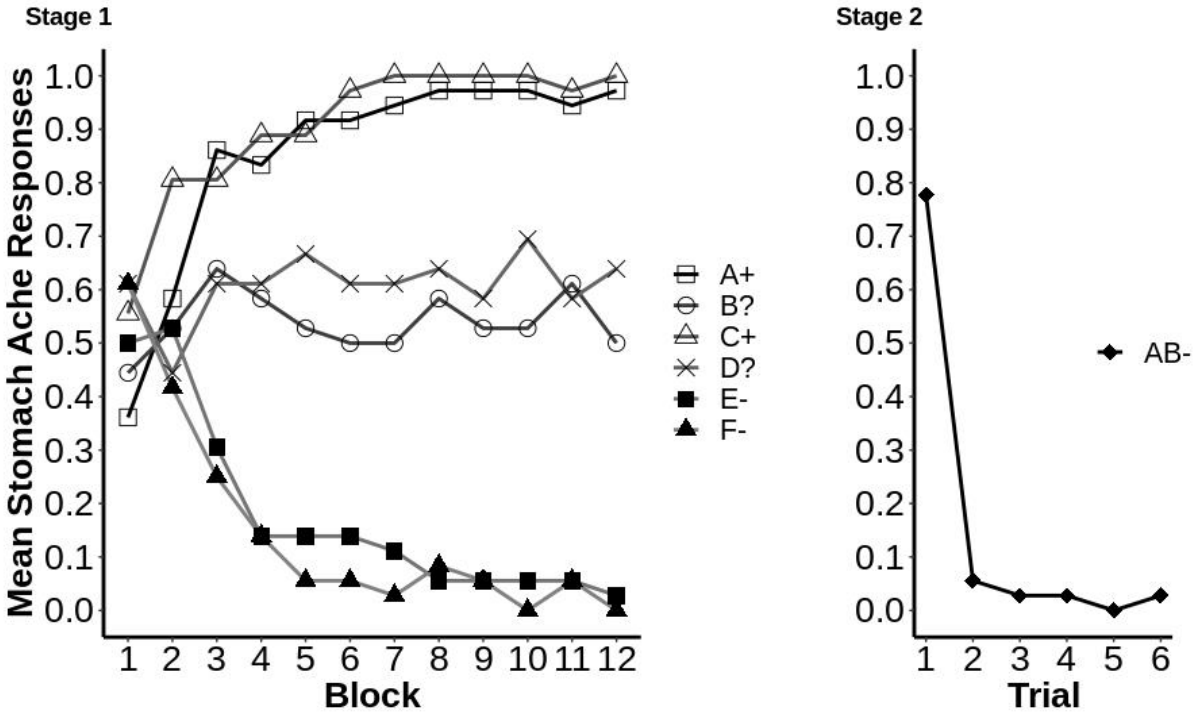

Figure 4. Experiment 2 training Stage 1 and Stage 2 data.

Descriptive statistics for the Experiment 2 Test stage are shown in Figure 5. Ratings for BC were significantly lower than for AD; $t(35) = 3.34$, $p = .002$, $BF = 17.83$, $SE = .53$, $d = .56$. Further testing revealed lower ratings for the single cue B compared to D; $t(35) = 5.06$, $p < .001$, $BF = 2.17 \text{x} 10^4$, $SE = .47$, $d = .84$. Conversely, there was evidence of no difference between the ratings assigned to A and C; $t(35) = .50$, $p = .62$, $BF = .03$, $SE = .22$, $d = .08$. These findings indicate that the differences between the compounds were specifically driven by learning about B during AB- trials. As with the novel cues in Experiment 1, the intermediate rating for D at test is consistent with participants holding no reliable theory about the causal status of concealed-outcome cues. Taken as a whole, these data are again consistent with theory protection (Spicer at al., 2020), as opposed to a prediction error account (Rescorla, 2001). This is because the novel and concealed-outcome cues had a smaller prediction error than the causal cues at the start of Stage 2 in both experiments, as indicated by the intermediate responses for B and D during Stage 1 (for the concealed-outcome cues) and the intermediate ratings for D during the Test stage (for both the novel cues in Experiment 1 and the concealed-outcome cues in Experiment 2).

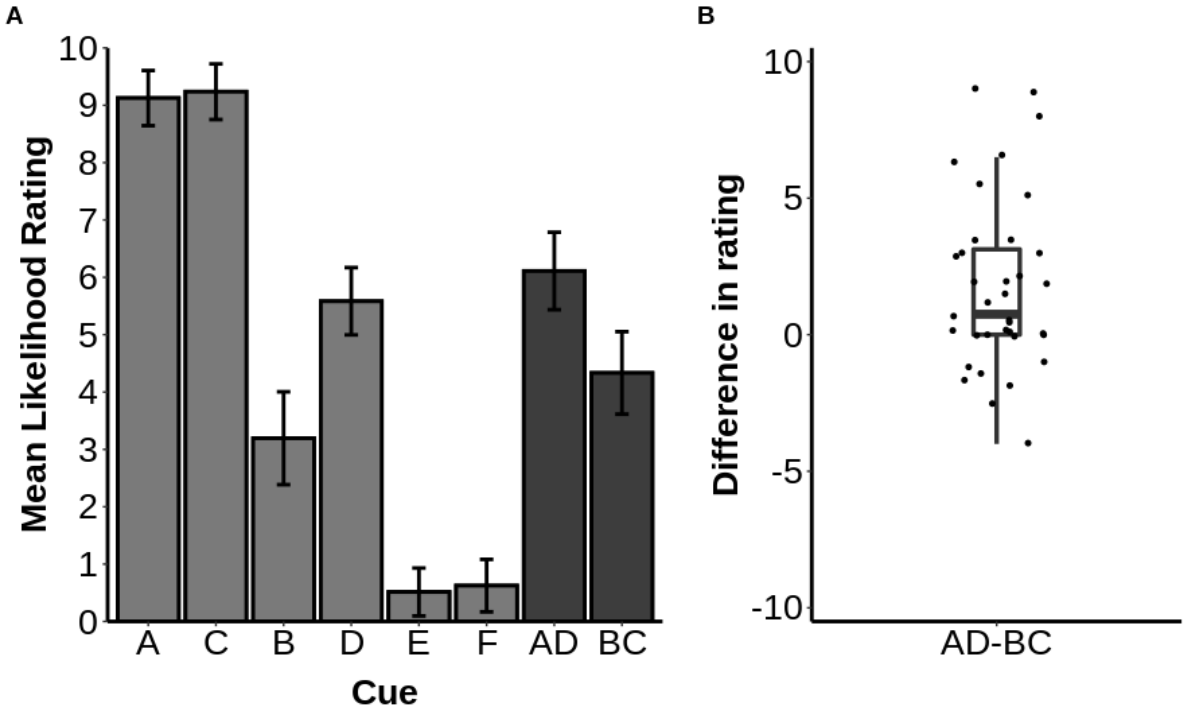

Figure 5. Panel A shows Experiment 2 Test stage ratings for single cues and compounds. Panel B is a plot showing inter-subject variability on the key AD-BC difference.

The results of Experiment 2 build on Experiment 1 by demonstrating that similar results can be observed even when B is not novel at the outset of AB- training. However, while the results of our first two experiments are consistent with theory protection, they can also be explained by Holmes et al.'s (2019) theory. Since participants had no information about B prior to AB- training in either Experiment 1 or Experiment 2, B might have entered Stage 2 with a moderately positive associative strength, and hence would be located at a point on Holmes et al.'s response function with a steep gradient (and thus responding would change rapidly in response to associative strength changes). There are three arguments one could use to support the idea that B has moderately positive associative strength at the end of Stage 1. First, B receives around 50% stomach-ache responses (and hence 50% no stomach ache responses) throughout Stage 1 of Experiment 2 – because participants are forced to guess whether or not it causes stomach ache. Second, given that B and D are treated identically to each other in Stage 1 (absent in Experiment 1; concealed outcome in Experiment 2), but D does not appear in Stage 2, the rating of D at Test may give us some

indication of the status of B at the end of Stage 1. Cue D received intermediate ratings in the Test phase of both experiments, suggesting that participants were unsure of whether or not D (and, by inference, B) would cause the outcome following Stage 1 training. Third, the idea that, in human learning at least, novel cues start with an associative strength greater than zero is consistent with previous work within prediction-error-based accounts; for example, it permits the Rescorla-Wagner model to accommodate the redundancy effect (see Spicer et al., 2021). In order to eliminate the Holmes et al. model's explanation of our results, B would need to have an uncertain causal status, but in a context where associative strength can be assumed to be close to zero (and thus associative updates would cause minimal changes in responding). The design of Experiment 3 fulfils this condition by presenting B in the absence of the outcome during Stage 1.

Experiment 3

Experiment 3 provides a further test of the theory protection and prediction error accounts, as well as testing the proposal of Holmes et al. (2019). Causal ambiguity was achieved by training B as predictive of the absence of the outcome during Stage 1. As stated in the introduction, this experiment followed the same abstract design as Haselgrove and Evans (2010), but using a scenario with chemicals instead of foods. Because our scenario permits inhibitory learning, we expected to observe the opposite result to both Rescorla (2001) and Haselgrove and Evans (2010). The design of Experiment 3 is shown in Table 1. In place of a novel cue (Experiment 1) or a cue with a concealed outcome (Experiment 2), here B was reliably followed by the absence of stomach ache. One notable aspect of this design is that theory protection predicts greater learning about a cue that should have no prediction error at the end of the first Stage 2 trial (B), and little to no learning about a cue with a large prediction error at the end of the first Stage 2 trial (A). This is the opposite of what would typically be expected if learning is governed by prediction error. As with the previous experiments, Rescorla's (2001) account predicts more learning about A than B during Stage 2 because of its larger prediction error. The Holmes et al. (2019) account also predicts a reduction in

responding to A that is at least as great as that for B. This is because the associative strength of both cues should change equally, but this should transfer to a smaller (or equal) change in responding to B because of its low associative strength at the start of AB- trials. Theory protection predicts more learning about B than A on AB- trials, just as in Experiments 1 and 2. This is because, following B- training, the status of B should still be somewhat ambiguous; it might be neutral with regard to stomach ache, or it might be preventative. Therefore, to protect their theory about cue A being causal, we expected participants to infer that B is preventative during AB- training (i.e. reducing uncertainty about B). The inclusion of B- and D- trials in Stage 1 obviated the need for E- and F- trials, as included in Experiments 1 and 2. As before, the Test stage contained two compounds, AD and BC, which permitted a comparison of learning about A and B. Based on theory protection, we predicted higher ratings for AD than for BC. Because Stage 1 training should have allowed participants to learn with confidence that B and D were not causes of stomach ache, we expected the size of the AD-BC difference to be smaller than in previous experiments.

Method

Participants

Forty psychology students from the University of Plymouth participated in this experiment, in return for course credit (35 female, 5 male; mean age = 22.03 SD = 7.55). This sample size has adequate power to detect effects somewhat smaller than those observed in Experiments 1-2. Specifically, the mean effect size for the AD-BC comparison across these two experiments is d = 0.73. At the current sample size, a 38% reduction in that effect size (to d = 0.45) would still result in adequate (80%) power. People who had previously taken part in similar experiments were excluded, to ensure participants were naive to the purpose of the experiment.

Materials

The materials used for Experiment 3 were the same as those used for Experiment 1.

Design

The experiment used a within-subjects design, as outlined in Table 1. During Stage 1, participants were presented with twelve blocks of training. The four trial types (A+, B-, C+, D-) appeared in a random order within each block. Each trial type was only presented once within each block. There were six trials in Stage 2, all with AB-. During the Test stage, participants were presented with two blocks of test cues. The six trial types (A, B, C, D, AD, BC) appeared in a random order within each block. Each trial type was only presented once within each block.

Procedure and analysis

Apart from the changes described above, the procedure for Experiment 3 was the same as for Experiments 1 and 2. The analyses were the same, except that the Bayesian priors for the compound test were updated on the basis of the mean result across Experiments 1 and 2. As explained above, in Stage 1 of Experiment 3, the range of outcomes that could be caused by cue B was reduced, so there was a theoretical basis for expecting the mean test difference to be smaller than the values observed in Experiments 1-2. Therefore, following the recommendations of Dienes (2011), a half-normal prior distribution was specified for the Bayesian t-test on the compounds. This was the positive portion of a normal distribution, with the full distribution having a mean of zero and a standard deviation set to the corresponding mean difference observed across Experiments 1-2.

Results and Discussion

The trial-level raw data and analysis script for this experiment will be available, upon publication of this manuscript, at https://osf.io/vqnbc/. Descriptive statistics for the training stages are shown in Figure 6. The data from Stage 1 indicate that participants learned sufficiently about the four

different trial types by the time training was complete. Similarly, the data from Stage 2 indicate that participants learned that the AB- compound was non-causal by the end of Stage 2. The response data for the first Stage 2 trial is consistent with the participants lacking confidence as to whether B was neutral or inhibitory.

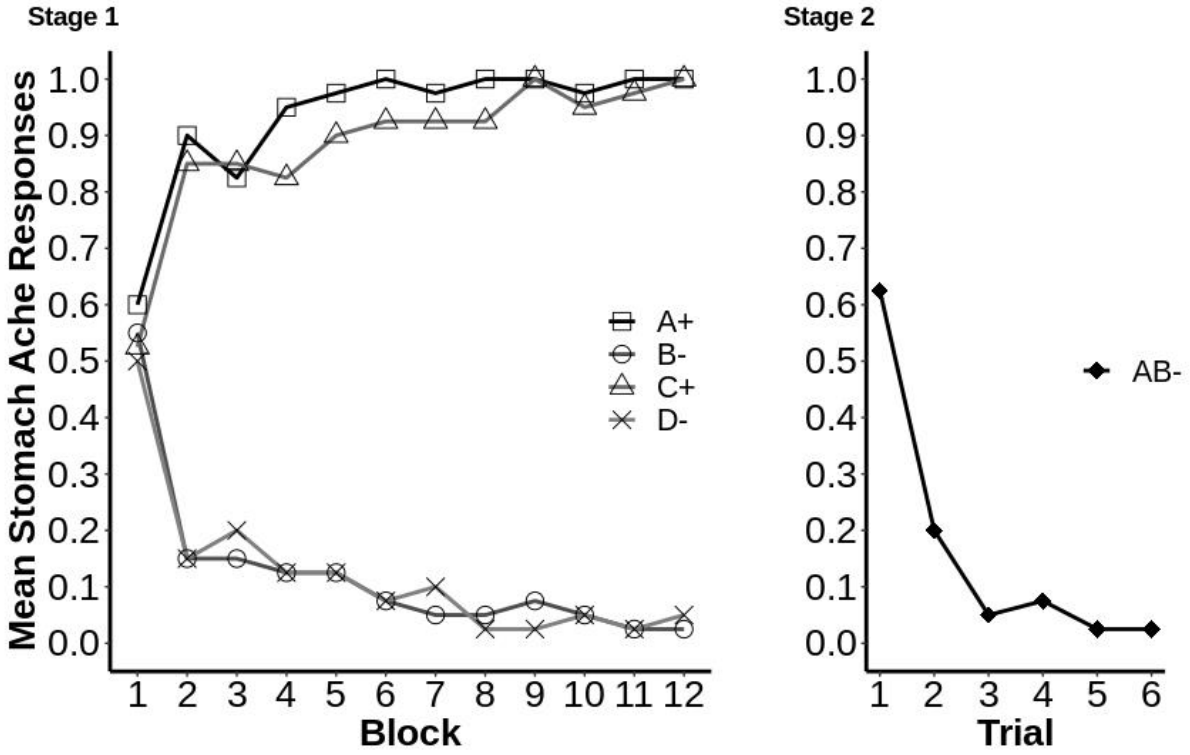

Figure 6. Experiment 3 training Stage 1 and Stage 2 data.

Descriptive statistics for the Test stage are shown in Figure 7. Ratings for BC were significantly lower than for AD; $t(39) = 3.37$, $p = .002$, $BF = 89.84$, $SE = .42$, $d = .53$. The eleven-point rating scale did not allow any difference between neutral and preventative cues to be detected, since both would be given a low rating. Therefore, there was no reason to expect ratings for B and D to differ $t(39) = .54$, $p < .59$, $BF = .05$, $SE = .32$, $d = .09$. As with Experiments 1-2, there was evidence of no difference between the ratings assigned to A and C; $t(39) = .50$, $p = .62$, $BF = .02$, $SE = .06$, $d = .21$. These data are again consistent with theory protection (Spicer at al., 2020) and are, like the results of Experiments 1 and 2, the opposite of the outcome anticipated by a prediction error account. The

results of Experiment 3 are also not permitted by Holmes et al.'s (2019) theory, and the opposite result to that observed by Haselgove and Evans (2010) using a different experimental scenario.

According to theory protection, participants should have lacked confidence about the causal status of B at the end of Stage 1 in spite of learning that it was not a cause, because it could have been neutral or preventative. However, the results of Experiment 3 contain no explicit evidence that this is so. This is because the response method used to predict the presence or absence of the outcome did not allow participants to express any lack of confidence about whether a cue was neutral or preventative. Although the responses made on the first Stage 2 trial might indicate some lack of confidence about whether a stomach ache would occur (see Figure 6), this could be due to either lacking causal confidence about B or some more general uncertainty about the novel combination of cues. Experiment 4 was intended to address this issue by including a measurement of participants' relative confidence about A and B at the end of Stage 1. Experiment 4 also tested whether participants had any sense that B might be to some extent inhibitory, as predicted.

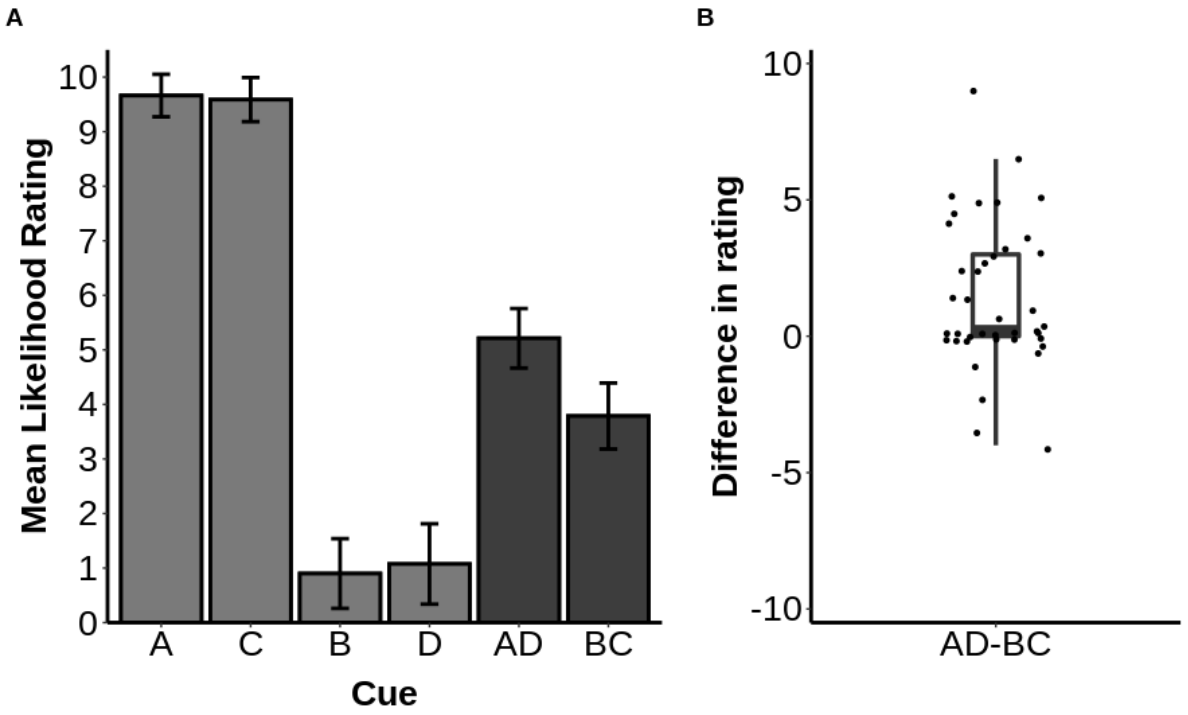

Figure 7. Panel A shows Experiment 3 Test stage ratings for single cues and compounds. Panel B is a plot showing inter-subject variability on the key AD-BC difference.

Experiment 4

In addition to replicating the results of Experiment 3, the aim of Experiment 4 was to assess whether there was a difference in participants' confidence about the causal status of A and B before the start of Stage 2. The design was identical to that of Experiment 3, except for the addition of two extra stages between Stage 1 and Stage 2. The full design of Experiment 4 is shown in Table 2. The first addition was a Probe Test, in which participants were asked to make ratings about cues A and B. Unlike the Test stage ratings, for which participants were asked to rate the likelihood of the outcome given a specific cue or compound, participants were instead asked what they thought each of cues A and B did. Participants were presented with a 21-point scale, ranging from -10 to +10, where +10 meant the cue causes the outcome, 0 meant the cue is neutral, and -10 meant the cue prevents the outcome. The second addition was a Forced Choice stage, in which participants were asked to choose which of those two ratings (i.e. the ratings assigned to A and B on the 21-point scale) they were most confident about. If participants were confident about the causal status of A, but comparatively uncertain about whether B was neutral or inhibitory, then they should have given a high rating to A on the positive end of the Probe Test scale, and an intermediate rating to B in the negative half of that scale. Furthermore, when asked which of these ratings they were most confident about during the Forced Choice stage, they should have chosen A. These findings would support theory protection.

Table 2. The design of Experiment 4

Experiment	Stage 1	Probe Test	Forced Choice	Stage 2	Test
4	A+ C+ B- D-	A B	A or B	AB-	AD BC A C B D

Method

Participants

Sixty-one psychology students from the University of Plymouth participated in this experiment, in return for course credit (51 female, 10 male; mean age = 21.02, SD = 6.01). This sample size has excellent power to detect the key AD versus BC comparison at the effect size observed in Experiment 3 (99% power at $d = 0.53$), excellent power to detect a medium-sized effect on the Probe Test (97% power at $d = 0.5$), and adequate power to detect a medium-to-large effect on the forced-choice test (80% power at $w = 0.36$). A larger sample size was chosen than in the previous experiments, in order to have adequate power for the novel forced-choice test. This choice was made a priori, with an optimum stopping rule of 60 participants tested. As per all previous experiments, we allowed a +/-5 range around that optimum for practical reasons. People who had previously taken part in similar experiments were excluded, to ensure participants were naive to the purpose of the experiment.

Materials

The materials used for Experiment 4 were the same as those used for Experiments 1 and 3.

Design

The experiment used a within-subjects design, as outlined in Table 2. Stage 1, Stage 2 and the Test stage used a design identical to Experiment 3. The Probe Test stage and Forced Choice stage were added between Stages 1 and 2. During the Probe Test, participants were presented with a single block containing only cues A and B. The cues appeared in a random order, and were only presented once within the block. The Forced Choice followed on immediately from the Probe Test. Participants were presented with a single trial, in which cues A and B were presented together on screen. The screen position of A and B was randomised and counterbalanced, so that an equal number of participants saw these cues in each of the two possible left-right configurations.

Procedure

The procedure for Experiment 4 was the same as for Experiment 3, except for the addition of the Probe Test and Forced Choice stages. Following the completion of Stage 1, an instruction screen was shown before the commencement of the Probe Test (see Appendices). For each trial during the Probe Test, the cues were visually presented in the centre of the screen. Text at the top of the screen stated 'Consider the following chemical:', with the cue presented below this. Underneath the cue, further text stated, 'What does this chemical do when ingested by your patient? (-10 = Definitely Prevents Stomach Ache; 0 = Definitely Does Nothing; 10 = Definitely Causes Stomach Ache)'. Participants were instructed to respond by clicking on a 21-point rating scale using their mouse pointer, to indicate what they believed the causal status of the cue to be. The rating scale was located in the lower part of the screen, with the 21-point scale running from left to right, in ascending numerical order. After participants made their response, a blank screen appeared for 0.4 secs, after which the next Probe Test trial was presented. Following the completion of the Probe Test, the Forced Choice test commenced. During the single Forced Choice trial, the cues were visually presented on either the left- or right-hand side of the screen. The cues were randomly pre-assigned for each participant to the left and right positions. Text at the top of the screen stated 'You gave these ratings to the two chemicals you were just asked about:', with the cues presented below this and the Probe Test ratings presented directly beneath each of the respective cues. Underneath the ratings, further text stated 'Which of these two ratings are you most confident (i.e. certain) about?' Participants were instructed to use their computer keyboard to respond (Z for 'Left Rating' and M for 'Right Rating'). After participants made their response, a blank screen appeared for 0.4 secs, after which an instruction screen for Stage 2 was presented (see Appendices). Following the presentation of these instructions, Stage 2 of the experiment commenced and the experiment continued, following the same procedure as Experiment 3.

Analysis

The analyses were the same as for Experiment 1, 2, and 3, except that the Bayesian priors for the Test phase compound test were updated on the basis of the results of Experiment 3 which was identical to Experiment 4 with the exception of the Probe and Forced Choice tests. Following the procedure recommended by Dienes (2011), a normal distribution was specified as the prior for the Bayesian t-test on the compounds, with the mean set to the corresponding Experiment 3 mean and the standard deviation set to half this value. As there was no suitable previous study on which to specify a plausible predicted effect size for the Probe Test, a uniform distribution was specified, with a lower limit of -10 and an upper limit of 10 (for the mean difference of the ratings from zero, and the mean difference between the unsigned ratings). These limits were chosen because 10 is the largest possible difference for these comparisons. The Forced Choice data were analysed with a traditional (null-hypothesis significance test) chi-square and a Bayesian contingency test.

Results and Discussion

The trial-level raw data and analysis script for this experiment will be available, upon publication of this manuscript, at https://osf.io/kwzdr/. The descriptive statistics for the training stages are shown in Figure 8. The data from Stage 1 indicate that participants learned sufficiently about the four different trial types by the time training was complete. Similarly, the data from Stage 2 indicate that participants learned that the AB- compound was non-causal by the end of Stage 2. As before, the intermediate proportion of stomach ache responses on the first Stage 2 trial supports the idea that participants are uncertain about the status of B.

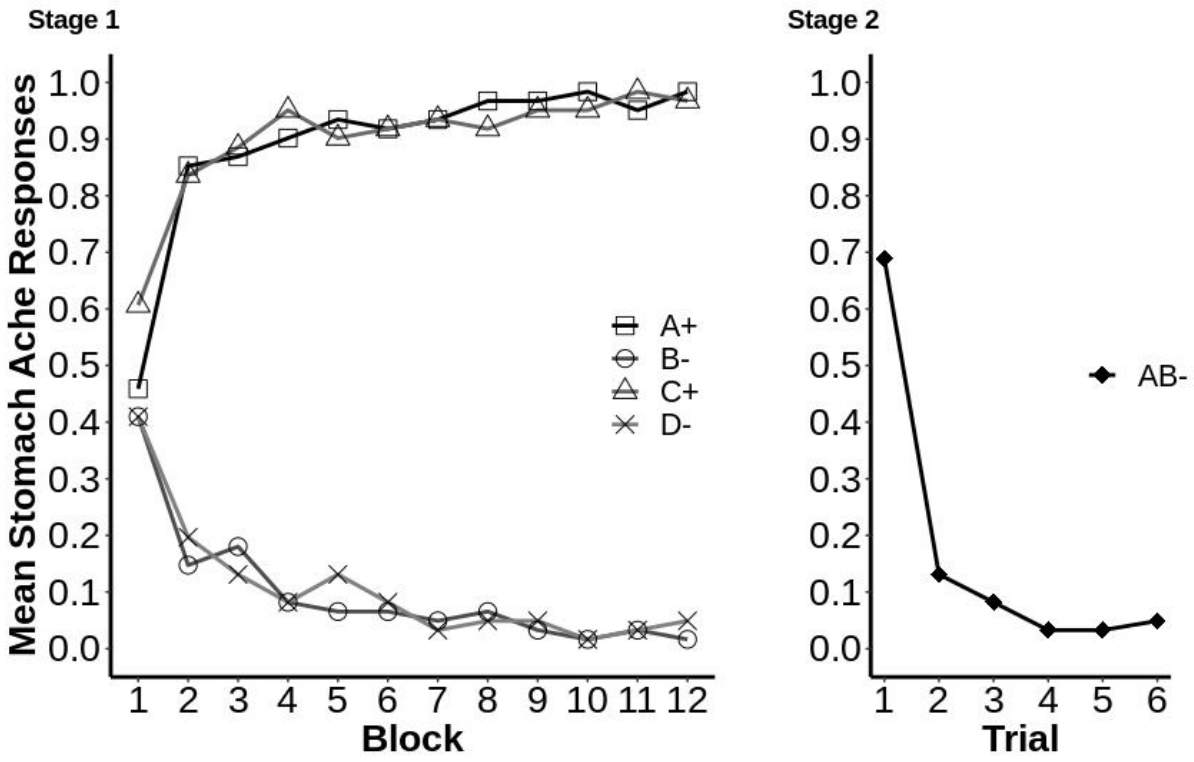

Figure 8. Experiment 4 training Stage 1 and Stage 2 data.

Descriptive statistics for the final Test stage are shown in Figure 9. As in Experiment 3, ratings for BC were significantly lower than for AD; $t(60) = 4.89$, $p < .001$, $BF = 6.57 \times 10^4$, $SE = .37$, $d = .63$. As in Experiment 3, the 11-point Test stage rating scale did not allow any difference between neutral and preventative cues B and D to be detected, $t(60) = .97$, $p < .34$, $BF = 0.04$, $SE = .18$, $d = .12$. There was also evidence of no difference between the ratings assigned to A and C; $t(60) = 1.02$, $p = .31$, $BF = 0.04$, $SE = .21$, $d = .13$, suggesting that nothing was learned about A during Stage 2. As a replication of Experiment 3, these findings provide further support to the theory protection proposal.

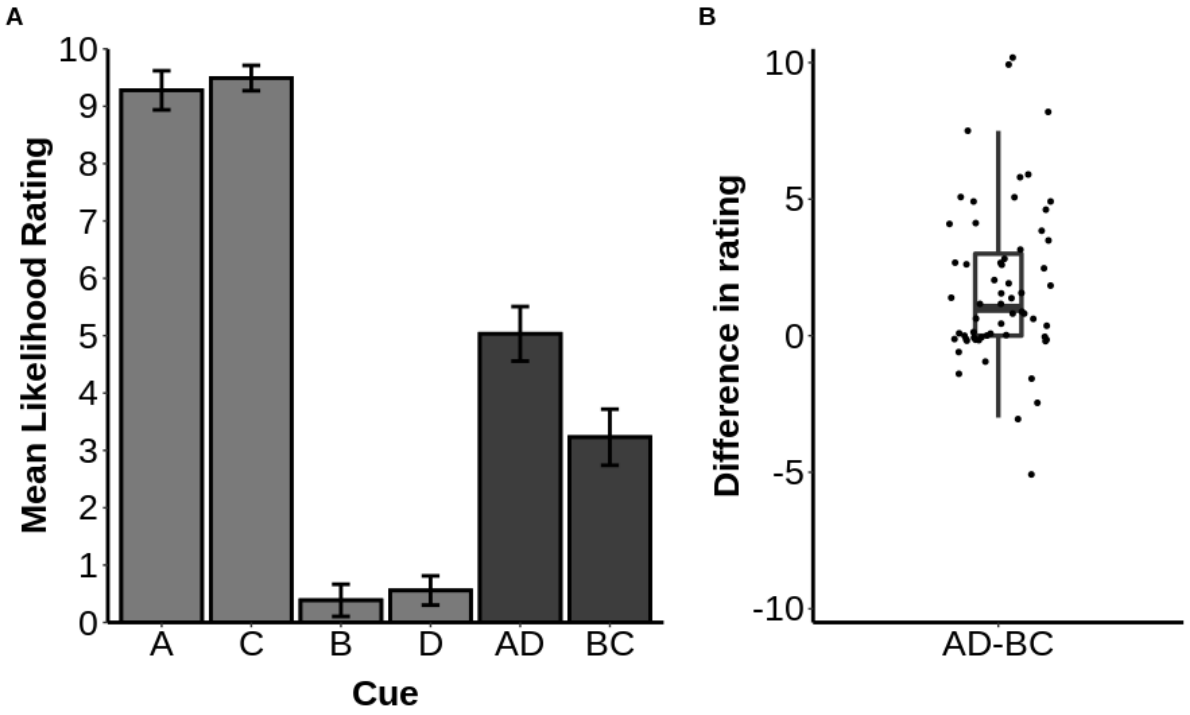

Figure 9. Panel A shows Experiment 4 Test stage ratings for single cues and compounds. Panel B is a plot showing inter-subject variability on the key AD-BC difference.

The descriptive statistics for the Experiment 4 Probe Test are shown in Figure 10. Ratings for A were significantly greater than zero; $t(60) = 13.05$, $p < .001$, $BF = 7.55 \times 10^{35}$, $SE = .63$, $d = 1.67$. Ratings for B were significantly below zero; $t(60) = 7.89$, $p < .001$, $BF = 2.34 \times 10^{12}$, $SE = .57$, $d = 1.01$. The ratings for cue A were significantly further from zero than the ratings for cue B (i.e. the unsigned difference from zero was greater for A); $t(60) = 8.56$, $p < .001$, $BF = 5.28 \times 10^{14}$, $SE = .53$, $d = 1.10$. The high positive mean rating for A, compared to the intermediate negative mean rating for B, suggests greater confidence about the causal status of A than of B. These data support the prediction that participants would not know whether B was neutral or inhibitory. The Forced Choice results further support this view, since 51 participants chose their rating for cue A as the one they were most confident about, while only 10 chose B. A chi-square test demonstrated that this was significantly different to chance; $X^2(60) = 27.56$, $p < .001$, $BF = 582$, $w = .73$. These findings indicate that participants lacked confidence about the causal status of B prior to the start of Stage 2, despite having the opportunity to learn that it was not a cause. We propose that this lack of knowledge facilitated subsequent learning about B, while confidence about the causal status of A

meant participants protected their theory about it, in spite of a larger prediction error. Deducing that B was preventative of stomach ache is consistent with B- trials in Stage 1, so participants could also protect their theory that B was not a cause.

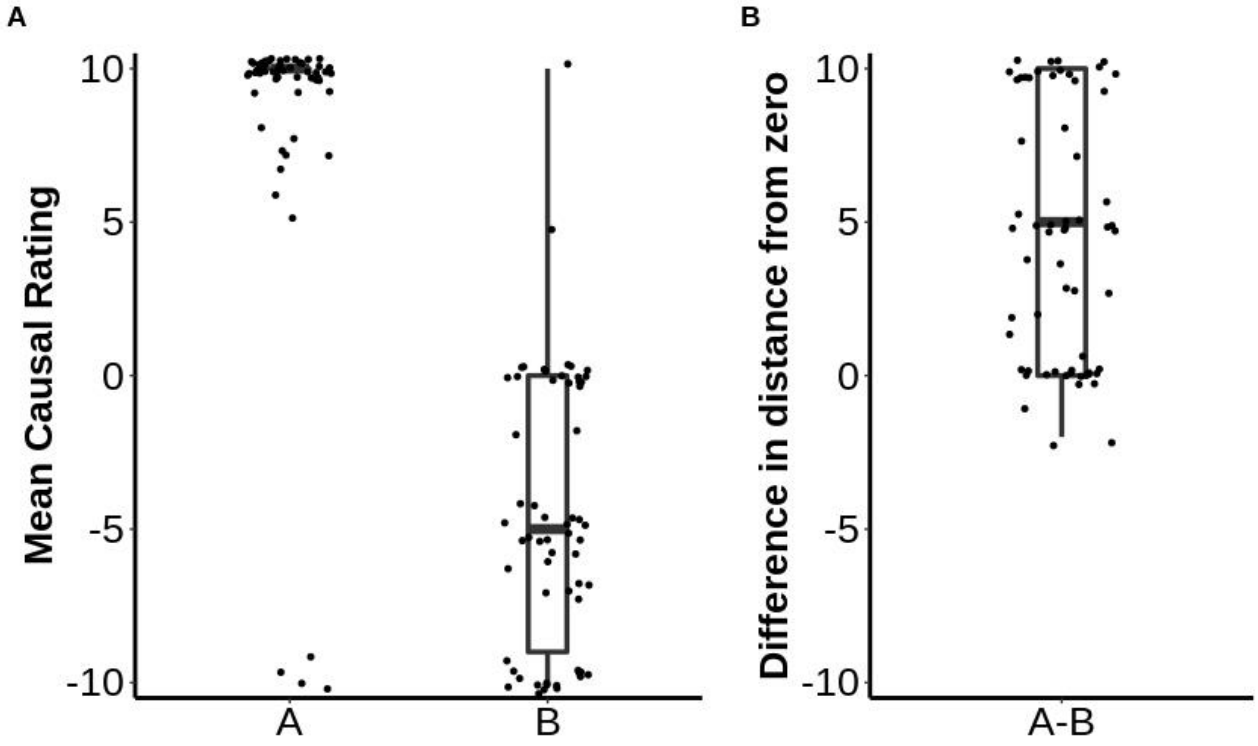

Figure 10. Panel A shows Experiment 4 Probe Test ratings for cues A and B. Panel B shows inter-subject variability on the difference between the unsigned differences from zero for these cues. Note that the left hand boxplot in panel A is masked by the concentration of individual participants providing that cue with a high rating.

General Discussion

The findings of these four experiments are consistent with the concept of theory protection in human associative learning (Spicer et al. 2020). It is therefore possible that humans protect existing theories about the causal status of cues, instead attributing unexpected outcomes to cues about which they do not hold a strong theory. These findings are not readily explainable by a prediction error account of learning (Rescorla, 2001). This is because in each experiment our results suggested greater learning about the cue with the smaller prediction error – and learning about a cue which

should have had no (or little) prediction error in Experiments 3 and 4. Furthermore, Experiment 4 provides evidence of a difference in participants' confidence before the compound training stage, supporting the suggestion that participants will protect their theory about a cue if they are confident about it.

Our results are not only the opposite of those observed in non-human animals (Rescorla, 2001), but also in human participants by Haselgrove & Evans (2010). We propose that their food allergy scenario imposed a floor on learning that both prevented inhibition and removed any potential for uncertainty about non-reinforced cues (i.e. B-). Conversely, in all four of our current experiments using a chemical allergy scenario, participants could learn that B was inhibitory during the AB- trials, accounting for the absence of the outcome that would have been caused by A alone. By extension, the current experiments provide evidence that results resembling theory protection can occur with inhibitory learning (AB- in Stage 2), as well as in excitatory learning, as previously demonstrated by Spicer et al. (2020). Consequently, the discrepancy between the results of the current Experiment 3 and Haselgrove and Evans' results may be well explained by differences in the level of uncertainty induced by different scenarios. The idea presented in the introduction, that excitatory learning might be governed by a theory protection process, but inhibition by prediction error is not supported by the current data.

The results of Experiments 3 and 4 are inconsistent with Holmes et al.'s (2019) model, which suggests a nonlinear mapping from learning to behaviour. Their model can account for previous results (Haselgrove and Evans, 2010; Rescorla, 2001; Spicer et al., 2020) because the cue with the greater change in responding during the compound conditioning phase had an associative strength closer to zero, where updates in associative strength result in minimal changes to responding in their model. However, in Experiments 3 and 4 here, the cue learned about the most is the one with the starting associative strength closer to zero. Even if A had a very high associative strength at the start

of AB- training, at best this would result in equal changes in responding for both cues. Consequently, Holmes et al.'s theory does not adequately predict the results of these experiments. However, we do not rule out the possibility that a modification of their model might capture our results. It should be possible to represent theory protection in a mathematical model and, as Jones et al. (2021) noted, a model of theory protection would share some properties with that proposed by Holmes et al.

Among other candidates for explaining our results are theories that propose changes in the amount of attention paid to cues. For instance, Pearce and Hall (1980) described a model in which attention declines for cues that are followed by predicted outcomes. As a result, their model predicts less learning once the relationship between a cue and its consequences are known. This has some conceptual similarity to theory protection, since both approaches predict most learning when there is uncertainty. However, a crucial difference is that Pearce and Hall's model predicts most learning when there is uncertainty about what the outcome will be, whereas the theory protection proposal predicts that there will be most learning when there is uncertainty about a cue. This difference in focus is important because, in Experiments 3 and 4 here, the outcome following B during Stage 1 was just as reliable as that following A, so Pearce and Hall's theory predicts that decreases in attention for B should have been just as large as those for A, assuming that predicting the presence versus absence of an outcome will result in equivalent declines in associability. This makes it hard to reconcile their theory with our results. Alternatively, theory protection allows for more subsequent learning about B than A because, despite the outcome of each trial being equally certain, participants could not be as certain about the status of B as for A. These theories therefore differ in their application to the current experiments, and our results are more consistent with theory protection.

An alternative theory based on attention was described by Mackintosh (1975). Mackintosh's theory proposes that attention is greatest for cues that are good predictors of outcomes. Specifically, the best available predictor of each trial's outcome receives a boost in associability, while others receive a decrease. Future learning is the product of this updated associability and a prediction error that is equivalent to that proposed by Bush and Mosteller (1951). Mackintosh's model (as originally devised) is unable to account for the results of Experiments 3 and 4, since it predicts no learning for cues with no prediction error. However, it is possible that another model containing a similar attentional process (Kruschke, 2001, 2006; Le Pelley, 2004) might provide a better fit for our results. It is also worth noting that all four experiments relied on extinction – an effect that has previously been demonstrated to result in little or no associative loss (Delamater, 1996) in a study with rats. At first blush, this seems like a simple associative explanation for our results – and one which is not unique to humans. However, the three Spicer et al. (2020) experiments showed equivalent results with excitation, so this interpretation would not apply to those experiments. Furthermore, the lack of associative loss only applies to A in the current experiments, while there is apparent associative loss for B.

A useful extension of the current research might be an investigation of whether theory protection is relevant in circumstances where there is limited scope for uncertainty about the causal status of cues. Le Pelley and McLaren (2001) trained participants in a similar manner to Experiment 3, except that B and D were explicitly trained as inhibitors of the outcome in Stage 1 (BE- DE- E+), along with two excitatory cues (A+ C+). In the next stage, one excitor and one inhibitor were paired in a compound (AB-). Similarly to our current findings, a compound test stage revealed greater learning about B than A, despite the latter having the greater prediction error during the compound training stage. This experiment differs from those presented in this paper in that participants should have been relatively confident about the causal status of both A and B at the end of Stage 1 training. One interpretation of

this result is that participants protected their theory about A being a cause of the outcome, while strengthening their theory that B prevented the outcome.

Le Pelley and McLaren proposed a role for associative history in explaining their findings, suggesting the APECS model (Le Pelley & McLaren, 2004; also see Le Pelley et al., 2000) as a possible implementation of this idea. APECS allows for generalisation between stimuli on the basis of whether they predict outcomes that are the same or different, with greater generalisation occurring when outcomes are the same. Recall that Le Pelley and McLaren (2004) pre-trained two cues in a compound that predicted an outcome (AB+) alongside two cues that predicted the absence of that outcome (CD-), and subsequently paired A and C in a compound that predicted the outcome (AC+). Because AC+ produces the same outcome as AB+, this generalisation produces more learning about A than C, since C was previously predictive of a different outcome. In other words, APECS suggests that selectivity in learning is affected by whether outcomes are the same or different. Instead, theory protection suggests that selectivity is affected by the causal status of previously learned cues being protected in the face of subsequently encountered prediction errors. Future research should test these two accounts.

The instructions in our experiments explicitly informed participants that some cues could be inhibitory. Whilst this information formed part of the experimental scenario, such instructions could influence human participants into learning that B is an inhibitor in each experiment – something not possible in non-human animals. Whilst it is true that instructions inevitably influence the type of learning observed in human predictive learning studies, we suggest that if participants are unaware that inhibition is possible then they will effectively be primed to learn that cues can only cause outcomes, rather than learning that cues can either cause or prevent outcomes. However, participants can be made aware that cues can prevent outcomes using pre-training (e.g. Le Pelley & McLaren, 2001), as well as instructional information. A useful follow up to these experiments

would be to either omit mention of preventative cues in the instructions, or to pre-train participants with inhibitors (using a unique set of cues that do not appear in the main experiment – unlike Le Pelley & McLaren) instead of providing this information via the instructions.

Despite the apparent inconsistency between a prediction error account and the results of the present experiments, we do not suggest that a role for prediction error in learning should be dismissed. In many cases, learning may be necessitated by a discrepancy between a predicted outcome and an actual outcome. This 'surprise' may therefore dictate whether learning takes place, and how much learning there will be. Meanwhile, the theories people have about the causal statuses of cues, and the confidence with which they hold those theories, may dictate which cues are learned about in order to predict future events more accurately. All four of the experiments in this paper introduce a surprising outcome omission in Stage 2, because a previously causal cue and an ambiguous cue are paired in a non-causal compound. In each experiment, the discrepancy between the stomach ache predicted following A as a result of Stage 1, and the absence of stomach ache following the Stage 2 AB- compound, presumably leads to learning about B. It is likely that, in the absence of surprise, no learning would have taken place.

To conclude, the experiments presented here are consistent with theory protection. It is possible that human participants maintain causal associations with respect to cues that have a known causal status, instead attributing unexpected outcomes to cues with a comparatively ambiguous status. This view is also supported by previous findings from experiments using similar compound testing procedures (Mitchell et al., 2008; Spicer et al., 2020). Our findings in Experiments 3 and 4 were the opposite of those found in rats and pigeons by Rescorla (2001), suggesting that further comparative studies between species might be valuable. However, they were also the opposite of human results reported by Haselgrove and Evans (2010), suggesting that experimental scenario affects learning by influencing uncertainty. Our findings in Experiments 3 and 4 were inconsistent with the prediction

of Holmes et al.'s (2019) model. Experiment 4 demonstrated a difference in confidence about cues that is consistent with theory protection. Nevertheless, we do not suggest that prediction error should be dismissed, and recognise that a number of prediction error accounts require further investigation, particularly those that incorporate a role for attentional processes (Kruschke, 2001, 2006; Le Pelley, 2004). Nonetheless, the results suggest that theory protection warrants further investigation.

References

Baguley, T., & Kaye, D. (2010). Book review: Understanding psychology as a science: An introduction to scientific and statistical inference. *British Journal of Mathematical and Statistical Psychology, 63*, 695–698.

Bartlett, F. C. (1932). *Remembering: A study in experimental and social psychology*. Cambridge, England: Cambridge University Press.

Bush, R. R., & Mosteller, F. (1951). A mathematical model for simple learning. *Psychological Review, 58*, 313–323.

Chan, Y. Y., Westbrook, R. F., & Holmes, N. M. (2021). Protecting the Rescorla-Wagner (1972) theory: a reply to Spicer et al. (2019). *Journal of Experimental Psychology: Animal Learning and Cognition*, *47*(2), 211–215.

Delamater, A. R. (1996). Effects of several extinction treatments upon the integrity of Pavlovian stimulus-outcome associations. *Animal Learning & Behavior*, *24*(4), 437-449.

Dienes, Z. (2011). Bayesian versus orthodox statistics: Which side are you on? *Perspectives on Psychological Science, 6*, 274–290.

Gluck, M. A., & Bower, G. H. (1988). From conditioning to category learning: an adaptive network model. *Journal of Experimental Psychology: General, 117*, 227-247.

Harman, G. (1986). *Change in view: Principles of reasoning*. Cambridge, MA: MIT Press.

Haselgrove, M., & Evans, L. H. (2010). Variations in selective and nonselective prediction error with the negative dimension of schizotypy. *The Quarterly Journal of Experimental Psychology*, *63*(6), 1127-1149.

Holmes, N. M., Chan, Y. Y., & Westbrook, F. (2019) A combination of common and individual error terms is not needed to explain associative changes when cues with different training histories are conditioned in compound: A review of Rescorla's compound test procedure. *Journal of Experimental Psychology: Animal Learning and Cognition, 45,* 242-256.

Jeffreys, H. (1961). *The Theory of Probability (3rd Ed.)*. Oxford: Oxford University Press.

Jones, P. M., Mitchell, C. J., Wills, A. J., & Spicer, S. G. (2021). Similarities and differences: Comment on Chan et al. *Journal of Experimental Psychology: Animal Learning and Cognition, 47*(2), 216–217.

Jones, P. M., Zaksaite, T., & Mitchell, C. J. (2019). Uncertainty and blocking in human causal learning. *Journal of Experimental Psychology: Animal Learning and Cognition, 45*, 111-124.

Kamin, L. J. (1969). Selective association and conditioning. In N. J. Mackintosh & W. K. Honig (Eds.), *Fundamental Issues in Associative Learning* (pp. 42–64). Halifax, Canada: Dalhousie University Press.

Kruschke, J. K. (2001). Toward a unified model of attention in associative learning. *Journal of mathematical psychology*, *45*(6), 812-863.

Kruschke, J. K. (2006). Locally Bayesian learning with applications to retrospective revaluation and highlighting. *Psychological review*, *113*(4), 677.

Le Pelley, M. E. (2004). The role of associative history in models of associative learning: A selective review and a hybrid model. *The Quarterly Journal of Experimental Psychology Section B*, *57*(3b), 193-243.

Le Pelley, M. E., Cutler, D. L., & McLaren, I. P. L. (2000). Retrospective effects in human causality judgment. In L. R. Gleitman & A. K. Joshi (Eds.), *Proceedings of the Twenty-Second Annual Conference of the Cognitive Science Society* (pp. 782–787). Hillsdale, NJ: Erlbaum.

Le Pelley, M. E., & McLaren, I. P. L. (2001). The mechanics of associative change. *Proceedings of the 27th Annual Conference of the Cognitive Science Society*. (pp. 534 –539). Hillsdale, NJ: Erlbaum.

LePelley, M. E., & McLaren, I. P. L. (2004). Associative history affects the associative change undergone by both presented and absent cues in human causal learning. *Journal of Experimental Psychology: Animal Behavior Processes*, *30*(1), 67.

Lubow, R. E., & Moore, A. U. (1959). Latent inhibition: the effect of nonreinforced pre-exposure to the conditional stimulus. *Journal of comparative and physiological psychology, 52*, 415-419.

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

Mitchell, C. J., Harris, J. A., Westbrook, R. F., & Griffiths, O. (2008). Changes in cue associability across training in human causal learning. *Journal of Experimental Psychology: Animal Behavior Processes, 34*, 423-436.

Pearce, J. M., & Hall, G. (1980). A model for Pavlovian conditioning: Variations in the effectiveness of conditioned but not of unconditioned stimuli. *Psychological Review, 87*, 532–552.

Peirce, J. W. (2007). PsychoPy - Psychophysics software in Python. *Journal of Neuroscience Methods, 162*, 8-13. www.psychopy.org. (Version 1.90.3).

R Core Team. (2018). *R: A language and environment for statistical computing*. www.r-project.org. (Version 3.5.3)

Rescorla, R. A. (2000). Associative changes in excitors and inhibitors differ when they are conditioned in compound. *Journal of Experimental Psychology: Animal Behavior Processes, 26*, 428–438.

Rescorla, R. A. (2001). Unequal associative changes when excitors and neutral stimuli are conditioned in compound. *Quarterly Journal of Experimental Psychology, 54B*, 53–68.

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

Spicer, S. G., Mitchell, C. J., Wills, A. J., and Jones, P. M. (2020). Theory protection in associative learning: humans maintain certain beliefs in a manner that violates prediction error. *Journal of Experimental Psychology: Animal Learning and Cognition*, *46*(2), 151.

Spicer, S.G., Wills, A.J., Jones, P.M., Mitchell, C.J. and Dome, L. (2021). Representing uncertainty in the Rescorla-Wagner model: Blocking, the redundancy effect, and outcome base rate. *Open Journal of Experimental Psychology and Neuroscience, 1,* 14-21.

Uengoer, M., Lotz, A., & Pearce, J. M. (2013). The fate of redundant cues in human predictive learning. *Journal of Experimental Psychology: Animal Behavior Processes*, *39*(4), 323.

Zaksaite, T., & Jones, P. M. (2019). The redundancy effect is related to a lack of conditioned inhibition: Evidence from a task in which excitation and inhibition are symmetrical. *Quarterly Journal of Experimental Psychology.*

Appendices

Experiment 1 Instructions (Training):

This study is concerned with the way in which people learn about relationships between events. In the present case, you should learn whether the consumption of chemicals used in drug research, leads to an allergic reaction.

Imagine that you are working in a drug research laboratory, studying chemicals for potential use in medication. You are trying to identify which chemicals cause the side effect of a stomach ache in your test patient.

To identify which chemicals they react to, the patient ingests specific chemicals and observes whether a stomach ache occurs or not. The results of these tests are shown to you on the screen one after the other.

You will then be asked to predict whether the patient suffers from stomach ache. For this prediction, please click on the appropriate response button. After you have made your prediction, you will be informed whether your patient suffered from stomach ache or not. Use this feedback to find out which chemicals cause a stomach ache in your patient.

You will always be told what your patient has ingested. Sometimes, they have only consumed a single chemical and other times they have consumed two different chemicals. Please look at the chemicals carefully.

At first you will have to guess the outcome because you do not know anything about your patient. But eventually you will learn which chemicals lead to stomach ache in this patient and you will be able to make correct predictions.

All of the chemicals being studied can be easily identified by a unique logo. Each logo is both a different shape and a different colour.

For all of your answers, accuracy rather than speed is essential. Please do not take any notes during the experiment. If you have any questions, please ask them now.

Please note, some chemicals will cause a stomach ache, while others will be neutral and will not cause a stomach ache. However, it is also possible for specific chemicals to actively PREVENT a stomach ache from occurring in your patient.

If you do not have any questions, please start the experiment by pressing the space bar.

Experiment 1 Instructions (Test):

Next, your task is to judge the probability with which specific chemicals cause stomach ache in your patient. Single chemicals and pairs of chemicals will be shown to you on the screen.

In this part of the experiment, you will receive no feedback about the actual reaction of the patient. Use the information that you have collected so far, to make your rating.

Press space bar to continue the experiment.

Experiment 4 Instructions (Probe Test):

Next, your task is to make ratings about two of the chemicals you have learned about. Firstly, you will be asked what these chemicals do when ingested by your patient (e.g. cause stomach ache, prevent stomach ache, neutral). You will be able to make your response using a rating scale.

Once you have made your ratings, you will be asked which of those ratings you feel most confident (i.e. certain) about. You will be able to make a response using a key press.

In this part of the experiment, you will receive no feedback about the actual reaction of the patient. Use the information that you have collected so far, to make your choices.

Press space bar to continue the experiment.

Experiment 4 Instructions (Stage 2):

Next, you will continue to learn about the chemicals used in drug research, as you did during the first part of the experiment. As before, the patient ingests specific chemicals and observes whether a stomach ache occurs or not. The results of these tests are shown to you on the screen one after the other.

You will then be asked to predict whether the patient suffers from stomach ache. For this prediction, please click on the appropriate response button. After you have made your prediction, you will be informed whether your patient suffered from stomach ache or not.

Press the space bar to continue.