Outcome predictability biases learning

Oren Griffiths[1], Chris J. Mitchell[2], Anna Bethmont[1] and Peter F. Lovibond[1]

[1] University of New South Wales, Sydney, Australia

[2] Plymouth University, Plymouth, United Kingdom

Keywords: Associative learning, reasoning, attention, causal models, induction

Please address correspondence to:

Oren Griffiths

University of New South Wales

Kensington, Sydney,

NSW, 2052,

Australia

Email: oren.griffiths@unsw.edu.au

Phone: +61-2-93851380

Abstract

Much of contemporary associative learning research is focused on understanding how and when the associative history of cues affects later learning about those cues. Very little work has investigated the effects of the associative history of *outcomes* on human learning. Three experiments extended the 'learned irrelevance' paradigm from the animal conditioning literature to examine the influence of an outcome's prior predictability on subsequent learning of relationships between cues and that outcome. All three experiments found evidence for the idea that learning is biased by the prior predictability of the outcome. Previously predictable outcomes were readily associated with novel predictive cues, whereas previously unpredictable outcomes were more readily associated with novel non-predictive cues. This finding highlights the importance of considering the associative history of outcomes, as well as cues, when interpreting multi-stage designs. Associative and cognitive explanations of this certainty matching effect are discussed.

When a cue repeatedly fails to reliably predict an outcome, this failure could, in principle, be attributed to the cue or to the outcome. One could learn that the cue is non-predictive or that the outcome is unpredictable. Much experimental work has been conducted examining how people learn about non-predictive cues, and what the consequences are for the subsequent processing of and learning about that cue (e.g. Le Pelley & McLaren, 2003). However, there has been relatively little investigation of how people learn that an outcome is unpredictable, and whether that has implications for subsequent learning involving that outcome. The present experiments address this gap. Although there is little empirical evidence within the human learning literature concerning our ability to learn that an outcome is unpredictable, this idea has been investigated in the animal conditioning literature, primarily within two paradigms: unconditioned stimulus (US) pre-exposure effects and learned irrelevance. These paradigms will be discussed in turn.

US pre-exposure effects

Unpredictable presentations of a US retard later learning about that US. This is termed the 'US pre-exposure' effect, and has been replicated in a number of species and learning paradigms, such as conditioned emotional response tasks with rats (Kamin, 1961) and eye-blink avoidance learning in humans (Taylor, 1956). This finding has been interpreted in a number of ways. One possibility is that subjects learn in the first phase of training that their behaviour and the outcome are uncorrelated, and therefore that the outcome is uncontrollable. This 'learned helplessness' then reduces motivation and impairs subsequent learning of the reliable cue-outcome relationship (Maier & Seligman, 1976). In their extensive review of the 'US pre-exposure' literature, Randich & LoLordo (1979) favoured an alternative explanation. They found that the majority of US pre-exposure effects could be accounted for by a process of blocking. Under this account, animals do not learn that the unsignalled US is unpredictable

during the pre-exposure phase. Instead, they learn an association between the experimental context and the US. This context-US association then blocks the subsequent learning of the relationships between a discrete cue and the US. Two predictions made by the context-blocking account, but not the learned helplessness account, are that: (i) the animals will come to fear the context in which the pre-exposure occurred, and (ii) manipulations that reduce fear of the context, such as overshadowing by a discrete cue, will also reduce the magnitude of interference with subsequent conditioning. Baker and colleagues found evidence consistent with both predictions (Baker & Mackintosh, 1979; Baker, Mercier, Gabel & Baker, 1981), and thus the US preexposure effect is typically considered to be a product of blocking by the context.

Interestingly, however, one key finding was inconsistent with the context blocking account. Baker et al (1981) found that exposure to unsignalled shocks in one context, context A, interfered with subsequent conditioning in a second context, context B, even though the animals did not display fear to context B (a similar result was observed with the addition/removal of a session-long auditory cue). This finding demonstrates that while context blocking goes some way to explain the US-preexposure effect, there may be another mechanism that allows the animals to transfer their knowledge about unpredictable shocks in context A to their subsequent learning about shock in context B. Baker et al (1981) concluded that it was likely that animals also learned that the unsignalled US was unpredictable, in a manner akin to learned helplessness, and that this learning interfered with the subsequent formation of associations involving the US.

Learned irrelevance

A second approach to examining the influence of unpredictability on subsequent learning uses a similar, but importantly different, manipulation. In ‘learned irrelevance’ studies, rats

are exposed to both the cue (e.g. auditory tone) and the outcome (e.g. electric shock) in an uncorrelated fashion. This cue is then paired with the outcome reliably, to assess the degree to which the initial manipulation of unpredictability affected subsequent learning. The typical finding is that animals' learning of the reliable cue-outcome relationship is impaired, relative to either (i) animals that were not previously exposed to the cues or the outcomes prior to conditioning or (ii) to animals that had been exposed to either the cue alone or the outcome alone prior to conditioning (Kremer, 1971; Overmier & Wielkiewicz, 1983). This impairment is typically attributed to animals learning about a feature of the cue (i.e. that the cue is irrelevant; but see Bonardi & Ong, 2003 for a different interpretation). A further possibility is that animals learn about a feature of the outcome (i.e. that the outcome is unpredictable).

In a series of cleverly designed studies, Matzel, Schachtman & Miller (1988) demonstrated that 'learned irrelevance' effects, unlike 'US pre-exposure' effects, were not readily attributed to animals learning to predict the outcome using the conditioning context as a cue. In their critical experiment, Matzel et al (1988; Experiment 1c) exposed animals to uncorrelated presentations of the target cue (a tone) and the target outcome (foot shock). However, to reduce the degree to which the shock was associated with the conditioning context, a second (non-target) cue was used to signal the shock throughout pre-exposure (an earlier experiment, 1b, demonstrated the efficacy of this manipulation in ameliorating conditioning to the context). A similar manipulation was used to control for the effect of pre-exposing the cue (also termed 'latent inhibition,' e.g. Lubow, 1959). Matzel et al. argued, therefore, that the slow learning observed when the cue and outcome were paired in the final stage of training could not be readily attributed to the influence of learning context-outcome associations, or of mere exposure to the cues. Rather, it appeared that the animals had learnt that the outcome could not be predicted by the cue, and it was this that slowed later learning.

In brief, Matzel et al (1988) found a reduced, but significant impairment in subsequent conditioning for the animals given this uncorrelated cue-outcome pre-exposure, relative to controls. Interestingly, just as Baker et al (1981) found that US-preexposure effects persisted across a context change, Matzel et al (1988) showed that impairments in conditioning following uncorrelated cue-outcome exposure persisted when a context change occurred between pre-training and conditioning. These impairments in learning also persist across a change in the type of learning: Baker and Mackintosh (1976, 1977) showed that uncorrelated cue-outcome exposure influences subsequent learning about the cue-outcome relationship both when the relationship to be learned is excitatory (learning that the cue predicts the presence of the outcome) and also when it is inhibitory (learning that the cue predicts the absence of the outcome). Taken together, this body of research suggests that animals are learning something additional about the cue (or about the outcome) during uncorrelated cue-outcome exposure that affects subsequent learning.

Crucially, Matzel et al (1988) attributed this impairment in subsequent cue-outcome learning to the animals learning that the cue was irrelevant (or non-predictive), rather than to the animals learning that the outcome was unpredictable. This approach is consistent with theories that account for cue competition effects by positing variations in 'associability,' or the degree to which a cue is associated with an outcome (e.g. Mackintosh, 1975; Pearce & Hall, 1980). The Mackintosh (1975) approach is particularly relevant here, as its fundamental prediction is addressed in the learned irrelevance design. Specifically, a cue that fails to predict outcomes of significance will subsequently receive less attention and will enter into associations less readily than more predictive cues. This prediction has been well validated in the animal literature (e.g. in 'blocking of unblocking' effects) and in human learning preparations (Mackintosh & Turner, 1971; Kruschke & Blair, 2000; Le Pelley, Beesley & Suret, 2007; Griffiths & Le Pelley, 2009).

Indeed, the learned irrelevance procedure developed in animal conditioning studies has been translated to an analogous preparation used to study human learning: the 'learned predictiveness' procedure (Lochmann & Wills, 2003; Le Pelley & McLaren, 2003). The typical finding is that cues that have been shown to be good predictors in the past are more readily associated with novel outcomes than are cues shown to have little predictive value. Much is now known about this phenomenon. For instance, recent experiments have shown that it is reflected in eye-gaze (Le Pelley, Beesley & Griffiths, 2011), that it is influenced by instruction (Mitchell, Griffiths, Lovibond & Seetoo, 2012) and outcome value (Le Pelley, Mitchell & Johnson, 2013), that it is evident in causal judgments (Le Pelley & McLaren, 2003), social evaluations (Le Pelley et al, 2010) and sequential reaction time tasks (Beesley & Le Pelley, 2010), and that it is attenuated in people who are high in schizotypal personality traits or who are currently experiencing positive symptoms of schizophrenia (Le Pelley et al, 2010; Morris, Griffiths, Le Pelley & Weickert, 2013).

So it is clear that, in the human and animal contexts, being exposed to uncorrelated presentations of a cue and outcome results in subsequent impairment of learning involving the previously irrelevant cue. What is less clear, however, is whether outcomes that have been shown to be unpredictable in the past will be learnt about more slowly than previously predictable outcomes. Gunther, Miller and Matute (1997) noted the similarity in factors that affect 'CS pre-exposure' effects and 'US pre-exposure' effects. They highlighted the possibility that, just as cue associability has been shown to influence subsequent learning, so too might parallel 'outcome associability' influence learning.

In sum, there is substantial evidence that humans (and other animals) encode whether a cue is predictive of an outcome or not (learned predictiveness/learned irrelevance). However, these data also leave open the possibility that the prior reinforcement history of an outcome (or US) may also subsequently shape learning involving that outcome. A novel

human learning task was constructed to examine the impact of manipulating outcome predictability on subsequent cue-outcome learning. It was predicted that outcomes that were predictable in the past will more readily enter into associations with novel cues than will outcomes shown to be unpredictable.

Experiment 1

The present experimental design extrapolated from Le Pelley & McLaren's (2003) study of the influence of cue predictiveness on subsequent learning to instead focus on the influence of outcome predictability on learning. The manipulation of outcome predictability was made within subjects in an initial training phase, and then its impact on subsequent learning was measured in a second training phase (again, within-subjects).

The present experiments used an allergist task, a common task used to assess human associative learning (Larkin, Aitken & Dickinson, 1998), in which participants were shown a fictional patient, Mr X, who ate different foods on each day. On some days he would experience an allergic reaction and on some he would not, and in this manner participants were required to learn which foods (cues) predicted Mr X's allergic reactions (outcomes). Mr X experienced two types of allergies in the present experiment: stomach reactions (cramping, bloating) and skin reactions (itchiness, swelling). For each participant, the values on one outcome dimension were predictable and the values on the other outcome dimension were not (e.g. skin reactions were predictable, but stomach reactions were not).

After participants learned the cue-outcome relationships in this first training phase, they were transferred to a second training phase in which the cues (foods) were changed but the outcomes (allergies) remained the same. Because a new set of cues was used in the second training phase, any learning of the cue-outcome relationships from the first training phase would not aid performance in the second phase; instead, participants needed to learn

new cue-outcome relationships involving the existing outcomes and the new cues. Moreover, in the second training phase, the cue-outcome contingencies were arranged such that both outcome dimensions were predictable. Thus, from an objective viewpoint, the relationships between the novel foods and the previously predictable outcome dimension (say, skin reactions) were just as strong as those between the novel foods and the previously unpredictable outcome dimension (e.g., the stomach reactions). For example, the novel food 'cherry' might have been followed by itchiness (a previously predictable skin reaction) and bloating (a previously unpredictable stomach reaction). If participants were merely sensitive to the cue-outcome contingencies during the second phase of training, learning of the cherry-itchiness association should proceed at the same speed as learning of the cherry-bloating relationship. If, however, participants learn that some outcomes are more predictable than others (that skin reactions are predictable and stomach reactions are not), then they may more readily associate cherry with itchiness (a previously predictable skin reaction) than with bloating (a previously unpredictable stomach reaction). This was the primary hypothesis addressed in Experiment 1.

Method

Participants. Fifty one undergraduate students from the University of New South Wales participated for course credit.

Design. The design of Experiment 1 is summarized in Table 1. On each trial in the first training phase, participants were shown either one or two foods, and were asked to predict Mr X's allergies on both allergy dimensions. On meals in which Mr X ate one food, he experienced one allergy (either a stomach reaction or a skin reaction). On meals in which Mr X ate two foods, he experienced two allergic reactions (both a stomach reaction and a skin reaction). The primary purpose of the first training phase was to arrange the cue-

outcome contingencies so that participants could learn that one outcome dimension was predictable (labelled *p* in Table 1; e.g. skin reactions), but that the other outcome dimension was unpredictable (labelled *u* in Table 1; e.g. stomach reactions). Each outcome dimension had two positive values and an absent value. That is, if the predictable outcome dimension was skin reactions, then the two positive values on this dimension were itchiness (labelled *p1* in Table 1) and swelling (labelled *p2* in Table 1). The absence of a skin reaction is labelled *pØ* in Table 1 ('no skin reaction'). Equivalently for the unpredictable outcome dimension, the two positive values on the unpredictable outcome dimension (e.g. bloating and cramping if stomach reactions are unpredictable) are labelled *u1* and *u2*, respectively, while the absence of a stomach reaction is labelled *uØ*.

[Table 1 about here.]

The first phase of training was arranged so that one cue (A) predicted a particular skin reaction (*p1*). On every trial in which cue A appeared, the outcome *p1* occurred (A-*p1* trials, *AXp1,u1* trials and *AX-p1,u2* trials). Moreover, outcome *p1* never occurred on a trial in which cue A was absent. This rendered the outcome value *p1* predictable (and cue A predictive). Similarly, a second cue, B, perfectly predicted the presence of the other value on the predictable outcome dimension, *p2*. On every trial in which cue B occurred, so too did outcome *p2* (B-*p2* trials, B-*p2,u1* trials and B*p2,u2* trials). Again, outcome *p2* never occurred on a trial in which cue B was absent. These two cues, A and B, were completely non-predictive of the second outcome dimension (e.g. stomach reactions). On the trials in which cue A occurs, there was just as likely to be no stomach reaction (*uØ*), bloating (*u1*) or swelling (*u2*). The same was true for cue B.

When seeking to examine the influence of outcome predictability on subsequent learning, it is important to minimize or negate the possible influence of context-outcome associations that may form and potentially block subsequent learning of the relationship

between that outcome and discrete cues. It is worthwhile noting that the training context in computer-based human associative learning tasks is likely not as salient a potential cue as the conditioning boxes used in animal conditioning tasks (and its associated processes: handling, transfer from home box, etc.). Nevertheless, to reduce the possible influence of associations with the context, we followed the approach used by Matzel et al (1988) in which a discrete cue signalled the unpredictable outcome dimension in the first training phase. A third cue, cue X, always preceded a value (*u1* or *u2*) on the unpredictable outcome dimension, but did not predict which value would occur. On half of the trials, cue X preceded the value *u1* and on the remaining half it preceded *u2*. However, if cue X was absent, then the absent value on the unpredictable outcome dimension was always correct. That is, if cue X did not occur, no stomach reaction would occur. This made cue X a better predictor of the unpredictable outcome dimension than the context, and, in combination with the likely benefit in salience discrete cues enjoy over diffuse contexts, should have resulted in cue X overshadowing the context.

In the second phase, a new set of cues was introduced (E, F, G, H and Y), but the outcomes remained the same as those used in phase one. In the second phase, both outcome dimensions were predictable. That is, the contingencies were arranged such that cues E, F, G and H each predicted a unique outcome value on each outcome dimension. For example, cue E was reliably followed by a value on the predictable outcome dimension, itchiness (outcome *p1*), and also by a second value on the unpredictable outcome dimension, bloating (*u2*). It was equally possible for participants to learn the association between cue E and its associated value on the predictable outcome dimension, *p1*, as it was for them to learn the association between cue E and its associated value on the unpredictable dimension, *u2*. All that differed between these outcome values was their prior signalled history; *p1* was reliably predictable in phase one, where *u2* was not. All cues E, F, G and H were similarly predictive of a single

value on the previously predictable outcome dimension (e.g. skin reactions) and a single value on the previously unpredictable outcome dimension (e.g. stomach reactions).

Note that another cue, Y, was present on every trial in phase two, such that each trial consisted of two foods followed by two allergic reactions. This cue was included for two reasons. First, even though the previously unpredictable outcome dimension was objectively predictable, it may not have been perceived as predictable, and consequently it may have been associated with the context rather than the discrete cues. This was undesirable because assessing learning to the context in human associative learning tasks is difficult. Thus, again following the logic of Matzel et al (1988), cue Y was as predictive of the outcomes as the diffuse context, but due to being a discrete cue, should minimize conditioning between the phase two context and the previously unpredictable outcome dimension. Moreover, because cue Y was a discrete cue, learning for cue Y was readily measurable at the end of training.

The addition of cue X (in phase one) and Y (in phase two) also served a secondary function. Namely, they allowed participants to learn that each outcome was associated with a unique cue: stomach reactions were predicted by cues A and B, whereas skin reactions were (partially) predicted by cue X. This was important because pilot testing revealed that participants experienced substantial difficulty mapping a single cause to more than one effect.

Procedure. Participants assumed the role of an allergist who needed to learn which allergic reactions a new patient, Mr. X, experiences after he is exposed to vegetables (Phase 1) and fruits (Phase 2). Either one or two food cues were shown on each trial. Each cue consisted of a coloured line drawing with a text label (e.g. apple). On trials where two foods were shown, the positions of the foods on the screen were randomly determined (either upper or lower). Participants were required to predict the allergic reactions that would occur after each meal. One reaction could be chosen from the skin reaction dimension (including no skin reaction, skin itchiness) and one from the stomach reaction dimension (including no stomach

reaction, stomach bloating), each of which included an 'absent' value (no skin reaction, no stomach reaction). The six allergic reactions were displayed as labelled buttons (e.g. "Stomach Bloating") on which a small image of the outcome was shown (e.g. a small image of Mr X suffering from stomach bloating). The skin reactions were always shown on the left of the screen and the stomach reactions on the right. Once participants had made each selection they were asked to assess their confidence in each prediction on a scale of 1 to 5 (1 was labelled 'not at all confident,' 5 was labelled 'very confident'). Correct outcome values were then circled onscreen and matching pictures for each of the correct outcome values (e.g. stomach bloating and skin itchiness) were shown.

The phase one trials were organized in blocks, whereby each block consisted of one repetition of each of the eight trial-types listed in Table 1. Each block consisted of both elemental trials (A, B and X alone trials) and compound trials (AX, BX trials). As noted earlier, cue A always preceded outcome *p1*; cue B always preceded outcome *p2*; and cue X preceded outcome *u1* on 50% of trials and *u2* on 50% of trials. If cues A and B were absent (X alone trials), no outcome was observed on the predictable outcome dimension (i.e. *pØ* occurred). If cue X was absent, then no outcome was observed on the unpredictable outcome dimension (i.e. *uØ* occurred).

The trial order was randomized within blocks. The transition between blocks was not signalled. Fifteen blocks of Phase 1 training were given before participants proceeded to phase two. The second phase was preceded by a brief instruction that participants would now be shown fruits, and that their job was to learn about Mr X's fruit allergies. The two training phases appeared similar to the participant, except that new foods (E, F, G, H and Y) were used in phase two (see Table 1). In the second phase, both outcome dimensions were predictable because cues E-H all reliably predicted particular values on both outcome dimensions. For example, cue E reliably predicted values *p1* and *u2*. By contrast, cue Y

occurred on every trial and was thus partially reinforced (50%) with respect to outcomes *p1*, *p2*, *u1* and *u2*. As in Le Pelley & McLaren's (2003) learned predictiveness experiment, five repetitions of each trial-type were shown in Phase 2.

After Phase 2, participants proceeded directly to the test phase. On each of the five trials in this phase, participants were shown one of the fruit cues (E, F, G, H or Y). The fruits were shown in a random order. Participants were asked to individually rate the likelihood that each fruit would lead to each of the six outcome values (*uØ, u1, u2* and *pØ, p1, p2*). Each food-allergy rating was made by manipulating an onscreen scrollbar with a continuous scale from 1 (labelled 'very unlikely') to 100 ('very likely'). The six ratings made for each food cue were made on the same screen, and each food allergy rating could be adjusted independently of the others. For example, if a participant desired, they could rate cue E's relationship with outcome *p1* as 100, and also rate its relationship with outcome *p2* as 100 (we examine whether participants did so below). The assignment of foods to cues was randomized for each participant within each phase. The food cues A, B, and X were always vegetables (eggplant, potato, and carrot) and cues E, F, G, H, and Y were always fruits (cherry, banana, peach, lemon, grapes, and apricots). The assignment of skin and stomach reactions to the 'predictable' or 'unpredictable' outcome roles was counterbalanced between participants. This variable did not lead to any significant main effects or interactions with our comparisons of interest, and is therefore not discussed further.

Results

Training performance and exclusions. Mean performance of participants indicated that they learned the contingencies presented to them in training. However, upon closer inspection there was considerable individual variability in training performance. Many people were able to learn the training contingencies rapidly and to a high degree of accuracy, but

others exhibited significant difficulty throughout both training phases. Because the manipulation of primary interest, outcome predictability, required participants to learn the phase one training contingencies well, we considered these two populations separately. A median-split was performed based upon average prediction accuracy across phase one. In calculating the average we only considered trials on which participants could make a correct prediction. For example, on an AX trial, the correct prediction for the predictable outcome was *p1*, but for the unpredictable outcome it was impossible to determine whether the outcome would be *u1* or *u2*. On these trials we considered the predictions on the predictable outcome (i.e. did the participant choose *p1*?) but not their prediction for the unpredictable outcome (*u1* or *u2*). The high performer group had 25 members, and the low performer group had 26. Because this performance criterion was based on phase one performance, training data from the second phase could be analyzed without risk of circularity. The phase one prediction accuracy of the high and low performing groups is depicted in Figure 1.

[Figure 1 about here.]

Prediction accuracy in phase two for the previously predictable outcome dimension and the previously unpredictable outcome dimension is plotted in the left-hand panel of Figure 2. The confidence ratings given by participants on these trials are plotted in the right hand panel. For both panels, the data are further separated into the first and last half of phase two, and by group (high and low performers). The hypothesis to be tested by these data is that if an outcome becomes more readily associable once it has been shown to be predictable, then participants will learn more readily about the predictable outcome dimension than about the unpredictable outcome dimension in phase two. This was the pattern observed in the high-performing group, but no such pattern was seen in the low performing group.

[Figure 2 about here]

To examine this pattern further, separate 2 (early or late training) x 2 (predictable or unpredictable outcome) ANOVAs were used to examine the prediction accuracy of the high and low performing groups. The high performing group is discussed first. There was a main effect of phase of training (early versus late) on accuracy, $F(1,23)=68.51$, $p < .001$, MSE = 0.03, $\eta_P^2=.75$, 95% CI [.21, .35]; averaged across outcome-type, high-performing participants increased their prediction accuracy across Phase 2. A main effect of outcome-type on prediction accuracy was also observed, $F(1,23)=4.78$, $p = .04$, MSE = 0.03, $\eta_P^2=.17$, 95% CI [+.00, .13], whereby accuracy for the predictable outcome was higher than for the unpredictable outcome. A non-significant trend towards an interaction between training phase and predictability of the outcome was also observed. The difference in prediction accuracy between the predictable and unpredictable outcomes tended to be larger late in training than early in training, $F(1,23)=3.00$, $p = .10$, MSE = 0.02, $\eta_P^2=.12$, 95% CI [-.01,.11]. Although this interaction was non-significant, an interested reader may be curious as to whether prediction accuracy for the predictable outcome dimension was greater than for the unpredictable outcome dimension at the termination of Phase 2 training; it was, $F(1,24)=5.44$, $p = .03$, MSE =0.03, $\eta_P^2=.19$, 95% CI [.01, .23].

The low performing group showed a significant main effect of training phase, $F(1,24) = 16.82$, $p < 0.001$, MSE = 0.04, $\eta_P^2=.41$, 95% CI [.08, .24], whereby their overall prediction accuracy increased across Phase 2. No main effect of outcome type was observed, $F < 1$, and no interaction between these variables was observed, $F(1,24) = 1.27$. The confidence ratings (right hand panel) were analyzed in an identical manner to the outcome prediction data. Both the high and low performing groups showed a significant main effect of training phase, minimum $F(1,24)=9.45$, $p < .01$, MSE = 0.34, $\eta_P^2=.28$, 95%[.12, .59], whereby their mean confidence increased across Phase 2. Neither group showed a main effect of outcome-type (previously predictable or unpredictable), maximum $F(1,24)=1.51$, or an interaction between

outcome-type and training phase, maximum $F(1,24)=2.80$. Note that the low performing group gave higher initial confidence ratings than the high performing group. It is unclear why this occurred. However, we note the close proximity of the low performers' mean ratings throughout training to the midpoint of the rating scale, perhaps indicative of a "don't know" response.

Test Phase ratings. Our primary dependent variable was the ratings participants gave at test for the associations they learned in phase two. Recall that in this phase, most cues (E-H) were reliably paired with a value from each outcome dimension. For example, cue E was always paired with outcome values *p1* and *u2*. This was also true for cues F, G and H. The exception to this rule was cue Y. For cues E-H there was a clear 'correct' value for each outcome dimension (e.g. *p1* and *u2* were correct for cue E), a clear incorrect value (e.g. *p2* and *u1* were incorrect for cue E) as well as the two 'outcome absent' values (*pØ* and *uØ*), which were never shown in Phase 2). The correct, incorrect and outcome-absent (hereafter labelled 'nil') ratings were averaged across the individual cues E-H, as these cues were treated identically. These values are shown in the left-hand panels of Figures 3 and 4 (which depict the high and low performing groups, respectively).

The division between 'correct' and 'incorrect' was inappropriate for the Y cue, however, because cue Y did not have a clear 'correct' and 'incorrect' response in Phase 2. This is because cue Y was equally often followed by *u1* and *u2* on the unpredictable dimension, and by outcomes *p1* and *p2* on the predictable outcome dimension (but was never followed by *pØ* or *uØ*). For this reason, and because cue Y was shown more frequently than cues E-H, our analyses of the non-predictive cue Y were separated from our analyses of cues E-H. The measure of participants' learning about the non-predictive cue Y was defined as the average of their ratings for both of the outcome values cue Y was associated with on each outcome dimension (i.e. the average of the ratings for *u1* and *u2* and, separately, the average

of their ratings for outcomes *p1* and *p2*). These "correct" values were compared to participants' ratings for the outcome-absent response on each dimension (*pØ* and *uØ*), as cue Y was never followed by this value. The "correct" and "nil" values are summarized in the right-hand panels of Figures 3 and 4 (high and low performing groups, respectively).

There is some evidence that participants treated the test ratings for individual outcome values as independently as intended. To examine this issue, we sought to measure whether participants summed their ratings within each outcome dimension to 100 (or some approximation of this number). The ratings for individual values within a dimension were summed for each individual (i.e. ratings for *p1*, *p2* and *pØ* were summed, as were ratings for *u1*, *u2* and *uØ*). The mean sum of ratings significantly exceeded 100 for both the previously predictable outcome dimension (M=127.91, t(50)=5.85, p<.001, d = 0.82, 95% CI [118.47, 137.34]) and the previously unpredictable dimension (M=127.72, t(50)=5.19, p<.001, d = 0.73, 95% CI [117.15, 138.30]). These group statistics, however, may obscure individual patterns of responding, such that many individuals may have reliably sought to sum their ratings to approximately 100 but this may not have been evident at the group level. To investigate this possibility, each individual participant's ratings for each cue (E, F, G, H and Y) on each outcome dimension (predictable, unpredictable) were investigated separately. Of the 51 participants, only 4 individuals reliably gave ratings that summed to within 20% (80-120) for each individual outcome dimension. Together, these analyses suggest that the majority of participants did not systematically treat the ratings for individual outcome values (e.g. *p1, p2, pØ*) as zero-sum dependent.

The inferential analyses of test ratings for predictive cues (E-H) and the non-predictive cue Y are reported separately for the high and low performing groups. The high performer group is discussed first.

High performers. Our primary hypothesis concerned whether participants would exhibit a "learned predictability" effect, whereby they would more readily associate a cue with a previously predicable outcome (e.g. *p1, p2*) than with a previously unpredictable outcome (*u1, u2*). Participants' ratings following cues E-H were entered into a 2 (outcome value: correct versus incorrect/nil) x 2 (outcome dimension: predictable versus unpredictable) repeated measures ANOVA. These data are summarized in the left-hand panel of Figure 3.

[Figure 3 about here.]

Overall there was a main effect of outcome value, whereby the correct outcome values were given higher ratings than the average of the two inappropriate responses (those labelled Nil and Incorrect in Figures 3 and 4), $F(1,23)= 15.60$, $p= .001$,MSE=190.26, $\eta_P^2=.40$, 95% CI[8.39,26.86]. There was no significant main effect of previously predictable versus unpredictable outcome dimension, $F < 1$, but there was a significant interaction between outcome value (correct versus incorrect and nil) and outcome dimension (previously predictable versus unpredictable), $F(1,23)=6.17$, $p = .021$, MSE=842.75, ηP2=.21, 95% CI[2.09,22.91]. The filled black columns of Figure 3 (Panel A) show that for the correct outcome values, high-performing participants gave higher ratings on the previously predictable outcome dimension than on the previously unpredictable outcome dimension. A simple effect analysis confirmed this impression, $F(1,23)=5.23$, $p = .03$, MSE = 752.79, ηP2=.18, 95% CI[1.67, 33.22]. That is, participants appeared to show a learned predictability effect whereby the most readily learned associations in Phase 2 were those involving the previously predictable outcome dimension.

The ratings for the nil and incorrect outcome values showed, if anything, the opposite pattern. Ratings for these values on the previously unpredictable outcome dimension were numerically higher than those on the previously predictable outcome dimension, $F(1,23)=4.65$, $p =.04$, MSE = 307.22, ηP2=.17, 95% CI[0.30,14.82].

Turning to the high-performing participants' ratings for the non-predictive cue Y (Panel B in Figure 3), these ratings show the reverse pattern to that seen for the predictive cues E-H (Panel A). The ANOVA used to analyze cue Y was very similar to that used for cues E-H, except that for cue Y the outcome values variable only had two values (correct or nil). For ratings of cue Y, there was no significant main effect of outcome value (correct or nil), $F(1,23)=2.91$, $p = .10$, MSE = 1293.34, but there was a significant main effect of outcome dimension (previously predictable versus unpredictable), $F(1,23)=5.56$, $p = .03$, MSE = 406.15, $\eta_P^2=.19$, 95% CI[1.17, 17.86]. Overall, the correct outcome values (e.g. the average of *p1* and *p2*) were rated significantly higher than the nil outcome value (e.g. *pØ*). A significant interaction between the two factors was observed, $F(1,23)=9.02$, $p < .01$, MSE = 1999.62, $\eta_P^2=.28$, 95% CI[8.36, 45.39]. A simple effect contrast revealed that the ratings for the correct values following cue Y were higher on the unpredictable outcome dimension than on the predictable outcome dimension, $F(1,23)=4.46$, $p = .046$, MSE = 842.754, ηP2=.16, 95% CI[0.36, 34.36]. This could be considered to be a "learned unpredictability" effect (we will return to this idea), as participants more readily associated the nonpredictive cue Y with the previously unpredictable outcome dimension than they did with the previously predictable dimension.

A second simple effect contrast revealed that participants' ratings for the nil outcome showed the opposite trend to that seen in the correct ratings. Ratings for the nil outcome were significantly higher on the previously predictable outcome dimension than on the unpredictable dimension, $F(1,23)=10.57$, $p < .01$, MSE = 1563.02, ηP2=.31, 95% CI[13.24, 59.54].

In summary, participants showed no overall propensity to associate previously predictable outcomes with novel cues. This was due, however, to the confluence of two patterns of performance: predictive cues were more readily associated with previously

predictable outcomes but the non-predictive cue (Y) was more readily associated with the previously unpredictable outcome dimension.

Low performers. The test ratings of the low-performing group are summarized in Figure 4. Notice that in this group, unlike the high-performing group, test ratings for the correct outcome values for the predictive cues E-H (left panel) were similar to their ratings for the incorrect outcome values. This reduced ability to distinguish between the correct and incorrect Phase 2 outcome values confirms that these participants did not learn the training contingencies well (they were classified as low performers based on their Phase 1 prediction accuracy, and this poor performance continued in Phase 2).

[Figure 4 about here.]

The data from these participants were entered into the same 2 x 2 (outcome value by outcome dimension) analyses used to examine the ratings of the high-performer group. For the predictive cues E-H, the low performing participants gave significantly higher ratings to the correct outcome value than to the nil or incorrect outcome values, $F(1,24)=33.22$, $p < .001$, MSE = 320.26, $\eta_p^2=.56$, 95% CI [11.25, 23.79]. There was no main effect of outcome dimension, $F < 1$, and the two factors did not significantly interact, $F(1,24)=2.00$, $p = .17$, MSE = 261.77.

To ascertain whether a learned predictability effect was observed in this group, a follow-up simple effect contrast compared their cue E-H ratings for the correct outcome value on the previously predictable dimension against the correct value on the previously unpredictable dimension. No significant difference was observed, $F(1,24)=1.38$, $p = .25$, MSE = 267.85. Similarly, with respect to the non-predictive cue Y's relationships with the outcomes, the low-performers' ratings did not reveal any influence of the predictability manipulation. No main effect of outcome predictability was observed, $F(1,24)=1.79$, $p = .20$, MSE = 281.23, although overall this group gave higher ratings for the correct outcome value

than the nil outcome value, $F(1,24)=25.47$, $p < .001$, MSE = 833.13, $\eta_P^2=.51$, 95% CI[16.88, 40.25]. Unlike the high performing group, in the low performing group there was no interaction between the outcome dimension variable (previously predictable versus unpredictable) and the outcome value variable (correct versus nil), $F(1,24)=2.26$, $p = .15$, MSE = 1173.12. Although no significant interaction was observed by between outcome dimension and value, a final simple effect contrast was conducted to test for a learned predictability (or unpredictability) effect. Ratings for the correct outcome on the predictable dimension were not significantly different to ratings for the correct outcome on the unpredictable dimension, $F < 1$. Overall, the low performer group's test ratings were unaffected by the outcome predictability manipulation, which is consistent with their inability to learn the cue-outcome relationships in phase one.

Discussion

The participants who were able to learn the phase one contingencies (which constituted the outcome predictability manipulation) demonstrated significant biases in their subsequent learning of a second set of associations between novel cues and familiar outcomes. These high-performing participants learned more rapidly about the phase two relationships involving previously predictable outcomes than the previously unpredictable outcomes. However, no overall learning bias for previously predictable outcomes was evident in their test ratings, when averaged across the training cues. Instead, two opposing learning biases were seen for the predictive cues E-H and the non-predictive cue Y. Specifically, their test ratings indicated that they more readily associated the previously predictable outcomes with the predictive cues E-H, than the previously unpredictable outcomes, but that this pattern was reversed for the non-predictive cue Y. We interpret these findings as evidence that participants' learning is affected by the prior predictability of the outcome, in a manner

consistent with, and reminiscent of, the learned predictiveness effect seen with cues (Le Pelley & McLaren, 2003). However, as in the literature discussed in the introduction (US pre-exposure and learned irrelevance effects), alternative explanations are possible. We withhold a full discussion of alternative accounts (e.g. context-blocking), and of the apparent modulating role of cue predictiveness, to the General Discussion.

While it is tempting to directly compare the ratings given to the predictive cues E-H with those given to cue Y, and thereby consider the effect of cue-predictiveness on the formation of associations, such comparisons are inappropriate. Cue predictiveness was not the only difference between cues E-H and cue Y. For example, cue Y was shown more frequently, and in conjunction with more outcome values (*p1, p2, u1, u2*), than were cues E-H. Thus, such direct conclusions concerning cue predictiveness cannot be drawn from these data.

Finally, there is an important caveat to all of these interpretations. Each of the effects discussed above was only observed in the performance of participants who learned the initial contingencies well. No significant learning biases were observed in the low-performing group. Although there is little risk of circularity in the definition of the two groups (based on overall first phase performance), the generality of these conclusions is nonetheless threatened by considering only the highest performing half of the sample. Further, as acknowledged earlier, we only defined these groups after the experiment was conducted (via a median split analysis). For this reason we conducted a conceptual replication of Experiment 1 in which participants were trained until they reached a pre-specified high level of performance.

Experiment 2

The present experiment replicated Experiment 1, but used a performance criterion to determine the end point of phases one and two. Participants continued to be shown training

trials in the first phase until they demonstrated perfect knowledge of the contingencies. In order to be maximally sensitive to differences in learning of the second phase contingencies, we sought to cease this phase prior to participants reaching asymptote, but after learning had been demonstrated. For this reason, phase two training was terminated when participants reached 75% accuracy. Second, we conducted Experiment 2 using eye-scanning equipment. Attentional biases are often evident in participant's gaze behaviour (e.g. Kruschke, Kappenman & Hetrick, 2005; Le Pelley, Beesley & Griffiths, 2011). If the biases towards and away from outcomes are comparable to biases amongst cues then differences in gaze toward outcome stimuli may be evident in the present experimental procedure. Finally, because the confidence ratings did not provide any additional explanatory power in Experiment 1 (and participants found them obstructive), we removed the confidence ratings in Experiment 2.

Method

Participants. Thirty four undergraduate students from the University of New South Wales participated in exchange for course credit.

Design and Procedure. The present experiment was very similar to Experiment 1 with a few exceptions. First, participants no longer made confidence ratings after generating their outcome predictions. Second, eye-gaze was recorded continuously throughout both phases of training. Third, participants did not complete a set number of training blocks in either training phase. Participants could complete between fourteen and thirty blocks of phase one training, and between four and eight blocks of phase two training. The amount of training depended upon how rapidly participants passed the performance criterion set for that phase. At the end of each training block in phase one (every eight trials) participants' number of correct responses in the last five training blocks (forty trials) was tallied. This algorithm allowed multiple responses to be considered correct for some predictions. For example, on an A-

p1,uØ trial, the algorithm would consider only *p1* and *uØ* to be correct responses. In contrast, on an AX-*p1,u1* trial, the algorithm would only score *p1* as a correct response (not *p2* nor *pØ*), but would allow either *u1* or *u2* to be scored as correct responses (not *uØ*). This is because the participant had no way of predicting whether outcome *u1* or *u2* would occur. If participants performed perfectly on the previous forty trials they were considered to have met the criterion. They were then given four additional training blocks to cement this knowledge before moving on to phase two. This criterion was only applied from the tenth training block, so fourteen blocks was the minimum a participant could experience prior to the second phase (for comparison, all participants completed fifteen blocks in phase one of Experiment 1). If a participant did not reach criterion they progressed to phase two after thirty blocks (240 trials) and their data were discarded.

The performance criterion used in phase two operated similarly to that used in the first phase. The phase two criterion algorithm tallied participants' responses after each block of four trials (there were four different trial-types in phase two), starting at the fourth block of training (trial number 16). If a participant responded accurately on 75% of the previous eight trials (12 correct responses from 16 outcome predictions), then they would proceed immediately to the test phase. Otherwise a maximum of eight blocks (32 trials) were administered.

Results

Training. As in Experiment 1, participants showed significant individual differences in their ability to learn the training contingencies. Fifteen participants (44%) failed to reach criterion performance in either phase one or two. For the nineteen participants who reached criterion performance, the mean number of trials before the criterion was passed was 116 trials (SD= 33) in Phase 1 and 24 trials in Phase 2 (SD=4.8).

Participants' prediction accuracy on the two outcome-types (previously predictable and unpredictable) in the first two and last two blocks of phase two training is shown in Figure 5. Note that the first two blocks of training were common to all participants, but the last two blocks of training occurred at different times for different participants. The participants who learned the phase two contingencies rapidly may have only experienced four blocks of training, so their first and last two blocks of training abut each other. A participant who learned more slowly may have seen seven blocks of phase two training, and their first and last two blocks of training would be separated by three training blocks.

[Figure 5 about here.]

The prediction accuracy data were entered into a 2 (early or late in training) by 2 (outcome type) multivariate ANOVA. Unsurprisingly there was a main effect of training phase, with participants showing greater accuracy late in training, $F(1,17)=145.89$, $p <.001$, MSE = 0.03, $\eta_P^2=.89$, 95% CI [0.39,0.55]. There was also a main effect of outcome-type, as prediction accuracy was greater for the previously predictable outcome than for the previously unpredictable outcome, $F(1,18)=4.85$, $p =.04$, MSE = 0.02, ηP2=.22, 95% CI [+.00, .15]. These two effects did not significantly interact, $F < 1$.

Gaze behaviour. Participant's gaze-behaviour was recorded continually across phase two. Specifically, we recorded the total time spent fixating upon each of the cues (E-Y) and outcomes (*uØ, u1, u2, pØ, p1, p2*) prior to generating a response. These values were tallied separately for each trial. These fixation times were then divided by the total time spent fixating anywhere on the screen during that trial. Calculating a proportion in this manner controls for inter-trial and inter-individual differences such as the duration of the trial, or the portion of the trial in which the participant is looking at the screen. The resultant proportional fixation times (hereafter referred to as dwell times) were then aggregated across cues types (predictive cues E-H versus non-predictive cue Y). The fixations to outcome stimuli were

classified in the same manner as outcomes were classified in Experiment 1 (i.e. correct, incorrect or nil). The mean gaze dwell times for the cues (panel A) and outcomes (panel B) shown during phase two are summarized in Figure 6.

[Figure 6 about here.]

Dwell times for the cue stimuli were entered into a 4 (training block) by 2 (predictive or nonpredictive cue) multivariate ANOVA. Averaged across both cue-types, a significant linear trend was seen in the data, with participants proportionally gazing at all cue-types less at the end of training than at the beginning, $F(1,18)=140.63$, $p < .001$, MSE= 0.03, $\eta_P^2=.86$, 95% CI[.05,.08]. A second main effect revealed that participants spent more time gazing at the predictive cues E-H than at the non-predictive cue Y, $F(1,17)= 10.67$, $p < .01$, MSE=0.02, ηP2=.28, 95% CI [.01,.03]. The linear trend contrast did not significantly interact with cue-type, $F < 1$.

Participants' gaze data for the outcome stimuli in phase two were entered into a 3 (outcome value: Correct, Incorrect, Nil) by 2 (outcome type: predictable or unpredictable) by 4 (training block) repeated measures ANOVA. One participant was removed due to incomplete gaze data. A linear trend contrast was used to examine the influence of training block on gaze. No significant linear influence of training block was observed, $F < 1$. Similarly, the main effect of outcome type (previously predictable or unpredictable) was not significant, $F(1,16)=4.49$, p = .05, MSE<0.01, $\eta_P^2=.15$, 95% CI[.00, .04]. There was a main effect of outcome value. Overall, participants gazed longer at the correct outcome values than at the average of the nil and incorrect values, $F(1,16)=23.70$, $p < .001$, MSE<0.01, $\eta_P^2=.59$, 95% CI[.04,09]. An interaction contrast showed that this bias toward the correct outcome values, over the incorrect and nil values, did not differ in magnitude between outcome types (predictable versus unpredictable), $F < 1$. That is, participants did not learn to selectively gaze at the correct outcomes on the previously predictable outcome dimension more rapidly

than they did for the previously unpredictable outcome dimension. A second interaction contrast revealed that the bias in gaze towards the correct outcome values increased in magnitude across the training blocks, $F(1,16)=58.49$, $p < .001$, MSE<0.01, ηP2=.74, 95% CI[.05, .09]. No other interactions were significant.

In summary, participants gazed more at the predictive cues (E-H) than at the non-predictive cue Y, and across training learned to gaze at the correct outcome values more than at the incorrect outcome values (both predictable and unpredictable).

Test phase ratings. The primary dependent variable, participants' mean likelihood ratings at test, are shown in Figure 7. These data were organized and analyzed in an identical manner to the test rating data in Experiment 1. To examine whether participants treated the test ratings for each outcome values as independently as intended, the same analysis of summed mean ratings per outcome dimension was conducted as that performed in Experiment 1. As in Experiment 1, the mean sum of ratings significantly exceeded 100 for both the previously predictable outcome dimension (M=120.31, $t(18)=2.51$, p=.02, d = 0.57, 95% CI [104.02,136.60]) and the previously unpredictable dimension (M=127.98, $t(18)=5.04$, p<.001, d = 1.15, 95% CI [116.81,139.12]).

The ratings following cues E-H were analysed separately to the ratings following the non-predictive cue Y, in exactly the same manner as was used in Experiment 1. Ratings for cues E-H were analyzed using a 3 (outcome value: correct versus incorrect/nil) by 2 (outcome dimension: previously predictable vs unpredictable) ANOVA. For the analysis of cue Y's ratings, the outcome value variable only had two values (correct and nil). Overall, there was no main effect of outcome dimension, $F < 1$, but there was a main effect of outcome type; reassuringly, the correct outcome values were given higher likelihood ratings than the nil and incorrect values overall, $F(1,18)=27.96$, $p < .001$, SEM = 592.51, η_P^2=.61, 95% CI[15.41,35.73]. The interaction between these variables was the crucial contrast; this was

significant, $F(1,18)=6.76$, $p = .02$, SEM=809.79, $\eta_P^2=.27$, 95% CI[2.83, 26.58]. Two further simple effect contrasts were conducted to aid interpretation of this interaction. The first simple effect contrast showed that people gave higher ratings to the correct outcome value on the previously predictable outcome dimension than to the correct value on the previously unpredictable dimension $F(1,18)=4.81$, $p = .04$, SEM = 783.24, $\eta_P^2=.21$, 95% CI[0.85, 39.00]. Thus the learned predictability effect seen in Experiment 1 was replicated in the present experiment. The second simple effect contrast examined whether ratings for the nil and incorrect outcome values combined showed the opposite pattern, in that the ratings for these values on the previously unpredictable outcome dimension were higher than the predictable outcome dimension. This effect was also significant, $F(1,18)=8.45$, $p < .01$, SEM= 202.35, $\eta_P^2=.32$, 95% CI[2.63, 16.34].

[Figure 7 about here.]

Similar to Experiment 1, participants appeared to give a pattern of ratings for the nonpredictive cue Y that were opposite to those they gave for the predictive cues E-H. Overall, there was no main effect of outcome dimension, $F < 1$, but there was a main effect of outcome value, $F(1,18)=23.86$, $p < .001$, SEM= 823.50, $\eta_P^2=.57$, 95% CI[18.32, 45.99], whereby the correct values were given higher ratings than the nil value overall. More importantly there was a significant interaction between these factors, $F(1,18)=9.59$, $p < .01$, SEM= 1281.00, $\eta_P^2=.35$, 95% CI[8.17, 42.67], that was opposite in direction to that seen for the predictive cues E-H. Two simple effect contrasts confirmed this. People gave higher ratings for the correct values on the previously unpredictable (*u1, u2*) outcome than for those on the previously predictable (*p1, p2*) outcome, $F(1,18)=6.33$, $p = .02$, SEM= 751.16, $\eta_P^2=.26$, 95% CI[3.69, 41.05], and showed the opposite pattern in their ratings of the nil outcome value, $F(1,18)=8.42$, $p = .01$, SEM= 915.24, $\eta_P^2=.32$, 95% CI[7.85, 49.10].

Discussion

The present experiment replicated the key findings of Experiment 1. That is, people learned more rapidly about the new associations involving previously predictable than about previously unpredictable outcome dimensions in phase two. Again, the opposite pattern was seen for the nonpredictive cue Y. Cue Y was more readily associated with values on the previously unpredictable outcome dimension than on the predictable outcome dimension. Contrary to Experiment 1, the present experiment did not divide participants on a *post hoc* basis. The use of a pre-defined performance criterion in the present experiment removes the risk of the important differences in test performance being entirely due to the application of *post hoc* classifications.

The present behavioural data were complemented by the addition of eye-gaze measures. Participants gazed longer at the predictive cues E-H than at the non-predictive cue Y in Phase 2. This is consistent with previous observations of cue predictiveness being reflected in gaze behaviour (e.g. Beesley & Le Pelley, 2010), but is also explicable in terms of the greater familiarity of cue Y than cues E-H. The predictable outcomes were not, however, gazed at longer overall than the predictable outcomes during Phase 2. Although Experiment 2 replicated the key findings of Experiment 1 in a group of people selected in a pre-defined and systematic manner, it was still somewhat unsatisfying in that many people were excluded due to poor training performance. A final experiment was conducted in which an effort was made to increase the number of participants who reached criterion-level performance.

Experiment 3

Many people were unable to learn the training contingences in Experiments 1 and 2, yet those participants who were able to learn the training contingencies did so relatively rapidly and

with high accuracy. It was not clear why some people were unable to learn the training contingencies. One possibility is that participants were not learning the elemental nature of the initial phase contingencies. To this end, the present experiment provided elemental pre-training with cues A, B and X (the Phase 1 cues) prior to phase one in the hope that this would encourage participants to treat the phase one cues elementally. Second, in Experiment 2 relatively few blocks of phase two training (8 blocks) were provided to participants. It is therefore possible that some participants would have passed the criteria if given additional practice. Thus additional phase two training was offered to participants in the present experiment.

Method

Participants. Fifty six undergraduate students from the University of New South Wales participated in exchange for course credit.

Procedure. The procedure of Experiment 3 was very similar to that of Experiment 2, with four differences. First, no eye-tracking data was collected in the present experiment. Second, the maximum number of phase two training blocks that were offered was increased from eight blocks to twenty blocks. Third, the performance criterion in phase one was relaxed slightly. Rather than requiring perfect performance on forty consecutive trials (eighty consecutive outcome predictions), one error was permitted. This change was made because a review of training data showed that some participants in Experiment 2 made a single error after a sustained period of perfect performance. As a consequence, these individuals had then been required to undergo (at least) another forty trials.

Fourth, and most importantly, an initial elemental pre-training phase (phase zero) was provided to all participants prior to phase one. The design of this pre-training phase is summarized in the left-most column of Table 1. Essentially, this training phase was identical

to the first phase, except that only elemental trials were shown (i.e. AX and BX trials were excluded). The same algorithm was used to calculate whether predictions on a trial were 'correct' or not as that used in Experiment 2 (i.e. a response of *p1* was required on a trial in which *p1* occurred, but a response of either *u1* or *u2* was considered to be correct on trials in which *u1* or *u2* occurred). Participants were required to make 39 or 40 consecutive accurate outcome predictions (that is, near perfect performance on twenty consecutive trials) before proceeding to Phase 1. The performance criterion was applied from the twentieth trial onwards (after five repetitions of each trial-type), and from then on was applied after every block of the four trial-types. Participants were provided 30 blocks of phase zero training before they were excluded if the performance criterion was not met. After reaching criterion performance, phase one began without a signal or break. Phases one, two and the test procedure were otherwise identical to Experiment 2.

Results

Training. Relative to Experiment 2, more participants passed the performance criteria in the present experiment. Only 4 people (7%) were unable to learn the training contingencies within the provided time, and their data have been excluded from further analyses. Three participants failed to learn the phase one contingencies to criterion with 30 repetitions of each trial-type (240 trials), and one failed to learn the phase two contingencies within 30 repetitions (120 trials). The remaining 52 participants (93%) all reached criterion performance in phases zero, one and two. The mean number of trials taken to reach criterion in phase zero was 45.64 (SEM=1.40, minimum possible= 40 trials[1]), the mean number of trials taken to reach criterion in phase one was 98.20 (SEM=6.72, minimum possible = 72) and the mean number of trials taken to reach criterion in Phase 2 was 28.86 (SEM=2.27,

[1] It was technically possible for participants to proceed after 36 trials, but this would require perfect performance from the very first trial.

minimum possible=16). Notably, just as in prior experiments, there was significant variation in the speed with which people learned the training contingencies. Many people learned the contingencies quite rapidly: 58% reached criterion at the first opportunity in phase zero, 58% reached criterion at the first opportunity in phase one and 33% reached criterion at the first opportunity in phase two. A sizeable proportion of participants, however, needed at least 50% more training than the minimum possible value ('extra training') in order to reach criterion on each phase: 8%, 21% and 60% needed extra training in phases zero, one and two, respectively.

Participants' prediction accuracy on the two outcome-types (previously predictable and unpredictable) in the first two and last two blocks of phase two training is shown in Figure 8. As in Experiment 2, the first two blocks of training were common to all participants, but the last two blocks of training occurred at different times for different participants. The participants who learned the phase two contingencies rapidly may have only experienced four blocks of training, so their first and last two blocks of training abut each other whereas for others there may be other blocks between their first and last two training blocks.

[Figure 8 about here.]

The prediction accuracy data were entered into a 2 (early or late in training) by 2(outcome type) multivariate ANOVA. Unsurprisingly there was a main-effect of training phase, with participants showing greater accuracy late in training, $F(1,50)=250.01$, MSE=.05, $p < .001$, $\eta_P^2=0.83$, 90% CI [.04,.04]. There was no significant main effect of outcome-type, $F < 1$, and no significant interaction between training phase and outcome type, $F < 1$.

Test phase ratings. The primary dependent variable, participants' mean likelihood ratings at test, is shown in Figure 9. To examine whether participants treated the test ratings for each outcome values as independently as intended, the same analysis of summed mean

ratings per outcome dimension was conducted as that performed in Experiments 1 and 2. Again the mean sum of ratings significantly exceeded 100 for both the previously predictable outcome dimension (M=117.46, t(51)=4.52, p<.001, d = 0.56, 95% CI [108.21, 127.50]) and the previously unpredictable dimension (M=123.79, t(51)=5.60, p<.001, d = 0.66, 95% CI [112.62, 134.95]).

The mean test rating data were organized and analyzed in an identical manner to the test rating data in Experiments 1 and 2, with the exception that one-tailed inferential analyses were used (as our hypotheses were confined to whether the key effects of Experiments 1 and 2 replicate).

The ratings following cues E-H were analyzed separately to the ratings following the non-predictive cue Y, in exactly the same manner as was used in Experiments 1 and 2. Ratings for cues E-H were analyzed using a 3 (outcome value: correct versus incorrect/nil) by 2 (outcome dimension: previously predictable vs unpredictable) ANOVA. For the analysis of cue Y's ratings, the outcome value variable only had two values (correct and nil). The ratings associated with cues E-H are discussed first.

[Figure 9 about here.]

Overall, there was no main effect of outcome dimension, F < 1, but there was a main effect of outcome type; reassuringly, the correct outcome values were given higher likelihood ratings than the nil and incorrect values overall, F(1,50)=111.32, p < .001, SEM = 908.75, η_P^2=.69, 90% CI[32.13, 44.27]. As in Experiments 1 and 2, the most important contrast is the interaction between outcome dimension (previously predictable, previously unpredictable) and the contrast comparing the correct outcome value with the other possible response options (incorrect, nil). This interaction was again significant: F(1,50)=7.02, p = .007, SEM = 685.01, η_P^2=.12, 90% CI[2.61, 11.58]. A simple effect contrast revealed that the central finding of Experiments 1 and 2 was replicated: for the predictive cues E-H participants gave

higher ratings to the correct values on the previously predictable outcome dimension than the correct values on the previously unpredictable outcome dimension, $F(1,50)=5.39$, $p = .01$, SEM = 575.26, $\eta_P^2=.10$, 90% CI[3.04, 18.80]. A second simple effect found that participants' average ratings for the two outcome values not associated with cues E-H (the incorrect and nil outcome values) did not significantly differ between the previously predictable and the previously unpredictable outcome dimension, $F(1,50)=2.73$, $p = .05$, SEM = 203.657, $\eta_P^2=.05$, 90% CI[-0.05,6.59]. When averaged across the previously predictable and unpredictable outcome dimensions, participants gave significantly higher ratings to the incorrect value than to the nil value, $F(1,50)=12.99$, $p < .001$, SEM = 685.014, $\eta_P^2=.21$, 90% CI[7.00,19.16] . This contrast did not interact with outcome type (previously predictable, unpredictable), $F < 1$.

Mean likelihood ratings for the outcome values associated with the non-predictive cue Y are shown in the right-hand panel of Figure 9. These data were analyzed in a similar manner to the data from cues E-H, except that the outcome value variable only had two values (correct, nil). No significant main effect of outcome dimension (predictable, unpredictable) was observed, $F(1,50)=2.40$, $p = .06$, SEM = 407.32, $\eta_P^2=.05$, 90% CI [-0.35, 9.03]. The main effect of outcome value (correct, nil) was significant, $F(1,50)=63.81$, $p < .001$, SEM = 1110.93, $\eta_P^2=.56$, 90% CI[29.18, 44.67] with the correct values given higher overall ratings than the nil values. Importantly, as in Experiments 1 and 2, these two contrasts significantly interacted, $F(1,50)=3.44$, $p = .04$, SEM = 835.43, $\eta_P^2=.06$, 90% CI[0.72, 14.15]. Two further simple effect contrasts clarified this interaction. People gave higher ratings to the correct values on the previously unpredictable outcome dimension (*u1, u2*) than the correct values on the previously predictable dimension (*p1, p2*), $F(1,50)=4.83$, $p = .02$, SEM = 745.07, $\eta_P^2=.09$, 90% CI[2.80, 20.74]. However, no significant differences in ratings to the nil outcomes were observed across the two outcome dimensions, $F < 1$.

Discussion

The present experiment replicated the central findings of Experiments 1 and 2, but did so using a training procedure in which almost everyone (93% of people) was able to learn well enough to attain near perfect performance within half an hour. Just as in Experiments 1 and 2, the prior predictability of an outcome appeared to strongly influence the degree to which that outcome entered into associations with novel cues. Again, the predictive status of the novel cues appeared to moderate this influence of prior predictability. Possible explanations of this effect are considered below. Finally, unlike in Experiments 1 and 2, no significant differences in learning rate (measured as the degree to which prediction accuracy increased over training blocks) were seen between the previously predictable and previously unpredictable outcome dimensions. It is unclear why no differences were seen in the present experiment. One possible explanation may be derived from the differences between Experiments 1 and 2 and the present experiment. Note that, in this experiment, participants were very highly trained on the phase one contingencies prior to engaging in phase two, so as to make all participants comparable to the 'high performers' in Experiment 1. This extra training may have allowed them to learn the relationships involving the previously predictable outcomes more rapidly in the present experiment, and that this afforded more resources to also learn about the previously unpredictable outcome values, thus reducing the performance difference between the two outcome dimensions.

General Discussion

Three experiments examined the degree to which prior experience of an outcome being unpredictable affected people's ability to learn associations involving that outcome. Overall, our initial hypothesis that previously predictable outcomes would be more readily associated

with novel cues was partially supported. In all experiments, prior experience of an outcome being unpredictable resulted in impaired learning of that outcome's relationship with a novel predictive cue (E-H). To our knowledge, this is the first such demonstration in a human population. Moreover, prior unpredictability appeared to impair the rate at which new predictive relationships were formed (as seen in the training prediction accuracy data in Experiments 1 and 2). Experiment 2 showed that prior predictability of outcomes did not influence the extent to which participants gazed at them later. Curiously, in a departure from our initial hypothesis, all three experiments found evidence that prior experience of an outcome behaving unpredictably had a facilitative effect on the formation of subsequent associations with a novel partially reinforced cue (Y). In summary, novel predictive cues were more readily associated with previously predictable outcome values, than previously unpredictable outcome values, whereas the opposite pattern was observed for novel non-predictive cues.

Importantly, the present experimental design affords confidence that the observed learning biases were not merely the product of a direct translation of the cue-outcome associations learned in phase one, or of some systematic generalization or response reassignment process, and instead reflect biases in new learning of cue-outcome associations. This is because novel cues were used in phase two, and there were no systematic relationships between the cues that were objectively predictive of particular outcome values in the first phase (e.g. A–*p1, B–p2*) and those that were predictive of particular outcome values in the second (E–*p1,u2;* H –*p2,u2.*). For example, knowing that potato predicted bloating in phase one provides no information as to whether cherry predicts bloating or swelling or perhaps no reaction. Thus, any biases seen in phase two as a result of phase one training cannot be mediated by discrete cues (but see below for a discussion of context driven effects), or by cue-outcome associations. The biases must instead be driven by some other

type of learning. This is an important finding, because learning biases mediated by a learned property of the outcome (i.e. its predictability) have previously been largely ignored in the human learning literature.

As noted earlier, we favour the view that the biased phase two learning is due to people encoding information about which outcomes were predictable (e.g. see Baker et al, 1981), and which were not, and using this information to guide subsequent learning. However, explanations that appeal to other forms of learning are also possible. We first consider explanations based on acquired equivalence and context-blocking, in turn, before returning to the possibility that people encode and use information about outcome predictability.

Acquired equivalence

One possible account of the present results is based on acquired equivalence (Miller & Dollard, 1941). Acquired equivalence (or mediated generalization) refers to the observation that two initially distinct stimuli can come to be treated as functionally identical by virtue of a shared association, in this case, a common antecedent (see Hall, Ray & Bonardi, 1993 for this effect in rats). This account suggests that, across phase one training, outcome values *u1* and *u2* came to be treated as the same outcome value, hereafter denoted *u12*, because both outcomes *u1* and *u2* were reliably preceded by the same cue (X). By the same mechanism, outcome values *p1* and *p2* came to be seen as more distinct due to their reliably different antecedents A and B ("acquired distinctiveness"). Thus, in the second phase, compound cues (EY, FY, GY, HY) are followed by three outcomes *p1, p2* and *u12*. Importantly, E and G are the best predictors of *p1*. Also, F and H are the best predictors of *p2*. Finally, Y is the best predictor of *u12*. Of course, if acquired equivalence between *u1* and *u2* hadn't taken place, E-H would have been the best predictors of these two outcomes. Thus, as a consequence of the

acquired equivalence process, E-H will become associated with *p1* and *p2*, whereas Y will become associated with *u1* and *u2* in Phase 2.

One aspect of the current data does not seem consistent with this account. Participants learned to predict outcomes *u1* and *u2* correctly across the Phase 2 trials (achieving 80% accuracy in Experiment 2). This in itself suggests that *u1* and *u2* were not treated as the same outcome. Furthermore, correct predictions with respect to *u1* and *u2* require that the participants know the relationship between cues E-H and these unpredictable outcomes; cue Y will not tell the participant which outcome is to be presented. However, partial (but not complete) acquired equivalence between *u1* and *u2* should be sufficient to ensure that 1) Y becomes more strongly associated with *u1/u2* than with *p1/p2* and, therefore, that 2) Y competes more successfully with E-H for association with *u1/u2* than it does for association with *p1/p2*. Because cues E-H and cue Y were treated very differently, it is not possible to compare them directly. However, within each cue type, the current results can be explained in terms of the idea that *u1* and *u2* were treated, to some extent, as the same outcome as a consequence of phase one training. This acquired equivalence learning could be the product of associative (e.g. Hall, Mitchell & Graham, 2003), attentional (Bonardi, Graham, Hall & Mitchell, 2005) and/or propositional processes (Smyth, Barnes-Holmes & Barnes-Holmes, 2008).

Context-blocking

A number of alternative explanations of the present data can be generated by adding an additional mechanism to the context-blocking account (discussed in the introduction) of US pre-exposure effects in animal conditioning. Recall that the US pre-exposure effect refers to the observation that initial provision of unsignalled shocks to an animal impairs subsequent learning about the relationship between shock and a valid predictor (the US pre-exposure

effect). This could be seen as evidence that animals encode and use information about the predictability of USs to guide the formation of associations. However, as Randich & LoLordo (1979) demonstrated in their review of US pre-exposure effects, the 'US pre-exposure' data are more consistent with the view that when animals encounter unsignalled USs, they associate them with the experimental context, and this context-US association blocks the ability of discrete cues to subsequently become associated with that US. Because the present data appear to indicate that prior predictability guides subsequent learning, it is also important to consider the role that contextual learning might play.

Applied to the present experiments, the context-blocking account suggests that participants may not have learned that outcomes *u1* and *u2* were unpredictable during Phase 1, but instead learned an association between these cues and the diffuse context cues (i.e. screen colour, room lighting, background noise, internal state). As argued above, we explicitly included the additional cue, X, to minimize this possibility (following Matzel et al, 1988). Cue X was (i) a better predictor of both unpredictable outcome values (*u1, u2*) than the context (0.5 contingency versus 0.33) during phase one, and (ii) was likely more salient to the participant (cue X was explicitly presented as a cue on each trial, the context was not). Nonetheless, it is possible that some conditioning accrued to the context for the unpredictable outcome values (*u1, u2*) in Phase 1, and that this conditioning would be greater than that for the predictable outcome values (*p1, p2*).

Under this contextual learning account, any cue shown in phase two in conjunction with the unpredictable outcome values (*u1, u2*) would be less readily associated with *u1* and *u2*, because the association between the context and *u1*/*u2* would block this learning. This account provides an explanation as to how the predictive cues (E-H) were more readily associated with the previously predictable outcomes (*p1, p2*) in phase two. It does not, however, explain why the non-predictive cue Y was more readily associated with the

previously unpredictable outcome values *u1* and *u2*, than with the previously predictable outcome values *p1* and *p2*. If outcomes *u1* and *u2* formed associations with the context in phase one, these associations should impair the formation of associations between any cue (including cue Y) and the outcomes *u1* and *u2* (relative to the previously predictable outcomes *p1* and *p2*). Thus, in order to provide a complete explanation of the present data set, the context blocking account needs an additional mechanism. There are a number of candidate mechanisms, including: the format of the response measure, inhibitory learning and differential overshadowing. We will consider these explanations in turn.

Format of response measure. The first possibility is that people's higher ratings for the unpredictable outcomes *u1* and *u2*, over *p1* and *p2*, for the non-predictive cue Y is a consequence of the nature of the test measure, rather than being reflective of learning itself. This explanation relies on the observation that ratings for all outcome values (*p1, p2, pØ, u1, u2, uØ*) were made on the same screen. Perhaps people felt that the ratings were dependent, and that their individual outcome value ratings must sum to 100. Data presented in detail for Experiment 1 (and briefly reported for Experiments 2 and 3) suggests that fewer than 8% of participants behaved in a manner consistent with this hypothesis. Nevertheless, it is possible that some participants felt compelled to sum their ratings to 100 (or some nearby value) and that this subjective constraint resulted in the translation of biased *learning* of the association between the predictive cues (E-H) and previously predictable outcome (*p1, p2*) to biased *responding* for cue Y (in the reversed direction).

One way to test this account is to examine the extent to which individuals constrained their responding in this manner, and the degree to which they showed biased test ratings for cue Y. Specifically, if this subjective constraint produced the higher ratings for the unpredictable outcome values (*u1, u2*) for cue Y, then one might expect that those participants who constrained their outcome ratings to sum to 100 (or thereabouts) would be

more likely to show greater bias in test ratings for cue Y at test. This hypothesis was tested using the data from Experiment 1. A measure of biased responding for cue Y for each individual (in the high performer group) was calculated by subtracting test ratings for the predictable outcomes *p1* and *p2* from their test ratings for outcomes *u1* and *u2*. This biased learning score was correlated with a measure of the degree to which participants constrained their response ratings to sum to 100. This latter measure of constrained responding was calculated by summing each participant's ratings for the three outcome values on each outcome dimension (e.g. p1+p2+ pØ or u1+u2+uØ) and for each cue (E-Y); yielding ten summed ratings scores. These summed scores were then averaged to yield a mean summed rating score (averaged across all cues and outcome dimensions). Finally, a 'deviance' score was calculated by taking the unsigned discrepancy between this mean summed ratings score and one hundred. Those participants that tended to constrain their test ratings such that they summed to 100 would have a deviancy score near 0, whereas those that gave independent test ratings for each of the outcome values would have higher discrepancy scores. This mean discrepancy score (M=20.13, SEM=5.96) was not significantly correlated with biased responding to cue Y, r=.06, $t < 1$. This result suggests that the tendency to associate cue Y with the unpredictable outcome values (*u1*, *u2*) was not associated with the tendency of some individuals to constrain their test ratings of individual outcome values to sum to 100.

Inhibition. A second possibility focusses on *uØ* and *pØ* rather than the presence of these outcomes. Thus, in phase 1, the context might become more strongly associated with *uØ* than with *pØ* (perhaps because X is a perfect predictor of *pØ* and blocks context-*pØ* learning). As a consequence, in phase 2, cue Y may enter into an inhibitory relationship with predicted (by the context), but never presented, *uØ*. An inhibitory Y-*uØ* relationship (Y predicts that the absence of "u" outcomes will not be observed) should be expressed on test as

a belief that Y predicts *u1/u2*. Future studies that use the present paradigm could directly test this prediction by testing cue Y's inhibitory properties via a summation test.

Differential overshadowing. The third possible account, which additionally assumes differential overshadowing, is based on the rapid learning of associations between predictive cues (e.g. E) and previously predictable outcome values (e.g. *p1*) at the beginning of phase two. These relationships (e.g. E-*p1*), may then differentially overshadow the non-predictive cue Y with respect to the two outcome dimensions. That is, cue E may overshadow cue Y more strongly on the previously predictable outcome dimension (*p1, p2*), than on the previously unpredictable outcome dimension (*u1, u2*). This would result in stronger associations between cue Y and previously unpredictable outcome values (e.g. *u2)* than between cue Y and previously predictable outcome values (e.g. *p1*); the observed result. The limitation of this account is that it can only explain the data if the influence of differential overshadowing (which impairs *Y-p1/p1* learning) was larger in magnitude than the effect of context blocking (which impairs *Y-u1/u2* learning). While possible, this ordering of magnitudes seems unlikely given that, under this account, the differential overshadowing (which needs to be the largest effect) is the product of differential context blocking (which needs to be the smallest effect).

An alternative explanation for the bias in learning for predictive cues E-H with previously predictable outcomes is that learned predictability has a more direct impact on associability, perhaps via outcome salience. This possibility, discussed below. When combined with differential overshadowing, it offers a complete explanation of the present data.

Predictability as a stimulus feature

The final explanation of the present data suggests that participants encoded the predictability of the outcomes during phase one, and used this information to guide subsequent learning

involving those outcomes. As noted in the Introduction, there is some evidence that rats are sensitive to the predictability of the US, and that this influences subsequent learning in a manner analogous to 'learned irrelevance' effects (e.g. Baker, 1979). Specifically, the observation that US preexposure effects occur across a change in context (but are removed by the provision of an overshadowing cue), was viewed by Baker et al (1981) as indicative that rats encode information about the unpredictability of a US and that this information impairs the subsequent formation of associations between that US and a discrete cue. One possibility is that predictable outcomes are higher in salience than unpredictable outcomes.

Such an explanation accords well with the present observation of weaker associations between previously unpredictable outcomes, than previously predictable outcomes, and novel predictive cues. It does not, however, explain how the previously unpredictable outcome was more readily associated with a novel, non-predictive cue (Y). Differential overshadowing provides one explanation. If cues E-H were rapidly associated with the perhaps more salient outcome values *p1* and *p2*, they may then have more readily overshadowed cue Y with respect to the predictable outcome dimension (values *p1*, *p2*, *pØ*), than to the unpredictable outcome dimension (values *u1*, *u2*, *uØ*). This would lead to a stronger association between cue Y and values *u1* and *u2*, than between cue Y and values *p1* and *p2*, as observed.

A second possibility is that, due to a bias for associations involving previously predictable outcomes, people learned the associations between the predictive cues and the previously predictable outcome first. Then, after these associations were learned, they sought to learn how to predict the previously unpredictable outcome. The question then becomes why the unpredictable outcome values were not also associated with the predictive cues. We speculate that this is due to inferences participants made during either training or test concerning the relationship between the cues and outcomes. We discuss two possible inferences below, but before doing so, it is worth noting that the 'nil' ratings in all three

experiments provide evidence that participants used inferential processes, as opposed to mere recollection of contingencies, when forming their test ratings. The 'nil' response values (pØ, uØ) were never shown in phase two (and thus never in conjunction with cues E-H and Y), and yet people gave high ratings to these outcomes values for some cues and not for others. This alone implies that participants are, at least partly, basing their test ratings on inferences drawn during either training or test, rather than reporting experienced contingencies.

One possibility is that people noticed that the outcome dimensions were independent (for instance, the presence of *p1* tells the learner nothing about the presence of *u2*). If they made the assumption that two independent outcomes were unlikely to be generated by the same cue then this would lead them to preferentially associate the previously unpredictable outcome with the only other cue available, the non-predictive cue Y. This kind of inference constitutes a 'Markov violation' (see Mayrhofer, Goodman, Waldmann & Tenenbaum, 2008 for a discussion) that has been shown previously in human reasoning (e.g. Rehder & Burnett, 2005). A related idea is that people might generally favour a 'one cause to one effect' causal mapping whereby cues are individually associated with outcomes (similar to the 'sparse and strong' generic assumptions posited by Lu, Yuille, Liljeholm, Cheng & Holyoak, 2008). The assumptions that participants do not readily associate independent outcomes with the same cue, or that people generally prefer individual cue-outcome mappings, remain to be empirically tested in this setting. Nonetheless, it is possible that people encoded the predictability of the outcomes in phase one, and more readily associated the previously predictable outcome with novel cues in phase two. Then, on the basis of an additional assumption (perhaps due to the statistical independence of the outcomes, or due to generic prior assumptions), they preferentially associated the previously unpredictable outcome with the only other cue available, the non-predictive cue Y.

Conclusion

The current experiments are the first to demonstrate that an outcome's prior reinforcement history influences the degree to which people subsequently associate that outcome with novel cues. Previously predictable outcomes were more readily associated with predictive cues than were previously unpredictable outcomes. The opposite bias was seen for non-predictive cues. This finding dovetails with prior observations that previously predictive cues are more readily associated with novel outcomes, than are previously nonpredictive cues. The means by which exposure to an unpredictable outcome affects subsequent learning remain unclear. Some accounts suggest that outcome unpredictability directly affects learning (via participants encoding stimulus predictability and using this information to guide subsequent learning), while others suggest that effect is indirect (via fostering differential contextual learning or acquired equivalence). Further work is required to discriminate between the possible mechanisms that may produce this effect.

References

Baker, A. G. (1979) Preexposure to the CS alone, US alone, or CS and US uncorrelated: Latent inhibition, blocking by context or learned irrelevance? *Learning and Motivation, 10,* 278-294.

Baker, A. G. & Mackintosh, N. J. (1976) Learned irrelevance and learned helplessness: Rats learn that stimuli, reinforcers, and responses are uncorrelated. *Journal of Experimental Psychology: Animal Behavior Processes, 2*, 130–141

Baker, A. G. & Mackintosh, N. J. (1977) Excitatory and inhibitory conditioning following uncorrelated presentations of the CS and UCS. *Journal of Experimental Psychology: Animal Learning and Behavior, 5,* 315–319

Baker, A. G. & Mackintosh, N. J. (1979) Preexposure to the CS alone, US alone, or CS and US uncorrelated: Latent inhibition, blocking by context or learned irrelevance? *Learning and Motivation, 10,* 278-294.

Baker, A. G., Mercier, M., Gabel, J. & Baker, R. A. (1981) Contextual conditioning and the US preexposure effect in conditioned fear. *Journal of Experimental Psychology: Animal Behavior Processes, 7,* 109-128.

Beesley T., Le Pelley M.E. (2010). The effect of predictive history on the learning of sub-sequence contingencies. *Quarterly Journal of Experimental Psychology, 63,* 108-135.

Bonardi, C., Graham, S., Hall, G. and Mitchell, C., 2005. Acquired distinctiveness and equivalence in human discrimination learning: evidence for an attentional process. *Psychonomic Bulletin & Review, 12,* 88-92

Bonardi, C. & Ong, S. Y. (2003) Learned irrelevance: A contemporary review. *The Quarterly Journal of Experimental Psychology B: Comparative and Physiological Psychology*, 56, 80-89.

Dickinson, A. & Mackintosh, N. J. (1979) Reinforcer specificity in the enhancement of conditioning by posttrial surprise. *Journal of Experimental Psychology: Animal Behavior Processes, 5,* 162-177.

Griffiths, O. & Le Pelley, M.E. (2009). Attentional changes in blocking are not a consequence of lateral inhibition. *Learning and Behavior, 37,* 27-41.

Gunther, L. M. Miller, R. R. & Matute, H. (1997) CSs and USs: What's the difference? . *Journal of Experimental Psychology: Animal Behavior Processes, 23,* 15-30

Hall, G., Mitchell, C. J., Graham, S. & Lavis, Y. (2003) Acquired equivalence and distinctiveness in human discrimination learning: Evidence for associative mediation. *Journal of Experimental Psychology: General, 132,* 266-276.

Hall, G., Ray, E. & Bonardi, C. (1993) Acquired equivalence between cues trained with a common antecedent. *Journal of Experimental Psychology: Animal Behavior Processes, 19,* 391-399.

Kamin, L. J. (1961) Apparent adaptation effects in the acquisition of a conditioned emotional response. *Canadian Journal of Psychology, 15,* 176-188.

Kremer, E. F. (1971). Truly random and traditional control procedures in CER conditioning in the rat. *Journal of Comparative and Physiological Psychology,* 76, 441–445.

Kruschke, J. K. & Blair, N. J. (2000) Blocking and backward blocking involve learned inattention. *Psychonomic Bulletin & Review, 7,* 636-645.

Kruschke J. K., Kappenman E. S., Hetrick W. P.(2005). Eye gaze and individual differences consistent with learned attention in associative blocking and highlighting. *Journal of Experimental Psychology: Learning, Memory, and Cognition, 31,* 830–845.

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

Le Pelley, M. E., & McLaren, I. P. L. (2003). Learned associability and associative change in human causal learning. *Quarterly Journal of Experimental Psychology, 56B*, 68-79

Le Pelley M.E., Wills A.J., Oakeshott S.M., McLaren I.P.L. (2005). The outcome specificity of learned predictiveness effects: Parallels between human causal learning and animal conditioning. *Journal of Experimental Psychology: Animal Behavior Processes, 31*, 226-236.

Le Pelley M.E., Beesley T., Suret M.B. (2007). Blocking of human causal learning involves learned changes in stimulus processing. *Quarterly Journal of Experimental Psychology, 60,* 1468-1476.

Le Pelley M.E., Schmidt-Hansen M., Harris N.J., Lunter C.M., Morris C.S. (2010). Disentangling the attentional deficit in schizophrenia: Pointers from schizotypy. *Psychiatry Research, 176*, 143- 149.

Le Pelley M.E., Beesley T. & Griffiths O. (2011). Overt Attention and Predictiveness in Human Contingency Learning. *Journal of Experimental Psychology: Animal Behavior Processes, 37*, 220-229.

Le Pelley, M. E., Mitchell, C. J., & Johnson, A. M. (2013). Outcome value influences attentional biases in human associative learning: Dissociable effects of training and instruction. *Journal of Experimental Psychology: Animal Behavior Processes, 39*, 39-55.

Lochmann, T. & Wills, A.J. (2003). Predictive history in an allergy prediction task. In Schmalhofer, F.,Young, R. M. & Katz, G. (Eds.) *Proceedings of EuroCogSci 03: The European Cognitive Science Congerence*. Mahwah, New Jersey: Lawrence Erlbaum Associates. pp. 217-222.

Lu, H., Yuille, A., Liljeholm, M., Cheng, P.W., & Holyoak, K.J. (2008). Bayesian generic priors for causal learning. *Psychological Review*, 115, 955-984.

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,* 415-9.

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

Mackintosh, N.J. & Turner, C. (1971) Blocking as a function of novelty of CS and predictability of UCS. *The Quarterly Journal of Experimental Psychology, 23*, 359-366.

Matzel, L. D., Schachtman, T. R. & Miller, R. R. (1988) Learned irrelevance exceeds the sum of CSpreexposure and US-preexposure deficits. *Journal of Experimental Psychology: Animal Behavior Processes, 14,* 311-319.

Mayrhofer, R., Goodman, N. D., Waldmann, M. R., & Tenenbaum, J. B. (2008). Structured correlation from the causal background. In *Proceedings of the Thirtieth Annual Conference of the Cognitive Science Society* (pp. 303-308)

Miller, N. E. & Dollard, J. (1941) *Social learning and imitation*. New Haven, CT: Yale University Press.

Mitchell C.J., Griffiths, O., Seetoo, J., Lovibond, P.F. (2012). Attentional mechanisms in learned predictiveness. *Journal of Experimental Psychology: Animal Behavior Processes, 38,* 191-202.

Morris, R., Griffiths, O., Le Pelley, M. E., & Weickert, T. W. (2013). Attention to irrelevant cues is related to positive symptoms in schizophrenia. *Schizophrenia Bulletin, 39,* 575-582.

Overmier, J. B., & Wielkiewicz, R. M. (1983). On unpredictability as a causal factor in "learned helplessness." *Learning and Motivation, 14*, 324–337.

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

Randich, A. & LoLordo, V. M. (1979) Associative and nonassociative theories of the UCS preexposure phenomenon: Implications for Pavlovian conditioning. *Psychological Bulletin, 86,* 523-548.

Rehder, B., & Burnett, R. (2005). Feature inference and the causal structure of categories. *Cognitive Psychology, 50*, 264-314.

Maier, S. F. & Seligman, M. (1976) Learned helplessness: Theory and evidence. *Journal of Experimental Psychology: General, 105,* 3-46.

Smyth, S., Barnes-Holmes, D., & Barnes-Holmes, Y. (2008). Acquired equivalence in human discrimination learning: The role of propositional knowledge. *Journal of Experimental Psychology: Animal Behavior Processes, 34*, 167-177.

Taylor, J. A. (1956) Level of conditioning and intensity of the adaptation stimulus. *Journal of Experimental Psychology*, 51, 127-130.

Author Note. This research was supported by an Australian Research Council Discovery project grant awarded to the first and fourth authors, #DP0774395. The authors acknowledge Prof. Ian McLaren for comments guiding the interpretation of this data.

Table 1. Design of Experiments 1 to 3.

Phase 0 (Experiment 3 only)	Phase 1	Phase 2	Test
A – p1 , uØ	A – p1 , uØ	EY – p1 , u2	Test E-Y for:
B – p2, uØ	B – p2, uØ	FY – p2, u1	pØ, p1, p2,
X – pØ , u1	X – pØ , u1	GY – p1 , u1	uØ, u1, u2
X – pØ , u2	X – pØ , u2	HY – p2 , u2	
	AX – p1 , u1		
	AX – p1 , u2		
	BX – p2, u1		
	BX – p2, u2		

Note: Letters A-Y denote foods, and the set of symbols [pØ, p1, p2, uØ, u1, u2] denote allergic reactions (or outcomes). The letter ‘p’ denotes values on the predictable outcome dimension, whereas the letter ‘u’ denotes values on the unpredictable outcome dimension. For example, if skin reactions were the predictable dimension, p1 would refer to ‘itchiness’ and p2 would refer to ‘swelling.’ The symbol Ø refers to the absence of an outcome on that dimension. For instance, if skin is the predictable dimension, then pØ refers to ‘no skin reaction.’ Equivalently, for the unpredictable outcome dimension, u1 might refer to ‘cramping,’ ‘u2’ to bloating and uØ to ‘no stomach reaction.’ The contingencies shown in the left-hand column (labelled Phase 0) were only presented in Experiment 3.

Figure 1.

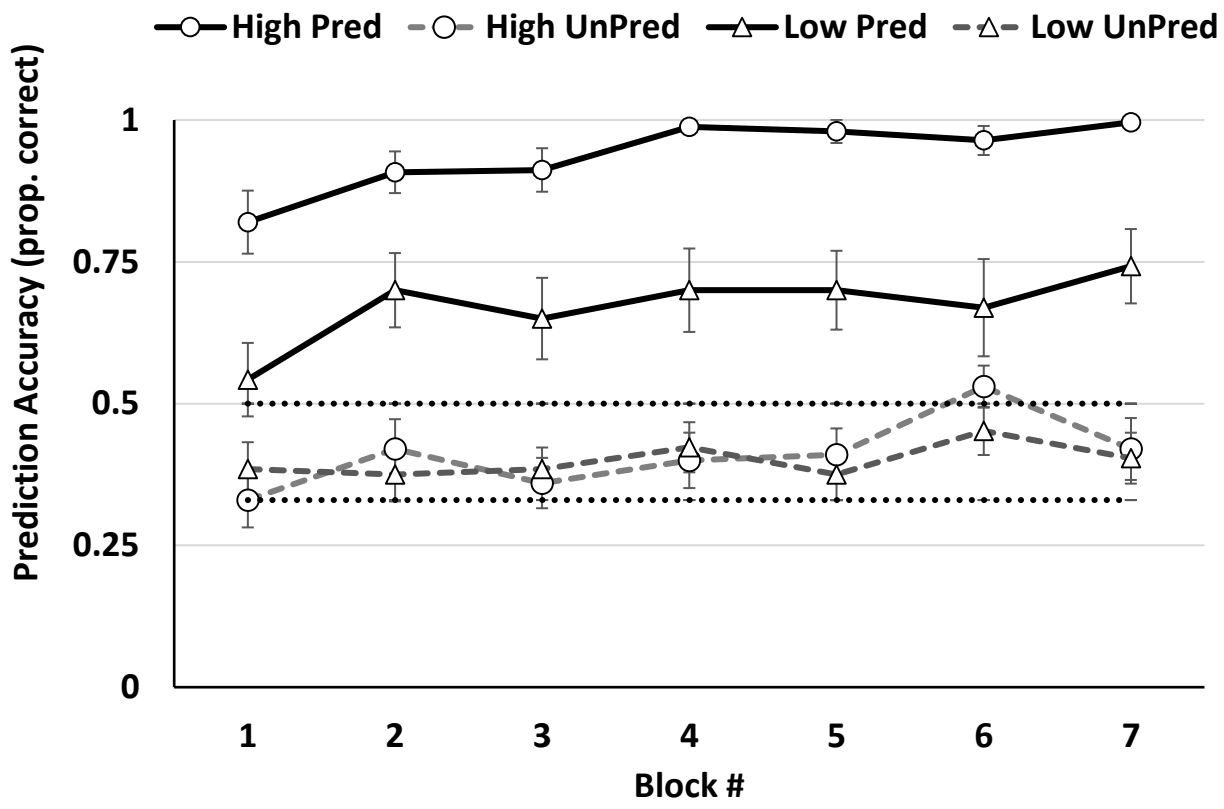

Figure 1. Mean prediction accuracy (proportion of correct outcome predictions) in Phase 1of Experiment 1 for the low- and high-performing groups. Prediction accuracy for the high performing group is indicated by the open circles, whereas the low performing group is indicated by open triangles. Prediction accuracy for outcomes that were possible to predict on each trial (labelled 'Pred', unbroken lines) are shown separately to those outcomes which were not possible to predict (labelled 'Unpred', broken lines) for both groups of participants. Each block of training included 16 trials (2 repetitions of each trial type). The dotted horizontal lines indicate 50% accuracy (the maximum achievable for the unpredictable outcomes values) and 33% accuracy (chance performance, assuming no knowledge of the predictable or unpredictable outcomes values). Error bars indicate SEM.

Figure 2.

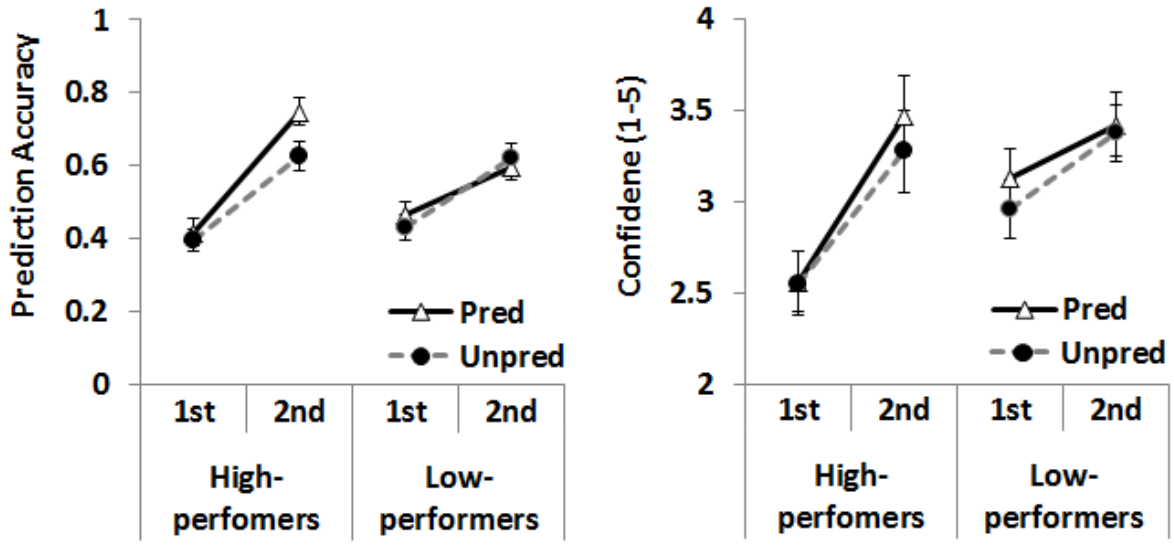

Figure 2. Mean prediction accuracy (left-hand panel) and confidence ratings (right-hand panel) in Phase 2 for the low- and high-performing groups. Prediction accuracy and mean confidence for the previously predictable outcomes are indicated by the empty triangles and solid lines. Prediction accuracy and mean confidence for the previously unpredictable outcomes is indicated by the filled circles and dashed lines. Error bars indicate SEM.

Figure 3.

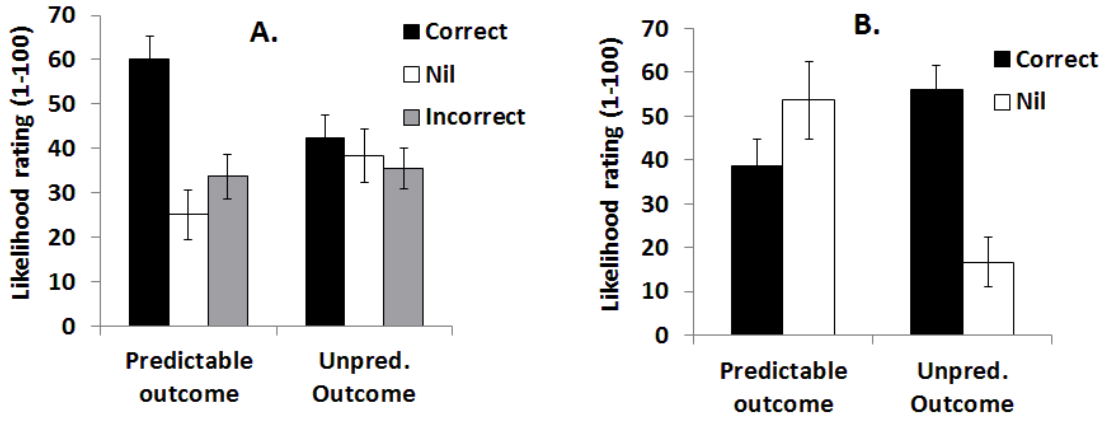

Figure 3. Mean likelihood ratings of the high-performing group at test in Experiment 1. Panel A refers to ratings following presentation of the predictive cues E-H, and panel B refers to ratings following the non-predictive cue Y. In both panels, filled black columns indicate ratings given for the 'correct' response (see text for definitions) and unfilled columns indicate ratings for the 'nil' outcome response. The additional grey columns in the left-hand panel refer to mean ratings given to the 'incorrect' outcome value. Error bars indicated SEM.

Figure 4.

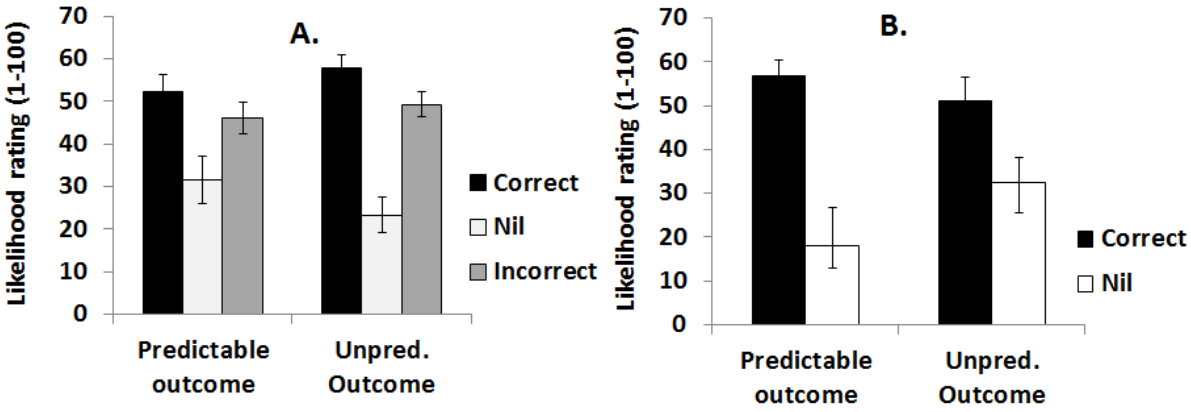

Figure 4. Mean likelihood ratings of the low-performing group at test in Experiment 1. Panel A refers to ratings following presentation of the predictive cues E-H, and panel B refers to ratings following the non-predictive cue Y. In both panels, filled black columns indicate ratings given for the 'correct' response (see text for definitions) and unfilled columns indicate ratings for the 'nil' outcome response. The additional grey columns in the left-hand panel refer to mean ratings given to the 'incorrect' outcome value. Error bars indicated SEM.

Figure 5.

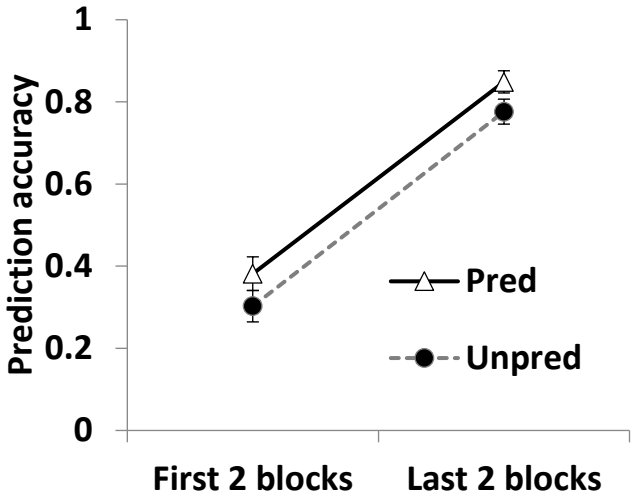

Figure 5. Prediction accuracy in Phase 2 of Experiment 2. Prediction accuracy for the previously predictable outcomes is indicated by the empty triangles, and prediction accuracy for the previously unpredictable outcomes is indicated by the filled circles. Error bars indicated SEM.

Figure 6.

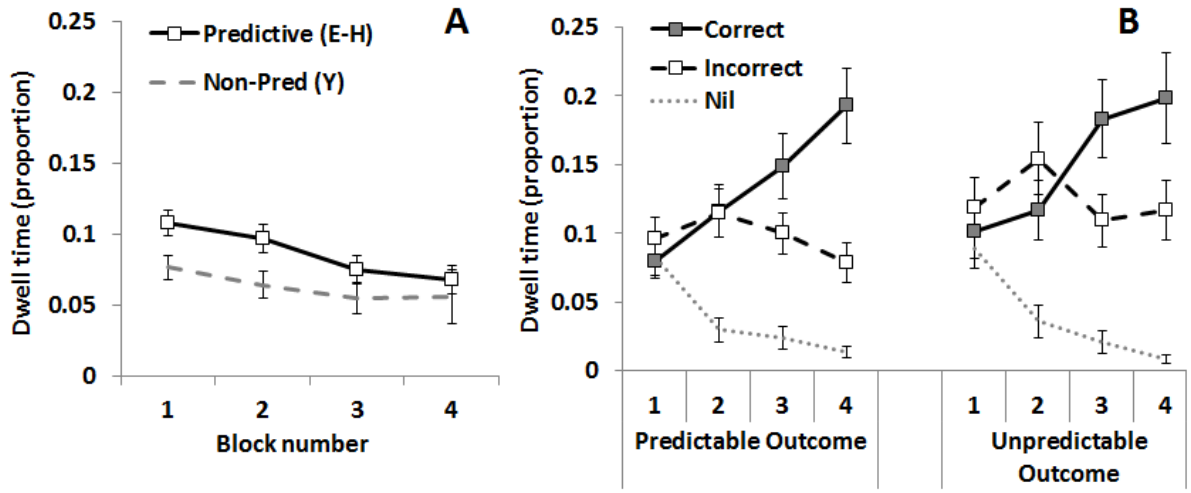

Figure 6. Mean dwell times for the cues (Panel A) and outcomes (Panel B) in Phase 2 of Experiment 2. For Panel A, the unbroken black line with unfilled squares refers to dwell times for the predictive cues E-H and the broken grey line refers to dwell times for the non-predictive cue Y. In Panel B, the solid lines with filled squares refer to dwell times for correct outcome values, the broken black lines with unfilled squares refers to dwell times for incorrect outcome values, and the dotted grey line refers to dwell times for the nil outcome values. Error bars indicate SEM.

Figure 7.

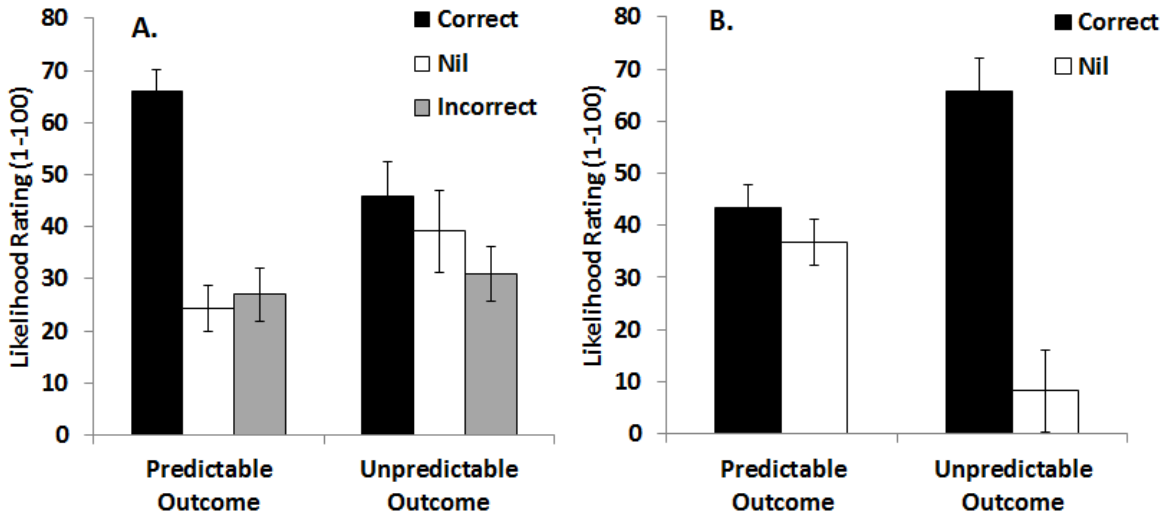

Figure 7. Mean likelihood ratings in Experiment 2 for the predictive cues E-H (Panel A) and for the non-predictive cue Y (Panel B). For both panels, the black filled columns indicate ratings for the correct outcome values, whereas the unfilled columns indicate ratings for the nil outcome values. For the left-hand panel only, the grey columns indicate mean ratings for the incorrect outcome value (there was no "incorrect" value for the non-predictive cue Y shown in the right-hand panel). Error bars indicated SEM.

Figure 8.

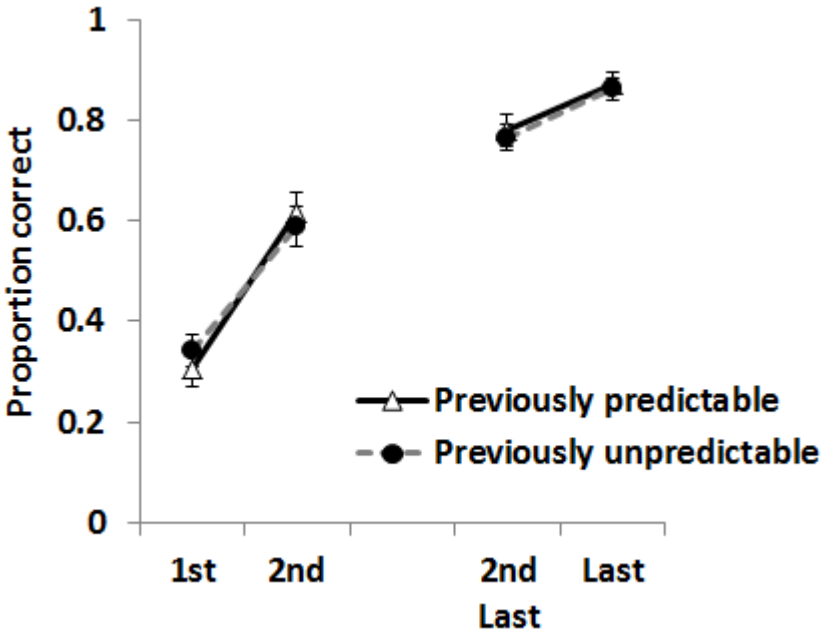

Figure 8. Prediction accuracy in Phase 2 of Experiment 3. Prediction accuracy for the previously predictable outcomes is indicated by the empty triangles, and prediction accuracy for the previously unpredictable outcomes is indicated by the filled circles. Error bars indicated SEM.

Figure 9.

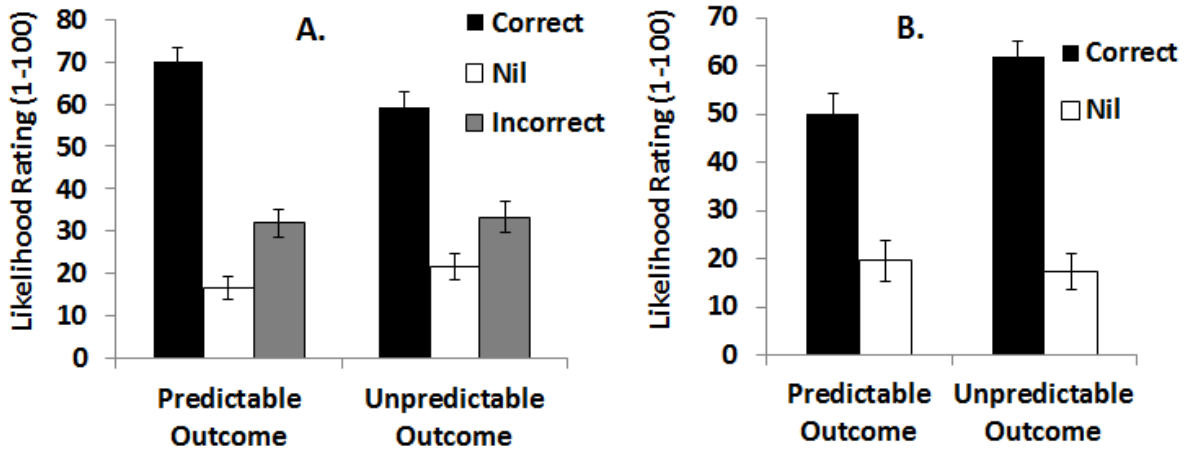

Figure 9. Mean likelihood ratings in Experiment 3 for the predictive cues E-H (Panel A) and for the non-predictive cue Y (Panel B). For both panels, the black filled columns indicate ratings for the correct outcome values, whereas the unfilled columns indicate ratings for the nil outcome values. For the left-hand panel only, the grey columns indicate mean ratings for the incorrect outcome value (there was no "incorrect" value for the non-predictive cue Y shown in the right-hand panel). Error bars indicated SEM.