Prediction error and overt attention to relevant and irrelevant cues:

Evidence for an interaction of two associability mechanisms

Dissertation
zur Erlangung des Doktorgrades
der Naturwissenschaften
(Dr. rer. nat.)

dem
Fachbereich Psychologie
der Philipps-Universität Marburg

vorgelegt von
David Torrents Rodas
aus La Torre d'Oristà, Katalonien, Spanien

Marburg, 2020

Vom Fachbereich Psychologie der Philipps-Universität Marburg (Hochschulkennziffer 1180) als Dissertation angenommen am ___.___.______

Erstgutachter: Prof. Dr. Harald Lachnit

Zweitgutachter: Prof. Dr. Mike Le Pelley

Tag der Disputation: ___.___.______

Acknowledgements

I want to thank the following people for contributing to this thesis, supporting me, or both.

Prof. Harald Lachnit gave me a great opportunity and has set the optimal conditions for learning. I am indebted with him, Dr. Stephan König, and Dr. Metin Üngör for sharing their knowledge and guiding me. Now I am certainly confused, but on a higher level.

Dr. Anna Thorwart has provided answers to all sorts of spontaneous questions. She and my other colleagues in AG Assoziatives Lernen, Martina Brähler-Haupt, Midhula Chandran, and Genisius Hartanto, have always been very friendly.

Inga Grössle, Hannah den Hartog, Tamara Herbst, Vanessa Neubauer, Carolin Schieferstein, Melanie Schwan, David Straub, Emily Ufken, and Laura Weber helped me with data collection.

My second supervisor, Prof. Winfried Rief, and my international mentor, Dr. Mike Le Pelley, kindly agreed to take part in the evaluation of this thesis.

The members of the RTG 2271 shared this journey with me. One of them even has the extra merit to stop by every day to have a nice little chat.

It is only fair to thank those who supported me “outside the academia”. Andrea, Dušica, and Golrang, who I was lucky to meet in Marburg. I a les persones “de casa”, totes elles estimades (en diferents graus i formes): la Cristina Martí, la Cristina Rabasa, l’Ikram, el Josep, la Laia, la Laura i la meva fillola Aina, la Núria i la Regina.

Finally, I want to acknowledge my grandmother, la Iaia, and her unconditional love.

Contents

Summary

Humans and other animals use cues in the environment to make predictions about important outcomes, thus preparing themselves to respond to those events. Prediction error refers to the extent to which an outcome is surprising in the presence of one or more cues. Within the research area of associative learning, some theories suggest that prediction error changes the amount of attention paid to cues. It was initially proposed that the attentional changes were driven by either relative or overall prediction error. In the first case, attention increases for cues generating less prediction error than other concurrent cues, otherwise attention decreases (Mackintosh, 1975). In the second case, the amount of attention paid to each cue is directly related to overall prediction error, i.e. how surprising the outcome is considering all the present cues (Pearce & Hall, 1980). Evidence for the role of relative prediction error comes from studies with pairs of cues including a component relevant to outcome prediction, together with an irrelevant component. Evidence for a role of overall prediction error comes from studies in which cues generating different amounts of prediction error are trained separately. Given that considering both relative and overall prediction error may account for a wider range of attentional changes, the two mechanisms were incorporated into hybrid models (e.g., Le Pelley, 2004). However, evidence for those models in humans is still scarce. The aim of the present thesis was to study the effect of a sudden rise in overall prediction error on overt attention to cues that were either relevant or irrelevant to outcome prediction, i.e. differing in terms of relative prediction error. Rather than considering sustained levels of prediction error, we focused primarily on sudden changes, because they are involved in important behavioral phenomena, such as the return of pathological anxiety. Each of the two studies included in the thesis started with a discrimination training, in which participants had to predict the occurrence of two possible outcomes. Participants' eye gaze showed that the relevant cues received more attention than the irrelevant cues. In a second stage, contingency reversal (Study I) or partial reinforcement (Study II) led to a rise in prediction error, as indicated by a drop in the accuracy of outcome predictions. The attentional preference for the relevant cues was temporarily weakened by contingency reversal, and it was completely lost following the introduction of partial reinforcement. In addition, both manipulations increased the amount of attention paid to both types of cues. The data were consistent with a combined effect of relative and overall prediction error, thus providing evidence for the hybrid models. In addition, the results have implications for understanding changes in attention to contextual cues.

Zusammenfassung

Menschen und andere Tiere nutzen Hinweisreize aus der Umwelt, um Vorhersagen über wichtige Ereignisse zu treffen, wodurch sie sich auf diese Ereignisse vorbereiten können. Der Vorhersagefehler bezeichnet das Ausmaß, in dem ein Ereignis bei Vorhandensein von einem oder mehreren Hinweisreizen überraschend ist. Einige Theorien aus dem Bereich des Assoziativen Lernens nehmen an, dass Vorhersagefehler das Ausmaß an Aufmerksamkeit auf Hinweisreize verändern. Ursprünglich wurde postuliert, dass Aufmerksamkeitsveränderungen entweder durch relative oder globale Vorhersagefehler gesteuert werden. Im ersten Fall steigt die Aufmerksamkeit auf Hinweisreize, die einen geringeren Vorhersagefehler als andere, gleichzeitig präsentierte Hinweisreize erzeugen – andernfalls sinkt die Aufmerksamkeit (Mackintosh, 1975). Im zweiten Fall steht die Aufmerksamkeit auf einen Hinweisreiz in direktem Zusammenhang mit dem globalen Vorhersagefehler, d. h. wie Überraschend ein Ereignis vor dem Hintergrund aller anwesenden Hinweisreize ist (Pearce & Hall, 1980). Belege für die Rolle des relativen Vorhersagefehlers stammen aus Studien, bei denen Paare von Hinweisreizen verwendet wurden, die sich jeweils aus einer für die Vorhersage relevanten Komponente und einer irrelevanten Komponente zusammensetzen. Belege für einen globalen Vorhersagefehler kommen aus Studien, bei denen Hinweisreize, die sich in der Größe des Vorhersagefehlers unterscheiden, separat trainiert werden. Da relative und globale Vorhersagefehler ein breites Spektrum an Aufmerksamkeitsveränderungen erklären, wurden beide Mechanismen in hybride Modelle integriert (z.B., Le Pelley, 2004). Allerdings ist die Befundlage im Humanbereich für diese Modelle noch gering. Das Ziel der vorliegenden Arbeit war es, den Einfluss eines plötzlichen Anstiegs des globalen Vorhersagefehlers auf die offene Aufmerksamkeit zu untersuchen, die sich auf Hinweisreize richtet, die für eine Vorhersage entweder relevant oder irrelevant waren, d. h. sich in Bezug auf den relativen Vorhersagefehler unterschieden. Wir konzentrierten uns auf plötzliche Änderungen des Vorhersagefehlers anstelle anhaltender Fehlerniveaus, da diese bei wichtigen Verhaltensphänomenen eine Rolle spielen, wie z. B. der Rückkehr pathologischer Angst. In zwei Studien erhielten Probanden zunächst ein Diskriminationstraining, bei dem das Auftreten von zwei möglichen Ereignissen vorhergesagt werden musste. Hierbei ergaben Blickbewegungsmessungen, dass relevante Komponenten mehr Aufmerksamkeit gewidmet wurde als irrelevanten Komponenten. In einer zweiten Phase führte ein Kontingenzwechsel (Studie I) oder partielle Verstärkung (Studie II) zu einem Anstieg des Vorhersagefehlers, welches sich in einer Abnahme der Vorhersagegenauigkeit widerspiegelte. Die Aufmerksamkeitspräferenz für relevante Reize wurde durch den Kontingenzwechsel vorübergehend geschwächt und ging bei partieller Verstärkung vollständig verloren. Zudem führten beide Manipulationen zu einer Erhöhung der Aufmerksamkeit auf relevante und irrelevante Reize. Die vorliegenden Daten sprechen für einen kombinierten Einfluss des relativen und globalen Vorhersagefehlers und liefern somit Evidenz für die hybride Modelle. Auch lassen sich aus den vorliegenden Ergebnissen Implikationen für unser Verständnis von Aufmerksamkeitsveränderungen bei kontextuellen Reize ableiten.

Proposed manuscripts

I. Torrents-Rodas, D., Koenig, S., Uengoer, M., & Lachnit, H. (2020). A rise in prediction error increases attention to irrelevant cues. *Biological Psychology*. Advance online publication. https://doi.org/10.1016/j.biopsycho.2020.108007

II. Torrents-Rodas, D., Koenig, S., Uengoer, M., & Lachnit, H. (2020). Evidence for an interaction of two attentional mechanisms during learning. Manuscript under review in *Quarterly Journal of Experimental Psychology*.

Introduction[1]

In order to function and thrive, humans and other animals must internally represent relationships between events in their environment. Specifically, they need to use sensory cues to make predictions about important outcomes, so that they can prepare themselves to respond to those events. As natural and social settings change, those predictions have to be updated according to experience (see Rief et al., 2015). Theories of associative learning describe this process as the development of associations between neural representations of cues and outcomes (Pearce & Bouton, 2001; Rescorla, 1988). An association grows when an outcome holds some surprise in the presence of one or more cues. Once the outcome is correctly predicted by any of the available cues, there is no reason for further adjusting the association. The extent to which an outcome is surprising on a given occasion can be quantified by the prediction error term:

$$\lambda - \Sigma V$$

where λ represents the magnitude of the outcome, e.g. set at 1, and determines the point at which the cues and the outcome become fully associated. ΣV represents the sum of strengths of the preexisting associations between all the present cues and the outcome. Before any of the cues is experienced together with the outcome for the first time, ΣV equals 0. This means that none of the cues predicts the outcome. The first co-occurrence entails a large prediction error, which will serve to strengthen the associations, thus making the outcome less surprising on the next occasion. Usually, learning is not completed after just one co-occurrence, because properties of the cues and the outcome also modulate the associative growth, yet the increase is in proportion to prediction error (Rescorla & Wagner, 1975; Schultz & Dickinson, 2000). Most behavior is assumed to be based on such associations. Organisms initiate responses appropriate to a given outcome when they experience a cue associated with it (Domjan, 2005).

Given the myriad of objects and beings present in most environments, organisms need to focus on those stimuli that are relevant to ongoing behavior. This task is carried out by a set of cognitive processes termed ‘attention’. The organism’s behavioral goals often interact with stimulus properties to determine

[1] The two studies of the present thesis originated from the Research Training Group (RTG) 2271 “Maintenance vs. change of expectations in the context of expectation violations”, funded by the German Research Foundation (Deutsche Forschungsgemeinschaft DFG), to explore why and when expectations are maintained or modified in the face of conflicting evidence. They were designed to contribute some answers to these questions by taking advantage of theories, concepts, and methods of the areas of associative learning and attention. In the following, expectations and expectation violations are reframed as predictions based on associations and prediction errors, respectively. Maintenance or change are thought of as depending, at least in part, on attentional processes, which by themselves depend on (different kinds of) prediction errors.

which cues are selected for further processing (Desimone & Duncan, 1995). Researchers in the associative tradition have been primarily interested in attention to the extent that it guides learning. The attentional properties of each cue involved in learning are quantified by an associability parameter, α, which modulates changes in associative strength driven by prediction error (e.g., Rescorla & Wagner, 1975). The higher the value of α, the faster the cue is associated with the outcome. Moreover, the associative change on a given trial is distributed among the cues in proportion to their α values. Thus, the speed of learning can be used to infer the extent to which a cue engages attention. In addition to the fixed effect of associability on learning, evidence from the fields of associative learning and cognitive psychology has shown that learning itself can change the attentional weight assigned to cues (for reviews, see Anderson, 2016; Awh, Belopolsky & Theeuwes, 2002; Gottlieb, 2012; Le Pelley et al., 2016). In particular, attentional theories of associative learning propose that α is modified by prediction error, with this variation playing a central role in important learning phenomena, such as competition between cues for associative strength (Holland & Maddux, 2010; Le Pelley, 2004; Pearce & Mackintosh, 2010).

The two most influential attentional theories of associative learning use completely different approaches to specify how the mechanism for changing α works. Mackintosh's (1975) theory suggests that attention increases for a cue generating less prediction error than other concurrent cues, otherwise attention decreases. On a given trial, the change in associability for each cue, e.g. S, is determined by comparing its prediction error with that of the remaining cues, T–S:

$$\alpha_{\mathrm{S}} \text{ increases, if } |\lambda - V_{\mathrm{S}}| < |\lambda - V_{\mathrm{T-S}}|$$

$$\alpha_{\mathrm{S}} \text{ decreases, if } |\lambda - V_{\mathrm{S}}| \geq |\lambda - V_{\mathrm{T-S}}|$$

Mackintosh's suggestion fits with the idea that people and animals tend to pay attention to those cues enabling correct predictions and ignore unreliable cues. On the other hand, the Pearce-Hall (1980) model proposes that attention is a direct function of overall prediction error, which is based on the prediction of the outcome considering all the cues, ΣV, present on a given trial. Then each of those cues, S, undergoes the same change in associability:

$$\alpha_{\mathrm{S}} = |\lambda - \Sigma V|$$

Bearing in mind that attention is assumed to facilitate learning, the suggestion made by Pearce and Hall has the reasonable implication that attention is maximal when a given cue is followed by a surprising outcome, i.e. before learning has taken place, and attention decreases as the cue comes to signal the outcome accurately.

Evidence providing selective support for Mackintosh's theory mainly comes from discrimination studies with pairs of cues including a component relevant to outcome prediction together with an irrelevant component. In a study conducted by Le Pelley and McLaren (2003), for instance, participants played the role of an allergist, who had to find out which allergic reaction a patient would suffer after eating different meals. Thus, the foods and reactions served as cues and outcomes, respectively. Based on a given pair of foods, on each trial participants selected the expected reaction out of two possibilities, before they received corrective feedback. Across trials, the foods were combined in a way that one food in each pair was consistently followed by a single allergic reaction, and thus was relevant to the prediction, whereas the other food was randomly followed by each of the two reactions, and thus was irrelevant. Under such contingencies, the relevant cues should come to generate less prediction error than the irrelevant cues and, according to Mackintosh's theory, an attentional preference should develop for the former cues. In order to test this prediction, participants received a second series of trials concerning a new patient, who ate the same foods as the previous one but suffered from different allergic reactions. Each pair consisted of a previously relevant food and a previously irrelevant one, but both of them were followed by a single reaction across trials, and thus generated a similar prediction error. A final test showed that the previously relevant foods became more strongly associated with the new allergic reactions than the previously irrelevant foods. This suggested that the attentional biases acquired during initial training transferred to the second series of trials, thus lending support to Mackintosh's theory. For an example in animal research from a related paradigm, see Mackintosh and Little (1969).

Selective evidence for the Pearce-Hall model typically comes from studies in which the cues are individually trained, i.e. presented alone in separate conditions or to different groups of participants. Take for instance a study by Griffiths, Johnson, and Mitchell (2011; based on a study conducted in rats by Hall & Pearce, 1982), who also used an allergist task. Participants were asked to predict the severity of an allergic reaction experienced by a patient each time she ate a particular food. Initially, they learned that the food produced a mild reaction. Then, one half of participants received a trial in which no reaction occurred, while the other half received a trial in which the reaction was once again mild. In the former group, the absence of an allergic reaction should have been surprising, i.e. linked to large prediction error, and thus, according to the Pearce-Hall model, attention to the food should have been reinstated, boosting further learning about the consequences of eating it. This prediction was confirmed on subsequent trials, in which the same patient experienced a much severe reaction. The severity predictions showed that participants who received the trial with no reaction were faster to learn about the change in severity than the others.

In animal research, abundant evidence exists for each of the two theories mentioned above (for reviews, see Le Pelley, 2004; Pearce & Mackintosh, 2010). This motivated the development of hybrid models (George & Pearce, 2012; Le Pelley, 2004; Pearce & Mackintosh, 2010) in which attention to cues is determined by an interaction of the associability mechanisms suggested by Mackintosh (1975) and Pearce and Hall (1980). For instance, by multiplying together the α values obtained according to each mechanism, Le Pelley's (2004) model allows to prioritize a cue generating less prediction error than other concurrent cues and, at the same time, to reduce attention to the whole set of cues as they come to produce a small overall prediction error.

The pertinence of the hybrid models to explain human behavior has not yet been established. Recently, Le Pelley et al. (2016) reviewed the existing evidence in humans for attentional changes driven by associability mechanisms. Similar to animal research, many of the reviewed studies inferred attentional effects from variations in the speed of learning (see, for instance, the above-mentioned study by Griffiths et al., 2011). In addition, some studies took advantage of eye-tracking, a measure of overt visual attention used in related research fields, such as cognitive psychology (e.g., Theeuwes & Belopolsky, 2012). The authors of the review concluded that, while there was ample evidence consistent with the associability mechanism put forward by Mackintosh (1975), support for the mechanism suggested by Pearce and Hall (1980) was scarce. However, they also noted that in most studies the experimental task enabled a full prediction of the outcomes, and thus overall prediction error could be reduced to a minimum across conditions (but see Le Pelley et al., 2010; Kattner, 2015; Livesey, Thorwart, De Fina, & Harris, 2011). This aspect of the design might have prevented attentional changes driven by the Pearce-Hall mechanism, because they typically become evident when conditions generating different amounts of overall prediction error are compared to each other. Thus, the authors noted that "it remains for future empirical work to establish more convincingly whether both of these mechanisms operate in humans" (Le Pelley et al., 2016, p. 1123).

Beesley, Nguyen, Pearson, and Le Pelley (2015) were the first to report human data consistent with an interaction of two associability mechanisms. They used a discrimination task with relevant and irrelevant cues, which were arranged in either deterministic or probabilistic pairs (within-subject or between-subjects manipulation in experiments 1 and 2, respectively). Each deterministic pair was systematically followed by its corresponding outcome, while each probabilistic pair was followed by its primary outcome on 70% of the trials and by another outcome on the remaining 30%. Given those contingencies, participants' outcome predictions were less accurate for the probabilistic pairs than the deterministic pairs. Thus, the former presumably generated larger overall prediction error than the latter. Consistent with the Pearce-Hall mechanism, the eye-tracking data revealed that the time spent looking (dwell time) at the probabilistic pairs was longer than the dwell time on the deterministic pairs. In addition, consistent with the Mackintosh mechanism, the dwell time on the relevant cues was longer than that on the irrelevant cues. Thus, by manipulating both overall and relative prediction error – between and within pairs, respectively, Beesley and colleagues showed changes in overt attention in accordance with the predictions of hybrid models (for more recent demonstrations from the same group, see Easdale, Le Pelley, & Beesley, 2019; Walker, Luque, Le Pelley, & Beesley, 2019).

The aim of the present thesis was to study the effect of a sudden rise in overall prediction error on overt attention to cues that were either relevant or irrelevant to outcome prediction, i.e. differing in terms of relative prediction error. In addition to account for a wider range of attentional changes, consideration of both relative and overall prediction error may be particularly useful for modelling real-life situations, which are characterized by dynamic predictive structures (Gottlieb, 2012; Yu & Dayan, 2005). Moreover, rather than considering sustained levels of prediction error, we focused primarily on sudden changes, because they are involved in important behavioral phenomena. Take, for instance, the presumed role of attention in context-specific learning. Certain situations are characterized by the occurrence of a new outcome that contradicts the originally learned association. Retrieval of the second association often depends on being present in the context where it was learned. To explain such situations, Bouton (1997) proposed that the sudden prediction error produced by the new outcome increases the amount of attention paid to the context. Contexts may be defined as background stimuli that are ignored because they are irrelevant to predict an outcome (Rosas, Callejas Aguilera, Ramos Álvarez, & Fernández Abad, 2006). This definition certainly fits with the irrelevant cues used in the discrimination tasks mentioned above. Therefore, we think that, by studying attentional changes to relevant and irrelevant cues driven by a sudden rise in prediction error, researchers may better understand context-specific learning. In turn, this could have implications for important behavioral phenomena linked to context effects, such as return of pathological anxiety or relapse to drug abuse.

Outline of the present thesis

In the two studies composing the present thesis, participants learned to predict the outcome of each trial by means of abstract cues presented on a screen and the feedback they received after each prediction. Participants earned a monetary reward for each correct guess (0.10 € in Study I, 0.08 € in Study II). The sudden rise in overall prediction error was induced by contingency reversal in Study I and by introducing partial reinforcement in Study II. Accuracy of outcome predictions was taken to measure overall prediction error, with the two being inversely related. Eye-tracking was used to measure changes in overt visual attention to the experimental cues while participants solved the learning task.

Study I: A rise in prediction error increases attention to irrelevant cues

We designed Study I to investigate the effect of a sudden rise in prediction error on the amount of attention paid to irrelevant cues. To this end, we used a discrimination task with contingency reversal. On each trial, participants had to press the correct mouse button (left or right) based on the present pair of cues. Thus, the correct buttons played the role of outcomes. Different cues were combined into pairs. The pair-outcome contingencies were arranged so that, across the trials of Stage 1, one cue in each pair was consistently followed by the same outcome, whereas the other cue was randomly followed by each of the two outcomes. Therefore, the former type of cue was relevant to predict the outcomes while the latter was irrelevant. At the beginning of Stage 2, one half of participants experienced a reversal of the contingencies trained before (reversal group). The other half of participants continued to experience the same contingencies throughout the task (control group).

Immediately after contingency reversal, the irrelevant cues should have generated a smaller prediction error than the relevant cues, because the former had already been paired with the reversed outcomes on half of the preceding trials, whereas the latter had never been paired with them. Thus, Mackintosh's theory predicted an increase in the amount of attention paid to the irrelevant cues, mirrored by a decrease in attention to the relevant cues. Moreover, the discrepancy between the outcomes predicted by the pairs of cues and those occurring on the reversal trials should have produced a large overall prediction error, which, according to the Pearce-Hall model, should have translated into an equal increase in attention to both types of cues. Thus, based on the predictions of both theories, we expected an increase in attention to the irrelevant cues following contingency reversal.

During Stage 1, as the accuracy of outcome predictions increased, a preference for the relevant cues became evident in each measure of attention (dwell time, first within-trial fixation, and fixation count). At the beginning of Stage 2, contingency reversal was followed by a drop in accuracy. This was indicative of a sudden rise in prediction error. Moreover, the frequency of the first fixations on the irrelevant cues increased to the detriment of those on the relevant cues, and the fixation count increased for both types of cues. Contingency reversal had no influence on the dwell times.

As predicted by both Mackintosh's theory and the Pearce-Hall model, the results suggested that a sudden rise in prediction error leads to an increase in overt attention to irrelevant cues. The data regarding the relevant cues showed a decrease in the first fixations, but also an increase in the fixation count, each finding being consistent with Mackintosh's theory and the Pearce-Hall model, respectively. It is noteworthy that the first fixations constitute a measure of selective attention, because only one stimulus can be fixated first on a given trial, which has been shown to be the stimulus receiving the highest attentional priority (Theeuwes, Vries, & Godjin, 2003). Similarly, Mackintosh's theory promotes selectivity, i.e. an increase in attention to the cue generating the smallest prediction error is typically accompanied by a decrease in attention to the rest. On the other hand, the fixation count involves the whole measurement interval and therefore allows participants to fixate on more than one stimuli. This aspect enables changes in attention in line with the suggestion made by Pearce and Hall, i.e. associability changes affecting each of the present cues equally. Given that the hybrid models (e.g., Le Pelley, 2004) incorporate both the associability mechanism suggested by Mackintosh and that suggested by Pearce and Hall, they are equipped to account for the results in both the first fixations and the fixation count, particularly if one allows each of these measures to be more sensitive to one of the two associability mechanisms, as mentioned above.

Researchers in the area of attention and associative learning typically use self-paced designs, where dwell times are measured on each trial from stimulus onset until participants indicate the expected outcome. For standardization purposes, the dwell time on each cue is then expressed as a proportion of the response time (e.g., Beesley et al., 2015; Kruschke, Kappenman, & Hetrick, 2005). When designing Study I, we reasoned that proportional dwell times could mask changes in absolute attention (e.g., if attention increased for each of the present cues equally, as predicted by the Pearce-Hall model) and thus opted for a fixed measurement interval. However, by the end of Stage 1, participants already spent most of the interval fixating on either of the cues. Thus, our choice might have limited the sensitivity of the dwell times to the effects of contingency reversal.

Study II: Evidence for an interaction of two attentional mechanisms during learning

Study I showed that a sudden rise in prediction error increased the amount of attention paid to irrelevant cues. By considering also the data regarding the relevant cues, we concluded that the whole pattern of results could be better explained by the hybrid models, compared to either Mackintosh's theory or the Pearce-Hall model. In Study II, we sought evidence for attentional changes consistent with the predictions of the hybrid models, i.e. compatible with an interaction of two associability mechanisms. In Stage 1, we used discrimination training with pairs of cues, each containing a relevant component and an irrelevant one. Across trials, each of the relevant cues was consistently paired with one outcome, while each of the irrelevant cues was paired with two outcomes. Therefore, as training went on, the relevant cues effectively predicted the outcome of each trial, whereas the irrelevant cues predicted the partial occurrence of either of the two outcomes. In such conditions, the relevant cues should generate a smaller prediction error than the irrelevant cues, and the overall amount of prediction error should be reduced gradually. During Stage 2, each pair of cues was randomly followed by each of the two outcomes (partial reinforcement). Therefore, as training went on, both the relevant and the irrelevant cues predicted the partial occurrence of either outcome. Given that only one 'full' outcome occurred on each trial, each type of cue should have generated a moderate and similar amount of prediction error, which should have been sustained throughout the stage.

We conducted simulations of the results predicted by Mackintosh’s theory, the Pearce-Hall model, and Le Pelley’s (2004) hybrid model. Each theory predicted a different pattern throughout the task. According to the changes in relative prediction error outlined above, Mackintosh’s theory predicted that the relevant cues should end up receiving more attention than the irrelevant cues during Stage 1, with such a difference diminishing during Stage 2. Based on the changes in overall prediction error, the Pearce-Hall model predicted that both types of cues should receive the same amount of attention, which should decrease during Stage 1, and then it should be reinstated by the introduction of partial reinforcement in Stage 2. Finally, according to an interaction of the two associability mechanisms, Le Pelley’s hybrid model predicted that the amount of attention paid to each type of cue should decrease during Stage 1, yet the decrease should be less pronounced for the relevant cues. During Stage 2, attention to each type of cue should increase and become similar.

We came up with a design that allowed us to use a fixed measurement interval and also avoid a ceiling effect on the attention paid to the cues, which presumably affected the sensitivity of the dwell time measure in Study I. The outcomes were a red circle and a green circle, each presented at the end of one half of the trials. Participants indicated which outcome they expected on each trial by pressing one of the mouse buttons. However, the correspondence between the outcomes and the buttons switched from trial to trial and was shown in a response trigger, in which the outcomes were represented as two colored dots next to each other. Participants pressed the button corresponding to the side (left or right) where the dot standing for the expected outcome was displayed. The trigger appeared for a brief interval (200 ms) in a known location on the screen, 4 s after the onset of the cues. Towards the end of the 4-s interval, and in order to avoid missing the trigger, we expected participants to move their eyes away from the cues and towards the location where the trigger was going to appear. We measured the amount of attention paid to the cues and to the location of the trigger by means of dwell times and fixation probabilities. The latter showed changes in the frequency of fixations on the different stimuli within the measurement interval, which was divided into 20 bins of 200 ms each.

As expected, the accuracy of outcome predictions showed an increase during Stage 1 and a drop upon the introduction of partial reinforcement at the beginning of Stage 2. The dwell time gradually decreased for each type of cue during Stage 1, yet it was longer for the relevant cues than the irrelevant cues. At the beginning of Stage 2, the dwell time on each type of cue increased and became similar. These attentional changes, which were also evident in the fixation probability, were consistent with the predictions based on Le Pelley’s hybrid model. In addition, the fixation probability showed that participants looked first on the cues and then on the location of the response trigger. Thus, the response trigger competed for attention with the cues, thereby preventing a ceiling effect. Moreover, as accuracy increased, participants shifted their gaze from the cues to the trigger location earlier within the trial. This shift was delayed by the drop in accuracy. Such a pattern was consistent with a modulation by overall prediction error, according to the mechanism suggested by Pearce and Hall (and incorporated to Le Pelley's model).

Contribution of the present thesis

We investigated the effect of a sudden rise in prediction error on attention to cues being either relevant or irrelevant to outcome prediction. The present thesis includes two studies, both starting with a discrimination training, in which the relevant cues received more attention than the irrelevant cues. In a second stage, contingency reversal (Study I) or partial reinforcement (Study II) led to a rise in prediction error, as indicated by a drop in the accuracy of outcome predictions. The attentional preference for the relevant cues was temporarily weakened by contingency reversal, and it was completely lost following the introduction of partial reinforcement. In addition, both manipulations increased the amount of attention paid to both types of cues. The results were consistent with a combined effect of two associability mechanisms, one giving attentional priority to the cue with the smallest relative prediction error, and the other allocating the same amount of attention to each cue, as a direct function of overall prediction error. Therefore, our data were explained by Le Pelley's (2004) hybrid model, which incorporates the associability mechanisms put forward by Mackintosh (1975) and Pearce and Hall (1980).

A recent empirical review concluded that hybrid models were not required to explain the existing data in humans (Le Pelley et al., 2016). Particularly problematic were findings showing an inverse relationship between attention allocation and overall prediction error, i.e. contrary to what the Pearce-Hall mechanism proposes (Kattner, 2015; Le Pelley, Turnbull, Reimers, & Knipe, 2010; Livesey, Thorwart, De Fina, & Harris, 2011). These findings were based on the speed at which the experimental cues became associated with new outcomes. Such a measure of attention allocation has been widely used in the associative tradition and is based on two assumptions. First, the more attention a cue receives, the faster it becomes associated with its outcome. Second, the associability of a cue transfers across different outcomes. In contrast to this strand of evidence, our results gave support to the hybrid models. In particular, we found that cues generating large overall prediction errors received increased attention, as indicated by eye-gaze behavior. In addition, we also replicated the more established finding that relevant cues receive more attention than irrelevant cues, i.e. consistently with the Mackintosh mechanism. Recently, a few previous studies have also provided evidence for the hybrid models by means of eye gaze (Beesley et al., 2015; Easdale et al., 2019; Walker et al., 2019). In order to explain the discrepancy between the data based on speed of learning and those based on eye gaze, in Study II, we suggested that the above-mentioned studies using speed of learning might have been more susceptible to the influence of participants' knowledge about the relational structure of the task (see Livesey, Don, Uengoer, & Thorwart, 2019). In addition, Beesley et al. (2015) showed that changes in overt attention consistent with the Mackintosh mechanism carried over into a second training stage with new outcomes, but changes in overt attention consistent with the Pearce-Hall mechanism did not. Thus, those authors suggested that associability changes driven by overall prediction error may not transfer to new learning situations. Such a suggestion departs from the formulations of the Pearce-Hall model and Le Pelley's hybrid model, in which the attentional properties of a cue exert an influence on learning that is independent of the outcome or the context. Although the present thesis did not study transfer of attention allocation, we acknowledge the theoretical importance of this question.

Our results may contribute to understanding changes in attention to contexts, a phenomenon with important behavioral implications. Once a cue has been established as a signal for a particular outcome, it may become associated with another one, generating a large prediction error on the first pairings with the latter. In such a situation, retrieval of the second association often depends on being present in the context where it was learned (Bouton, 1993). This context-dependence of learning has been implicated in the return of pathological anxiety after successful therapy. Fear-learning models postulate that anxiety disorders are sustained by associations between an innocuous object and a traumatic event (for a detailed account, see Mineka & Zinbarg, 2006). Accordingly, exposure therapy is based on confronting the feared object in absence of the traumatic event, with the aim to establish a new inhibitory association that will counteract the original one. However, despite an initial reduction, fear may reappear outside the therapy setting (Craske & Mystkowski, 2006). Bouton (1997) suggested that the large prediction error experienced on the first pairings with the second outcome drives attention to the context, thus making learning context-specific. Accordingly, research on the attentional changes following a sudden prediction error may improve our understanding of the mechanisms underlying context-dependent learning. Contexts may be defined as stimuli that are irrelevant to predict an outcome, and therefore end up being ignored (Aristizabal, Ramos-Álbarez, Callejas-Aguilera, & Rosas, 2016; Rosas et al., 2006). The irrelevant cues used in our studies fit with such a description, as they received less and less overt attention during the first stage. In line with Bouton's suggestion, we showed that a sudden rise in prediction error increased attention to those cues. Moreover, Rosas, Todd, and Bouton (2013) suggested that Mackintosh's theory or the Pearce-Hall model could be used to specify the attentional mechanisms involved in context effects, but they also noted that no empirical study had pursued this idea. Our hypotheses were based on specific predictions from those theories, as well as Le Pelley's (2004) hybrid model. And we found that the changes did not only involve a shift of attention towards the irrelevant/contextual cues (see Vadillo, Orgaz, Luque, & Nelson, 2016), but also an increase in attention to the relevant cues.

In Study II, we came up with a new method for measuring allocation of overt attention to cues. Specifically, we used a fixed measurement interval that ended with the presentation of a response trigger competing for attention with the cues. That is, given the brief duration of the trigger, participants moved their eyes in advance towards the location where it was set to appear. By contrast, self-paced designs use measurement intervals lasting until participants make a response. The dwell time on each cue is then expressed as a proportion of the response time, a correction intended to get rid of variations unrelated to attention. However, relative dwell times may leave out meaningful absolute changes in attention. Our method allowed to use absolute dwell times in a way that is not confounded by response time. Moreover, the competition for attention by the response trigger prevented potential ceiling effects.

In conclusion, the present thesis showed the effects of a sudden rise in prediction error on overt attention to relevant and irrelevant cues. The data provided evidence for the hybrid models, as they were consistent with an interaction of two associability mechanisms, each based on either relative or overall prediction error. The results also have implications for understanding changes in attention to contextual cues, which have been related to important phenomena, such as return of pathological anxiety.[2]

[2] Within the framework of the RTG 2271, our results show that previous expectation violations influence the allocation of overt attention to the cues involved in a situational exposition. On the one hand, those cues previously associated with expectation maintenance are given attentional priority. On the other hand, the amount of attention paid to the whole set of cues decreases as expectations become consolidated. Strong expectation violations increase attention to all the cues and weaken the pre-existing biases.

References

Anderson, B. A., Laurent, P. A., & Yantis, S. (2011). Value-driven attentional capture. *Proceedings of the National Academy of Sciences, USA, 108*, 10367–1037, https://doi.org/110.1073/pnas.1104047108

Aristizabal, J. A., Ramos-Álbarez, M. M., Callejas-Aguilera, J. E., & Rosas, J. M. (2016). Attention to irrelevant contexts decreases as training increases: Evidence from eye-fixations in a human predictive learning task. *Behavioural Processes, 124*, 66–73, https://doi.org/10.1016/j.beproc.2015.12.008

Awh, E., Belopolsky, A. V., & Theeuwes, J. (2012). Top-down versus bottom-up attentional control: A failed theoretical dichotomy. *Trends in Cognitive Sciences, 16*, 437–443, https://doi.org/10.1016/j.tics.2012.06.010

Beesley, T., Nguyen, K. P., Pearson, D., & Le Pelley, M. E. (2015). Uncertainty and predictiveness determine attention to cues during human associative learning. *Quarterly Journal of Experimental Psychology, 68*, 2175–2199, https://doi.org/10.1080/17470218.2015.1009919

Bouton, M. E. (1993). Context, time, and memory retrieval in interference paradigms of Pavlovian learning. *Psychological Bulletin, 114*, 80–99, https://doi.org/10.1037/0033-2909.114.1.80

Bouton, M. E. (1997). Signals for whether versus when an event will occur. In M. S. Fanselow & M. E. Bouton (Eds.), *Learning, motivation, and cognition: The functional behaviorism of Robert C. Bolles* (p. 385–409). Washington, DC: American Psychological Association.

Craske, M. G., & Mystkowski, J. L. (2006). Exposure therapy and extinction: Clinical studies. In M. G. Craske, D. Hermans, & D. Vansteenwegen (Eds.), *Fear and learning: From basic processes to clinical implications* (p. 217–233). American Psychological Association. https://doi.org/10.1037/11474-011

Desimone, R., & Duncan, J. (1995). Neural mechanisms of selective visual attention. *Annual Review of Neuroscience, 18*, 193–222, https://doi.org/10.1146/annurev.ne.18.030195.001205

Domjan, M. (2005). Pavlovian conditioning: A functional perspective. *Annual Review of Psychology, 56*, 179–206, https://doi.org/10.1146/annurev.psych.55.090902.141409

Easdale, L. C., Le Pelley, M. E., & Beesley, T. (2019). The onset of uncertainty facilitates the learning of new associations by increasing attention to cues. *Quarterly Journal of Experimental Psychology, 72*, 193–208, https://doi.org/10.1080/17470218.2017.1363257

George, D. N., & Pearce, J. M. (2012). A configural theory of attention and associative learning. *Learning & Behavior, 40*, 241–254, https://doi.org/10.3758/s13420-012-0078-2

Gottieb, J. (2012). Attention, learning, and the value of information. *Neuron, 76*, 281– 295, https://doi.org/10.1016/j.neuron.2012.09.034

Griffiths, O., Johnson, A. M., & Mitchell, C. J. (2011). Negative transfer in human associative learning. *Psychological Science, 22*, 1198–1204, https://doi.org/10.1177/0956797611419305

Hall, G., & Pearce, J. M. (1982). Restoring the associability of a preexposed CS by a surprising event. *The Quarterly Journal of Psychology, 34B*, 127–140. https://doi.org/10.1080/14640748208400881

Holland, P. C., & Maddux, J. M. (2010). Brain systems of attention in associative learning. In C. J. Mitchell & M. E. Le Pelley (Eds.), *Attention and associative learning: From brain to behaviour* (pp. 305–349). Oxford, UK: Oxford University Press.

Kattner, F. (2015). Transfer of absolute and relative predictiveness in human contingency learning. *Learning & Behavior, 43*, 32–43, https://doi.org/10.3758/s13420-014-0159-5

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, https://doi.org/10.1037/0278-7393.31.5.830

Le Pelley, M. E. (2004). The role of associative history in models of associative learning: A selective review and a hybrid model. *Quarterly Journal of Experimental Psychology, 57B*, 193–243, https://doi.org/10.1080/02724990344000141

Le Pelley, M. E., & McLaren, I. P. L. (2003). Learned associability and associative change in human causal learning. *The Quarterly Journal of Experimental Psychology, 56B*, 68–79, https://doi.org/10.1080/02724990244000179

Le Pelley, M. E., Mitchell, C. J., Beesley, T., George, D. N., & Wills, A. J. (2016). Attention and associative learning in humans: An integrative review. *Psychological Bulletin, 142*, 1111–1140, https://doi.org/10.1037/bul0000064

Le Pelley, M. E., Trunbull, M. N., Reimers, S. J., Knipe, R. L. (2010). Learned predictiveness effects following single-cue training in humans. *Learning & Behavior, 38*, 126–144, https://doi.org/10.3758/LB.38.2.126

Livesey, E. J., Don, H. J., Uengoer, M., & Thorwart, A. (2019). Transfer of associability and relational structure in human associative learning. *Journal of Experimental Psychology: Animal Learning and Cognition, 45*, 125–142, https://doi.org/10.1037/xan0000197

Livesey, E. J., Thorwart, A., De Fina, N. L., & Harris, J. A. (2011). Comparing learned predictiveness effects within and across compound discriminations. Journal of Experimental Psychology: *Animal Behavior Processes, 37*, 446–465, https://doi.org/10.1037/a0023391

Mackintosh, N. J. (1975). A theory of attention: Variations in the associability of stimuli with reinforcement. *Psychological Review, 82*, 276–298, https://doi.org/10.1037/h0076778

Mackintosh, N. J., & Little, L. (1969). Intradimensional and extradimensional shift learning by pigeons. *Psychonomic Science, 14*, 5–6, https://doi.org/10.3758/BF03336395

Mineka, S., & Zinbarg, R. (2006). A contemporary learning theory perspective on the etiology of anxiety disorders: It's not what you thought it was. *American Psychologist, 61*, 10–26. https://doi.org/10.1037/0003-066X.61.1.10

Pearce, J. M., & Bouton, M. E. (2001). Theories of associative learning in animals. *Annual Review of Psychology, 51*, 111–139, https://doi.org/10.1146/annurev.psych.52.1.111

Pearce, J. M., & Hall, G. (1980). A model for pavlovian learning: Variations in the effectiveness of conditioned but not of unconditioned stumuli. *Psychological Review, 87*, 532–552, https://doi.org/10.1037/0033-295X.87.6.532

Pearce, J. M., & Mackintosh, N. J. (2010). Two theories of attention: A review and a possible integration. In C. J. Mitchell & M. E. Le Pelley (Eds.), *Attention and associative learning: From brain to behaviour* (p. 11–40). Oxford, UK: Oxford University Press.

Rescorla, R. A. (1988). Pavlovian conditioning: It's not what you think it is. *American Psychologist, 43*, 151–160, https://doi.org/10.1037/0003-066X.43.3.151

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

Rief, W., Glombiewski, J. A., Gollwitzer, M., Schubö, A., Schwarting, R., & Thorwart, A. (2015). Expectancies as core features of mental disorders. *Current Opinion in Psychiarty, 28*, 378–385, https://doi.org/10.1097/YCO.0000000000000184

Rosas, J. M., & Callejas-Aguilera, J. E. (2006). Context switch effects on acquisition and extinction in human predictive learning. *Journal of Experimental Psychology: Learning, Memory, and Cognition, 32*, 461–474, https://doi.org/10.1037/0278-7393.32.3.461

Rosas, J. M., Todd, T. P., & Bouton, M. E. (2013). Context change and associative learning. *Wiley Interdiciplinary Reviews. Cognitive Science, 4*, 237–44, https://doi.org/10.1002/wcs.1225

Schultz, W., & Dickinson, A. (2000). Neuronal conding of prediction errors. *Annual Review of Neuroscience, 23*, 473–500, https://doi.org/10.1146/annurev.neuro.23.1.473

Theeuwes, J., & Belopolsky, A. V. (2012). Reward grabs the eye: Oculomotor capture by rewarding stimuli. *Vision Research, 74*, 80–85, https://doi.org/10.1016/j.visres.2012.07.024

Theeuwes, J., De Vries, G. J., & Godijn, R. (2003). Attentional and oculomotor capture with static singletons. *Pereception & Psychophysics, 65*, 735–746, https://doi.org/10.3758/BF03194810

Vadillo, M. A., Orgaz, C., Luque, D., & Byron Nelson, J. (2016). Ambiguity produces attention shifts in category learning. *Learning and Memory, 23*, 134–40, https://doi.org/10.1101/lm.041145.115

Walker, A. R., Luque, D., Le Pelley, M. E., & Beesley, T. (2019). The role of uncertainty in attentional and choice exploration. *Psychonomic Bulletin & Review, 26*, 1911–1916, https://doi.org/10.3758/s13423-019-01653-2

Yu, A., & Dayan, P. (2005). Uncertainty, neuromodulation, and attention. *Neuron, 46*, 681–692, https://doi.org/10.1016/j.neuron.2005.04.026

A rise in prediction error increases attention to irrelevant cues

David Torrents-Rodas [a], Stephan Koenig [a, b], Metin Uengoer [a], and Harald Lachnit [a]

[a] Faculty of Psychology, Philipps-Universität Marburg, Marburg, Germany

[b] Faculty of Psychology, Universität Koblenz-Landau, Landau, Germany

Corresponding author: David Torrents-Rodas, Faculty of Psychology, Philipps-Universität Marburg, Gutenbergstr. 18, 35032 Marburg, Germany. torrents@uni-marburg.de

Abstract

We investigated whether a sudden rise in prediction error widens an individual's focus of attention by increasing ocular fixations on cues that otherwise tend to be ignored. To this end, we used a discrimination learning task including cues that were either relevant or irrelevant for predicting the outcomes. Half of participants experienced contingency reversal once they had learned to predict the outcomes (reversal group, *n* = 30). The other half experienced the same contingencies throughout the task (control group, *n* = 30). As participants' prediction accuracy increased, they showed a decrease in the number of fixations directed to the irrelevant cues. Following contingency reversal, participants in the reversal group showed a drop in accuracy, indicating a rise in prediction error, and fixated on the irrelevant cues more often than participants in the control group. We discuss the results in the context of attentional theories of associative learning.

Keywords: associative learning, attention, eye-tracking, irrelevant cues, prediction error

Introduction

Attention plays an adaptive role by weighting sensory inputs according to their inherent or acquired biological significance (Gottlieb, 2012). Accordingly, the ability of a stimulus or cue to attract our attention is not only based on its physical salience, but it also changes as we learn about its consequences (Anderson, 2015). This idea is at the heart of attentional theories of associative learning. The studies conducted in this area have traditionally inferred the amount of attention directed to cues by measuring their associability – i.e. the speed at which a cue becomes associated with a given outcome upon repeated pairings (for a review, see Le Pelley, 2004). Thus, the relationship between learning and attention operates in both directions. The significance of a cue for outcome prediction determines the amount of attention it receives, and the more attention the cue receives, the faster it becomes associated to the outcome. The mechanisms driving attentional changes rely on prediction error, which is the discrepancy between the outcome magnitude predicted by the cue on the basis of previous experience and the actual occurrence (or omission) of the outcome. A fully expected outcome generates no prediction error, whereas a surprising outcome generates a large prediction error. As we will explain below, the specific ways in which prediction error is thought to drive the attentional changes differ across theories.

Not all the cues that are present in natural or social environments are equally relevant for the prediction of significant outcomes. The effects of this circumstance on behavior and attention have been extensively studied by means of discrimination learning paradigms using relevant and irrelevant cues. In a typical experiment, a cue compound (e.g., AX) is paired with outcome 1 (o1) on some trials, whereas another cue compound (e.g., BX) is paired with outcome 2 (o2) on other trials. Thus, cue A is consistently followed by o1, and cue B is consistently followed by o2. Conversely, cue X is followed by o1 on half of the trials, while, on the other half, it is followed by o2. Therefore, A and B are relevant for predicting the outcomes, whereas X is irrelevant. As participants learn to solve the discrimination, they will direct more attention to the relevant cues than to the irrelevant cues. This finding has been consistently reported in both human and animal research (for reviews, see Le Pelley, Mitchell, Beesley, George, & Wills, 2016; Pearce & Mackintosh, 2010).

There is little controversy over the finding that irrelevant cues receive less attention than relevant cues when the outcomes of a discrimination are *fully* predictable. Yet, environments are ever-changing. With contingencies between cues and outcomes shifting over time, humans and animals should possess attentional and behavioral mechanisms enabling them to adapt to such situations (Rief et al., 2015). The present study was conducted to investigate whether a sudden change in the environmental contingencies widens an individual's attentional focus by increasing attention to irrelevant cues. An experimental manipulation allowing researchers to model this environmental feature is contingency reversal, where initially two cues (or sets of cues) each become associated with a specific outcome, and later the cue-outcome relations are switched. This switch is assumed to create a large prediction error between the strong anticipation of the outcome based on a given cue and the actual occurrence of the other outcome.

To the best of our knowledge, only one previous study has focused on the effects of contingency reversal on attention to irrelevant cues. In each of two experiments, Vadillo, Orgaz, Luque, and Nelson (2016) asked participants to solve a discrimination task with four relevant cues (A, B, C, and D), each followed by one of two possible outcomes (o1 and o2). Each of these cues appeared together with the same low-salience irrelevant cue (X), which was followed by o1 on one half of the trials, and by o2 on the other half (i.e., AX – o1, BX – o2, CX – o1, DX – o2). After an initial series of trials, the outcomes of two of the cue compounds were reversed, whereas the outcomes of the other two remained unchanged (i.e., AX – *o2*, BX – *o1*, CX – o1, DX – o2). A further group experienced the same contingencies throughout the task. Attentional allocation was measured by means of a dot-probe paradigm. Reaction times from trials where the probe appeared on a relevant cue were subtracted from reaction times from trials where the probe appeared on the irrelevant cue, thus yielding a score of attentional preference. The values of this score were positive, therefore indicating an attentional bias towards the relevant cues. However, compared to participants in the control group, participants who experienced contingency reversal on AX and BX trials exhibited a reduction in both the accuracy of outcome predictions and the attentional score. On CX and DX trials, where contingencies stayed the same in both groups, such reductions were not observed. In the first experiment, contingency reversal only led to a marginally significant effect on the attentional score. Thus, Vadillo et al. suggested that participants might have come to ignore the irrelevant cue to the extent that it limited the reversal effect. Accordingly, in the second experiment, they used two irrelevant cues that switched constantly from trial to trial during the initial stage. After contingency reversal, only one of the irrelevant cues continued to be present. Taken together, the results from the two experiments indicated that a rise in prediction error, induced by contingency reversal, shifted attention away from the relevant cue and towards the irrelevant cue. As a possible interpretation, Vadillo et al. suggested that the shift in attentional preference protected the initially acquired associations from interference (see also Kruschke, 2001; Rosas, Callejas-Aguilera, Ramos-Álvarez, & Fernández-Abad, 2006).

Following Vadillo et al. (2016), in the present study we used contingency reversal to induce a sudden rise in prediction error. Our participants had to press one of two mouse-buttons in order to get a monetary reward. Thus, the outcome to be predicted on each trial was the correct button – either left (o1) or right (o2), and it depended on which of the relevant cues (A, B, C, or D) was presented on a screen. Each of these cues appeared together with an irrelevant cue (X or Y). During an initial stage, participants received a series of trials with the following cue-outcome contingencies: AX – o1, BX – o2, CY – o1, and DY – o2. At the outset of a second stage, we reversed the outcomes following A and B for one half of participants, whereas we kept all the cue-outcome contingencies for the other half. Based on the idea that continuity contributed to reduce the salience of the irrelevant cue in the first experiment of Vadillo et al., at the beginning of Stage 2, we changed the combinations of the relevant and the irrelevant cues (reversal group: AY – *o2*, BY – *o1*, CX – o1, and DX – o2, control group: AY – o1, BY – o2, CX – o1, and DX – o2). Moreover, the relevant and the irrelevant cues were represented by shape and color singletons, respectively. Each pair of cues was embedded in a search display with low salience distractors. Research using this "additional singleton" display has demonstrated that color singletons show a high probability to capture attention due to their physical salience (e.g., Theeuwes, 1991, 1992; Theeuwes, Vries, & Godjin, 2003). Therefore, we expected some salience-based interference by the irrelevant cue in the deployment of attention to the relevant cue, which could increase the chances of finding an effect of prediction error on the irrelevant cues. As an alternative approach to that used by Vadillo and his colleagues, we used eye-tracking to yield more elaborate measures of attentional allocation. Because eye-tracking enables researchers to index overt attention, a substantial number of studies on attention and learning has incorporated this method over the last decade (for a review, see Le Pelley et al., 2016).

As mentioned above, attentional theories of associative learning consider prediction error as the central force driving changes in attention. Yet, the two most influential of these theories (Mackintosh, 1975; Pearce & Hall, 1980) postulate opposing mechanisms by which prediction error influences attention. On the one hand, Mackintosh's (1975) theory proposes that associability increases when a given cue is a better predictor of the trial outcome than all the other accompanying cues (i.e., when it generates less prediction error). Otherwise, when the cue is equally or less predictive than the rest, associability decreases. On the other hand, the Pearce-Hall (1980) model suggests that cues followed by unexpected outcomes (i.e., generating a large prediction error) sustain higher associability than reliable cues. Moreover, changes in associability affect each of the present cues equally. Although Mackintosh's (1975) theory and the Pearce-Hall (1980) model make opposing predictions in a number of situations (e.g., Hogarth, Dickinson, Austin, Brown, & Duka, 2008; Le Pelley & McLaren, 2003), both predict that a rise in prediction error should increase attention to irrelevant cues. This prediction has been widely neglected in empirical research.

Going back to our study, at the end of the first stage, participants should come to anticipate o1 primarily on the basis of A or C, and o2 on the basis of B or D, while X and Y should just weakly contribute to the expectation of each of the two outcomes. Therefore, after contingency reversal on AY and BY trials – and until participants update their predictions according to the new contingencies, prediction error should be larger for A and B than for Y. Following Mackintosh (1975), this should lead to an increase in attention to Y. On the other hand, according to Pearce and Hall (1980), the strong prediction error generated on AY and BY trials should increase attention to Y as well as A and B. Thus, we expected that the rise in prediction error would be accompanied by an increase in attention to Y in the reversal group. We additionally wanted to test whether the expected attentional increase to Y would generalize to X, which should not be associated with prediction error. If this were the case, the amount of attention directed to X by participants in the reversal group should be larger than that directed to any of the irrelevant cues by participants in the control group.

Method

Participants

The study was approved by the ethics committee of the Faculty of Psychology at the Philipps-Universität Marburg (AZ: 2018-25k). Ninety-one students participated in exchange of 10 € or course credits. Based on their performance, they also received monetary reward (up to 9.60 €). All participants reported absence of visual impairment (< 1.50 diopters) or used soft contact lenses. We excluded two participants whose eye-tracking recordings were affected by excessive head movements.

Given that contingency reversal was the critical manipulation in our experiment, we only included participants who actually learned the contingencies trained during Stage 1. Figure 1 depicts the distribution density of the proportion of correct responses (accuracy) in the last block of Stage 1. Inspection of Figure 1 revealed that, while most participants exhibited maximal accuracy (1.00, good learners), for 29 participants accuracy was at chance level (0.50, poor learners). We excluded those poor learners from further analyses. The cutoff was determined at an accuracy of 0.70, according to the local minimum observed in Figure 1. The final sample consisted of 60 participants (age: *M* = 23.29 years, *SD* = 3.31, 11 male).

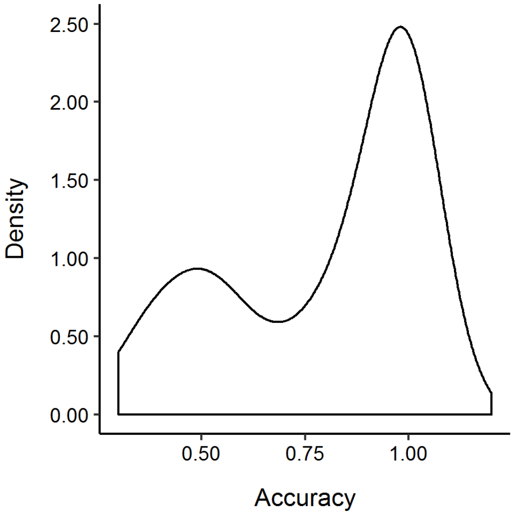

Figure 1. Density plot of response accuracy in the last block of Stage 1.

Stimuli and procedure

Table 1 shows the experimental design. Participants had to press one of two mouse-buttons in order to get a monetary reward. Thus, the outcomes of the discrimination were the correct buttons – either left (o1) or right (o2). On each trial, the outcome could be predicted on the basis of the relevant cue (A, B, C, or D) presented on the screen. Each relevant cue was combined with one of two irrelevant cues (X or Y), which could not be used to make a correct prediction, because they were followed by each of the two outcomes on the same number of trials throughout the task. We used four different cue compounds during Stage 1: AX and CY (followed by o1) and BX and DY (followed by o2). Participants received four blocks of training with these cue-outcome contingencies, each block including four presentations of each compound. At the beginning of Stage 2, we changed the combinations of the relevant and the irrelevant cues. In addition, we reversed the outcomes following A and B for one half of participants (reversal group: AY – o2, BY – o1, CX – o1, and DX – o2), whereas we kept all the cue-outcome contingencies for the other half (control group: AY – o1, BY – o2, CX – o1, and DX – o2). Stage 2 also consisted of four blocks of training, each of them including four presentations of each of the corresponding compounds. The trial order within each block was determined randomly, with the restriction that no more than two trials with the same outcome occurred consecutively. Although there was no break between the two stages, we added an extra trial at the beginning of Stage 2 (AY for half of participants within each group, BY for the other half, total number of trials in the experiment = 129). This was the first trial in which participants in the reversal group experienced contingency reversal. Because the attention-measurement interval on each trial preceded the outcome presentation, we did not include this trial in the analysis of the eye-tracking data.

Table 1. *Experimental design.*

	Stage 1	Stage 2
Reversal group	AX – o1, BX – o2, CY – o1, DY – o2	AY – o2, BY – o1, CX – o1, DX – o2
Control group		AY – o1, BY – o2, CX – o1, DX – o2

The relevant cues were represented by four dark-grey letters (G, H, M, S) inside a dark-grey shape singleton, and the irrelevant cues by two colored (blue or green) circles, i.e. color singletons. Both types of cues were embedded in a search display, which also contained four dark-grey circles. The screen background was light grey. All the stimuli were 31 mm in diameter, and were distributed in polar angles of 0°, 60°, 120°, 180°, 240°, and 300° (see also Koenig, Kadel, Uengoer, Schubö, & Lachnit, 2017; Koenig, Uengoer, & Lachnit, 2017). Within the search display, the relevant and the irrelevant cues were presented on different hemifields, but they did not appear directly opposite to each other. The positions varied randomly from trial to trial. Across participants, we counterbalanced the assignment of each letter and each colored circle to each of the relevant and the irrelevant cues, respectively. Figure 2 illustrates such a trial which began with a 2-s presentation of a fixation cross. Then, the display containing the cues appeared, but for clicking one of the two mouse-buttons participants had to wait until an auditory trigger was presented 2 s later. The response interval ended after 2 s, followed by an auditory feedback indicating whether the response was correct ("correct: 10 cents") or not ("incorrect: 0 cents"). In addition, a small dot appeared on the left or the right side of the relevant cue, indicating the correct button (2 more s).

Specific warning messages appeared instead of the feedback when participants made the response before (“response too early!”) or after (“response too late!”) the appropriate interval. In order to delay learning of the new contingencies after reversal, and thus extend the number of trials in which the cues would presumably generate prediction error, we omitted the feedback in one quarter of the trials (four out of the 16 trials in each block, randomly determined). A blank screen with a random duration between 1.5 and 4.5 s followed every trial.

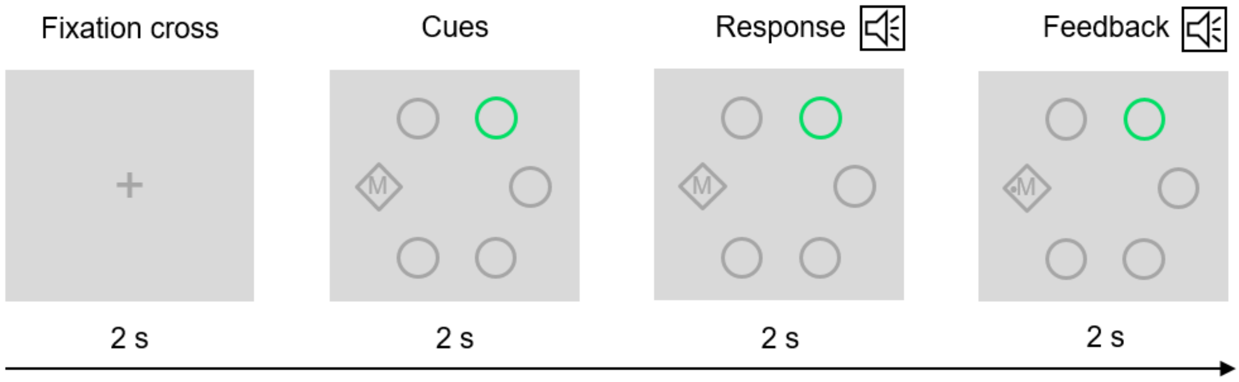

Figure 2. Structure of one trial. A fixation cross preceded the presentation of the relevant and the irrelevant cues, i.e. gray letter inside a diamond and colored circle, respectively. Participants indicated the expected outcome by clicking one of the mouse-buttons when they heard a tone. An auditory feedback indicated whether the response was correct or not. A small dot presented on the left or the right side of the relevant cue indicated the correct button.

Upon their arrival in the laboratory, participants signed the informed consent and read the instructions of the experiment, which gave specific details about the sequence of events within a trial. The instructions also asked participants to attend exclusively to the shapes, the color, and the letter, presented on each trial, and then to decide which button to press in order to earn a reward. They also indicated that, at the outset of the experiment, participants could only guess which button to press. Yet, over the course of the experiment, participants could learn which button to press after each combination of letter, color, and shape. After reading the instructions, participants did a minimum of six practice trials with stimuli that were similar to those used in the experiment. Then, we conducted the calibration procedure by means of a nine-point grid until subsequent validation confirmed a maximal error of < 0.5°. Sampling of the left or the right eye was counterbalanced across participants. The experiment started immediately after calibration. The visual stimuli were presented on a 22” CRT screen (Vision Master Pro 514, Iiyama; Tokio, Japan). Forehead and chin rests kept the head in a fixed position, with an eye-to-screen distance of 78 cm. An infrared video-based eye tracker (Eyelink 2000, SR-Research; Mississauga, Canada) recorded the eye movements, sampling the gaze position at a frequency of 1000 Hz. Presentation software (version 16.1; Neurobehavioral Systems, Inc.) controlled stimulus timing and response recording.

Data analysis

We used custom MATLAB software (The MathWorks, Inc., 2012) for signal conditioning of eye position traces (Koenig, 2010; Koenig & Lachnit, 2011). Ocular fixations were detected using a velocity-based algorithm with a threshold of 30°/s. Fixations were scored as on-stimulus if they deviated less than 60 mm from its center (and were closer to that stimulus than to any other). The analysis interval within each trial was limited to the 2 s between the onset of the cues and the presentation of the response trigger. We used the proportion of correct responses (accuracy) to infer the amount of prediction error associated with the different cues, with a low accuracy indicating a large prediction error. The data from those trials in which no response occurred during the appropriate interval were excluded from the accuracy analyses (2%). The attentional measures focused on the discrimination cues, both the irrelevant and relevant ones, and included the dwell time (with 0 ms on those trials in which the cues were not fixated), the frequency of first within-trial fixations, and the fixation count. We excluded those trials in which a) participants did not fixate on the cross before the onset of the cues (2%), b) participants did not move their eyes (1%), or c) an artefact had a duration above percentile 90 (380 ms, < 1%). The data were averaged across trials, separately by cue (irrelevant vs relevant), group (reversal vs control), trial type (reversal vs non-reversal)[3]

[3] For simplicity, the trials presenting A or B – whose outcomes were switched at the beginning of Stage 2 in the reversal group – are labeled “reversal trials”. The trials presenting C or D – whose outcomes remained unchanged throughout the task in both groups – are labeled “non-reversal trials”.

and block. For accuracy and each of the attentional measures, we conducted six separate ANOVAs, half of them for the reversal trials, and the other half for the non-reversal trials. Within each trial type, one the ANOVAs included the blocks immediately preceding and following contingency reversal, while the other two focused on either Stage 1 or Stage 2. Where appropriate, the Greenhouse-Geisser (1959) correction was applied to the degrees of freedom. The significance level was set at .05. Partial eta squared (η^2_p) and Cohen's *d* were used as measures of effect size.

Results

Accuracy

Figure 3 shows the accuracy of outcome predictions. Participants in each group learned to predict the outcomes during Stage 1. Accordingly, a Group (reversal vs control) × Block (1 – 4) ANOVA for the reversal trials revealed a significant main effect of block, $F(3, 174) = 50.62$, $p < .001$, $\eta^2_p = .47$ [linear trend: $F(1, 58) = 173.41$, $p < .001$, $\eta^2_p = .75$]. Neither the main effect of group, $F(1, 58) = 3.29$, $p = .08$, $\eta^2_p = .05$, nor the interaction, $F(3, 174) = 1.78$, $p = .15$, $\eta^2_p = .03$, were significant. For the non-reversal trials, an ANOVA with the same factors also revealed a significant main effect of block, $F(3, 174) = 35.62$, $p < .001$, $\eta^2_p = .38$ [linear trend: $F(1, 58) = 93.79$, $p < .001$, $\eta^2_p = .62$]. Neither the main effect of group, $F < 1$, nor the interaction, $F(3, 174) = 2.18$, $p = .09$, $\eta^2_p = .04$, were significant.

At the beginning of Stage 2, participants in the reversal group experienced contingency reversal on the trials presenting A or B. Not surprisingly, they showed a drop in accuracy from the last block of Stage 1 to the first block of Stage 2. However, a drop in accuracy also appeared on the non-reversal trials. On the other hand, participants in the control group did not experience any contingency change and therefore their accuracy continued to be high. For the reversal trials, a Group (reversal vs control) × Block (Stage 1, Block 4 vs Stage 2, Block 1) ANOVA revealed a significant interaction, $F(1, 58) = 36.81$, $p < .001$, $\eta^2_p = .39$, with a decrease in accuracy across blocks in the reversal group, $t(29) = 7.39$, $p < .001$, $d = 1.38$, but not in the control group, $t < 1$. Moreover, accuracy did not differ between groups in the last block of Stage 1, $t < 1$, but was lower in the reversal group than the control group in the first block of

Stage 2, $t(58) = 6.85$, $p < .001$, $d = 1.80$. A Group (reversal vs control) × Block (Stage 1, Block 4 vs Stage 2, Block 1) ANOVA for the non-reversal trials also revealed a significant interaction, $F(1, 58) = 21.34$, $p < .001$, $\eta^2_p = .27$, again, with a decrease in accuracy in the reversal group, $t(29) = 6.62$, $p < .001$, $d = 1.19$, but not in the control group, $t < 1$. Accuracy did not differ between groups in the last block of Stage 1, $t < 1$, but was lower in the reversal group than the control group in the first block of Stage 2, $t(58) = 5.29$, ps $< .001$, $d = 1.35$. In each ANOVA, the main effects of group and block were significant, Fs$(1, 58) > 17.62$, ps $< .001$, η^2_ps $> .23$.

During Stage 2, accuracy appeared to be lower in the reversal group than the control in each trial type. For the reversal trials, a Group (reversal vs control) × Block (1 – 4) ANOVA revealed a significant interaction, $F(3, 174) = 13.33$, ps $< .001$, $\eta^2_p = .19$, with the reversal group showing a lower accuracy than the control group from the first to the third block, ts$(58) > 2.12$, ps $< .04$, ds > 0.57 [fourth block: $t(58) = 1.88$, $p = .07$, $d = 0.47$]. For the non-reversal trials, a Group (reversal vs control) × Block (1 – 4) ANOVA also revealed a significant interaction, $F(3, 174) = 5.77$, $p < .001$, $\eta^2_p = .09$, with the between-group difference being limited to the first block, $t(58) = 5.29$, $p < .001$, $d = 1.35$ [remaining blocks: ts$(58) < 1.97$, ps $> .05$, ds < 0.49]. Each ANOVA also yielded significant main effects of group and block, Fs$(1, 58) > 11.08$, ps $< .01$, η^2_ps $> .16$.

Contingency reversal led to a drop in accuracy, which unexpectedly generalized to the non-reversal trials.

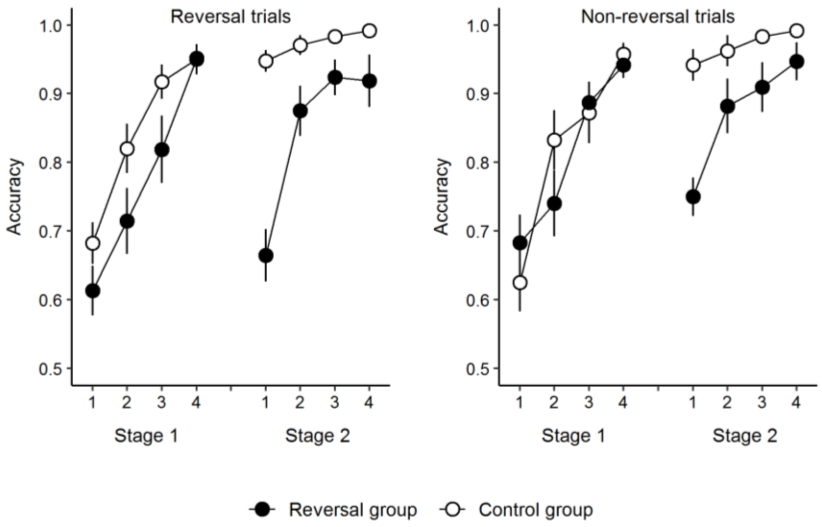

Figure 3. Mean accuracy of outcome predictions across blocks, separated by group (reversal vs control) for the reversal and non-reversal trials (left and right panels, respectively). Error bars represent SEM. Participants in the reversal group experienced contingency reversal at the beginning of Stage 2.

Dwell time

Figure 4 shows the within-trial dwell time on the discrimination cues. During Stage 1, the irrelevant cues were fixated for less time than the relevant cues. Moreover, the dwell time on the irrelevant cues decreased, while that on the relevant cues increased. This pattern was obtained for each trial type (reversal and non-reversal). A Cue (irrelevant vs relevant) × Group (reversal vs control) × Block (1 – 4) ANOVA for the reversal trials revealed a Cue × Block interaction, $F(3, 174) = 8.14$, $p < .001$, $\eta^2_p = .12$, with the linear trends across blocks being significant for both types of cues, but moving towards opposite directions, Fs(1, 59) > 8.45, ps < .01, η^2_ps > .12. In each block, the dwell time was shorter on the irrelevant cues than the relevant cues, ts(59) > 10.84, ps < .001, ds > 1.40. For the non-reversal trials, an ANOVA with the same factors also revealed a Cue × Block interaction, $F(3, 174) = 9.39$, $p < .001$, $\eta^2_p =$ .14. Again, the linear trends across blocks were significant for both types of cues, and moved towards opposite directions, Fs(1, 59) > 10.01, ps < .01, η^2_ps > .14. In each block, the dwell time was shorter on the irrelevant cues than the relevant cues, ts(59) > 10.20, ps < .001, ds > 1.32. In each ANOVA, the main effect of cue was significant, *F*s(1, 58) > 315.16, *p*s < .001, η^2_ps > .84, but the other main effects and interactions were non-significant, *F*s < 3.39, *p*s > .07 η^2_ps < .06.

Contingency reversal did not appear to increase the dwell time on the irrelevant cues. For the reversal trials, a Cue (irrelevant vs relevant) × Group (reversal vs control) × Block (Stage 1, Block 4 vs Stage 2, Block 1) ANOVA revealed a Cue × Block interaction, *F*(1, 58) = 5.50, *p* = .02, η^2_p = .09, indicating that the dwell time on the irrelevant cues remained similar across trials, *t*(59) = 1.21, *p* = .23, *d* = 0.16, while the dwell time on the relevant cues increased, *t*(59) = 2.79, *p* < .01, *d* = 0.36. The main effects of cue and block were also significant, *F*s(1, 58) > 6.23, *p*s < .02, η^2_ps > .09. The main effect of group and the remaining interactions were non-significant, *F*s(1, 58) < 3.06, *p*s > .08 η^2_ps < .06. For the non-reversal trials, an ANOVA with the same factors revealed a Cue × Group interaction, *F*(1, 58) = 5.24, *p* = .03, η^2_p = .08, indicating that the dwell time on the irrelevant cues was longer in the reversal group than the control group, *t*(58) = 2.84, *p* < .01, *d* = 0.73, whereas the dwell time on the relevant cues was similar in both groups, *t*(58) = 1.49, *p* = .14. The Cue × Group × Block interaction was non-significant, *F* < 1. Therefore, the between-group difference in the dwell time on the irrelevant cues cannot be related to contingency reversal. In addition, the main effect of cue was significant, *F*(1, 58) = 321.41, *p* < .001, η^2_p = .85, but the other main effects and interactions were non-significant, *F*s(1, 58) < 3.44, *p*s > .06 η^2_ps < .06.

During Stage 2, the irrelevant cues were fixated for less time than the relevant cues. For the reversal trials, a Cue (irrelevant vs relevant) × Group (reversal vs control) × Block (1 – 4) ANOVA revealed a significant main effect of cue, *F*(1, 58) = 358.42, *p* < .001, η^2_p = .86, with a shorter fixation time on the irrelevant cues than the relevant cues, but the other main effects and interactions were non-significant, *F*s < 2.51, *p*s > .07 η^2_ps < .05. For the non-reversal trials, an ANOVA with the same factors revealed a Cue × Block interaction, *F*(3, 174) = 3.73, *p* = .01, η^2_p = .06. However, in each block, the dwell time was shorter on the irrelevant cues than the relevant cues, *t*s(59) > 15.79, *p*s < .001, *d*s > 2.03. In addition, the main effect of cue was significant, *F*(1, 58) = 356.02, *p* < .001, η^2_p = .86, but the other main effects and interactions were non-significant, *F*s < 3.13, *p*s > .08 η^2_ps < .06.

Throughout the task, participants fixated for less time on the irrelevant cues than the relevant cues. Contrary to our expectations, contingency reversal did not led to an increase in the fixation time on the irrelevant cues.

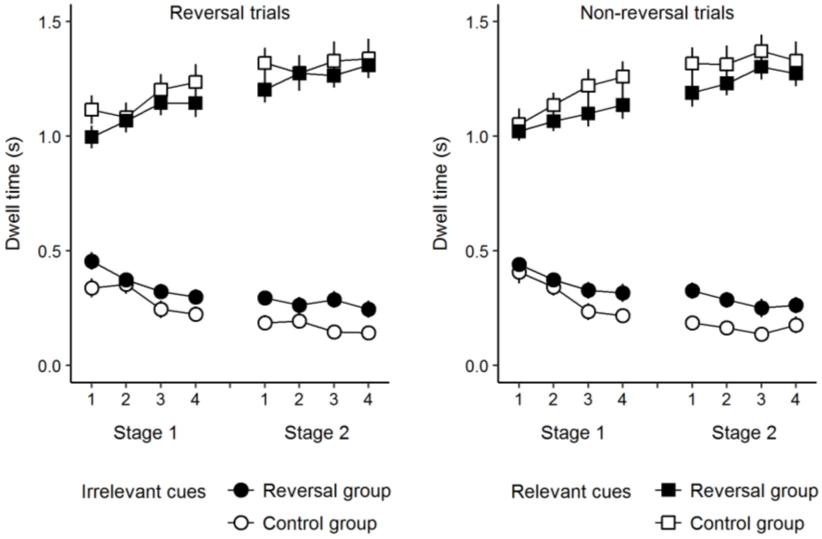

Figure 4. Mean within-trial dwell time across blocks, separated by cue (irrelevant vs relevant) and group (reversal vs control) for the reversal and non-reversal trials (left and right panels, respectively). Error bars represent SEM.

First fixation

Figure 5 shows the frequency of first within-trial fixations on the discrimination cues. The first fixations landed equally often on each type of cue in the first block of Stage 1. Yet, across the remaining blocks of the stage, the frequency of the first fixations on the irrelevant cues decreased in favor of that on the relevant cues. This pattern was present in each group and each trial type. Accordingly, for the reversal trials, a Cue (irrelevant vs relevant) × Group (reversal vs control) × Block (1 – 4) ANOVA revealed a Cue × Block interaction, $F(3, 174) = 18.27$, $p < .001$, $\eta^2_p = .24$, with a similar frequency of first fixations for each type of cue in the first block, $t < 1$, but a lower frequency for the irrelevant cues than the relevant cues from the second to the fourth block, *t*s(59) > 4.42, *p*s < .001, *d*s > 0.57. Moreover, the linear trends were significant for both types of cues but moved towards opposite directions, *F*s(1, 59) > 39.08, *p*s < .001, η^2_ps > .39. The main effects of cue and block were also significant, *F*s > 2.95, *p*s < .04, η^2_ps > .04, but the main effect of group and the remaining interactions were non-significant, *F* < 1. For the non-reversal trials, a Cue (irrelevant vs relevant) × Group (reversal vs control) × Block (1 – 4) ANOVA also revealed a Cue × Block interaction, *F*(3, 174) = 11.01, *p* < .001, η^2_p = .16, again, with a similar frequency for each type of cue in the first block, *t* < 1, and a lower frequency for the irrelevant cues than the relevant cues from the second to the fourth block, *t*s(59) > 5.75, *p*s < .001, *d*s > 0.74. The linear trends were significant for both types of cues but moved towards opposite directions, *F*s(1, 59) > 15.99, *p*s < .001, η^2_ps > .21. The main effects of cue and block were also significant, *F*(1, 58) = 32.85, *p* < .001, η^2_p = .36, but the other main effects and interactions were non-significant, *F* < 1.

Following contingency change, the reversal group showed an increase in the frequency of first fixations on the irrelevant cues, together with a decrease in the frequency of first fixations on the relevant cues. The effect did not generalize from the reversal to the non-reversal trials. In the control group, the first fixations continued the pattern observed during Stage 1. For the reversal trials, a Cue (irrelevant vs relevant) × Group (reversal vs control) × Block (Stage 1, Block 4 vs Stage 2, Block 1) ANOVA revealed a three-way interaction, *F*(1, 58) = 18.25, *p* < .001, η^2_p = .24, together with a significant main effect of cue, *F*(1, 58) = 74.87, *p* < .001, η^2_p = .56. The other main effects and interactions were non-significant, *F*s < 2.23, *p*s > .14, η^2_ps < .04. In order to interpret the three-way interaction, we conducted a Group (reversal vs control) × Block (Stage 1, Block 4 vs Stage 2, Block 1) ANOVA for each type of cue. For the irrelevant cues, the ANOVA revealed a significant interaction, *F*(1, 58) = 13.42, *p* < .001, η^2_p = .19, with a similar frequency between groups in the last block of Stage 1, *t* < 1, but a higher frequency in the reversal group, compared to the control group, in the first block of Stage 2, *t*(58) = 3.69, *p* < .001, *d* = 0.94. Moreover, the frequency increased across the two blocks in the reversal group, *t*(29) = 2.58, *p* = .02, *d* = 0.47, but remained similar in the control group, *t*(29) = 2.64, *p* = .01, *d* = 0.48. The main effects of group and block were non-significant, *F*s(1, 58) < 2.72, *p*s > .10, η^2_ps < .05. For the relevant cues, the Group × Block ANOVA also revealed a significant interaction, *F*(1, 58) = 22.36, *p* < .001, η^2_p = .28, with a similar frequency between groups in the last block of Stage 1, $t < 1$, but a lower frequency in the reversal group, compared to the control group, in the first block of Stage 2, $t(58) = 3.53$, $p < .001$, $d = 0.90$. The frequency decreased across the two blocks in the reversal group, $t(29) = 3.66$, $p < .001$, $d = 0.67$, but remained similar in the control group, $t(29) = 2.99$, $p < .01$, $d = 0.55$. The main effects of group and block were non-significant, Fs < 1.63, $p > .20$, $\eta^2_p < .03$. The Group (reversal vs control) × Block (Stage 1, Block 4 vs Stage 2, Block 1) ANOVA for the non-reversal trials revealed a significant main effect of cue, with lower frequencies for the irrelevant cues than the relevant cues, $F(1, 58) = 33.85$, $p < .001$, $\eta^2_p = .37$. The main effect of block and the interaction were non-significant, Fs(1, 58) < 3.37, ps > .07, η^2_ps < .06.

In the remaining blocks of Stage 2, the frequency of first within-trial fixations was similar between groups. For the reversal trials, a Cue (irrelevant vs relevant) × Group (reversal vs control) × Block (1 – 4) ANOVA revealed a three-way interaction, $F(3, 174) = 5.55$, $p < .01$, $\eta^2_p = .09$, together with a significant main effect of cue, $F(1, 58) = 137.33$, $p < .001$, $\eta^2_p = .70$. The other main effects and interactions were non-significant, Fs < 2.33, ps > .07, η^2_ps < .04. In order to interpret the three-way interaction, we conducted a Group (reversal vs control) × Block (1 – 4) ANOVA for each type of cue. For the irrelevant cues, the ANOVA revealed a significant interaction, $F(3, 174) = 3.95$, $p < .01$, $\eta^2_p = .06$, indicating that the between-group difference did not extend beyond the first block, ts < 1. Moreover, the frequency decreased across blocks in the reversal group, linear trend: $F(1, 29) = 17.79$, $p < .001$, $\eta^2_p = .38$, but remained similar in the control group, linear trend: $F < 1$. The main effects of group and block were non-significant, Fs(3, 174) < 2.58, ps > .05, η^2_ps < .05. For the relevant cues, the Group × Block ANOVA also revealed a significant interaction, $F(3, 174) = 6.24$, $p < .001$, $\eta^2_p = .10$, again, indicating that the between-group difference did not extend beyond the first block, ts < 1. Moreover, the frequency increased across blocks in the reversal group, linear trend: $F(1, 29) = 17.00$, $p < .001$, $\eta^2_p = .37$, but remained similar in the control group, linear trend: $F < 1$. The main effects of group and block were non-significant, Fs(3, 174) < 1.80, ps > .14, η^2_ps < .03. The Cue (irrelevant vs relevant) × Group (reversal vs control) × Block (1 – 4) ANOVA for the non-reversal trials revealed a Cue × Block interaction, $F(3, 174) = 6.24$, p

$< .001$, $\eta^2_p = .10$. However, the frequency was lower for the irrelevant cues than the relevant cues in each block, *t*s(59) > 5.07, *p*s < .001, *d*s = 0.65. The main effect of cue was significant, $F(1, 58) = 85.43$, $p < .001$, $\eta^2_p = .60$, but the other main effects and interactions were non-significant, *F*s < 1.53, *p*s > .20, η^2_ps < .03.

From the second block of Stage 1, the irrelevant cues received fewer first within-trial fixations than the relevant cues. Contingency reversal, attenuated this effect by increasing the frequency of first fixations on the irrelevant cues, to the detriment of that for the relevant cues. The effect was limited to the block immediately following reversal and did not generalize to the non-reversal trials.

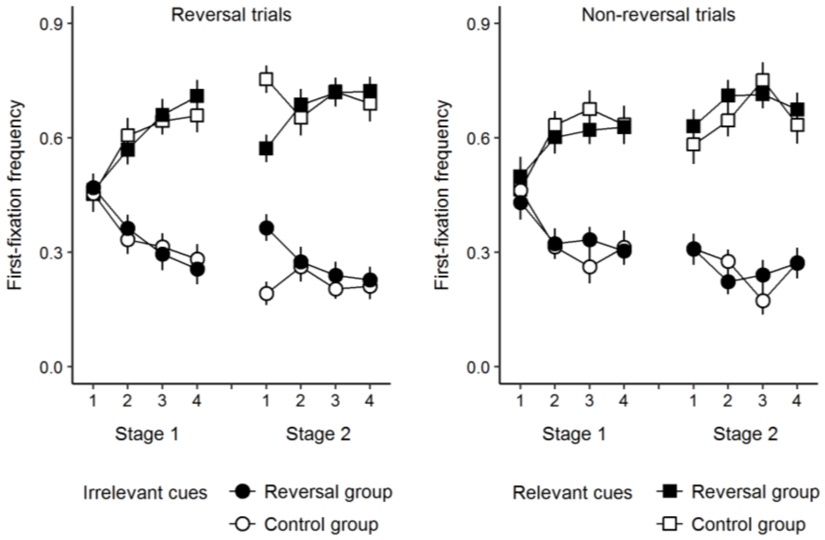

Figure 5. Mean frequency of the first within-trial fixations across blocks, separated by cue (irrelevant vs relevant) and group (reversal vs control) for the reversal and non-reversal trials (left and right panels, respectively). Error bars represent SEM.

Fixation count

Figure 6 shows the count of within-trial fixations on the discrimination cues. During Stage 1, the fixation count was lower and decreased faster for the irrelevant cues than the relevant cues. This pattern was observed in each group and each trial type. Accordingly, for the reversal trials, a Cue (irrelevant vs relevant) × Group (reversal vs control) × Block (1 – 4) ANOVA revealed a Cue × Block interaction, *F*(3, 174) = 12.15, $p < .001$, $\eta^2_p = .17$, with the fixation count decreasing across blocks for the irrelevant cues, *F*(3, 177) = 16.91, $p < .001$, $\eta^2_p = .22$ [linear trend: *F*(1, 59) = 41.33, $p < .001$, $\eta^2_p = .41$], but remaining similar for the relevant cues, *F*(3, 177) = 2.17, $p = .09$, $\eta^2_p = .04$. Moreover, in each block, the fixation count was lower for the irrelevant cues than the relevant cues, *t*s(59) > 5.03, *p*s < .001, *d*s > 0.63. The main effects of cue and block were also significant, *F*s > 17.47, *p*s < .001, η^2_ps > .17, but the main effect of group and the remaining interactions were non-significant, *F*s < 1.91, *p*s > .12, η^2_ps < .04. For the non-reversal trials, a Cue (irrelevant vs relevant) × Group (reversal vs control) × Block (1 – 4) ANOVA also revealed a Cue × Block interaction, *F*(3, 174) = 7.11, $p < .001$, $\eta^2_p = .11$, again, with the fixation count decreasing across blocks for the irrelevant cues, *F*(3, 177) = 11.45, $p < .001$, $\eta^2_p = .16$ [linear trend: *F*(1, 59) = 26.77, $p < .001$, $\eta^2_p = .31$], but remaining similar for the relevant cues, *F*(3, 177) = 1.52, $p = .22$, $\eta^2_p = .03$. In each block, the fixation count was lower for the irrelevant cues than the relevant cues, *t*s(59) > 5.28, *p*s < .001, *d*s > 0.68. The main effects of cue and block were also significant, *F*s > 8.82, *p*s < .001, η^2_ps > .13, but the main effect of group and the remaining interactions were non-significant, *F*s < 1.85, *p*s > .17, η^2_ps < .04.

After contingency change, the reversal group showed a larger fixation count than the control group. This appeared to be true for both types of cues and in both trial types. For the reversal trials, a Cue (irrelevant vs relevant) × Group (reversal vs control) × Block (Stage 1, Block 4 vs Stage 2, Block 1) ANOVA revealed a significant Group × Block interaction, *F*(1, 58) = 7.27, $p < .01$, $\eta^2_p = .11$, with a similar fixation count between groups in the last block of Stage 1, $t < 1$, but a larger fixation count in the reversal group, compared to the control group, in the first block of Stage 2, *t*(58) = 2.51, $p < .02$, $d = 0.66$. Moreover, the fixation count increased across the two blocks in the reversal group, *t*(29) = 2.53, $p = .02$, $d = 0.45$, but remained similar in the control group, *t*(29) = 1.16, $p = .26$, $d = 0.20$. These results, together with the fact that the three-way interaction was non-significant, *F*(1, 58) = 2.88, $p = .10$, $\eta^2_p = .05$, indicated that contingency reversal increased the fixation count for both the irrelevant and the relevant cues. Given our hypotheses, we conducted planned comparisons for each type of cue. For the irrelevant cues, the two groups showed a similar fixation count in the last block of Stage 1, $t < 1$, but the reversal group showed a larger fixation count than the control group in the first block of Stage 2, $t(58) = 2.87$, $p < .01$, $d = 0.74$. Across blocks, the fixation count increased in the reversal group, $t(29) = 2.05$, $p = .049$, $d = 0.37$, and showed no significant change in the control group, $t(29) = 2.04$, $p = .051$, $d = 0.37$, though it tended to a decrease. For the relevant cues, the two groups showed a similar fixation count in both the last block of Stage 1, $t < 1$, and the first block of Stage 2, $t(58) = 1.84$, $p = .07$, $d = 0.46$. However, the fixation count increased across blocks in the reversal group, $t(29) = 2.48$, $p = .02$, $d = 0.45$, and showed no significant change in the control group, $t < 1$. All in all, from the last block of Stage 1 to the first block of Stage 2, the reversal group showed an increase in the fixation count for each type of cue, while the control group showed no significant change. Going back to the ANOVA, the main effect of cue and the Cue × Group interaction were also significant, Fs(1, 58) > 4.54, ps < .04, η^2_ps > .07. The main effect of group was non-significant, $F(1, 58) = 3.53$, $p = .07$, $\eta^2_p = .06$, as were the main effect of block and the Cue × Block interaction, Fs(1, 58) < 1.85, ps > .17, η^2_ps < .04. For the non-reversal trials, a Cue (irrelevant vs relevant) × Group (reversal vs control) × Block (Stage 1, Block 4 vs Stage 2, Block 1) ANOVA revealed a significant main effect of cue, $F(1, 58) = 139.75$, $p < .001$, $\eta^2_p = .71$, with a lower fixation count for the irrelevant cues than the relevant cues. The main effect of group was non-significant, $F(1, 58) = 3.29$, $p = .08$, $\eta^2_p = .05$, as were the other main effects and interactions, Fs(1, 58) < 1.10, ps > .30, η^2_ps < .02.

In general, during Stage 2, the reversal group showed a larger fixation count than the control group. For the reversal trials, a Cue (irrelevant vs relevant) × Group (reversal vs control) × Block (1 – 4) ANOVA revealed a three-way interaction, $F(3, 174) = 3.27$, $p = .02$, $\eta^2_p = .05$, together with significant main effects of cue, group, and block, Fs > 5.31, ps < .03, η^2_ps > .08. The remaining interactions were non-significant, Fs < 1.80, ps > .14, η^2_ps < .04. In order to interpret the three-way interaction, we conducted a Group (reversal vs control) × Block (1 – 4) ANOVA for each type of cue. For the irrelevant cues, the ANOVA revealed a significant interaction, $F(3, 174) = 3.05$, $p = .03$, $\eta^2_p = .05$, with the reversal group showing a larger fixation count than the control group in the first and third blocks, ts(58) > 2.45, ps

< .02, *d* > 0.64, but not in the second and fourth, *t*s(58) < 1.54, *p*s > .13, *d* < 0.56.The main effects of group and block were also significant, *F*s > 4.82, *p*s < .04, η^2_ps > .07. For the relevant cues, a Group × Block ANOVA revealed a significant main effect of group, *F*(1, 58) = 4.71, *p* = .03, η^2_p = .08, with a larger fixation count in the reversal group than the control group. The main effect of block was also significant, *F*(1, 58) = 11.34, *p* < .001, η^2_p = .16. The interaction was non-significant, *F* < 1. For the non-reversal trials, a Cue (irrelevant vs relevant) × Group (reversal vs control) × Block (1 – 4) ANOVA revealed a significant main effect of group, *F*(1, 58) = 4.23, *p* = .04, η^2_p = .07, with a larger fixation count in the reversal group than the control group. The Cue × Block interaction was also significant, *F*(3, 174) = 8.05, *p* < .001, η^2_p = .12, as were the main effects of cue and block, *F*s > 5.47, *p*s < .01, η^2_ps > .08. The remaining interactions were non-significant, *F*s < 2.34, *p*s > .04, η^2_ps < .04.

The irrelevant cues received fewer fixations than the relevant cues, with the difference increasing during the task. In the first block of Stage 2, contingency reversal was followed by an increase in the fixation count for both types of cues. A larger fixation count in the reversal group, compared to the control group, was evident during most blocks of Stage 2. The difference between groups largely generalized to the non-reversal trials.

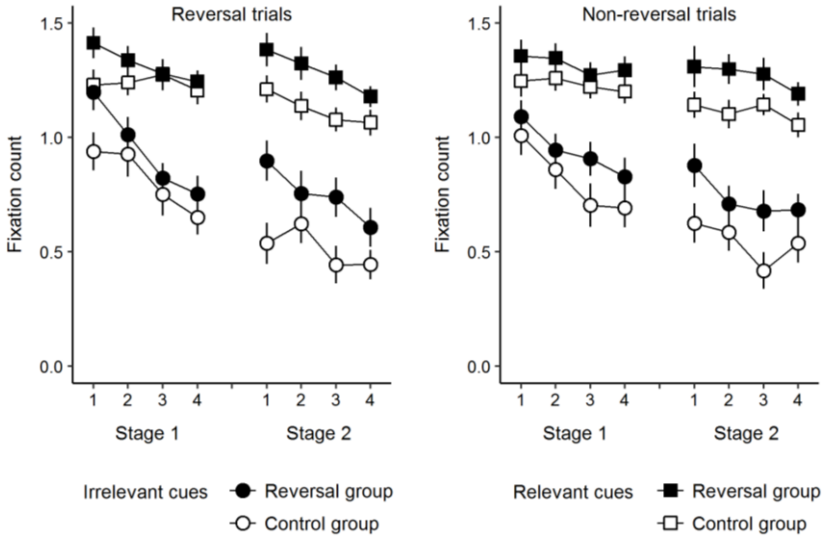

Figure 6. Mean count of within-trial fixations across blocks, separated by cue (irrelevant vs relevant) and group (reversal vs control) for the reversal and non-reversal trials (left and right panels, respectively). Error bars represent SEM.

Excluded participants

Figure 7 shows the attentional data from the 29 excluded participants, who did not learn to predict the outcomes (age: *M* = 22.97 years, *SD* = 2.68, 11 men; accuracy in the last block of Stage 1: *M* = 0.49, *SD* = 0.10). In order to see how deficient learning influenced attentional allocation during Stage 1, we conducted a Cue (irrelevant vs relevant) × Block (1 – 4) ANOVA for each attentional measure. The dwell time on the irrelevant cues was shorter than that on the relevant cues, $F(1, 28) = 26.10$, $p < .001$, $\eta^2_p = .48$. However, unlike the pattern observed among participants who learned to predict the outcomes, the difference between cues did not increase across blocks, Cue × Block: $F < 1$. The first within-trial fixations landed equally often on each type of cue, $F < 1$. This pattern contrasts with the one observed among the other participants, in which a clear preference for fixating first on the relevant cues developed during Stage 1. For the fixation count, main effect of block was significant, $F(3, 84) = 3.00$, $p = .04$, $\eta^2_p = .10$, but no significant linear or quadratic trends were evident, *F*s (1, 28) < 2.39, *p*s > .13, η^2_ps < .08. The other main effects and interactions in the ANOVAs were non-significant, *F*s < 2.44, *p*s > .06, η^2_ps < .09.

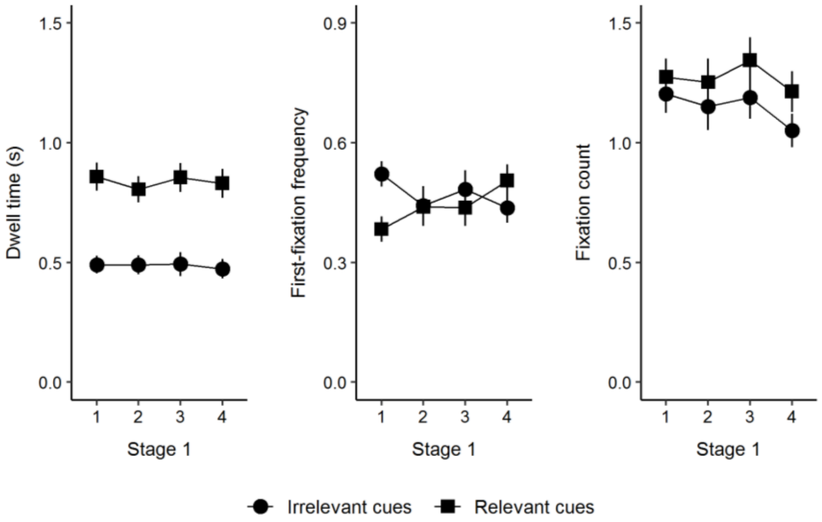

Figure 7. Attentional data for the participants who did not learn to predict the outcomes during Stage 1. The data focus on the fixations directed to the irrelevant and the relevant cues during the within-trial interval, across blocks. Error bars represent SEM.

Discussion

We investigated whether a rise in prediction error would lead to an increase in attention to cues that were irrelevant for solving a discrimination task. The accuracy of outcome predictions was taken as an indicator of prediction error, assuming an inverse relationship between them. As participants became more and more accurate in predicting the outcomes during Stage 1, they showed a decrease in attention to the irrelevant cues. At the beginning of Stage 2, half of participants experienced a reversal of the outcomes associated with two relevant cues, and therefore they showed a drop in accuracy that largely generalized to non-reversal trials. Moreover, contingency reversal led to the expected increase in attention to the irrelevant cues, with a rise in the first within-trial fixations and the fixation count. These results are consistent with the predictions of attentional theories of associative learning. According to Mackintosh's (1975) theory, the associability of a cue – a proxy for attention – increases on those trials where the cue predicts the outcome better than other stimuli, i.e. generates less prediction error, otherwise it decreases. After contingency reversal, our participants experienced outcomes different from those expected on the basis of the relevant cues. Thus, until participants learned the new contingencies, the irrelevant cues were – although imperfect – the best available predictors. Therefore, the amount of attention directed to those cues should have increased to the expense of the attention paid to the relevant cues. On the other hand, the model by Pearce and Hall (1980) suggests that associability increases to the extent that the cues presented on a given trial are followed by a surprising outcome. Thus, the large prediction error resulting from contingency reversal should have increased the amount of attention directed to both the relevant and the irrelevant cues.

Contingency reversal had no influence on the dwell time on either type of cue. This might be related to a particular aspect of our task. In most studies in the area of attention and associative learning, the trial duration is self-paced, and the interval for measuring attention lasts from stimulus onset to the time when participants make a response. The dwell time is then expressed in proportion of response time for standardization purposes (e.g., Beesley, Nguyen, Pearson, & Le Pelley, 2015; Kruschke, Kappenman, & Hetrick, 2005). We reasoned that the proportional dwell time might leave out meaningful changes in absolute attention (i.e., if prediction error produced a similar increase in attention to all the present cues, as predicted by the Pearce-Hall model). In order to circumvent this potential limitation, and standardize the overall fixation time across trials and participants, we used a fixed interval. This choice could have limited the sensitivity of dwell time. Moreover, on the non-reversal trials, in the blocks immediately preceding and following contingency change on the trials presenting A or B, the reversal group showed a longer dwell time on the irrelevant cues than the control group. This could be taken to indicate that our experimental groups showed some baseline differences in this attentional measure.

The present study was partly inspired by Vadillo et al. (2016), who used a dot-probe paradigm to measure attentional allocation while participants solved a discrimination task. A score on attentional preference was computed by subtracting the reaction times recorded when the probe appeared on a relevant cue from the reaction times recorded when the probe appeared on an irrelevant cue. The score was positive, thus showing a bias towards the relevant cue. In a subsequent stage, half of participants experienced contingency reversal, while the other half did not. Relative to the latter participants, the former then showed a decrease both in response accuracy and the attentional score. This was interpreted as evidence for an attentional shift from the relevant cue to the irrelevant one resulting from prediction error. In agreement with this suggestion, we found that the increase in the first fixations on the irrelevant cues after contingency reversal was accompanied by a decrease in the first fixations on the relevant cues. It is interesting to note that Vadillo and colleagues presented the probe 250 ms after stimulus onset. Therefore, both the reaction times to the probe and the first fixations measured early attentional processing. Vadillo et al. showed that the decrease in the attentional score did not extend to trials where the same irrelevant cue was presented together with relevant cues whose outcomes remained unchanged. Similarly, we found no effect on the first fixations on an irrelevant cue that was not involved in contingency reversal. Thus, although the study by Vadillo et al. and ours used different attentional measures, the results from both studies suggest that a sudden rise in prediction error (indicated by a drop in accuracy) is associated with a fast and stimulus-specific attentional shift from the relevant to the irrelevant cues. Easdale, Le Pelley, and Beesley (2019) showed how a shift in attention driven by sudden prediction error may facilitate learning. Half of participants in their study experienced a deterministic relationship between pairs of relevant and irrelevant cues, on the one hand, and outcomes, on the other. The other half of participants experienced a probabilistic relationship, i.e. each pair was followed by its corresponding outcome on 80% of the trials and by an alternative outcome on the remaining 20%. Thus, only participants in the former group achieved a maximum accuracy in outcome predictions. In a subsequent stage, some of the outcomes were switched, so that the irrelevant cues became relevant and vice versa. Given the differences in initial training, the deterministic group showed a drop in accuracy that was accompanied by a shift in attention from the previously relevant cues to those previously irrelevant, as measured by eye-tracking. These changes were absent in the probabilistic group. Importantly, the former group learned the new contingencies faster than the latter.

Our data showed an interesting dissociation between the first within-trial fixations and the fixation count regarding the relevant cues. Contingency reversal led to a decrease in the first fixations on those cues, to the advantage of the irrelevant ones, but also to an increase in the fixation count for both types of cues. Mackintosh's theory is able to account for the pattern observed in the first fixations, including the attentional preference for the relevant cues that existed throughout the task. On the other hand, the effect on the fixation count fits better with the role of prediction error postulated in the Pearce-Hall model. This specific link between theories and measures can be understood in terms of differences in competition for attention and selectivity. Mackintosh (1975) implemented changes in associability in a way that an increase in attention to the best predictor should be accompanied by a decrease in attention to the other cues. Thus, his theory emphasizes competition for attention between cues presented concurrently. This characteristic fits well with the fact that the first within-trial fixation can be directed to only one cue, which is thought to be the stimulus with the largest attentional signal (winner-takes-all; Theeuwes, Vries, & Godjin, 2003). On the other hand, Pearce and Hall (1980) suggested that the associability changes were similar for each of the cues present on a given trial. Since the fixation count had a wider temporal scope, participants were able to pay attention to more than one stimulus per trial. Thus, it allowed to distribute the effect of prediction error across different cues, in line with the suggestion from Pearce and Hall. The observation of different response patterns at different times within trials underscores the importance of taking into account within-trial time in the evaluation of associative learning theories (see Lachnit et al., 2013). All in all, each theory anticipates part of our results, but neither can explain the whole pattern.

Given the fundamental differences between Mackintosh's theory and the Pearce-Hall model, empirical studies have traditionally sought to support one theory while disproving the other. Yet, abundant evidence has accumulated in favor of each (for reviews, see Le Pelley, 2004; Pearce & Mackintosh, 2010). Such a state of affairs motivated the development of hybrid models (George & Pearce, 2012; Le Pelley, 2004; Pearce & Mackintosh, 2010), in which the attentional changes taking place during learning result from an interaction between two associability processes, one characterized by

Mackintosh's (1975) suggestion and the other by that of Pearce and Hall (1980). For instance, in Le Pelley's hybrid model (2004), the individual prediction errors of the concurrent stimuli are compared with each other, with the cue generating the smallest prediction error potentially getting more attention than the rest. On the other hand, overall prediction error, which is derived from the outcome prediction based on the whole set of cues, exerts an effect on attention in proportion to its magnitude, so that it tends to produce a gain when the outcome is surprising and a loss when the outcome is correctly anticipated. The tendencies resulting from both processes multiply together to determine attention deployment. Le Pelley's hybrid model is thus equipped to explain both the weakening of attentional preference for the relevant cues, and the increased attention to both types of cues, following contingency reversal, particularly if one allows the first fixations and the fixation count to be each more sensitive to one of the two associability process (as suggested above). Moreover, the model also accounts for the attentional changes occurring throughout the task, which were characterized by both the exploitation of the relevant cues as accuracy increased and the exploration of the irrelevant cues following the drop in accuracy (see Beesley et al., 2015). Our results add to recent evidence from both animal and human studies lending support to hybrid models (Beesley et al., 2015; Easdale et al., 2019; Haselgrove, Esber, Pearce, & Jones, 2010; Luque, Vadillo, Le Pelley, & Beeseley, 2017; Torrents-Rodas, Koenig, Uengoer, & Lachnit, 2020; Walker, Luque, Le Pelley, & Beesley, 2019).

In addition to our primary aim, we wanted to explore whether the expected increase in attention would generalize to an irrelevant cue that was not associated with prediction error, i.e. present on non-reversal trials. We presumed that accuracy would be relatively high on those trials. Yet, the drop in accuracy following contingency reversal was evident in both trial types. Having said that, the first within-trial fixations showed different patterns in each trial type, indicating that the increase in attention did not generalize to the irrelevant cue present on the non-reversal trials. Conversely, the effect on the fixation count was evident for the cues presented in both trial types. This latter finding is in line with the suggestion that, once a cue generates a strong prediction error, attention increases for all the contexts constituting the learning task, even those in which prediction error does not occur (Rosas & Callejas-

Aguilera, 2006). However, considering the whole set of results, we cannot draw a clear conclusion on this question. The observed generalization of accuracy across trial types could result from acquired equivalence, an effect by which two cues that are followed by the same outcome subsequently show retarded discrimination learning, i.e. increased generalization, when they come to be paired with different outcomes (e.g., Hall, Mitchell, Graham, & Lavis, 2003). In Stage 1 of our experiment, A and C were followed by o1, and B and D were followed by o2. Thus, on the basis of the outcomes trained during Stage 1, the cues can be thought of as belonging to two categories. According to acquired equivalence, at the outset of Stage 2, participants would tend to expect the same outcome for the cues belonging to the same category. Given that the outcomes actually differed within categories, participants would then have made wrong predictions in both the reversal and the non-reversal trials. Generalized accuracy could also be explained by the change in cue combinations at the beginning of Stage 2 (i.e., AX, BX, CY, and DY in Stage 1 vs AY, BY, CX, and DX in Stage 2). The new cue combinations came to signal the contingency change on the reversal trials. According to a rule-based generalization account (e.g., Lachnit, Kinder, & Reinhard, 2002; Lachnit & Lober, 2001; Lachnit, Lober, Reinhard, & Kinder, 2001), this would have led participants to generalize the response switch to the non-reversal trials, because those trials also showed new cue combinations.

Previous research has shown that the attentional changes resulting from learning may be automatic, i.e. outside voluntary control (for reviews, see Anderson, 2015; Awh, Belopolsky, & Theeuwes, 2012). In a typical study, a cue becomes associated with a reward or a task-relevant outcome during initial training. The cue is then presented as a distractor in a second task, so that paying attention to it interferes with the goal of responding efficiently to the target stimulus. In this context, increased response times to the target (Anderson, Laurent, & Yantis, 2011a; Le Pelley, Vadillo, & Luque, 2013; Luque, Molinero, Jevtovic, & Beesley; 2020), or increased capture of the first within-trial fixations by the distractor (Anderson & Yantis, 2012; Theeuwes & Belopolsky, 2012), are taken as evidence for an automatic bias towards the distractor driven by previous learning. These effects are reminiscent of those produced by physically salient stimuli (Theeuwes, 1991, 1992; Theeuwes, Vries, & Godjin, 2003).

Recently, Koenig et al. (2017) suggested that cues generating sustained prediction error may also exert attentional effects automatically (see also Luque et al., 2017). In the present study, the attentional changes following contingency reversal were presumably linked to a sudden rise in prediction error. Therefore, one might wonder whether those attentional changes had an automatic component. In this regard, it is noteworthy that the increase in the frequency of first within-trial fixations reflected an early-processing effect, whereas the larger fixation count was based on the whole measurement interval. It has been shown that automatic effects occur quickly after stimulus onset, whereas goal-directed processes come into play later (see Failing & Theeuwes, 2018). Thus, one could speculate that the effect on the first fixations reflects an automatic component, whereas that on the fixation count results from an effort to regain accuracy in predicting the outcomes. However, no incompatibility existed between the task goal and paying attention to the irrelevant cues, and therefore our task did not allow to distinguish between automatic and voluntary deployment of attention. Finally, the strong physical salience of the irrelevant cues might have added to the effects of prediction error to weaken the attentional priority awarded to relevant cues (Anderson, Laurent, & Yantis, 2011a; Folk, Remington, & Johnston, 1992).

The present study strengthened the evidence showing that a sudden rise in prediction error leads to an increase in attention to cues that otherwise tend to be ignored because they do not predict significant events. These results were anticipated by both Mackintosh's (1975) theory and the Pearce-Hall (1980) model. In addition, hybrid models (Le Pelley, 2004; Pearce & Mackintosh, 2010) went further in explaining the whole set of findings, because they accounted also for the data regarding the relevant cues and the attentional changes observed throughout the task. We should exercise some caution in generalizing the present results, because the manipulations intended to increase the salience of the irrelevant cues (i.e., using color singletons and changing the cue combinations) might have interacted with prediction error to produce the attentional effects.

Funding

Our study was supported by the Deutsche Forschungsgemeinschaft (DFG, German Research Foundation), project number 290878970-GRK 2271 project 5, as well as project number 222641018–SFB/TRR 135 project B4.

References

Anderson, B. A. (2015). The attention habit: how reward learning shapes attentional selection. *Annals of the New York Academy of Sciences, 1369,* 24–39, https://doi.org/10.1111/nyas.12957

Anderson, B. A., Laurent, P. A., & Yantis, S. (2011a). Value-driven attentional capture. *Proceedings of the National Academy of Sciences, USA, 108*, 10367–1037, https://doi.org/110.1073/pnas.1104047108

Anderson, B. A., Laurent, P. A., & Yantis, S. (2011b). Learned value magnifies salience-based attentional capture. *PLoS One, 6*, 1–6, https://doi.org/10.1371/journal.pone.0027926

Anderson, B. A., & Yantis, S. (2012).Value-driven attentional and oculomotor capture during goal-directed, unconstrained viewing. *Attention, Perception, & Psychophysics, 74*, 1644–1653, https://doi.org/10.3758/s13414-012-0348-2

Awh, E., Belopolsky, A. V., & Theeuwes, J. (2012). Top-down versus bottom-up attentional control: a failed theoretical dichotomy. *Trends in Cognitive Sciences, 16*, 437–443, https://doi.org/10.1016/j.tics.2012.06.010

Beesley, T., Nguyen, K. P., Pearson, D., & Le Pelley, M. E. (2015). Uncertainty and predictiveness determine attention to cues during human associative learning. *Quarterly Journal of Experimental Psychology, 68*, 2175–99, https://doi.org/10.1080/17470218.2015.1009919

Easdale, L. C., Le Pelley, M. E., & Beesley, T. (2019). The onset of uncertainty facilitates the learning of new associations by increasing attention to cues. *Quarterly Journal of Experimental Psychology, 72*, 193–208, https://doi.org/10.1080/17470218.2017.1363257

Failing, M., & Theeuwes, J. (2018). Selection history: how reward modulates selectivity of visual attention. *Psychonomic Bulletin & Review, 25*, 514–538, https://doi.org/10.3758/s13423-017-1380-y

Folk, C. L., Remington, R. W., & Johnston, J. C. (1992). Involuntary covert orienting is contingent on attentional control settings. *Journal of Experimental Psychology: Human Perception and Performance, 18*, 1030–44. https://doi.org/10.1037/0096-1523.18.4.1030

George, D. N., & Pearce, J. M. (2012). A configural theory of attention and associative learning. *Learning & Behavior, 40*, 241–254, https://doi.org/10.3758/s13420-012-0078-2

Gottieb, J. (2012). Attention, learning, and the value of information. *Neuron, 76*, 281–295, https://doi.org/10.1016/j.neuron.2012.09.034

Greenhouse, S. W., & Geisser, S. (1959). On methods in the analysis of profile data. *Psychometrika, 24*, 95–112. https://doi.org/10.1007/BF02289823

Hall, G., Mitchell, C., 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, https://doi.org/10.1037/0096-3445.132.2.266

Haselgrove, M., Esber, G. R., Pearce, J. M., & Jones, P. M. (2010). Two kinds of attention in Pavlovian conditioning: evidence for a hybrid model of learning. Journal of Experimental Psychology: Animal Behavior Processes, 36, 456–470, https://doi.org/10.1037/a0018528

Hogarth, L., Dickinson, A., Austin, A., Brown, C., & Duka, T. (2008). Attention and expectation in human predictive learning: the role of uncertainty. *Quarterly Journal of Experimental Psychology, 61*, 1658–68, https://doi.org/10.1080/17470210701643439

Koenig, S. (2010). *Modulation of saccadic curvature by spatial memory and associative learning* (Doctoral dissertation). Retrieved from http://archiv.ub.uni-marburg.de/diss/z2010/0636/

Koenig, S., Kadel, H., Uengoer, M., Schubö, A., & Lachnit, H. (2017). Reward draws the eye, uncertainty holds the eye: associative learning modulates distractor interference in visual search. *Frontiers in Behavioral Neuroscience, 11*, 1–15, https://doi.org/10.3389/fnbeh.2017.00128

Koenig, S., & Lachnit, H. (2011). Curved saccade trajectories reveal conflicting predictions in associative learning. *Journal of Experimental Psychology: Learning, Memory and Cognition, 37*, 1164–77, https://doi.org/10.1037/a0023718

Koenig, S., Uengoer, M., & Lachnit, H. (2017). Attentional bias for uncertain cues of shock in human fear conditioning: evidence for attentional learning theory. *Frontiers in Human Neuroscience, 11*. 1–13, https://doi.org/10.3389/fnhum.2017.00266

Kruschke, J. K. (2001). Toward a unified model of attention in associative learning. *Journal of Mathematical Psychology, 45*, 812–863, https://doi.org/10.1006/jmps.2000.1354

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, https://doi.org/10.1037/0278-7393.31.5.830

Lachnit, H., Kinder, A., & Reinhard, G. (2002). Are rules applied in Pavlovian electrodermal conditioning with humans general or outcome specific? *Psychophysiology, 39*, 380–387, https://doi.org/10.1017/S0048577201393125

Lachnit, H., & Lober, K. (2001). What is learned in patterning discrimination? Further tests of configural accounts of associative learning in human electrodermal conditioning. *Biological Psychology, 56*, 45–61, https://doi.org/10.1016/S0301-0511(00)00087-9

Lachnit, H., Lober, K., Reinhard, G., & Kinder, A. (2001). Evidence for the application of rules in Pavlovian electrodermal conditioning with humans. *Biological Psychology, 56*, 151–166, https://doi.org/10.1016/S0301-0511(01)00067-9

Lachnit, H., Thorwart, A., Schultheis, H., Koenig, S., Lotz, A., & Uengoer, M. (2013). Indicators of early and late processing reveal the importance of within-trial-time for theories of associative learning. *PLoS One, 8*, e66291. https://doi.org/10.1371/journal.pone.0066291

Le Pelley, M. E. (2004). The role of associative history in models of associative learning: a selective review and a hybrid model. *Quarterly Journal of Experimental Psychology, 57B*, 193–243, https://doi.org/10.1080/02724990344000141

Le Pelley, M. E., & McLaren, I. P. L., (2003). Learned associability and associative change in human causal learning. *The Quarterly Journal of Experimental Psychology, 56B*, 68–79, https://doi.org/10.1080/02724990244000179

Le Pelley, M. E., Mitchell, C. J., Beesley, T., George, D. N., & Wills, A. J. (2016). Attention and associative learning in humans: an integrative review. *Psychological Bulletin, 142*, 1111–40, https://doi.org/10.1037/bul0000064

Le Pelley, M. E., Vadillo, M., & Luque, D. (2013). Learned predictiveness influences rapid attentional capture: evidence from the dot probe task. *Journal of Experimental Psychology: Learning, Memory, & Cognition, 39*, 1888–1900, https://doi.org/10.1037/a0033700

Luque, D., Molinero, S., Jevtovic, M., Beesley, T. (2020). Testing the automaticity of an attentional bias towards predictive cues in human associative learning. *Quarterly Journal of Experimental Psychology, 73*, 762–780, https://doi.org/10.1177/1747021819897590

Luque, D., Vadillo, M. A., Le Pelley, M. E., & Beeseley, T. (2017). Prediction and uncertainty in associative learning: examining controlled and automatic components of learned attentional biases. *Quarterly Journal of Experimental Psychology, 70*, 1485–1503, https://doi.org/10.1080/17470218.2016.1188407

Mackintosh, N. J. (1975). A theory of attention: variations in the associability of stimuli with reinforcement. *Psychological Review, 82*, 276–98, https://doi.org/10.1037/h0076778

Pearce, J. M., & Hall, G. (1980). A model for pavlovian learning: variations in the effectiveness of conditioned but not of unconditioned stumuli. *Psychological Review, 87*, 532–552, https://doi.org/10.1037/0033-295X.87.6.532

Pearce, J. M., & Mackintosh, N. J. (2010). Two theories of attention: a review and a possible integration. In C. J. Mitchell & M. E. Le Pelley (Eds.), *Attention and associative learning: from brain to behaviour* (pp. 11–40). Oxford, UK: Oxford University Press.

Rief, W., Glombiewski, J. A., Gollwitzer, M., Schubö, A., Schwarting, R., & Thorwart, A. (2015). Expectancies as core features of mental disorders. *Current Opinion in Psychiarty, 28*, 378–385, https://doi.org/10.1097/YCO.0000000000000184

Rosas, J. M., & Callejas-Aguilera, J. E. (2006). Context switch effects on acquisition and extinction in human predictive learning. *Journal of Experimental Psychology: Learning, Memory, and Cognition, 32*, 461–474, https://doi.org/10.1037/0278-7393.32.3.461

Rosas, J. M., Callejas-Aguilera, J. E., Ramos-Álvarez, M. M., Fernández-Abad, M. J. (2006). Revision of retrieval theory of forgetting: what does make information context-specific? *International Journal of Psychology and Psychological Therapy, 6*, 147–66.

Torrents-Rodas, D., Koenig, S., Uengoer, M., & Lachnit, H. (2020). *Evidence for an interaction of two attentional mechanisms during learning*. Manuscript under review.

Theeuwes, J. (1991). Cross-dimensional perceptual selectivity. *Perception & Psychophysics, 50*, 184–193, https://doi.org/doi.org/10.3758/BF03212219

Theeuwes, J. (1992). Perceptual selectivity for color and form. *Perception & Psychophysics, 51*, 599–606, https://doi.org/doi.org/10.3758/BF03211656

Theeuwes, J., & Belopolsky, A. V. (2012). Reward grabs the eye: oculomotor capture by rewarding stimuli. *Vision Research, 74*, 80–85, https://doi.org/10.1016/j.visres.2012.07.024

Theeuwes, J., De Vries, G. J., & Godijn, R. (2003). Attentional and oculomotor capture with static singletons. *Pereception & Psychophysics, 65*, 735–746, https://doi.org/10.3758/BF03194810

Vadillo, M. A., Orgaz, C., Luque, D., & Byron Nelson, J. (2016). Ambiguity produces attention shifts in category learning. *Learning and Memory, 23*, 134–40, https://doi.org/10.1101/lm.041145.115

Walker, A. R., Luque, D., Le Pelley, M. E., & Beesley, T. (2019). The role of uncertainty in attentional and choice exploration. *Psychonomic Bulletin & Review, 26*, 1911–1916, https://doi.org/10.3758/s13423-019-01653-2

Evidence for an interaction of two attentional mechanisms during learning

David Torrents-Rodas [a], Stephan Koenig [a, b], Metin Uengoer [a], and Harald Lachnit [a]

[a] Faculty of Psychology, Philipps-Universität Marburg, Marburg, Germany

[b] Faculty of Psychology, Universität Koblenz-Landau, Landau, Germany

Corresponding author: David Torrents-Rodas, Faculty of Psychology, Philipps-Universität Marburg, Gutenbergstr. 18, 35032 Marburg, Germany. torrents@uni-marburg.de

Abstract

We sought to provide evidence for an interaction of two attentional mechanisms during associative learning. Participants' eye movements were recorded as they predicted the outcomes following different pairs of cues. Across the trials of an initial stage, a relevant cue in each pair was consistently followed by one of two outcomes, while an irrelevant cue was equally followed by either of them. Thus, the relevant cue should have been associated with small relative prediction errors, compared to the irrelevant cue. In a later stage, each pair came to be followed by one outcome on a random half of the trials and by the other outcome on the remaining half, and thus there should have been a rise in the overall prediction error. Consistent with an attentional mechanism based on relative prediction error, an attentional advantage for the relevant cue was evident in the first stage. On the other hand, in accordance with a mechanism linked to overall prediction error, the attention paid to both types of cues increased at the beginning of the second stage. These results showed up in both dwell times and within-trial patterns of fixations, and they were predicted by a hybrid model of attention.

Keywords: attention, associative learning, eye-tracking, discrimination learning, partial reinforcement

Introduction

Both animals and humans use environmental cues to predict the occurrence of significant outcomes and to engage in the appropriate responses. According to theories of associative learning, the underlying process is based on the development of associations between the neural representations of the cues and those of their related outcomes (Pearce & Bouton, 2001). Attention enhances the processing of cues being significant for adaptation (e.g., those with bright colors or abrupt onsets). In models of associative learning, the features of a cue influencing attention are captured by an "associability" parameter, α, modulating the speed at which the cue-outcome associations are formed. In recent years, research using eye-tracking has contributed to validate the analogy between associability and attention by showing that changes in overt visual attention closely match the predictions made for associability (for a review, see Le Pelley, Mitchell, Beesley, George, & Wills, 2016).

Attentional theories of associative learning postulate that associability is not a fixed property of cues, or merely based on their physical features. Instead, associability changes in accordance to prediction error, which is the discrepancy between the magnitude of an outcome and the extent to which it is predicted by one or more cues. The overall prediction error experienced on a trial is given by the term

$$\lambda - V_{\mathrm{T}} \tag{1}$$

where λ indicates the magnitude of the outcome, e.g. taking the values of 1 or 0 to code for its presence or absence, and V_{T} represents the strength of the association between the cues and the outcome (e.g., Pearce & Hall, 1980; Rescorla & Wagner, 1972). Typically, V_{T} is 0 at the outset of learning, indicating that the cues have never been experienced together with the outcome. But through repeated pairings, V_{T} is driven towards the maximum value, usually set to 1. The specific way in which prediction error is assumed to change associability depends on the theory under consideration.

In his attentional theory, Mackintosh (1975) suggested that the associability of a cue S increases when it signals an outcome more accurately than the other present stimuli (T-S). Otherwise, when S is equally or less accurate than T-S, its associability decreases. Mackintosh expressed this idea with the following rules:

$$\Delta\alpha_S \text{ is positive, if } |\lambda - V_S| < |\lambda - V_{T\text{–}S}| \tag{2.1}$$

$$\Delta\alpha_S \text{ is negative, if } |\lambda - V_S| \geq |\lambda - V_{T\text{–}S}| \tag{2.2}$$

Thus, the change of associability, Δα, is based on a comparison of the prediction error of each cue, e.g. S, with that of all the other cues, T–S. The cue producing the smallest relative prediction error gets higher associability than the rest. Strong support for Mackintosh's theory comes from discrimination paradigms where some cues are relevant for predicting the outcomes, while others are irrelevant (e.g., George & Pearce, 1999; Le Pelley, Beesley, & Griffiths, 2011; Le Pelley & McLaren, 2003; Le Pelley, Vadillo, & Luque, 2013; Lochmann & Wills, 2003). Take for instance the study conducted by Le Pelley et al. (2011), who asked participants to solve a discrimination where the occurrence of different sounds could be predicted on the basis of nonsense words. On each trial, participants saw a pair of words before hearing one of two possible sounds. The pairs were arranged so that, across trials, one word was consistently followed by the same sound, whereas the other word was followed by each of the two sounds on an equal number of occasions. These contingencies rendered the words in the former class relevant for the prediction of the sounds – eventually becoming strongly associated with a specific sound and generating small prediction errors, as inferred from the percentage of correct predictions, while those words in the latter class were irrelevant – presumably acquiring weak associations with both sounds and generating large prediction errors throughout the task. Eye-tracking data showed that participants paid more attention to the relevant words than to the irrelevant ones. Moreover, the attentional bias persisted during a subsequent discrimination where the same words were paired with two new sounds. Importantly, in the second discrimination the words in each pair were equally relevant for predicting the sounds. A final test of associability also indicated that the previously relevant words acquired stronger associations with the sounds of the second discrimination, compared to the previously irrelevant words.

In contrast to Mackintosh's theory, the Pearce-Hall (1980) model proposed that decreasing prediction error leads to a reduction in associability. Thus, the associability of a cue is directly related to the prediction error generated by all the present cues on the most recent trial:

$$\alpha_S = |\lambda^{n-1} - V_T{}^{n-1}| \qquad (3)$$

Evidence in favor of the Pearce-Hall model typically comes from research using partial reinforcement schedules (e.g., Hogarth, Dickinson, Austin, Brown, & Duka, 2008; Kaye & Pearce, 1984; Koenig, Kadel, Uengoer, Schubö, & Lachnit, 2017; Koenig, Uengoer, & Lachnit, 2017, 2018; Swan & Pearce, 1988). For instance, Hogarth et al. (2008) asked participants to predict the occurrence of a tone based on different visual patterns. One of those patterns was followed by the tone on every trial, i.e. continuously reinforced cue. Another visual pattern was followed by the tone on a randomly selected half of the trials, i.e. partially reinforced cue. And, finally, a third pattern was never paired with the tone, i.e. non-reinforced cue. As indicated by trial-by-trial ratings, most participants acquired differential expectancies of the tone that matched the actual contingencies. Specifically, participants came to expect the tone during the continuously reinforced cue and to expect its omission during the non-reinforced cue. Thus, prediction error should have been at a minimum in both of these conditions. On the other hand, participants indicated intermediate tone expectancies during the partially reinforced cue. Yet, since the outcome of a particular trial was either the occurrence of the tone or its absence, the intermediate expectancies always bore some degree of discrepancy with the outcome. Therefore, this condition should have generated a relatively large prediction error. The time that participants spent looking at each type of cue (dwell time) was compared to the dwell time on a control cue that was present on every trial. These relative dwell times showed that the partially reinforced cue received more attention than both the continuously reinforced cue and the non-reinforced cue (for related findings using speed of learning to measure associability, see Griffiths, Johnson, & Mitchell, 2011).

Despite the difference in the mechanisms of associability change suggested by Mackintosh (1975) and Pearce and Hall (1980), ample evidence exists for each of them (for a review, see Pearce &

Mackintosh, 2010). Thus, Le Pelley (2004, 2010) developed a hybrid model in which both mechanisms interact to determine changes in associability. On a given trial, the individual prediction error of each cue is compared to the combined prediction error of all the other stimuli, thereby the "Mackintosh associability" increases for the cue with the smallest relative prediction error, while it decreases for the rest. In addition, the "Pearce-Hall associability" is obtained from the prediction error produced by the summed associative strengths of all the stimuli present. For each cue, these two values are multiplied to determine the overall associability. The attentional pattern observed on a given study would be consistent with Mackintosh's mechanism, if the differences in prediction error concern cues that are presented simultaneously, as in a discrimination with relevant and irrelevant cues. Conversely, the Pearce-Hall mechanism would prevail, if those differences are associated with stimuli presented on separate trials, as with continuously vs partially reinforced cues. Beesley et al. (2015) termed the attentional patterns corresponding to each mechanism as 'exploitation' versus 'exploration'. Attentional exploitation operates when the learning agent can use one of the environmental cues to predict the outcome accurately, in which case that cue is selected by attention. Attentional exploration is engaged in situations where it is uncertain whether the present cues are followed by the outcome, and thus the amount of attention paid to potentially informative cues is kept high. Notwithstanding the wider scope of the hybrid model, only few empirical studies have focused on it (Haselgrove, Esber, Pearce, & Jones, 2010; Kattner, 2015; Le Pelley, Trunbull, Reimers, & Knipe, 2010) and, among those, even less have measured overt attention (Besleey et al., 2015; Easdale, Le Pelley, & Beesley, 2019). Thus, the aim of the present study was to obtain further evidence for changes in overt attention consistent with an interaction of the two associability mechanisms.

Our design included a discrimination stage followed by a partial reinforcement stage. Cues A, B, X, and Y were presented in pairs throughout the experiment, and the participants' task was to predict which of two outcomes, o1 or o2, would follow each pair. During discrimination training, pairs AX and AY were followed by o1, while BX and BY were followed by o2. Given that A and B were consistently paired with o1 and o2 respectively, they were relevant for predicting the outcomes. Therefore, they should be associated with a small prediction error. Conversely, X and Y were paired with both outcomes, each on half of the trials, and were thus irrelevant for making the predictions. Those latter cues should be associated with a relatively large prediction error. During partial reinforcement, each of the four pairs was followed by either o1 or o2, on an equal number of randomly distributed trials. Thus, in this stage participants should not be able to make consistently accurate predictions, and prediction error should become large for the pairs and the individual cues.

Figure 1 shows simulations of the predicted changes in associability based on the above-mentioned theories (see the Appendix for details). Mackintosh’s (1975) theory (upper panel) predicts that, during discrimination, the relevant cues should receive more attention than the irrelevant ones, and that this differential allocation should decrease during partial reinforcement. The Pearce-Hall model (Pearce, Kaye, & Hall 1981; middle panel) predicts that both types of cues should receive a similar amount of attention throughout the experiment. Moreover, attention should decrease during discrimination, and it should be reinstated by partial reinforcement. Finally, according to Le Pelley’s (2004, 2010) hybrid model (lower panel), attention should decrease for both types of cues during discrimination, yet the relevant cues receive more attention than the irrelevant ones. During partial reinforcement, the amount of attention paid to both types of cues should increase at the outset, and become similar in the course of training. We designed our experiment to test the predictions derived from the three theoretical accounts. Changes in overt attention were measured by recording the dwell times on the pairs of cues as participants solved the task.

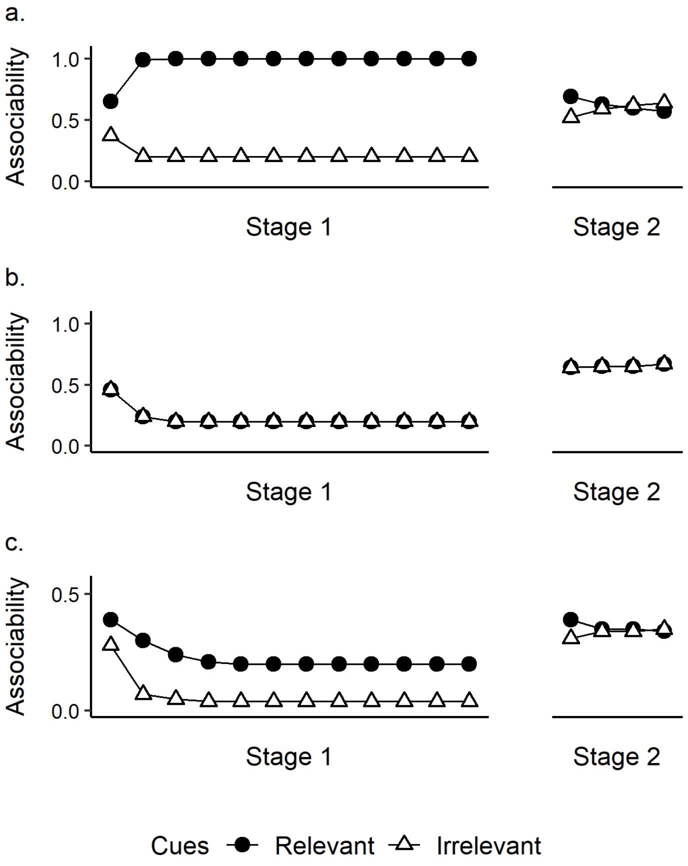

Figure 1. Simulation results of attentional changes for the present experiment according to a) Mackintosh's (1975) theory, b) the Pearce-Hall model (Pearce et al., 1981), and c) Le Pelley's (2004, 2010) hybrid model. In Stage 1, discrimination training comprised stimulus pairs containing one cue being relevant for predicting the outcome (o1 or o2) and another cue being irrelevant. In Stage 2, each pair of cues was randomly followed by either o1 or o2 (partial reinforcement). Note that the values obtained in the simulation based on the hybrid model are generally lower than those obtained in simulations based on the other models. However, our predictions were concerned with the relative amount of attention paid to each type of cue and the attentional changes across stages. For additional information, see the Appendix.

Methods

Participants

Our study was approved by the ethics committee of the Faculty of Psychology at Philipps-Universität Marburg (AZ: 2018-25k). Twenty-six university students participated in exchange of 8.00 € or course credits. Depending on their performance, they also earned a monetary reward ranging from 3.92 to 9.20 €. Participants either reported absence of visual impairment (< 1.5 diopter) or used soft contact lenses. Given that our hypotheses relied on the condition that prediction error would be small at the end of discrimination training, we excluded two participants with an accuracy in outcome prediction below 0.60 in the last two blocks of discrimination. Thus, the final sample consisted of 24 participants ($M_{Age} = 22.04$ years, $SD = 3.32$, 6 men).

Stimuli and procedure

Four pairs of cues were used throughout the experiment. Participants' task was to predict the outcome of each trial based on the cues presented on the screen, earning 8 cents for each correct prediction. Table 1 shows the experimental design. During an initial discrimination stage (Stage 1), the pairs AX and AY were followed by o1, while BX and BY were followed by o2. Thus, A and B were consistently paired with one of the outcomes, thereby becoming relevant for making the prediction, whereas X and Y were paired with both outcomes, each on half of the trials, and therefore were irrelevant. During a subsequent partial reinforcement stage (Stage 2), each of the pairs was followed by o1 on a random half of the trials, and by o2 on the other half. The experiment was divided into blocks (24 during Stage 1, and eight during Stage 2) each including each of the four pairs once in random order. There was no break between the two stages. We excluded the attentional data of the first trial in Stage 2, because during its attentional measurement interval participants had yet to experience prediction error. Throughout partial reinforcement, the trials where the outcomes had changed relative to the discrimination stage and those where the outcomes remained the same were evenly distributed, so that each block contained two instances of each.

Table 1. Experimental design.

Stage 1: discrimination	Stage 2: partial reinforcement
AX – o1	AX – o1/o2
AY – o1	AY – o1/o2
BX – o2	BX – o1/o2
BY – o2	BY – o1/o2

Figure 2 shows an example of the trials used in the experiment. Each trial started with a fixation cross that lasted for 2 s. Then, three gray circles (34 mm in diameter) appeared around the center of the screen, distributed at 180 mm from each other (center-to-center). Two of the circles constituted the cues, each enclosing a specific pattern of gray dots. The other circle was empty. After 4 s, a response trigger was presented inside the empty circle for 0.2 s, together with a brief tone. The response trigger consisted of two small colored dots showing each of the two possible outcomes (green or red) next to each other, whereby the order of the colors varied across trials (see below). Once the trigger disappeared, participants had 1.8 s to indicate the expected outcome by pressing the mouse button (left or right) corresponding to the position where the dot representing the outcome had been shown. The response trigger competed for attention with the cues and prevented participants from keeping looking at them throughout the whole attentional measurement interval, thus avoiding a potential ceiling effect. Finally, the correct outcome (a big dot that was either green or red) was presented for 2 s at the center of the empty circle. In addition, an auditory feedback indicated whether the response was correct (‘correct: 8 cents’) or not (‘incorrect: 0 cents’). If participants made the response after the appropriate interval, a warning message was displayed on the screen following the outcome (‘response too late!’) and the auditory feedback was omitted. A blank screen with a random duration between 1.5 and 4.5 s followed every trial. The position of the gray circles varied randomly across trials, taking any of the following polar angles: 0°, 60°, 120°, 180°, 240°, and 300°. The position of the colored dots within the response trigger (left or right) was determined randomly across the experiment, with four trials of each alternative every two consecutive blocks. The correspondence between the gray dot patterns and the cues they represented was counterbalanced across participants.

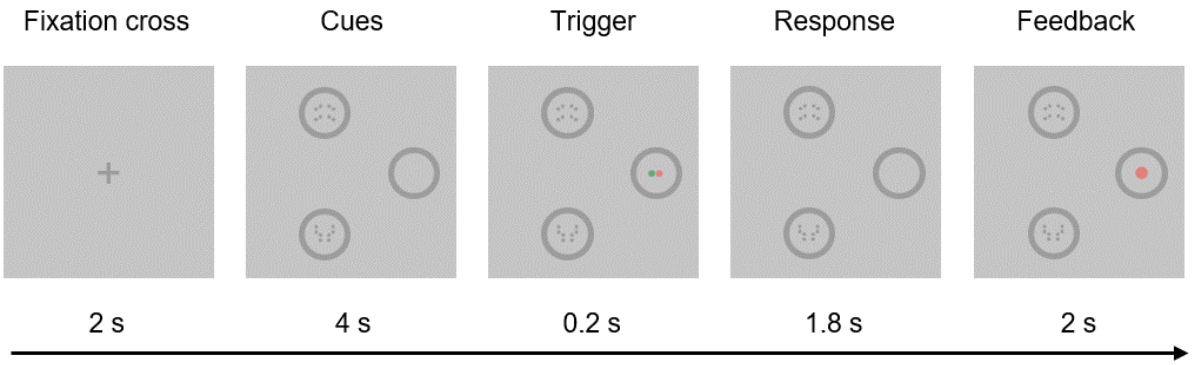

Figure 2. Trial outline.

Participants read the instructions of the task after signing the informed consent. The instructions asked participants to pay attention to the dot patterns and to look then at the empty circle, in order to make sure that they would not miss the response trigger. In addition, the instructions stated that over the course of the experiment participants could learn to predict the outcomes on the basis of the gray patterns. Then, a minimum of eight practice trials were run, using stimuli that were similar to those used later in the experiment. The eye movements were recorded with an infrared video-based eye tracker (Eyelink 2000, SR-Research; Mississauga, Canada), sampling the gaze position at a frequency of 1000 Hz. Before the experiment started, the eye tracker was calibrated by means of a nine-point grid, keeping the maximal error below 0.5°. Sampling of the left or the right eye was counterbalanced across participants. The visual stimuli were presented on a 22” CRT screen (Vision Master Pro 514, Iiyama; Tokio, Japan). Forehead and chin rests kept participants’ head in a fixed position, with an eye-to-screen distance of 78 cm. Stimulus timing and response recording was controlled by Presentation software (version 16.1; Neurobehavioral Systems, Inc.).

Measures and data analysis

We used the proportion of correct outcome predictions (accuracy) as a measure of the amount of prediction error generated by the cues, with a low accuracy corresponding to a large prediction error. The data from those trials in which no response occurred during the appropriate interval were excluded from the accuracy analyses (< 0.01%). We used custom MATLAB software (The MathWorks, Inc., 2012) for conditioning the signal of eye position traces (Koenig, 2010). Ocular fixations were detected using a velocity-based algorithm. Fixations were scored as on-stimulus, if they deviated less than 60 mm from its center. The analysis interval was limited to the 4 s between the onset of the stimulus and the presentation of the response trigger. We excluded the attentional data from those trials in which a) participants did not fixate on the cross before the onset of the cues (0.02 %), and b) an artefact had a duration above percentile 90 (> 380 ms, 0.01%). Our primary attentional measure was the dwell time on each type of stimulus (relevant cue, irrelevant cue, and empty circle, where the response trigger was set to appear). The dwell times were log-transformed to normalize their distributions. In addition, we measured the probability of fixating on each type of stimulus during the measurement interval, across 20 bins of 200 ms each.

We averaged the data from every two trials showing the same pair of cues, resulting in 12 analysis epochs during discrimination and eight during partial reinforcement. For accuracy and dwell time, we conducted separate ANOVAs for discrimination, partial reinforcement, and the transition between the two stages – i.e. the last epoch of discrimination and the first epoch of partial reinforcement. For fixation probabilities, we conducted separate ANOVAs for the first and last epochs of discrimination, the last epoch of discrimination and the first epoch of partial reinforcement, and the first and last epochs of partial reinforcement. The factors included in each ANOVA are indicated in the results section. Where appropriate, the Greenhouse-Geisser (1959) correction was applied to the degrees of freedom. The significance level was set at .05. For the multiple comparisons between time bins, the significance value was adjusted through the Benjamini-Hochberg procedure (Thissen, Steinberg & Kuang, 2002). Partial eta squared (η^2_p) and Cohen's *d* were used as measures of effect size.

Results

Accuracy

Figure 3 shows the accuracy of outcome prediction across epochs. During discrimination, participants' accuracy increased following an asymptotic pattern, suggesting that prediction error was reduced to a minimum. This was confirmed by a one-way ANOVA showing a significant effect of epoch, $F(11, 253) = 23.08$, $p < .001$, $\eta^2_p = .50$, together with significant linear and quadratic trends, Fs$(1, 23) > 45.58$, ps $< .001$, η^2_ps $> .66$.

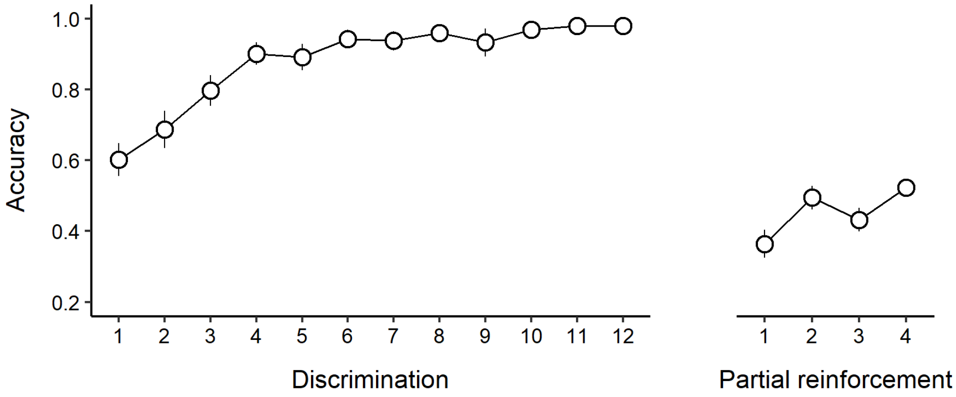

Figure 3. Mean values of the accuracy of outcome prediction across epochs. Error bars represent SEM.

The introduction of partial reinforcement resulted in a loss of accuracy. Thus, from the last epoch of discrimination to the first epoch of partial reinforcement, accuracy dropped below chance level (0.50), $t(23) = 12.33$, $p < .001$, $d = 2.57$. Then, throughout partial reinforcement accuracy showed an increase, albeit it did not go above the chance level, suggesting that the cues continued to generate large prediction errors. A one-way ANOVA showed a significant effect of epoch, $F(3, 69) = 5.08$, $p < .01$, $\eta^2_p = .18$, with a significant linear trend, $F(1, 23) = 5.90$, $p = .02$, $\eta^2_p = .20$.

Fixation dwell time

Figure 4a shows the dwell times on the relevant and irrelevant cues across epochs. The dwell times decreased during discrimination, particularly in the early epochs. However, the decrease was less pronounced for the relevant than for the irrelevant cues. This pattern was confirmed by means of a Cue (relevant vs irrelevant) × Epoch (1 – 12) ANOVA, which revealed a significant interaction, $F(11, 253) = 3.62$, $p < .01$, $\eta^2_p = .14$. Thus, although the dwell times on each type of cue were similar in Epochs 1 and 2, *ts*(23) < 1.78, *ps* > .09, *ds* < 0.37, throughout the remaining epochs the dwell times were longer on the relevant cues than on those irrelevant, *t*s(23) > 2.16, *p*s < .05, *d*s > 0.45. Across epochs, the decrease was significant for each type of cue, *F*s(11, 253) > 14.15, *p*s < .001, η^2_ps > .38, with significant linear and quadric trends, *F*s(1, 23) > 15.00, *p*s < .001, η^2_ps > .39. In addition, the main effects of cue and epoch were also significant, *F*s > 12.42, *p*s < .01, η^2_ps > .35.

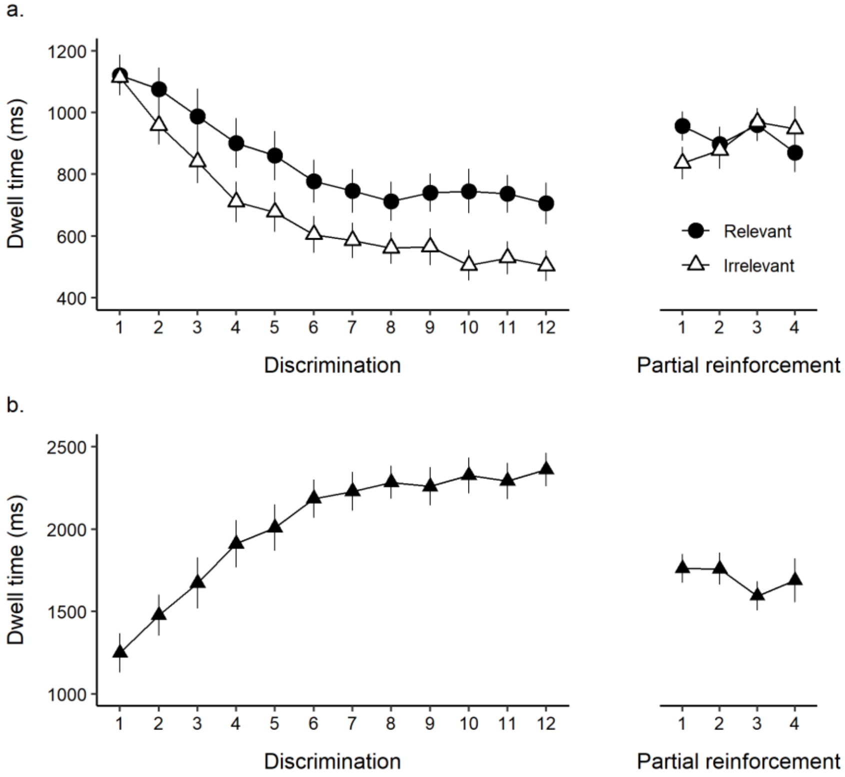

Figure 4. Mean dwell times across epochs on a) the discrimination cues and b) the position of the response trigger. Error bars represent SEM.

Upon the introduction of partial reinforcement, the dwell times on each type of cue showed a similar increase. Thus, the dwell times continued to be longer on the relevant cues than on those irrelevant. Accordingly, a Cue (relevant vs irrelevant) × Epoch (Discrimination, Epoch 12 vs Partial reinforcement, Epoch 1) ANOVA showed significant main effects of both cue, $F(1, 23) = 9.99$, $p < .01$, $\eta^2_p = .30$, and epoch, $F(1, 23) = 24.65$, $p < .001$, $\eta^2_p = .52$. In addition, the interaction effect was non-significant, $F(1, 23) = 3.02$, $p = .10$, $\eta^2_p = .12$.

After the first epoch of partial reinforcement, the dwell times on each type of cue became similar and did not substantially change throughout the end of the stage. This was confirmed by a Cue (relevant vs irrelevant) × Epoch (1 – 4) ANOVA, which revealed a significant interaction, $F(3, 69) = 3.19$, $p = .03$, $\eta^2_p = .12$. Thus, although in the first epoch the dwell times were longer on the relevant than on the irrelevant cues, $t(23) = 2.36$, $p = .03$, $d = 0.48$, they were similar during the remaining epochs (however, in the last epoch there was a trend towards longer dwell times on the irrelevant than on the relevant cues, $t(23) = 1.93$, $p = .07$, $d = 0.35$). On the other hand, the dwell times on each type of cue remained relatively constant throughout partial reinforcement, Fs$(3, 69) < 1.79$, ps $> .18$, η^2_ps $< .08$. Finally, none of the main effects were significant, Fs$(3, 69) < 1.11$, ps $> .32$, η^2_ps $< .05$.

Figure 4b shows the dwell times on the position where the response trigger was set to appear. These dwell times followed a pattern inverse to that of the dwell times on the cues. During discrimination, the dwell times showed an asymptotic increase. This was confirmed by a one-way ANOVA revealing a significant effect of epoch, $F(11, 253) = 30.30$, $p < .001$, $\eta^2_p = .57$, together with significant linear and quadratic trends, Fs$(1, 23) > 28.91$, ps $< .001$, η^2_ps $> .55$. From the last epoch of discrimination to the first epoch of partial reinforcement, there was a drop in the dwell times, $t(23) = 4.74$, $p < .001$, $d = 0.98$. Finally, as confirmed by a one-way ANOVA, the dwell times remained unchanged throughout the last epochs of partial reinforcement, $F(3, 69) = 1.66$, $p = .20$, $\eta^2_p = .07$.

Fixation probability

Figure 5 shows the pattern of fixations occurring within trials. We focused on the data of the first and last epochs of discrimination (panels a and b) and the first and last epochs of partial reinforcement (panels c and d). Overall, participants started fixating on the discrimination cues and then proceeded to look at the position of the response trigger.

In order to analyse the changes in the fixation patterns that resulted from learning, we conducted three separate ANOVAs on the fixation probabilities. The first epoch of discrimination was compared to the last one by means of a Cue (relevant vs irrelevant) × Epoch (Discrimination, Epoch 1 vs Discrimination, Epoch 12) × Bin (1 – 20) ANOVA. At the outset of discrimination, the fixation probabilities were similar for both types of cues. However, by the end of the stage, the probabilities were higher for the relevant cues than for those irrelevant. This was confirmed by a significant Cue × Epoch interaction, $F(1, 23) = 6.71$, $p = .02$, $\eta^2_p = .23$, with similar fixation probabilities in the first epoch, $t < 1$, but not in the last one, $t(23) = 2.37$, $p = .03$, $d = 0.43$. Across epochs, the probabilities decreased for each type of cue, *t*s(23) > 8.77, *p*s < .001, *d*s > 1.78. Moreover, there was a change in the shape of the distribution of the probabilities. Accordingly, a significant Epoch × Bin interaction, $F(19, 437) = 14.02$, $p < .001$, $\eta^2_p = .38$, pointed to an increase in the probabilities between 200 and 600 ms, *t*s(23) > 2.30, *p*s < .04, *d*s > 0.53, and a decrease in the last 3 s, *t*s(23) > 3.11, *p*s < .01, *d*s > 0.65. The main effects of epoch and bin were also significant, *F*s > 137.35, *p*s < .001, η^2_ps > .85, but none of the remaining effects were, *F*s < 2.78, *p*s > .10, η^2_ps < .11.

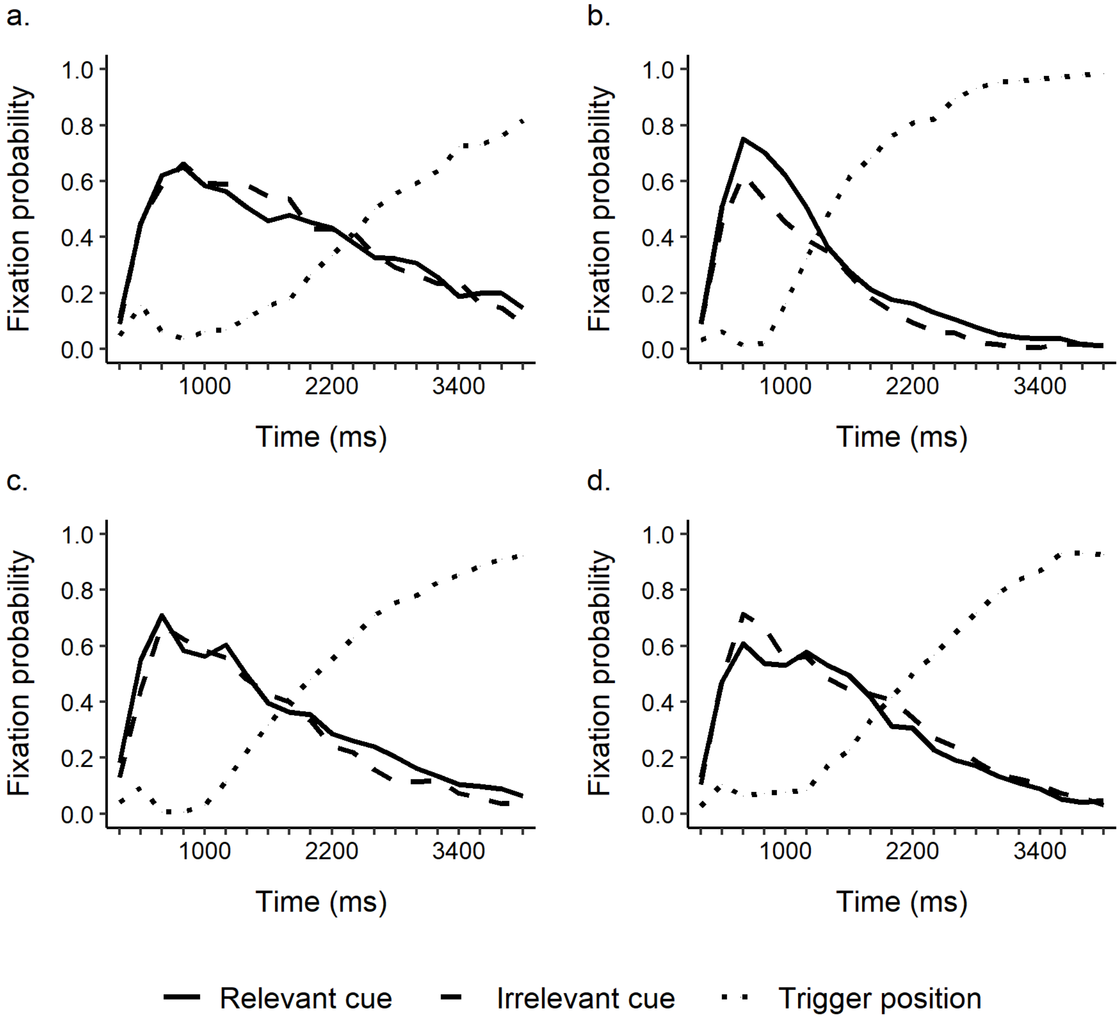

Figure 5. Probabilities of fixating on each of the stimuli (i.e., relevant cue, irrelevant cue, and trigger position) within trials. Panels a and b depict the data from the first and last epochs of discrimination, and panels c and d the data from the first and last epochs of partial reinforcement.

From the last epoch of discrimination to the first epoch of partial reinforcement, the fixation probabilities increased for both types of cues. A Cue (relevant vs irrelevant) × Epoch (Discrimination, Epoch 12 vs Partial reinforcement, Epoch 1) × Bin (1 – 20) ANOVA revealed a significant Epoch × Bin interaction, $F(19, 437) = 6.06$, $p < .001$, $\eta^2_p = .21$. Thus, the increase in the probabilities occurred during specific time bins, namely in the first 200 ms, then between 1000 and 1200 ms, and finally in the last 2600 ms, ts(23) > 2.52, ps < .03, ds > 0.75. In addition, the probabilities continued to be higher for the relevant cues than for those irrelevant, as indicated by a main effect of Cue, $F(1, 23) = 7.87$, $p = .01$, $\eta^2_p = .26$. The main effects of epoch and bin were also significant, Fs > 23.77, ps < .001, η^2_ps > .50, but none of the other effects were Fs < 1.24, ps > .30, η^2_ps < .06.

In the last epoch of partial reinforcement, the difference in the fixation probabilities between the two types of cues appeared to reverse. Consistently, a Cue (relevant vs irrelevant) × Epoch (Partial reinforcement, Epoch 1 vs Partial reinforcement, Epoch 4) × Bin (1 – 20) ANOVA showed a significant Cue × Epoch interaction, $F(1, 23) = 7.12$, $p = .01$, $\eta^2_p = .24$. Thus, while in the first epoch the probabilities were higher for the relevant than for the irrelevant cues, $t(23) = 2.37$, $p = .03$, $d = 0.43$, the reverse was true in the last epoch, $t(23) = 2.15$, $p = .04$, $d = 0.41$. Accordingly, both types of cues showed changes in opposite directions in the fixation probabilities, ts(23) > 3.71, ps < .01, ds > 0.74. Among the remaining effects, only the main effect of bin was significant, $F(19, 437) = 149.77$, $p < .001$, $\eta^2_p = .87$ (all the other Fs < 1.35, ps > .24, η^2_ps < .06).

Finally, we analysed the changes in the probabilities of fixating on the position of the response trigger. For simplicity, we averaged the probabilities across time bins. A one-way ANOVA showed a significant epoch effect, $F(3, 69) = 20.37$, $p < .001$, $\eta^2_p = .47$. Thus, the average probability increased from the first epoch of discrimination to the last one, $t(23) = 11.00$, $p < .001$, $d = 2.21$, and then it decreased from the last epoch of discrimination to the first epoch of partial reinforcement, $t(23) = 4.57$, $p < .001$, $d = 0.91$. This pattern was opposed to the changes observed in the fixation probabilities of the discrimination cues. No change was observed between the first and last epochs of partial reinforcement, $t < 1$.

Discussion

In the present study, participants' eye movements were recorded as they predicted the outcomes following different pairs of cues. During an initial discrimination, one of the cues in each pair was relevant for making correct predictions, because it signaled the same outcome on every trial. The other cue was irrelevant because it was equally associated with each of the two possible outcomes. In a subsequent partial reinforcement stage, each pair was followed by one outcome on a random half of the trials, and by the other outcome on the remaining half. Thus, it was no longer possible to make accurate predictions. Our results revealed that outcome predictions became fully accurate during the initial discrimination, suggesting that they were based on the relevant cues, whose prediction errors should have been reduced to a minimum. Accordingly, an attentional preference emerged for the relevant cues in both the dwell times and the fixation probabilities. Moreover, there was a decrease in the dwell times on each type of cue, and a change in the pattern of fixations within trials, so that the fixation probabilities increased for a brief period after cue onset (200 – 600 ms), but for most of the interval (last 3 s) they decreased. Accuracy dropped to chance level immediately following the introduction of partial reinforcement, and it stayed at that level throughout the stage, thus suggesting sustained large prediction errors. Consistently, both the dwell times and the fixation probabilities showed an increase in attention for the two types of cues. In addition, the attentional preference for the relevant cues was no longer apparent beyond the first epoch of partial reinforcement.

The present findings are in full agreement with the predictions of the hybrid model by Le Pelley (2004, 2010), which assumes an interaction of two associability mechanisms. One of them, based on Mackintosh's (1975) theory, should give attentional preference to the relevant cues during discrimination, as they should be associated with smaller relative prediction errors than the irrelevant cues. The other mechanism, put forward in the Pearce-Hall (1980) model, should determine the amount of attention paid to each cue as a direct function of the prediction error produced by all the present cues, with a decrease in attention during discrimination and an increase following partial reinforcement. Neither Mackintosh's

(1975) theory nor the Pearce-Hall (1980) model by itself is able to fully explain the observed pattern. Regarding the amount of attention paid to the relevant cues, Mackintosh's theory did not anticipate a decrease during discrimination, or an increase following partial reinforcement. On the other hand, the Pearce-Hall model did not predict an attentional preference for the relevant cues during discrimination.

Early studies on the hybrid model indicated a discrepancy between the animal and human domains. Using rats, Haselgrove et al. (2010) conducted four experiments in which the animals were trained with cues that were either continuously or partially reinforced with food. In two of the experiments, the partially reinforced cues were presented alone and non-reinforced on half of the trials and, on the other half, they were arranged in pairs with the continuously reinforced cues, so that it was possible to compare the relative prediction errors produced by each type of cue. This manipulation led to changes in associability consistent with the mechanism suggested by Mackintosh (1975). In the other two experiments, each cue was presented separately, and therefore the continuously and partially reinforced cues only differed in terms of overall prediction errors. This produced associability changes in line with the mechanism put forward by Pearce and Hall (1980). Thus, Haselgrove et al. provided support for the hybrid model albeit their conclusion was based on a comparison across experiments.

In the human domain, Le Pelley et al. (2010) noted that earlier studies had used designs favoring the associability pattern predicted by Mackintosh's (1975) theory. Therefore, they aimed to test the prediction of the hybrid model that a procedure emphasizing differences in overall prediction error would show associability changes consistent with the Pearce-Hall (1980) model. In four related experiments, participants learned a discrimination in which all the cues were presented separately. Some of them were consistently paired with one of two possible outcomes, whereas others were paired with both, each on a random half of the trials. Thus, the latter cues should have produced larger overall prediction errors than the former. The changes in associability resulting from the initial training were tested in a subsequent discrimination in which the same cues were paired with two new outcomes. Contrary to the prediction, learning of the second discrimination was faster for the cues that had been previously associated with small overall prediction errors. Therefore, the authors concluded that the hybrid model might not apply to humans (for similar findings, see Kattner, 2015; Livesey, Thorwart, De Fina, & Harris, 2011, Experiments 3 and 4).

First evidence for the hybrid model in humans came from a study conducted by Beesley et al. (2015). Participants were trained with different pairs of cues, which were associated with their respective outcomes on either a deterministic or a probabilistic way. The probabilistic pairs were followed by a primary outcome on 67% of the trials and by another outcome on the remaining 33%, and thus should have produced larger overall prediction errors than the deterministic pairs, which were consistently followed by just one outcome. In addition, each pair contained a relevant cue that was (mostly) associated with one outcome, and an irrelevant cue that was equally associated with two outcomes. Thus, within pairs, the former cue should have produced smaller relative prediction errors than the latter. The authors found that the probabilistic pairs were fixated for longer times than the deterministic pairs and, within the latter, the same was true for the relevant cues compared to the irrelevant ones. However, while the advantage in overt attention for the relevant cues transferred to a subsequent task, the advantage for the probabilistic pairs did not. In fact, in line with the studies mentioned in the previous paragraph, learning was faster for the deterministic pairs than the probabilistic ones. In this regard, Beeseley et al. suggested that the associability changes based on the Pearce-Hall (1980) mechanism might not generalize to new learning situations. This suggestion fits well with the findings reported by the only human study that has shown facilitated learning for inconsistently reinforced cues (Griffiths et al., 2011). Instead of using a subsequent learning task with new outcomes, the authors of that study tested associability changes by increasing the magnitude of the original outcome and then analysing the speed at which participants updated their predictions accordingly.

Our results are in agreement with those of Beesley et al. (2015), in that they showed changes in overt attention consistent with an interaction of two associability mechanisms (see also Easdale, et al., 2019; Walker, Luque, Le Pelley, & Beesley, 2019). Thus, the previously reported inverse association between overall prediction error and speed of subsequent learning is at variance with the patterns of overt attention. Recently, it has been suggested that participants may acquire knowledge about the relational structure of learning tasks in a cue-specific way (Livesey, Don, Uengoer, & Thorwart, 2019). This type of knowledge could explain the discrepancy between overt attention and subsequent rate of learning. For instance, in the first stage of the study conducted by Le Pelley et al. (2010) participants might have learned whether each particular cue was followed by only one outcome or two. Associability was then measured in a discrimination that involved pairing each of the cues with just one outcome. If participants applied the relational knowledge acquired in the first stage, learning would have been relatively easy for the cues previously followed by one outcome (i.e., those producing small overall prediction error), because for those cues the relational structure was congruent across stages. This effect would be unrelated to changes in associability. A similar case can be made for between-subjects manipulations (e.g., Experiment 2 in Beesley et al., 2015) by referring to the formation of abstract rules (e.g., Lachnit & Lober, 2001; Lachnit, Lober, Reinhard, & Kinder, 2001).

In the present study, we tested the predictions of Le Pelley's hybrid model against those of Mackintosh's theory and the Pearce-Hall model. However, our results may also be explained by other implementations of the hybrid model (e.g., Pearce & Mackintosh, 2010). Moreover, instead of the associability mechanism suggested by Mackintosh (1975), the attentional preference for the relevant cue could be accounted for the different associative strengths of the relevant and the irrelevant cues. Thus, the pattern of overt attention observed throughout the experiment would result from a function combining the influence of the differences in associative strength with the influence of overall prediction error (i.e., the Pearce-Hall mechanism; see Koenig, Kadel, et al., 2017; Koenig, Uengoer, et al., 2017). Finally, Esber and Haselgrove (2011) put forward a model that can explain attentional patterns suggestive of two separate associability mechanisms by means of a single process. According to their model, the amount of attention paid to cues is an additive function of separate associative strengths that they developed with different outcomes, including associations with the absence of an expected outcome. To our understanding, this model is also equipped to explain most of our results, with the exception of the complete loss of differential allocation of attention between the relevant and the irrelevant cues observed during partial reinforcement.

Studies on the influence of associative learning on overt attention typically rely on self-paced designs, where, on each trial, the eye gaze is recorded from cue onset to the time when participants make the response (e.g., Easdale et al., 2019; Kruschke, Kappenman, & Hetrick, 2005). Then, the dwell times are often corrected by response times, as the two are confounded with each other (Koenig, Kadel, et al., 2017), and the response times may vary across participants and trials for reasons unrelated to attention. However, in studies where both absolute and corrected dwell times are reported, it is not uncommon to find different results in each of them (e.g., Jones & Zaksaite, 2018). Moreover, corrected dwell times are useful for comparing attentional changes between cues, but do not show the absolute changes undergone by each cue. In the present study, the cues appeared for 4 s on each trial, before a very brief presentation of the response trigger on a known location of the screen. Thus, towards the end of the interval, the response trigger competed for attention with the cues. This allowed us to use a fixed measurement interval and still avoid a ceiling effect, which might have occurred if participants had no reason to look at another location after fixating on the relevant cue (e.g., Torrents-Rodas, Koenig, Uengoer & Lachnit, 2020). Indeed, the fixation probabilities showed that participants looked at the cues during the first part of the interval and then moved their eyes towards the position where the response trigger was set to appear. Importantly, this within-trial shift was modulated by learning, and it was in accordance with the associability mechanism suggested by Pearce and Hall (1980). Thus, by the end of discrimination, when prediction accuracy was high, participants fixated on the cues for a short initial interval and then directed their eye gaze towards the location of the response trigger. The introduction of partial reinforcement, which was associated with an increase in prediction error, significantly extended the interval in which participants fixated on the cues, to the detriment of the amount of attention paid to the trigger location. To the best of our knowledge, this measurement method has been used for the first time in the present study. We believe that it provides a standardized way to measure dwell times across trials and participants and to track fixation patterns within trials (for an example of within-trial fixation patterns in intervals of variable duration in the field of categorization, see Blair, Watson, Walshe, & Maj, 2009). Monitoring within-trial time might well be important in the development and evaluation of theories of associative learning (Lachnit et al., 2013).

By tracking participants' gaze, the present study provided evidence for an interaction of two associability mechanisms in determining attentional changes during learning. One of these mechanisms is consistent with a comparison of the individual prediction errors associated to each of the present cues, thus selecting the cue with the best predictive accuracy. The other mechanism is related to the prediction error produced by all the cues, thus increasing attention – or keeping it at a high level – if the outcome cannot be accurately predicted. Our results add to recent findings in human participants that are consistent with the predictions of hybrid models of attention. Future research should further elucidate to what extent the attentional changes accounted for each mechanism are transient or generalize to new learning situations.

Funding

Our study was supported by the Deutsche Forschungsgemeinschaft (DFG, German Research Foundation), project number 290878970-GRK 2271 project 5, as well as project number 222641018–SFB/TRR 135 project B4.

Declaration of conflicting interests

The authors declare that there is no conflict of interest.

References

Beesley, T., Nguyen, K. P., Pearson, D., & Le Pelley, M. E. (2015). Uncertainty and predictiveness determine attention to cues during human associative learning. *Quarterly Journal of Experimental Psychology, 68*, 2175 – 2199. https://doi.org/10.1080/17470218.2015.1009919

Blair, M. R., Watson, M. R., Walshe, R. C., & Maj, F. (2009). Extremely selective attention: eye-tracking studies of the dynamic allocation of attention to stimulus features in categorization. *Journal of Experimental Psychology: Learning, Memory, & Cognition, 35*, 1196 – 1206. https://doi.org/10.1037/a0016272

Easdale, L. C., Le Pelley, M. E., & Beesley, T. (2019). The onset of uncertainty facilitates the learning of new associations by increasing attention to cues. *Quarterly Journal of Experimental Psychology, 72*, 193 – 208. https://doi.org/10.1080/17470218.2017.1363257

Esber, G. R., & Haselgrove, M. (2011). Reconciling the influence of predictiveness and uncertainty on stimulus salience: a model of attention in associative learning. *Procedings of the Royal Society, 278*, 2553 – 61. https://doi.org/10.1098/rspb.2011.0836

George, D. N., & Pearce, J. M. (1999). Acquired distinctiveness is controlled by stimulus relevance not correlation with reward. *Journal of Experimental Psychology: Animal Behavior Processes, 25*, 363 – 373. https://doi.org/10.1037/0097-7403.25.3.363

Greenhouse, S. W., & Geisser, S. (1959). On methods in the analysis of profile data. *Psychometrika, 24,* 95 – 112. https://doi.org/10.1007/BF02289823

Griffiths, O., Johnson, A. M., & Mitchell, C. J. (2011). Negative transfer in human associative learning. *Psychological Science, 22*, 1198 – 1204. https://doi.org/10.1177/0956797611419305

Haselgrove, M., Esber, G. R., Pearce, J. M., & Jones, P. M. (2010). Two kinds of attention in Pavlovian conditioning: evidence for a hybrid model of learning. *Journal of Experimental Psychology: Animal Behavior Processes, 36*, 456 – 470. https://doi.org/10.1037/a0018528

Hogarth, L., Dickinson, A., Austin, A., Brown, C., & Duka, T. (2008). Attention and expectation in human predictive learning: the role of uncertainty. *Quarterly Journal of Experimental Psychology, 61*, 1658 – 1668. https://doi.org/10.1080/17470210701643439

Jones, P. M., & Zaksaite, T. (2018). The redundancy effect in human causal learning: No evidence for changes in selective attention. *Quarterly Journal of Experimental Psychology, 71*, 1748 – 1760. https://doi.org/10.1080/17470218.2017.1350868

Kattner, F. (2015). Transfer of absolute and relative predictiveness in human contingency learning. *Learning & Behavior, 43*, 32 – 43. https://doi.org/10.3758/s13420-014-0159-5

Kaye, H., & Pearce, J. M. (1984). The strength of the orienting response during pavlovian conditioning. *Journal of Experimental Psychology: Animal Behavior Processes, 10*, 90 – 109. https://doi.org/10.1037/0097-7403.10.1.90

Koenig, S. (2010). *Modulation of saccadic curvature by spatial memory and associative learning* (Doctoral dissertation). Retrieved from http://archiv.ub.uni-marburg.de/diss/z2010/0636/

Koenig, S., Kadel, H., Uengoer, M., Schubö, A., & Lachnit, H. (2017). Reward draws the eye, uncertainty holds the eye: associative learning modulates distractor interference in visual search. *Frontiers in Behavioral Neuroscience, 11*, 1 – 15. https://doi.org/10.3389/fnbeh.2017.00128

Koenig, S., Uengoer, M., & Lachnit, H. (2017). Attentional bias for uncertain cues of shock in human fear conditioning: evidence for attentional learning theory. *Frontiers in Human Neuroscience, 11*, 1 – 13. https://doi.org/10.3389/fnhum.2017.00266

Koenig, S., Uengoer, M., & Lachnit, H. (2018). Pupil dilation indicates the coding of past prediction errors: evidence for attentional learning theory. *Psychophysiology, 55*, 1 – 12. https://doi.org/10.1111/psyp.13020

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. https://doi.org/10.1037/0278-7393.31.5.830

Lachnit, H., Thorwart, A., Schultheis, H., Koenig, S., Lotz, A., & Uengoer, M. (2013). Indicators of early and late processing reveal the importance of within-trial-time for theories of associative learning. PLoS One, 8, e66291. https://doi.org/10.1371/journal.pone.0066291

Lachnit, H. & Lober, K. (2001). What is learned in patterning discrimination? Further tests of configural accounts of associative learning in human electrodermal conditioning. *Biological Psychology, 56*, 45 – 61. https://doi.org/10.1016/S0301-0511(00)00087-9

Lachnit, H., Lober, K., Reinhard, G., & Kinder, A. (2001). Evidence for the application of rules in Pavlovian electrodermal conditioning with humans. *Biological Psychology, 56*, 151 – 166. https://doi.org/10.1016/S0301-0511(01)00067-9

Le Pelley, M. E. (2004). The role of associative history in models of associative learning: a selective review and a hybrid model. *Quarterly Journal of Experimental Psychology, 57B*, 193 – 243. https://doi.org/10.1080/02724990344000141

Le Pelley, M. E. (2010). The hybrid modeling approach to conditioning. In N. A. Schmajuk (Ed.), *Computational Models of Conditioning* (pp. 71 – 107). Cambridge: Cambridge University Press.

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, https://doi.org/10.1037/a0021384

Le Pelley, M. E., & McLaren, I. P. L. (2003). Learned associability and associative change in human causal learning. *The Quarterly Journal of Experimental Psychology, 56B*, 68 – 79. https://doi.org/10.1080/02724990244000179

Le Pelley, M. E., Mitchell, C. J., Beesley, T., George, D. N., & Wills, A. J. (2016). Attention and associative learning in humans: an integrative review. *Psychological Bulletin, 142*, 1111 – 1140. https://doi.org/10.1037/bul0000064

Le Pelley, M. E., Trunbull, M. N., Reimers, S. J., Knipe, R. L. (2010). Learned predictiveness effects following single-cue training in humans. *Learning & Behavior, 38*, 126 – 144. https://doi.org/10.3758/LB.38.2.126

Le Pelley, M. E., Vadillo, M., & Luque, D. (2013). Learned predictiveness influences rapid attentional capture: evidence from the dot probe task. *Journal of Experimental Psychology: Learning, Memory, & Cognition, 39*, 1888 – 1900. https://doi.org/10.1037/a0033700

Livesey, E. J., Don, H. J., Uengoer, M., & Thorwart, A. (2019). Transfer of associability and relational structure in human associative learning. *Journal of Experimental Psychology: Animal Learning and Cognition, 45*, 125 – 142. https://doi.org/10.1037/xan0000197

Livesey, E. J., Thorwart, A., De Fina, N. L., & Harris, J. A. (2011). Comparing learned predictiveness effects within and across compound discriminations. *Journal of Experimental Psychology: Animal Behavior Processes, 37*, 446 – 465, https://doi.org/10.1037/a0023391

Lochmann, T. & Wills, A. J. (2003). Predictive history in an allergy prediction task. In F. Schmalhofer, R. M. Young, & G. Katz (Eds.), *Proceedings of EuroCogSci 03* (pp. 217 - 222). Mahwah, NJ: Erlbaum.

Mackintosh, N. J. (1975). A theory of attention: variations in the associability of stimuli with reinforcement. *Psychological Review, 82*, 276 – 298. https://doi.org/10.1037/h0076778

Pearce, J. M., & Bouton, M. E. (2001). Theories of associative learning in animals. *Annual Review of Psychology, 51*, 111 – 139. https://doi.org/10.1146/annurev.psych.52.1.111

Pearce, J. M., & Hall, G. (1980). A model for pavlovian learning: variations in the effectiveness of conditioned but not of unconditioned stumuli. *Psychological Review, 87*, 532 – 552. https://doi.org/10.1037/0033-295X.87.6.532

Pearce, J. M., Kaye, H., & Hall, G. (1981). Predictive accuracy and stimulus associability: development of a model for Pavlovian learning. In M. L. Commons, R. J. Herrnstein, & A. R. Wagner (Eds.), *Quantitative analyses of behavior: acquisition* (Vol. 3, pp. 241 – 255). Cambridge, MA: Ballinger.

Pearce, J. M., & Mackintosh, N. J. (2010). Two theories of attention: a review and a possible integration. In C. J. Mitchell & M. E. Le Pelley (Eds.), *Attention and associative learning: from brain to behaviour* (pp. 11 – 40). Oxford, UK: Oxford University Press.

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

Swan, J. A., & Pearce, J. M. (1988). The orienting response as an index of stimulus associability in rats. *Journal of Experimental Psychology: Animal Behavior Processes, 14*, 292 – 301. https://doi.org/10.1037/0097-7403.14.3.292

Thissen, D., Steinberg, L., & Kuang, D. (2002). Quick and easy implementation of the Benjamini-Hochberg procedure for controlling the false positive rate in multiple comparisons. *Journal of Educational and Behavioral Statistics, 27*, 77 – 83. https://doi.org/10.3102/10769986027001077

Torrents-Rodas, D., Koenig, S., Uengoer, M., & Lachnit, H. (2020). *A rise in prediction error increases attention to irrelevant cues.* Manuscript under review.

Walker, A. R., Luque, D., Le Pelley, M. E., & Beesley, T. (2019). The role of uncertainty in attentional and choice exploration. *Psychonomic Bulletin & Review, 26*, 1911 – 1916. https://doi.org/10.3758/s13423-019-01653-2

Appendix: Simulation details

Table A1 lists the equations we used for computing the simulation results presented in Figure 1. For each model, we determined the associability, α_S, and the change in associative strength, ΔV_S, for each of the cues presented on a given trial. The error term, $\lambda - V$, was determined by the discrepancy between the magnitude of the outcome, λ, and the associative strength of one or more cues, V. The changes in associative strength were computed separately for each of the two outcomes. For each outcome, the λ parameter was set at 1 on the trials in which the outcome was present and at 0 on the trials in which the outcome was absent. The initial value of all V was set at 0. The β parameter, modulating the rate of learning according to characteristics of the outcome, was set at .3. The α_S values were allowed to range between .2 and 1. In the simulation based on Mackintosh's (1975) theory and that based on the Pearce-Hall model (Pearce et al., 1981), we set the starting values of α_S at .5. In Le Pelley's (2004, 2010) hybrid model, α_S is the product of $\alpha_{S \text{ Mackintosh}}$ and $\alpha_{S \text{ Pearce-Hall}}$ (see below). Thus, in the simulation based on that model, the starting values of each of those two α_S were set at .7 ($.7 \times .7 \approx .5$). Additional simulations, using a wide range of parameters, left the order of predictions from each model unchanged.

Given that Mackintosh (1975) did not provide exact rules for computing the associability value, we based Equation A1.1 on a suggestion by Le Pelley (2004; see also Beesley, Nguyen, Pearson, & Le Pelley, 2015). Thus, the prediction error of the target cue S was subtracted from the prediction error of all the other cues, i.e. T–S. The obtained value was multiplied by the parameter μ, which determines the magnitude of associability change and was set at .7. As indicated in Equation A1.2, the change in associative strength in Mackintosh's theory is determined on the basis of the individual prediction error of each cue, $\lambda - V_S$.

Equation A2.1 was taken from the description of the Pearce-Hall model presented by Pearce et al. (1981). In contrast to the original version (see Equation 3), this equation includes parameter γ, which determines whether associability depends exclusively on the most recent trial ($\gamma = 1$) or is weighted by the events that took place on earlier trials ($\gamma < 1$). We set γ at .7. In the Pearce-Hall model, a given cue may develop separate associations with both an outcome and the absence of that outcome, i.e. excitatory associative strength, $V_{ex\,\mathrm{S}}$, and inhibitory associative strength, $V_{in\,\mathrm{S}}$ (see Equations A2.2 and A2.3). The latter increases on trials where the magnitude of the outcome is less than what was expected based on the net associative strength of all the present cues, $V_{net\,\mathrm{T}}$ (i.e. the sum of $V_{ex\,\mathrm{T}}$ and $V_{in\,\mathrm{T}}$).

In the simulation based on the hybrid model (Le Pelley, 2004, 2010), we computed associability by multiplying the values obtained in Equations A1.1 and A2.1, i.e. the associability values determined according to the mechanisms suggested by Mackintosh (1975) and by Pearce et al. (1981). In addition, we took the equation for determining changes in associative strength (Equation A3.2) from the study conducted by Beesley, Nguyen, Pearson, and Le Pelley (2015). This equation is simpler than the one presented by Le Pelley (2004), in that it excludes the individual error term from the computation.

Table A1. Equations used in the simulations presented in Figure 1.

Models and equations	
Mackintosh's (1975) theory	
$\alpha_\mathrm{S} = \alpha_\mathrm{S}^{n-1} + \mu \times (\lvert\lambda^{n-1} - V_{\mathrm{T-S}}^{n-1}\rvert - \lvert \lambda^{n-1} - V_\mathrm{S}^{n-1}\rvert)$	(A1.1)
$\Delta V_\mathrm{S} = \alpha_\mathrm{S} \times \beta \times (\lambda - V_\mathrm{S})$	(A1.2)
Pearce-Hall model (Pearce, Kaye, & Hall 1981)	
$\alpha_\mathrm{S} = \gamma \times \lvert\lambda^{n-1} - V_{net\,\mathrm{T}}^{n-1}\rvert + (1 - \gamma) \times \alpha_\mathrm{S}^{n-1}$	(A2.1)
$\Delta V_{ex\,\mathrm{S}} = \alpha_\mathrm{S} \times \beta \times \lambda$	(A2.2)
$\Delta V_{in\,\mathrm{S}} = \alpha_\mathrm{S} \times \beta \times (V_{net\,\mathrm{T}} - \lambda)$	(A2.3)
Le Pelley's (2004, 2010) hybrid model	
$\alpha_\mathrm{S} = \alpha_{\mathrm{S\ Mackintosh}} \times \alpha_{\mathrm{S\ Pearce\text{-}Hall}}$	(A3.1)
$\Delta V_\mathrm{S} = \alpha_\mathrm{S} \times \beta \times (\lambda - V_\mathrm{T})$	(A3.2)

Hiermit versichere ich, dass ich die vorliegende Dissertation:

> “Prediction error and overt attention to relevant and irrelevant cues: Evidence for an interaction of two associability mechanisms”

selbständig, ohne unerlaubte Hilfe angefertigt und mich dabei keiner anderen als der von mir ausdrücklich bezeichneten Quellen und Hilfen bedient habe.

Die Dissertation wurde in der jetzigen oder einer änlichen Form noch bei keiner anderen Hochschule eingereicht und hat noch keinen sonstingen Prüfungszwecken gedient.

Marburg, ___.___._______

David Torrents Rodas