Theory protection in associative learning: humans maintain certain beliefs in a manner that violates prediction error

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

Plymouth University

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 Peter M. Jones, School of Psychology, Plymouth University, Plymouth, PL4 8AA, United Kingdom. Email: peter.m.jones@plymouth.ac.uk

Author contributions

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. Consulted on analysis and write up as PhD supervisor.

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

Andy J. Wills: Consulted on analysis and write up as PhD supervisor.

Abstract

Three experiments were conducted to investigate a possible role for certainty in human causal learning. In these experiments, human participants were initially trained with a set of cues, each of which was followed by the presence or absence of an outcome. In a subsequent training stage, two of these cues were trained in a causal compound, and the change in associative strength for each of the cues was compared, using a procedure based on Rescorla (2001). In each experiment, the cues differed in both their causal certainty (on the part of participants) and size of their prediction error (with respect to the outcome). The cue with the larger prediction error was always the cue with the more certain causal status. According to established prediction error models (Bush & Mosteller, 1951; Rescorla & Wagner, 1972; Rescorla, 2001), a larger prediction error should result in a greater updating of associative strength. However, the opposite was observed, as participants always learned more about the cue with the smaller prediction error. A plausible explanation is that participants engaged in a form of theory protection, in which they were resistant to updating their existing beliefs about cues with a certain causal status. Instead, participants appeared to attribute outcomes to cues with a comparatively uncertain causal status, in an apparent violation of prediction error. The potential role of attentional processes (Mackintosh, 1975; Pearce & Hall, 1980) in explaining these results is also discussed.

Keywords: associative learning, prediction error, theory protection, redundancy effect, uncertainty

Prediction error is one of the most common components in theories of how associations between cues and outcomes are learned (e.g. Bush & Mosteller, 1951; Mackintosh, 1975; Rescorla & Wagner, 1972). According to such theories, learning occurs when humans and non-human animals encounter surprising outcomes. Specifically, associations are formed when there is a discrepancy (i.e. error) between an expected outcome and an experienced outcome. For example, if a person expects that a type of food is safe to eat, but they experience an allergic reaction after eating that food, then learning will take place and expectations about that food will update accordingly. If they eat the same food on subsequent occasions and an allergic reaction occurs again, then further learning will take place. This learning will continue until the asymptote of learning is reached. At this point, no more learning can take place, because there is no longer any error between what is expected and what is experienced. In other words, if a person correctly anticipates that a food will cause an allergic reaction then there will be little or no learning, because there is little or no error. Consequently, the amount of learning is greatest when there is a large error and smallest when there is a small error. Much of the research supporting prediction error as a determinant of learning stems from animal conditioning research (e.g. Pavlov, 1927; Kamin, 1969; Rescorla, 2001).

Not all kinds of prediction error are the same. According to Bush and Mosteller's (1951) theory, learning about each cue is determined by the discrepancy between the outcome that occurs and the outcome that was predicted by that cue alone. This is referred to as individual prediction error. However, following the observation that knowledge about one cue can interfere with learning about another (e.g. the blocking effect, Kamin, 1969), Rescorla and Wagner (1972) suggested that learning was instead governed by overall prediction error; that is, the discrepancy between the outcome that occurs and an aggregate prediction derived from all the cues that are present. Subsequently, Rescorla (2001) provided evidence that learning is governed by a mixture of individual prediction error and overall prediction error. In other words, when more than one cue is

present, learning can be greater (but not smaller) for those cues whose individual causal status is most discrepant from the outcome. For example, in one of the experiments (where the cues are represented by different letters and the presence or absence of the outcome is represented by a + or - respectively), Rescorla initially trained rats with two excitatory cues (A+ and C+). The rats were subsequently trained with a compound containing one previously excitatory cue and a neutral cue, with no outcome (AB-). A compound testing procedure was then used to compare learning about A and B on the AB- trials. Each test compound contained both a previously trained excitatory cue and a neutral cue that was absent from the A+ C+ training phase (AD and BC). In the absence of the AB- stage, both compounds would have resulted in equal responding at test, since both would have contained cues with equivalent status. Any difference between responses for AD and BC during the test must therefore have been a consequence of learning during the AB- trials. This procedure was required because associative strengths do not necessarily map onto experimental test responses via a linear function (e.g. Gluck & Bower, 1988), making it hard to directly compare any change in associative strength for individual cues that are trained from different starting strengths. The procedure thus provided a sophisticated method for assessing differential changes in the underlying associative strength of cues A and B during the AB- training stage. According to individual prediction error (e.g. Bush & Mosteller, 1951), cues presented in compound each have their own error term. This means that they enter into cue-outcome associations in the same manner as cues presented individually, as long as the salience of the cues and the learning rate parameter remain unchanged. Under an individual error account, cue A would be learned about more than cue B on presentation of the AB- compound, because of the larger prediction error for A at the start of the AB- trials (since A was learned as a cause of the outcome, but nothing was previously known about the causal status of B). According to overall prediction error, however, learning about A and B on AB- trials should have been the same. This is because cues presented together would account for an equal share in any change in associative strength, as a consequence of the aggregated error. In this

experiment, the results supported learning via individual error, with the test compound containing A (AD) eliciting less responding than the compound containing B (BC), indicating a greater decrease in associative strength for the former. Rescorla's (2001) proposal of a hybrid of individual and overall prediction error was subsequently challenged by Holmes, Chan and Westbrook (2019), who showed that an overall prediction error alone can account for these findings if the function mapping associative strength to responding is appropriately modified. The logic of this is that a non-linear mapping function allows equal changes in associative strength to result in unequal changes in responding, as well as unequal summation of individual cues into compounds.

Although Rescorla (2001) demonstrated a greater change in responding (and potentially a greater change in learning) for cues with a larger prediction error, this seems at odds with what one might expect in human causal learning. During A+ trials participants would presumably form a belief that A causes the outcome. On the other hand, B should be somewhat ambiguous after B- training because there is no information available to indicate whether B is neutral or inhibitory. To produce a result that is analogous to Rescorla's, participants would need to change their existing beliefs about A rather than attributing a surprising outcome to B, which is ambiguous. Perhaps a more intuitive prediction would be for human participants to maintain their belief about the causal status of A, and instead attribute the absence of the outcome (during the AB- trials) to B, resulting in more learning about B. If true, this would run contrary to Rescorla's proposal. Even a modified response function (Holmes et al., 2019) should not permit this, since a cue with a smaller prediction error should not be learned about more than a cue with a larger prediction error, assuming both cues are equally salient. In the case of two cues being trained in compound, the idea that the cue undergoing a larger change in associative strength would result in the smaller change in responding seems implausible. However, the view that human participants might maintain existing associations, in spite of a large prediction error, does seem plausible. In fact, such a view is consistent with theoretical approaches

seen in other cognitive fields. For example, the suggestion that humans should resist updating their beliefs has parallels with theories about schemata (e.g. Bartlett, 1932), in terms of humans accommodating new information into existing frameworks. Also, the processing of new cue-outcome associations, in a manner that fits with existing causal associations, is comparable to confirmation bias (Wason, 1960). Additionally, attribution of an unexpected outcome to one cue, without beliefs about other cues needing to be updated, could be considered efficient in a manner that is similar to the concept of the cognitive miser (Fiske & Taylor, 1984). Despite the influence of these theories in fields such as decision making, there is currently no formal theory of associative learning that implements this kind of process. A possible reason is the apparent lack of an obvious variable to govern learning. However, certainty about the causal status of cues provides a promising candidate.

A testable prediction that formed the rationale for the three current experiments was that greater certainty about cues should lead to reduced learning (i.e. reduced causal updating). In other words, the more participants are certain about the causal status of a cue, the more resistant they should be to updating its causal status through subsequent learning. By extension, if participants are uncertain about the causal status of a cue, this should facilitate the updating of beliefs about that cue via learning. It is possible to test this account against a simple prediction error account (i.e. individual prediction error, overall prediction error, or a mixture of both), by training two cues that are known to differ in their causal certainty (from the perspective of participants) in a compound, in order to compare the amount of learning for each cue. In the absence of a formal theory of certainty and learning, we propose a simple theory protection account for cases in which two cues are trained in compound. According to this account, human participants should engage in theory protection by showing little or no learning about the cue with the most certain (i.e. least ambiguous) causal status, while showing a greater amount of learning for the cue with the least certain (i.e. most ambiguous)

causal status. The current experiments utilised Rescorla's compound testing procedure (Rescorla, 2000; 2001) so that differences in learning about cues trained in compound could be measured. The aim of the first experiment reported here was to pit the theory protection account against a prediction error account. To achieve this, cues were chosen that had specific properties. Firstly, one cue needed to have a more certain causal status than the other, so that the theory protection account would predict greater learning for the less certain cue. Secondly, the cue with the more certain causal status needed to be judged as a less likely cause of the outcome prior to the compound conditioning phase, so that a prediction error account would predict greater learning for this cue. Thirdly, both cues should be familiar to participants because novel stimuli are often characterised as being more salient than familiar stimuli (e.g. Lubow & Moore, 1959).

Two cues with these properties are found in the redundancy effect (Uengoer, Lotz, & Pearce, 2013). A typical redundancy effect design involves presenting participants with a training stage incorporating blocking (A+ AX+) and a simple discrimination (BY+ CY-). Blocking (e.g. Kamin, 1969) occurs when learning about a cue (X) is apparently restricted by the simultaneous presence of another cue (A) that has also been trained separately. Cue Y, from the simple discrimination, is referred to as an uncorrelated cue, because it appears in both a causal compound and a non-causal compound. The redundancy effect is the robust finding that X is rated as a more likely cause of the outcome than Y during a subsequent test phase (Jones & Zaksaite, 2018; Jones, Zaksaite, & Mitchell, 2019; Uengoer, Dwyer, Koenig, & Pearce, 2019; for analogous results in non-human animals, see Jones & Pearce, 2015; Pearce, Dopson, Haselgrove, & Esber, 2012). Crucially, there is evidence from Jones et al. (2019) that participants' certainty about the causal status of X and Y differs at test. Participants were asked to make confidence ratings during the test stage of a redundancy effect experiment, in addition to outcome likelihood ratings. After participants had rated the likelihood of the outcome for specific cues, they were asked to rate how confident they were

about their likelihood ratings. The mean confidence ratings for the blocked cue (X) were significantly lower than those for the uncorrelated cue (Y), suggesting that participants were less certain about the causal status of X than Y. A further experiment by Jones et al. (2019) showed that the likelihood ratings of blocked cues, but not uncorrelated cues, were dependent on the overall proportion of training trials on which the outcome occurred. This suggests that participants' beliefs about X were more labile than for Y, further supporting the idea that participants are less certain about the causal status of blocked cues than uncorrelated cues. It is worth noting that this theoretical view, in which participants are uncertain about the causal status of blocked cues, is established within the learning literature. For example, uncertainty about the causal status of blocked cues is supported by several studies investigating the effects of manipulating assumptions about the outcome (e.g. Lovibond, Been, Mitchell, Bouton, & Frohardt 2003; Beckers, De Houwer, Pineno, & Miller, 2005; Vandorpe, De Houwer, & Beckers, 2007). In terms of the likelihood ratings provided by participants, cue X is typically assigned ratings in the middle of an 11-point likelihood scale, while Y is given lower likelihood ratings, suggesting that participants learn that Y is unlikely to be a cause of the outcome. On the basis of the aforementioned confidence data, the intermediate likelihood ratings given to X support the view that participants are uncertain about the status of X, despite encountering it in a causal compound (AX+). In other words, participants might give such intermediate likelihood ratings when they are uncertain because they lack the confidence to assign either a high or a low rating. Nevertheless, this difference in likelihood ratings suggests that, if X and Y were combined in a subsequent XY+ training phase, greater learning for Y than for X would be consistent with a prediction error account.

The experiments in this paper were designed to test a prediction error account (e.g. Rescorla, 2001) against the theory protection account, using Rescorla's compound testing procedure. All three experiments were based on the redundancy effect, with training in the first phase of each

experiment incorporating blocked and uncorrelated cues. Subsequently, a cue with a more certain causal status was trained in compound with a cue with a more uncertain causal status. The cue with the larger prediction error at the start of the compound training phase was always the cue with the more certain causal status. According to a prediction error account, participants should learn more (or equally) about the cue with larger prediction error irrespective of how certain its causal status is. Conversely, according to the theory protection account, learning should be greatest for the cue with less certain causal status, despite that cue having the smaller prediction error.

Experiment 1

As with previous redundancy effect studies using human participants (e.g. Uengoer et al., 2013), this experiment used a food allergy scenario. Participants were required to learn whether an allergic reaction would occur, on the basis of different single foods or pairs of foods being eaten. They were presented with a fictional scenario, in which they played the role of a medical doctor, trying to ascertain which foods cause a stomach ache in a test patient. During the training trials, participants were presented with a series of food cues and were asked to predict whether or not the fictional patient would experience a stomach ache after eating these foods. After participants made their prediction, they were then provided with feedback as to whether or not a stomach ache occurred. Following training, participants were tested by being asked to make a likelihood rating indicating how likely they thought a stomach ache would be after the patient ate specific foods.

The design of the experiment is shown in Table 1. This design contained two blocked cues (W and X) and two uncorrelated cues (Y and Z). Following this training phase, one blocked cue and one uncorrelated cue were trained together and paired with the outcome (XY+). If learning in this phase is the result of individual prediction error, in a similar manner to Rescorla (2001), there should be more learning about Y than X as it will have the greater error at the start of the XY+ stage.

Alternatively, if learning is determined by overall prediction error, there should be equal learning about X and Y. Finally, if learning is determined by the theory protection account, then participants should be resistant to changing their beliefs about Y since they have already learned that it is not a cause of the outcome with relative certainty (Jones et al., 2019). However, they should readily attribute the outcome to the causally uncertain X during XY+ trials. Consequently, X should be learned about more than Y.

Table 1. The design of the experiments

Experiment	Stage 1	Stage 2	Test
1	A+ AX+ BY+ CY- D+ DW+ EZ+ FZ-	XY+	XZ WY A B C D E F X Y W Z
2	AX+ BY+ CY- DW+ EZ+ FZ-	XY+	XZ WY A B C D E F X Y W Z
3	A+ AX+ BY+ CY- D+ DW+ EZ+ FZ-	XC+	XF WC A B C D E F X Y W Z

Learning about X and Y was compared using a final discrimination of a similar kind to that used by Rescorla (2001). As well as being asked for likelihood ratings for each individual cue, participants were asked to rate the likelihood of the outcome for two compounds, XZ and WY. Each of these compounds contained one cue that had been blocked in Stage 1, and one that had been an uncorrelated cue in Stage 1. In the absence of Stage 2 training, these two compounds should be assigned the same likelihood ratings at test. Consequently, any difference between these compounds must necessarily be the result of the XY+ training in Stage 2, and would indicate a different amount of learning about X and Y during that stage. Therefore, if learning is governed by an individual prediction error term, participants should rate the likelihood of the outcome as being higher for WY than XZ. If learning is governed by an overall error term, then there should either be no difference between ratings of XZ and WY, or a higher rating for WY if a modified response function (i.e.

Holmes et al.) is assumed. However, if learning is determined by the theory protection account, then XZ should be assigned higher ratings than WY at test.

Method

Participants

Thirty-six psychology students from the University of Plymouth participated in this experiment, in return for course credit (30 female, 6 male; mean age = 20.2, SD = 3.1). This sample size has adequate power to detect medium-sized within-subjects effects (83% power at d = 0.5). 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 all tested in the same lab at Plymouth University. The experiment was conducted using Viglen Genie desktop computers, running the Windows 10 operating system. The computers all used 22-inch Phillips LED displays, with participants at a typical distance (of approximately 40 – 80 cm) from the screen. The experiment was designed and executed in the Psychopy desktop application version 1.83.04 (Peirce, 2007; www.psychopy.org), with the output generated as individual CSV files for each participant. 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 ten individual cue types were represented on screen as photographs of fruits: apple, banana, cherry, kiwi, mango, orange, peach, pear, plum and strawberry. All the fruits were presented within a white square. The dimensions of each cue (including the white square) were 300 x 300 pixels, with a screen resolution of 1920 x 1080 pixels. For each participant, the foods were randomly assigned to each cue (A, B, C, D, E, F, X, Y, W and Z). 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 291 x 332 pixels. All experimental text, including instructions, was white. A black background was used throughout the experiment. Study information sheets, consent forms and debrief forms were all printed on paper.

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 eight trial types (A+, AX+, BY+, CY-, D+, DW+, EZ+, FZ-) 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 XY+. During the Test stage, participants were presented with two blocks of test cues. The twelve trial types (A, B, C, D, E, F, X, Y, W, Z, XZ and WY) appeared in a random order within each block. Each trial type was only presented once 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 Uengoer et al. (2013) and were as follows:

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 certain foods leads to stomach ache or not.

Imagine that you are a medical doctor. One of your patients often suffers from a stomach ache after eating. To identify which foods they react to, the patient eats specific foods 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 always be told what your patient has eaten. Sometimes, he has only consumed a single kind of food and on other times he has consumed two different foods. Please look at the foods carefully.

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 actually suffered from stomach ache. Use this feedback to find out what foods cause a stomach ache in your patient. At first you will have to guess the outcome because you do not know anything about your patient. But eventually you will learn which foods lead to stomach ache in this patient and you will be able to make correct predictions.

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

For each trial during the training stages, the cues were presented visually 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 eats the following:', with the stimuli presented below this. Underneath the stimuli, 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. This stage consisted of a previously unseen compound XY 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 "Stomach Ache" as the outcome.

After the completion of Stage 2, a further instruction screen was shown before commencement of the Test stage:

Next, your task is to judge the probability with which specific foods cause stomach ache in your patient. Single foods and pairs of foods 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 the space bar to continue the experiment.

For each trial during the Test stage, the cues were 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 eats the following:', with the stimuli presented below this. Underneath the stimuli, 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 eleven-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 black screen appeared for 0.4 secs, after which the next 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; www.r-project.org). The difference between the XZ and WY 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, there was no requirement for Bonferroni corrections. Bayesian t-tests were also conducted, using the procedure recommended by Dienes (2011) and implemented as R code by Baguley and Kaye (2010). As there was no suitable previous study on which to specify a plausible predicted effect size, a uniform distribution was specified, with a lower limit of -5 and an upper limit of 5 (in terms of the mean difference between ratings). The motivation for this was that previous human experiments utilising this compound testing procedure (e.g. Mitchell, Harris, Westbrook, & Griffiths, 2008) produced mean differences in compounds lower than 5 (where 5 would be considered a reasonable limit to any observed difference when an 11-point rating scale is used). For Bayesian t-tests conducted on single cues, the lower limit was set to -10 and the upper limit was set to 10, since these are the largest mean

differences in either direction permitted by an 11-point scale. In keeping with accepted conventions (e.g. Jeffreys, 1961), a Bayes factor of over three was set as the level providing evidence for a difference, while a Bayes factor of less than one third was set as the level providing evidence for no difference. Values between these levels were accepted as being inconclusive.

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/4xbkp/. The descriptive statistics for the Experiment 1 training stages are shown in Figure 1 and Figure 2. The Figure 1 data indicate that participants learned sufficiently about the eight different trial types by the time first training stage was complete. Similarly, Figure 2 indicates that participants learned that the XY+ compound was causal by the end of the second training stage, after giving it an intermediate rating on the first trial.

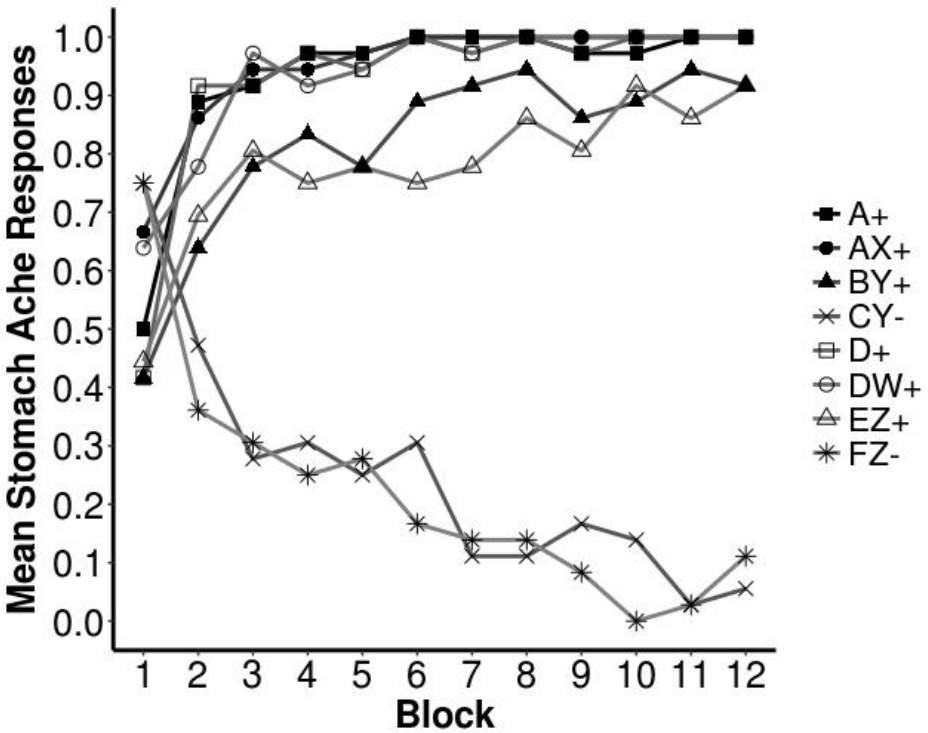

Figure 1. Experiment 1 Stage 1 data.

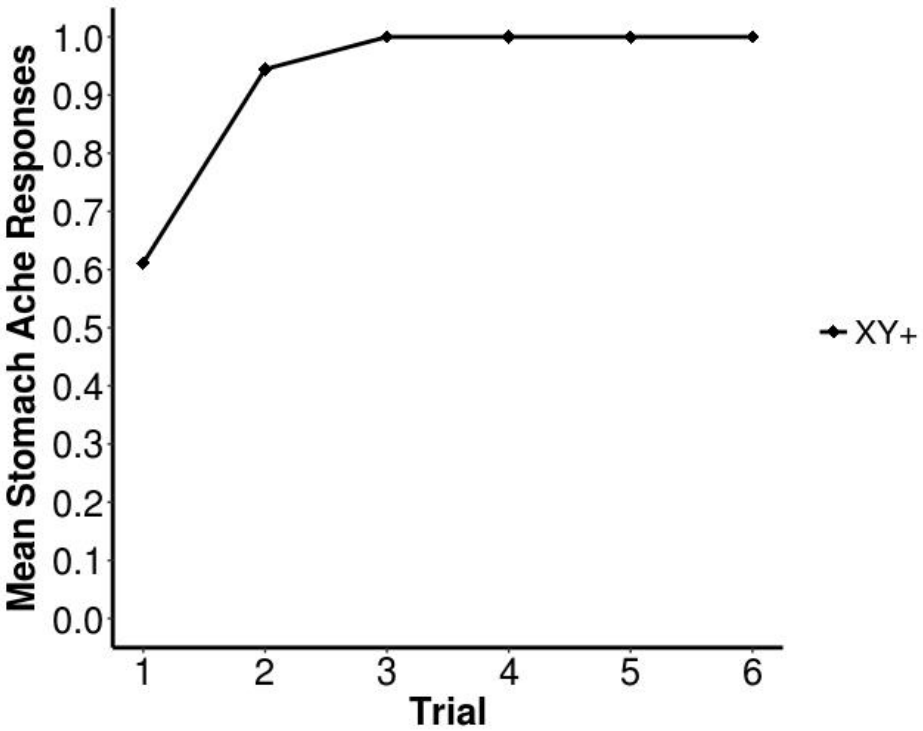

Figure 2. Experiment 1 Stage 2 data.

Descriptive statistics for the Experiment 1 test stage are shown in Figure 3. Ratings for XZ were significantly higher than for WY; $t(35) = 2.99$, $p = .005$, $BF = 15.62$, $d = .50$. Further testing revealed higher ratings for the single cue X compared to W; $t(35) = 3.96$, $p < .001$, $BF = 215.66$, $d = .66$. Conversely, there was no difference between the ratings assigned to Y and Z; $t(35) = 0.53$, $p = .597$, $BF = .08$, $d = .09$, and therefore no evidence that participants learned about Y during the second stage of the experiment. Taken as a whole, these findings show that the differences between the compounds were specifically driven by learning about X during XY+ trials. These data are consistent with the theory protection account, as opposed to a prediction error account (Rescorla, 2001). Despite the theoretical implications of this result, the crucial next step was to ensure its generality. Experiment 2 was intended as an extension of Experiment 1, to increase the generality of the findings. The design of Experiment 2 was modified, to vary the type of causally-uncertain cue incorporated into the design.

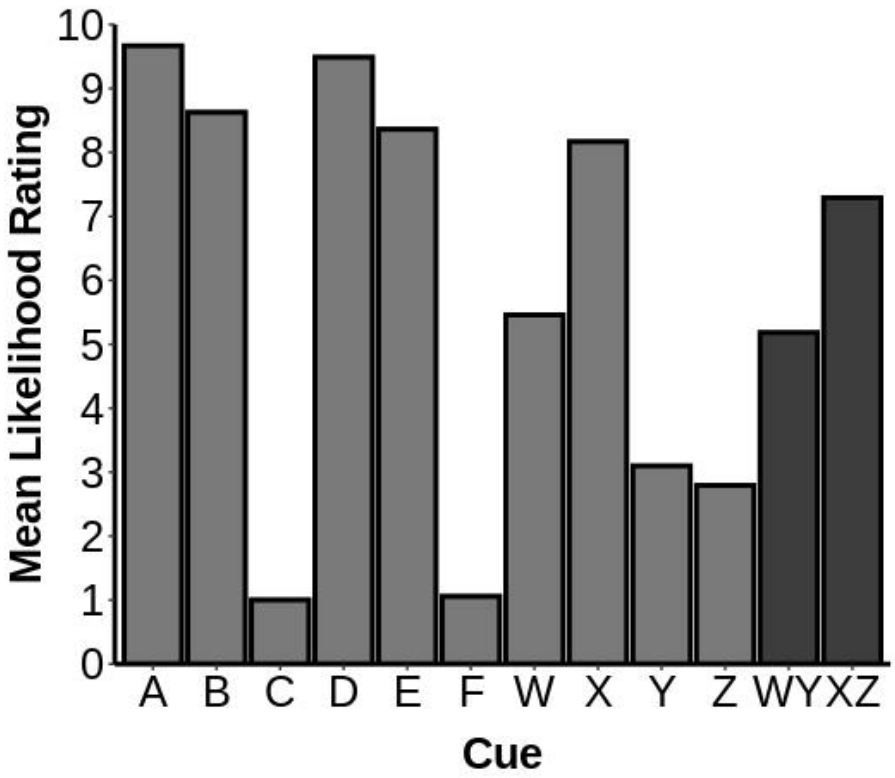

Figure 3. Experiment 1 Test stage ratings for all single stimuli and the two compound cues.

Experiment 2

There were two key aims for Experiment 2. The first was to extend the generality of the findings of Experiment 1 by using a different kind of causally-uncertain cue. To achieve this, the experimental design was modified so that it incorporated an 'overshadowed' cue (cf. Waldmann, 2001) from a two-item compound, rather than a blocked cue. The logic was that, like blocked cues, overshadowed cues are causally ambiguous from the perspective of the participant. This is supported by Jones et al. (2019), who reported that participants were less certain about their causal judgements of overshadowed cues than of uncorrelated cues. Presumably this is because, when two cues are presented in compound and the outcome is present (e.g. AX+, without A+ training), participants do not know which of the two cues is the cause of the outcome. However, unlike blocked cues, compound trials containing a pair of overshadowed cues do at least allow participants to infer that at least one, or both, of the cues must be a cause of the outcome. Accordingly, Jones et

al. observed higher causal ratings for overshadowed cues than for blocked cues. Learning during Stage 2 should again be greater for X (the overshadowed cue) rather than Y, because of the difference in certainty between these two cues. However, the higher causal ratings given to overshadowed cues, compared to blocked cues, meant that there was a theoretical basis for expecting less learning during the XY+ stage, since the prediction error for X would be slightly smaller.

The design of the experiment is shown in Table 1. All details were identical to Experiment 1, except that the A+ and D+ trials were omitted from Stage 1. Following Experiment 1, we expected causal ratings during the test stage to be higher for XZ than for WY. However, on the basis of the higher causal ratings reported by Jones et al. (2019) for overshadowed cues than for blocked cues, we expected the size of this effect to be somewhat smaller than in Experiment 1.

Method

Participants

Forty participants were recruited from the University of Plymouth, in return for a small monetary payment (32 female, 8 male; mean age = 26.7, SD = 10.4). This sample size has adequate power to detect medium-sized within-subjects effects (87% power at d = 0.5). 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, except that Psychopy version 1.85.1 was used (Peirce, 2007; www.psychopy.org).

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 (AX+, BY+, CY-, DW+, EZ+, FZ-) 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 exactly the same as in Experiment 1.

Procedure and analysis

The procedure for Experiment 2 was the same as for Experiment 1. The analyses were the same, except that the Bayesian priors were updated on the basis of the results of Experiment 1. As outlined above, there was a theoretical basis for expecting the Experiment 2 mean differences to be smaller than the values observed in Experiment 1 (because of the higher ratings given to overshadowed cues than blocked cues). Therefore, following the recommendations of Dienes (2011), a half-normal prior distribution was specified for each Bayesian t-test, with a mean of zero and a standard deviation set to the mean differences observed in Experiment 1.

Results

The trial-level raw data and analysis script for this experiment will be available, upon publication of this manuscript, at https://osf.io/8ceub/. The descriptive statistics for the Experiment 2 training stages are shown in Figure 4 and Figure 5. The Figure 4 data indicate that participants learned sufficiently about the six different trial types by the time first training stage was complete. Similarly, the Figure 5 data indicate that participants learned that the XY+ compound was causal by the end of the second training stage.

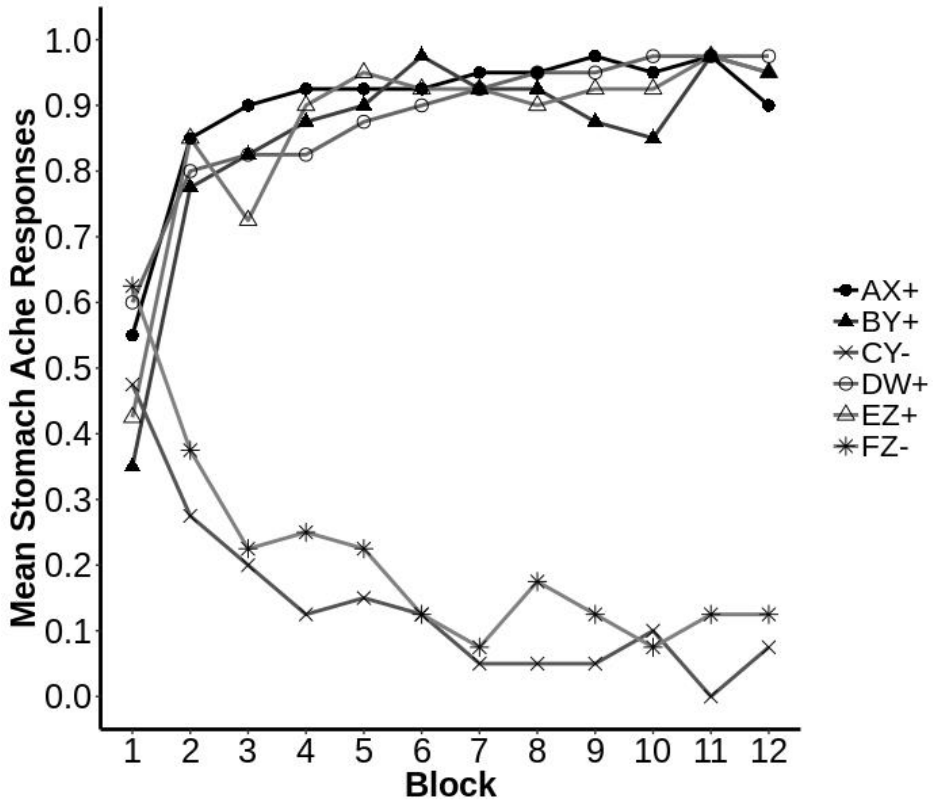

Figure 4. Experiment 2 Stage 1 training data.

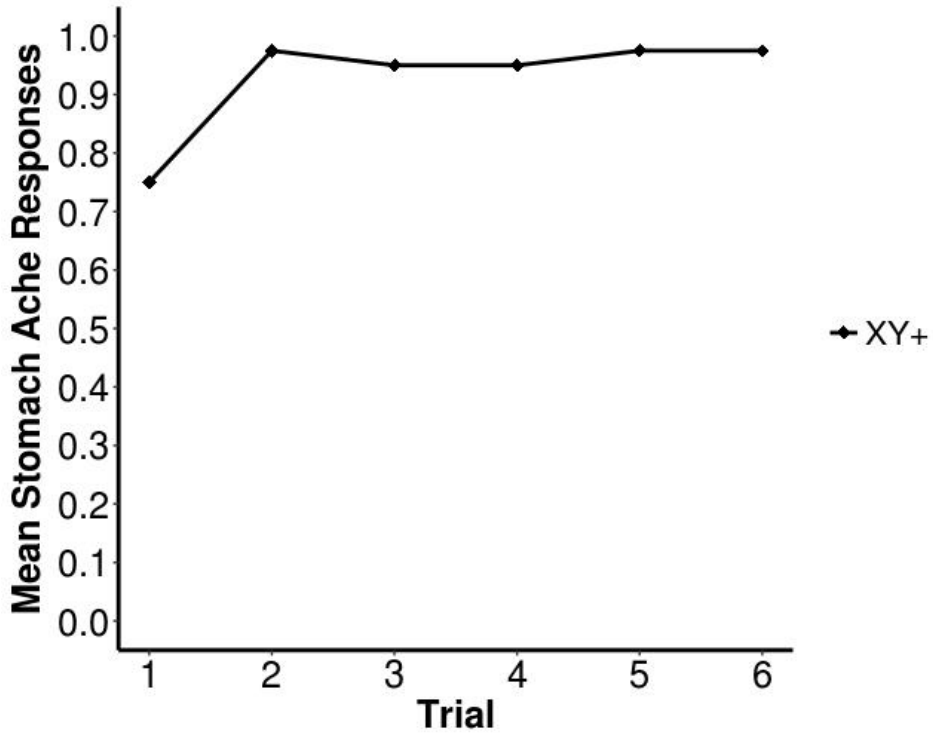

Figure 5. Experiment 2 Stage 2 training data.

The descriptive statistics for the Test stage are displayed in Figure 6. Ratings for XZ were significantly higher than for WY; $t(39) = 2.37$, $p = .023$, $BF = 6.42$, $d = .37$. This finding supports the prediction that the compound containing X would receive higher ratings. The difference between the compounds was again reflected in higher ratings for X compared to W; $t(39) = 2.37$, $p = .023$, $BF = 4.6$, $d = .38$. There was no evidence for a significant difference between Y and Z, despite the ratings for Y appearing slightly lower; $t(39) = 1.57$, $p = .124$, $BF = 1.13$, $d = .25$. As with Experiment 1, this finding supports the theory protection account, rather than a prediction error account (Rescorla, 2001). Although the ratings for W appeared to be slightly higher than for A or D, there was no evidence for a significant difference between either A and W, $t(39) = 1.17$, $p = 0.249$, $BF = .58$, $d = .19$; or D and W, $t(39) = 1.72$, $p = 0.093$, $BF = 1.09$, $d = .27$.

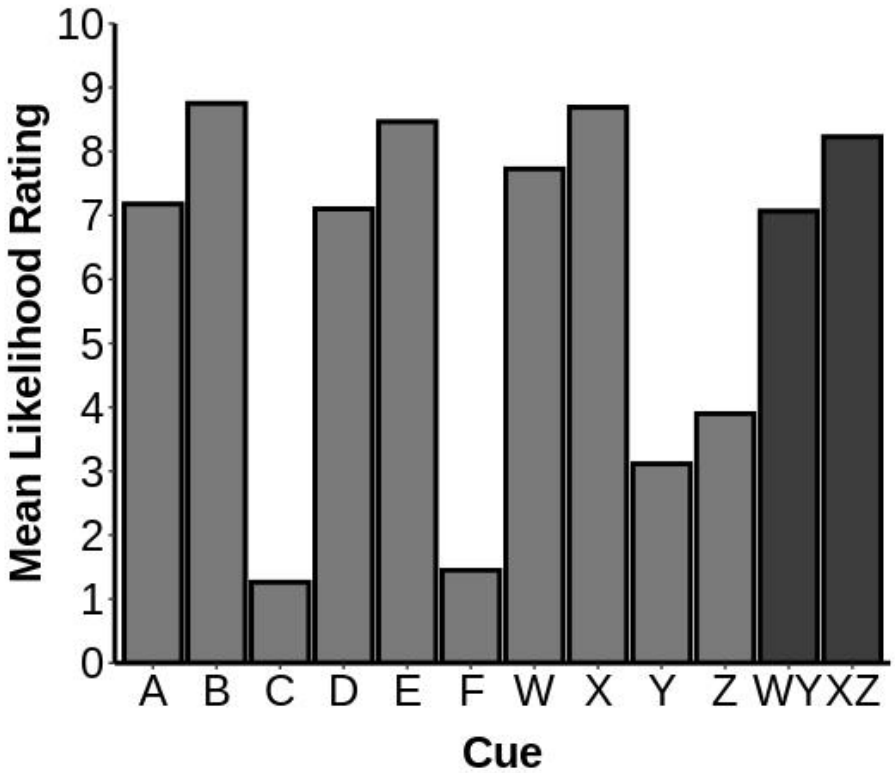

Figure 6. Experiment 2 Test stage ratings for all single stimuli and the two compound cues.

The results of Experiment 2 are consistent with those from Experiment 1; in both cases, participants learned more about the causally-uncertain cue than the causally-certain cue when the two were

trained in compound, despite the latter presumably having the larger prediction error. However, one limitation of these experiments is that the number of trials featuring the two critical cues during Stage 1 was not matched. Cue Y was presented to participants twice as often as X, because it was included in both BY+ and CY- trials. This might be important because, according to some theories of attention (e.g. Mackintosh, 1975; Pearce & Hall, 1980), learning rate is influenced by the amount of prior exposure to each cue. Another limitation is that the causal status of Y could be considered somewhat ambiguous, since its occurrence during Stage 1 was followed equally often by stomach ache and no stomach ache. Accordingly, Rescorla and Wagner's (1972) model predicts that Y will maintain some associative strength during the first stage of Experiment 1. It does so because C is predicted to become an inhibitor for the outcome, protecting Y from extinction on CY- trials. In fact, Rescorla and Wagner's model predicts that Y should have more associative strength than X in many circumstances. Although the opposite result has been observed numerous times (e.g. Uengoer et al., 2013), this is a reason to treat assumptions about the causal status of Y with caution. In Experiment 3 we overcame these problems by comparing learning about a blocked cue with learning about a different causally-certain cue; one that was presented the same number of times as the blocked cue, and that was never presented with the outcome during Stage 1.

Experiment 3

The design of Experiment 3 is shown in Table 1. Having extended the generality of the findings of Experiment 1 by comparing learning about Y to a different causally-uncertain cue in Experiment 2, our next objective was to check that the effect would persist if a different causally-certain cue was used in Stage 2. Training during Stage 1 was the same as for Experiment 1, but Stage 2 differed in Experiment 3, in that participants received XC+ training. The inclusion of C in this compound was motivated by the high confidence ratings observed by Jones et al. (2019) for this cue compared to X, indicating a greater degree of causal certainty about C. It was also motivated by an assumption

that the prediction error for C on the CX+ trials would be large. Evidence for the latter assumption is provided by the consistently low causal ratings assigned to C in previous experiments (Experiments 1 and 2, and all experiments reported by Jones et al., 2019; and Uengoer et al., 2013), and the fact that C was always presented without the outcome during Stage 1. The Test stage contained two compounds, XF and WC, that permitted a comparison of learning about X and C according to the same logic as the previous experiments. If learning is determined by prediction error, participants should learn more about C than X during Stage 2, and causal ratings should be higher for the WC compound than for the XF compound during the Test stage. Alternatively, if participants in Experiments 1 and 2 learned most about X during Stage 2 because of its uncertain causal status, that effect should persist in Experiment 3, leading to higher causal ratings for XF than for WC at test.

Method

Participants

Forty Psychology students from the University of Plymouth participated in this experiment, in return for course credit (37 female, 3 male; mean age = 22.6, SD = 7.0). This sample size has adequate power to detect medium-sized within-subjects effects (87% power at d = 0.5). 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 3 were the same as those used for Experiments 1 and 2, except that Psychopy version 1.85.2 was used (Peirce, 2007; www.psychopy.org).

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 eight trial types (A+, AX+, BY+, CY-, D+, DW+, EZ+, FZ-) 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 XC+. During the Test stage, participants were presented with two blocks of test cues. The twelve trial types (A, B, C, D, E, F, X, Y, W, Z, XF and WC) appeared in a random order within each block. Each trial type was only presented once within each block.

Procedure and analysis

The procedure for Experiment 3 was the same as for Experiments 1 and 2. The analyses were the same, except that the Bayesian priors were updated on the basis of the results of Experiment 1. There was a theoretical basis for expecting the Experiment 3 mean differences to be about the same as the values observed in Experiment 1, in that the abstract design of the first training stage was the same. Therefore, following Dienes (2011), a normal distribution was specified as the prior for each Bayesian t-test, with each mean set to the corresponding Experiment 1 mean and each standard deviation set to half this value.

Results

The trial-level raw data and analysis script for this experiment will be available, upon publication of this manuscript, at https://osf.io/jqrb6/. The descriptive statistics for the Experiment 3 training stages are shown in Figure 7 and Figure 8. Data from Stage 1 indicate that participants learned sufficiently about the eight different trial types by the time training was complete. Similarly, Figure 8 shows that participants learned that the XC+ compound was causal by the end of Stage 2, after giving it an intermediate rating on the first trial.

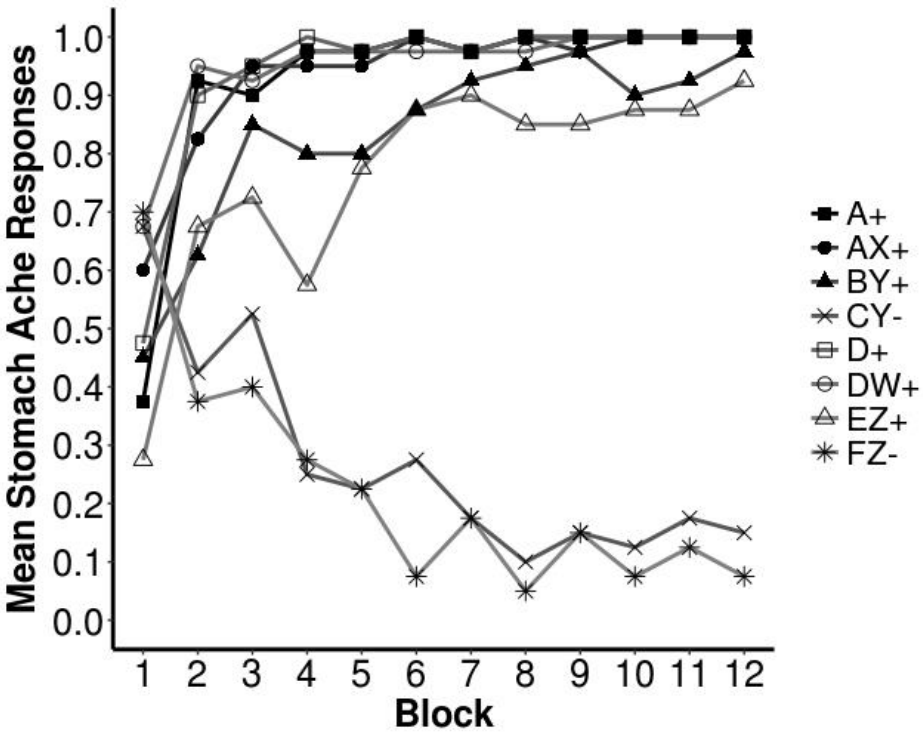

Figure 7. Experiment 3 Stage 1 data.

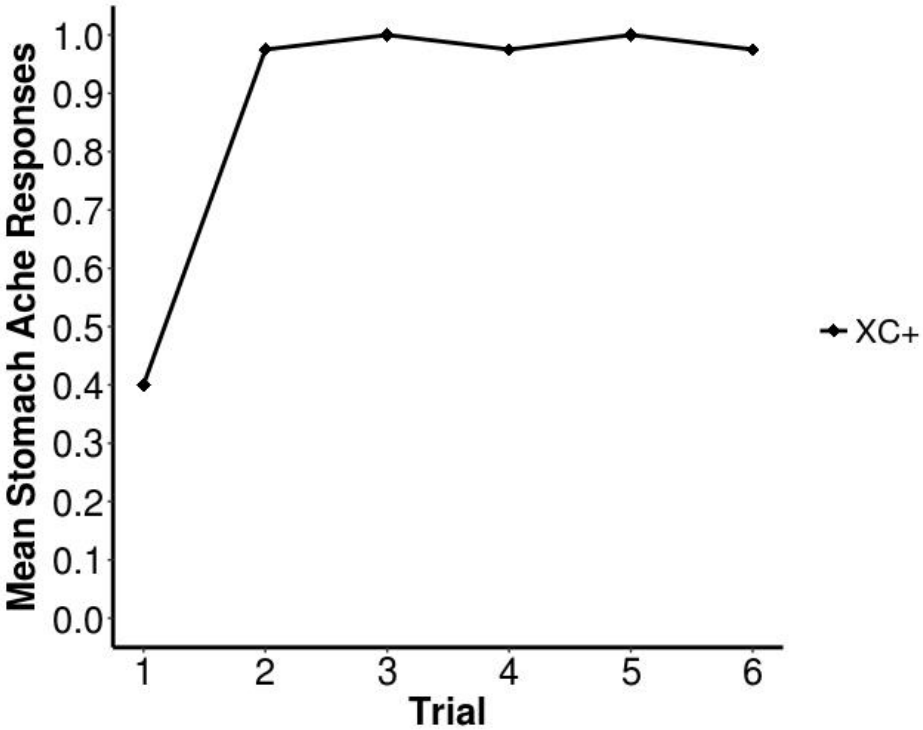

Figure 8. Experiment 3 Stage 2 data.

Descriptive statistics for the Test stage are displayed in Figure 9. Causal ratings were significantly higher for XF than for WC; $t(39) = 2.39$, $p = .022$, $BF = 6.66$, $d = .38$. This finding supports the prediction that the compound containing X would be rated higher at test. As expected, the difference between the compounds was driven by higher ratings for X compared to W; $t(39) = 3.42$, $p = .001$, $BF = 128.57$, $d = .54$. Although the ratings for C looked a little higher than for F, there was no evidence for a significant difference; $t(39) = 1.93$, $p = .061$, $BF = .70$, $d = .31$. These results are consistent with theory protection, with participants learning more about the causally-uncertain blocked cue (X) than the causally-certain discriminative cue (C). As with the previous experiments, this finding is incompatible with a prediction error account (Rescorla, 2001).

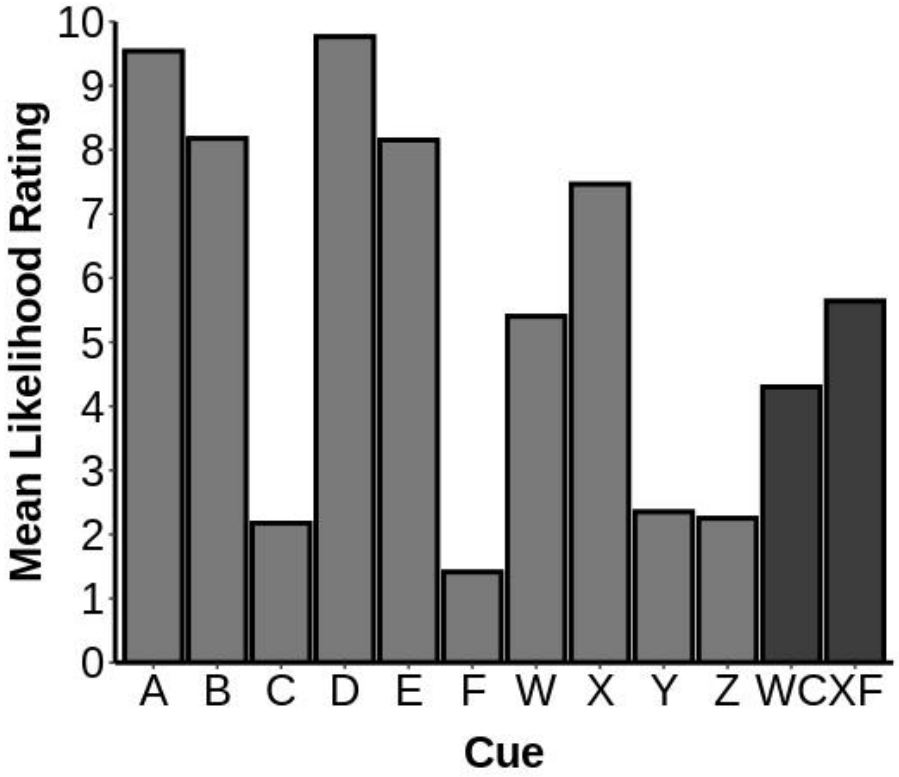

Figure 9. Experiment 3 Test stage ratings for all single stimuli and the two compound cues.

General Discussion

The three experiments in this paper provide evidence that causal learning in humans is, at least in part, driven by differing certainty about the causal status of cues. In each experiment, participants apparently resisted updating their beliefs about cues with a known causal status, instead attributing unexpected outcomes to cues with a comparatively ambiguous causal status. This occurred despite the ambiguous cues having the smaller prediction error at the start of the second training stage in all three experiments. The results of these experiments are opposite to those that might be expected on the basis of prediction error accounts of learning (Bush & Mosteller, 1951; Rescorla & Wagner, 1972; Rescorla, 2001). A theoretical implication of particular interest is the idea that humans act to maintain beliefs about known causal relationships. This idea is intuitive, in terms of learning being a process of acquiring information about the things we are unsure about and incorporating this information alongside existing knowledge.

The theory protection account provides an alternative explanation for previous experiments using compound testing procedures in humans. For example, Mitchell et al. (2008) initially gave participants A+ C+ training, and then subsequently trained cue A and a novel cue B in a causal compound (AB+). A compound test (AD versus BC) revealed that participants had learned more about the novel cue B during AB+ trials. This effect was subsequently replicated with a larger outcome (AB++) in the second training stage, again showing more learning about B than about A. Although these data are consistent with a prediction error account, they are also consistent with the theory protection account, showing that these accounts need not make opposing predictions in all cases. This is because B would have both the greater prediction error and the more uncertain causal status at the start of the second training stage, as nothing had yet been learned about it. In another experiment, Le Pelley, Oakeshott and McLaren (2001) initially trained two cues as excitors (A+ C+), and two other cues (B and D) as inhibitors (BE- DE- E+). One cue of each type was used in the second training phase (AB+); the results of a subsequent compound test revealed that

participants had learned more about cue A than cue B, despite the latter presumably having a greater prediction error. Although participants should have been certain about the causal status of both A and B following initial training, these findings are still compatible with the theory protection account because participants appeared to maintain (and strengthen) their causal belief that A was a cause of the outcome but B was not.

While the results of the present experiments appear to be inconsistent with a prediction error account, they might be accommodated if an additional process is invoked. One obvious candidate is the modification of attention to cues as a result of experience. For example, Mackintosh's (1975) model of learning is compatible with the results of Experiments 1 and 2. According to Mackintosh's model, more attention will be paid to cues that are better predictors of outcomes rather than poorer predictors, resulting in more learning about these cues. Rather than this attentional process being an alternative to prediction error, it is instead suggested to operate alongside individual prediction error, with attention determining associability. In Experiments 1 and 2, X might be considered more predictive of the outcome than Y during Stage 1. This is because X was consistently followed by the outcome, whereas Y was followed by the outcome on BY+ trials but not CY- trials. If participants paid more attention to X than to Y as a result, this could have led to better learning about X than Y during Stage 2. Experiment 3, however, is harder to reconcile with Mackintosh's model. Cue C was a better predictor of the absence of the outcome than Y on CY- trials, so the model predicts that participants should have learned to pay attention to C during Stage 1. Although X was consistently paired with the outcome on AX+ trials, it was a poorer predictor of that outcome than A (because of the separate A+ trials) and should therefore have suffered at least some decline in attention during Stage 1. Consequently, greater learning for X than for C during Stage 2 cannot have been the result of changes in attention that occurred during Stage 1. Mackintosh's model can only be reconciled with the results of Experiment 3 if we assume that the relevant changes in attention occurred during

Stage 2. Before the first XC+ trial, C had only been paired with the absence of the outcome. When it was subsequently presented in the XC+ compound, it was therefore a poor predictor of the outcome. This could have resulted in a rapid decline in the associability of C relative to X, leading to more learning about X than C from the second trial onwards, in spite of C having the larger prediction error. However, this account is not readily supported by the XC+ training data (see Figure 8). If these data are taken at face value, almost all the learning about the XC+ compound appears to have taken place during the first Stage 2 trial, before any update in associability could have influenced learning. This interpretation should be treated with caution, however, since high causal ratings for XC+ from the second trial onwards do not necessarily imply that learning was complete. To investigate this point further, an obvious next step would be to conduct a similar experiment to Experiment 3, but with only one XC+ training trial in Stage 2. We have conducted two pilot experiments using this procedure but the results were inconclusive. The data and analyses for these two unpublished experiments will be available online, upon publication of this manuscript, at https://osf.io/8f3a7/ and https://osf.io/9dvcf/.

An additional way in which Mackintosh's (1975) theory could be reconciled with these results is if one assumes that cues given high and low causal ratings differ in their associability changes. For example, it may be that cues given low causal ratings acquire associability at a slower rate than those given high causal ratings. If this were the case, then it would be possible for X to have higher associability than C at the start of Stage 2. The most obvious way to test this would be to train participants with one cue that is a cause of an outcome, which also has a certain causal status, alongside another cue that does not cause that outcome, but which is less predictive and has a less certain causal status. If both cues were then trained in a compound that does not cause the outcome in Stage 2, Mackintosh's model would predict a more substantial decrease in associative strength for the previously predictive causal cue. The theory protection account, on the other hand, would

predict more learning about the causally-uncertain cue. A series of experiments testing this idea would be an obvious follow up to the present experiments. Additionally, both of these potential explanations based on Mackintosh's model (the rapid decline in the associability of C, and the difference in associability changes for cues given high and low causal ratings) could also be tested by examining overt attention to cues during training. It should be possible to detect whether there is a difference in the attention paid to these cues at the start of Stage 2 by tracking eye movements. Furthermore, if there is a rapid shift in attention from one cue to another, it should be possible to detect this from eye-tracking. Similarly, it should be possible to look for differences in certainty consistent with Stage 2 learning if learning is governed by a theory protection process. Although the corresponding differences in certainty have been observed elsewhere (Jones et al., 2019), we did not measure certainty in the present experiments. Of course, investigating certainty and attention using the suggested methods may reveal that both have a role to play, rather than these processes being mutually exclusive. For example, it may be that variations in certainty about the causal status of cues result in differences in the attention paid to such cues.

An alternative view of how attention changes as a result of experience was provided by Pearce and Hall (1980). They proposed that animals pay attention to cues that are followed by surprising outcomes, and that cues that are followed by predicted outcomes suffer a decline in attention. Conceptually, this model has something in common with the theory protection account. Both accounts suggest that there is less learning about cues which are reliable predictors of outcomes. However, the mechanisms are quite different. In Pearce and Hall's model, learning is outcome-directed, while in the theory protection account it is cue-directed. More specifically, according to Pearce and Hall (1980), decreases in the associability of cues are the consequences of certainty about outcomes, with greater certainty leading to reduced associability for any cues present. This results in less learning on subsequent trials. However, the theory protection account suggests that

learning is lowest for cues that have a known causal status. This means that future learning is driven by knowledge about individual cues themselves, rather than being driven by knowledge about the outcomes they are paired with. This difference in focus means that the two theories make differing predictions for the present experiments. For example, by the end of Stage 1 of Experiment 3, Pearce and Hall's model predicts a decline in attention for all cues because the outcome is predictable on every trial. The training data from this experiment suggest that, if anything, the presence or absence of the outcome was better predicted on trials containing X than trials containing C. As a result, the associability of each cue should have been low at the outset of Stage 2, with a possible advantage for C. This is inconsistent with the observation of greater learning for X than for C in Stage 2. The theory protection account predicts more learning about X than about C because the causal status of X was uncertain at the end of Stage 1, even though the outcome was predictable on AX+ trials.

At present, the evidence for the theory protection account is enough to warrant further investigation. However, while the results of the current experiments are not consistent with a prediction error account (Rescorla, 2001), nor with the attentional model proposed by Pearce and Hall (1980), they could be accounted for by Mackintosh's (1975) attentional model. A useful next step would be to test the theory protection account against Mackintosh's model. A further point for future investigation is whether the proposed theory protection account is a general principle, or whether there are situations in which people readily update their beliefs about previously learned cue-outcome associations in a manner akin to prediction error, even in the presence of causally-ambiguous cues. There is also the issue of whether a comparable process to theory protection occurs in non-human animals. It is tempting to conclude that there is no such process in rats and pigeons, on the basis of Rescorla's results. However, the designs of the current experiments were not the same as Rescorla's. It is possible that comparable experiments to those reported here using non-human animals would produce similar results. Nonetheless, the three experiments in this paper

suggest that causal certainty could play an important role in determining the extent to which humans learn about cues. It appears that causally-uncertain cues are learned about more readily than causally-certain cues, in a manner that is inconsistent with a prediction error account.

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.

Beckers, T., De Houwer, J., Pineno, O. & Miller, R. R. (2005). Outcome additivity and outcome maximality influence cue competition in human causal learning. *Journal of Experimental Psychology: Learning, Memory, and Cognition, 31*(2), 238 – 249.

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

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

Fiske, S. T., & Taylor, S. E. (1984). *Social cognition (2nd Ed.)*. Reading, MA: Addison-Wesley.

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

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., & Pearce, J. M. (2015). The fate of redundant cues: Further analysis of the redundancy effect. *Learning & Behavior 43*, 72-82.

Jones, P. M., & Zaksaite, T. (2018). The redundancy effect in human causal learning: no evidence for changes in selective attention. *Quarterly Journal of Experimental Psychology, 71*, 1748-1760.

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.

Le Pelley, M. E., Oakeshott, S. M., & McLaren, I. P. L. (2001). The mechanics of associative change. *Proceedings of the 27th Annual Conference of the Cognitive Science Society*.

Lovibond, P. E., 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 and Cognition, 31*, 133-142.

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.

Pavlov, I. P. (1927). *Conditioned reflexes* (G. V. Anrep, Trans.). Oxford: Oxford University Press.

Pearce, J. M., Dopson, J. C., Haselgrove, M., & Esber, G. R. (2012). The fate of redundant cues during blocking and a simple discrimination. *Journal of Experimental Psychology: Animal Behaviour Processes, 38*, 167-179.

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.

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

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.

Uengoer, M., Dwyer, D. M., Koenig, S., & Pearce, J. M. (2019). A test for a difference in the associability of blocked and uninformative cues in human predictive learning. *Quarterly Journal of Experimental Psychology*, *72*, 222–237.

Uengoer, M., Lotz, A., Pearce, J. M., (2013). The fate of redundant cues in human predictive learning. *Journal of Experimental Psychology: Animal Behaviour Processes, 39*, 323-333.

Vandorpe, S., De Houwer, J., & Beckers, T. (2007). Outcome maximality and additivity training also influence cue competition in causal learning when learning involves many cues and events. *Quarterly Journal of Experimental Psychology, 60*, 356-368.

Waldmann, M. R. (2001). Predictive versus diagnostic causal learning: Evidence from an overshadowing paradigm. *Psychonomic Bulletin & Review, 8*, 600-608.

Wason, P. C. (1960). On the failure to eliminate hypotheses in a conceptual task. *Quarterly Journal of Experimental Psychology, 12*, 129-140.