Brief communication

Surprise-induced enhancements in the associability of Pavlovian cues facilitate learning across behavior systems

Inmaculada Márquez[1], Gabriel Loewinger[2], Juan Pedro Vargas[1], Juan Carlos López[1], Estrella Díaz[1*], Guillem R. Esber[3*]

[1]Animal Behavior and Neuroscience Lab, Department of Experimental Psychology, Faculty of Psychology, University of Seville, 41018, Spain

[2] Department of Biostatistics, Harvard School of Public Health, Boston, USA.

[3]Brooklyn College, City University of New York, USA

IM's current address: Centro de Investigación en Conducta y Cognición Comparada, Universidad de Guadalajara, Centro Universitario de la Ciénega, Av. Universidad 1115, Col. Lindavista, Ocotlán, Jalisco, 47820, México

*These authors should be considered joint senior authors.

Corresponding author: Guillem R. Esber (gesber@brooklyn.cuny.edu)

Keywords: associability, salience, attention, behavior systems, surprise, uncertainty, Pearce-Hall model, counterconditioning

Acknowledgements

This research was supported by Spain's Ministry of Science, Innovation and Universities grant PID2019-110739GB-I00 (ED and JPV) and by the National Institute on Drug Abuse grants 5R00DA036561 and 1R15DA051795 (GE). We declare no conflict of interest, financial or otherwise, related to the current research. We thank Felipe Cabrera for helpful comments on a previous version of the MS.

Author contributions

Inmaculada Marquez: Methodology, Investigation, Validation, Data curation, Visualization, Writing- Original draft preparation; **Gabriel Loewinger:** Formal analysis; **Juan Pedro Vargas**: Project Administration, Funding acquisition; **Juan Carlos López:** Conceptualization, Project Administration, Funding acquisition, Software, Visualization, Methodology, Supervision, Writing- Reviewing and Editing; **Estrella Díaz:** Conceptualization, Project Administration, Funding acquisition, Methodology, Supervision, Writing- Reviewing and Editing**; Guillem R. Esber:** Conceptualization, Writing- Original draft preparation

Author ORCIDs

Inmaculada Márquez: https://orcid.org/0000-0001-7914-6907
Gabriel Loewinger: https://orcid.org/0000-0002-0755-8520
Juan Pedro Vargas: https://orcid.org/0000-0002-4358-5737
Juan Carlos López: https://orcid.org/0000-0001-8106-5006
Estrella Díaz: https://orcid.org/0000-0001-9465-9460
Guillem R. Esber: https://orcid.org/0000-0001-8874-4255 564

Abstract

Surprising violations of outcome expectancies have long been known to enhance the *associability* of Pavlovian cues; that is, the rate at which the cue enters into further associations. The adaptive value of such enhancements resides in promoting new learning in the face of uncertainty. However, it is unclear whether associability enhancements reflect increased associative plasticity within a particular behavior system, or whether they can facilitate learning between a cue and any arbitrary outcome, as suggested by attentional models of conditioning. Here, we show evidence consistent with the latter hypothesis. Violating the outcome expectancies generated by a cue in an appetitive setting (feeding behavior system) facilitated subsequent learning about the cue in an aversive setting (defense behavior system). In addition to shedding light on the nature of associability enhancements, our findings offer the neuroscientist a behavioral tool to dissociate their neural substrates from those of other, behavior system- or valence-specific changes. Moreover, our results present an opportunity to utilize associability enhancements to the advantage of counterconditioning procedures in therapeutic contexts.

In an ever-changing world, brain mechanisms have evolved to modulate the associability of Pavlovian cues in order to meet the learning demands of the environment (e.g., Mitchell & Le Pelley, 2010). One form of modulation is captured by the so-called *uncertainty principle*, according to which a cue's associability increases whenever its consequences are surprising (Pearce & Hall, 1980; Pearce et al., 1982). In support of this notion, cues that predict an outcome inconsistently (i.e., partial reinforcement) are subsequently learned about more rapidly than cues that predict the outcome consistently (i.e., continuous reinforcement; Haselgrove et al., 2010, Collins & Pearce, 1985). Similarly, repeated confirmation of outcome expectancies decreases a cue's associability (Pearce & Hall, 1979; Griffiths et al., 2011; Mackintosh & Turner, 1971), whereas a sudden violation of those expectancies restores it (Hall & Pearce, 1982; Dickinson et al., 1976; Holland, 1984). Surprise-induced associability enhancements have been documented both in appetitive (e.g., Holland, 1984) and aversive (e.g., Dickinson et al., 1976) procedures as well as across phylogenetically distant species (e.g., rats: Kaye & Pearce, 1984; pigeons: Collins & Pearce, 1985; humans: Hogarth et al., 2008; Russo et al., 2019), suggesting they might constitute a widespread, if not universal property of learning systems.

While these findings have fostered important neurobiological discoveries (reviewed in: Holland & Schiffino, 2016; Roesch et al., 2012; Holland & Maddux, 2010) and theoretical developments (Dayan et al., 2000; Le Pelley, 2004; Courville et al., 2006; Pearce & Mackintosh, 2010; Esber & Haselgrove, 2011), the nature and scope of associability enhancements remains poorly understood. On the one hand, such enhancements might reflect a state of heightened associative plasticity involving a specific association or behavior system (e.g., feeding, mating, defense, etc.; Timberlake, 1993; 1994; Cabrera et al., 2019). Such a labile state would facilitate the updating of associative representations involving the cue and outcomes within that behavior system. It follows from this view that a surprise-induced associability enhancement by a food-

predictive cue, for instance, should translate as more rapid learning between that cue and food-related outcomes (including food omission), but not necessarily outcomes related to other behavior systems, such as the presence of a sexual partner (Domjan & Gutiérrez, 2019). On the other hand, associability enhancements might arise from increased attentional processing of the cue, as assumed by attentional models of associative learning (Mackintosh, 1975; Pearce & Hall, 1980). If so, those enhancements should manifest as faster learning regardless of the nature of the outcome and the behavior system engaged. Since studies on associability modulation have traditionally employed a single reinforcer or reinforcer class (thus engaging a single behavior system), this fundamental issue remains unresolved.

To decide between these alternatives, we violated the outcome expectancies generated by a cue in an appetitive setting (feeding system) and tested the associability of the cue in an aversive setting (defense system). To achieve this, we modified a serial prediction task (Wilson et al., 1992) that has been extensively used to investigate the neural substrates of surprise-induced associability changes in rats (e.g., Holland & Gallagher, 1993, 2006; Chiba et al., 1995; Bucci & MacLeod, 2007; Esber et al., 2015). In the original task, a light stimulus is initially followed by a tone that is partially reinforced with food (L→T→food, L→T→nothing). After developing an expectancy of the tone during light presentations, animals in the Surprise condition experience the unexpected omission of the tone on nonreinforced trials (L→T→food, L→nothing), whereas control subjects continue to receive the initial training (L→T→food, L→T→nothing). The omission of the tone is intended to increase the associability of the light without fundamentally changing its predictive or incentive properties (which, if anything, should decrease during tone omission). This increase in associability is typically revealed in a subsequent test in which the light is paired with food (L→food) and more rapid learning is observed in Surprise than control animals. The fact that greater associability can be detected

days after the end of the Surprise phase rules out transient increases in arousal and suggests more enduring changes in the mnemonic representation of the cue.

Here, we tested the associability of the light by pairing it with foot shock in order to determine whether associability increases can be expressed across behavior systems. Our results disconfirmed the hypothesis that associability enhancements reflect heightened plasticity within a particular behavior system. Rather, they are consistent with the view that such enhancements result from increased attentional processing of the cue (Mackintosh, 1975; Pearce & Hall, 1980; Pearce et al., 1982). Our procedure will provide neuroscientists with a tool to dissociate the neural bases of associability changes from those of other, behavior system- or valence-specific changes that a cue representation may undergo during learning. In the clinical setting, our findings suggest the possibility of administering associability-boosting treatments to bolster counterconditioning-based interventions.

Methods

Subjects. Thirty-four experimentally naïve, male Wistar rats were used in the study, run in three cohorts that included animals from all three groups. They were obtained from the Animal Production and Experimentation Center at the University of Seville. Upon arrival, rats were acclimated to the colony room for two weeks with free access to food and water. The colony room was maintained on a 14:10 light/dark cycle schedule at a constant temperature of 21ºC. Rats were housed individually in standard clear-plastic tubs (35×20×20 cm) with woodchip bedding. At the start of the experiment, they were 7–9 weeks old and weighed 230–280 g. One week prior to the beginning of the study, they were food deprived by progressively restricting their diet until they reached 90% of their original body weight and maintained at that weight thereafter. Once training began, they were fed a restricted amount immediately after the experimental sessions. They had free access to water in their home chambers at all times. All procedures and methods were carried out in accordance with the European Directive

2010/69/EU for the maintenance and use of laboratory animals and following Spanish regulations (R.D. 53/02013). The protocol was approved by the Ethics Committee for Animal Research of the University of Seville (Protocol Number: CEEA-US2015-27/4).

Apparatus. Rats were trained in four identical, modular conditioning chambers (31.8 × 25.4 × 34.3 cm, Med Associates, Inc.) enclosed in a ventilated light- and sound-attenuating cubicle (63.5×41.9×49.4 cm, Med Associates, Inc.). An extractor fan was fitted on the right wall of the cubicle and produced a ~60-dB background noise in the conditioning chamber. The side walls of the conditioning chambers were made of aluminum, while the front and back walls and the roof were made of transparent acrylic plastic. The floor consisted of 0.4 mm-diameter steel bars oriented perpendicular to the front wall and spaced 1.4 cm apart as measured from their centers. This floor grid was connected to a shock dispenser capable of delivering a foot shock unconditioned stimulus (US). Each conditioning chamber housed a 6-W white jewel lamp mounted 20 cm above the floor on the center panel of the left wall. Illumination of this lamp provided the visual stimulus used during behavioral training. A speaker was mounted 20 cm above the floor on the left panel of the left wall. This speaker was connected to a tone generator set to deliver a 1500-Hz, 80-dB tone which served as the auditory stimulus used during training. Each chamber also housed a recessed food cup located 2 cm above the floor on the center panel of the right wall. This food cup was equipped with an infrared sensor for detecting head entries and connected to a pellet dispenser capable of delivering 45-g sucrose pellets (DietTM; Mlab Rodent Tablet-45mg; St Andrews University). The chambers remained dark throughout the experimental session except during presentations of the light stimulus. In the same experimental room was a computer running Med PC IV software (Med Associates Inc., St. Albans, VT, USA) on Windows OS which controlled and automatically recorded all experimental events via a Fader Control Interface.

Behavioral Procedure. Rats initially received a single session of magazine training in which a pellet was delivered in the food cup once every minute for a total of 30 minutes. They were then randomly assigned to three groups (Figure 1, table). In the first, Appetitive serial conditioning phase, the No-surprise and Surprise groups received Pavlovian magazine-approach training with a serial compound consisting of a 10-s light immediately followed by a 10-s tone. On half the trials, two pellets were delivered immediately after the termination of the tone (light→tone→food, light→tone→nothing). This training was intended to establish the light as a predictor of the tone. A Naïve group also received partial reinforcement training with the tone, but the latter was not preceded by the light (tone→food, tone→nothing). In each session, 10 trials were presented in pseudorandom order (reinforced and nonreinforced), with the constraint that no more than 2 reinforced trials could occur in succession. The mean intertrial interval (ITI) was 300 s. In this and the remainder of the phases, the total number of magazine head-entries during the cues was taken as a measure of appetitive conditioning. That is, we summed the number of head entries for each animal and for each cue across all trials in a session, and calculated the group means based on those sums. Training continued for 10 sessions conducted over a period of 5 days, with two daily sessions run at 8 am and 3 pm.

In the next, Surprise phase, the No-surprise and Naïve groups continued to receive the same training for an additional 4 sessions (conducted over 2 days as in the first phase). In the Surprise group, however, the tone was omitted on nonreinforced trials (light→tone→food, light→nothing) in order to boost the associability of the light (Wilson et al., 1992). All other procedural details remained the same in this phase.

On the next day, at 9 am, all groups received a threat conditioning session in which a single presentation of the light was followed by a 0.25-mA foot shock (L→shock). We arrived at the use of a single 0.25 mA shock after piloting our experimental design with two 0.5-mA and one 0.5-mA shocks and finding a floor effect; that is, almost complete suppression during the final test in all three groups. This single threat conditioning trial was preceded and followed by a

300-s period. No responses were recorded on this session. Later in the day, at 4 pm, rats were placed back in the conditioning chambers to receive an appetitive session consisting of 5 reinforced trials with the tone. The mean ITI was 300 s. The purpose of the latter session was to extinguish contextual threat conditioning and provide a baseline of magazine approach to the tone across the groups ahead of the final test.

On the following day, at 9 am, all groups received a suppression test consisting of 4 trials with the light and the tone presented simultaneously and reinforced with the delivery of two pellets (LT→food). The purpose of this test was to measure threat conditioning to the light by assessing the extent to which it was capable of suppressing magazine approach during the tone relative to the tone baseline taken on the previous day. If the unexpected omission of the tone in the Surprise group enhances the light's associability, and if associability changes can cross behavior systems, then greater suppression of responding to the tone should be observed in that group relative to the No-surprise group. The Naïve group provided a positive control for associability since those animals experienced the light as a novel stimulus on the threat conditioning session. Thus, we expected threat conditioning to be the strongest in Naïve rats.

Statistical analysis. Analyses were conducted in R version 3.6.1. Generalized Linear Models (GLMs) were conducted using the *stats* package, Generalized Linear Mixed Models (GLMMs) were conducted using *lme4* package. To assess magazine approach performance in the Appetitive Serial Conditioning and Surprise phases, we collapsed (summed responses) across the last two sessions of each phase in order to probe asymptotic behavior. Before running any statistical inference, we selected to proceed with this contrast (the sum of the last two sessions) since the progression of responding across sessions was not of interest. We conducted all analyses with a single Generalized Linear Mixed Effects Model, adopting a Poisson as the conditional distribution of our outcome given the random effects and the covariates. We included a random intercept for each rat. We proceeded with analyses that adopted a count

distribution for the outcome because all responses were head entries and thus treating the outcome as continuous (e.g., using ANOVA, repeated measures ANOVAs or t-tests) would not yield valid inference given the sample sizes in this study. For the Surprise phase, animals in the Surprise group received 5 trials in a session, whereas animals in the No-Surprise and Naïve groups received 10. To compare rates of responding between conditions with different numbers of trials, we included in the model an offset of the log number of trials that each animal received. For all statistical analyses in which post-hocs were necessary, we used Bonferroni corrections to account for multiple comparisons. Post-hocs analyses were conducted with the *glht* function in the *multicomp* package in R. All statistical tables are shown in the Supplementary Materials. Data as well as code to reproduce statistical analyses and tables are available at the github repository: https://github.com/gloewing/marquez-et-al-2021.

To assess magazine approach performance on the Tone baseline and Suppression test phases, we proceeded by collapsing across trials. We opted not to conduct a repeated measures analysis and no analyses of that kind were ever inspected. This was motivated by the fact that the progression of responding across trials within the test day was not of interest and thus the associated loss in statistical power from the increase in parameters we would need to estimate was not justifiable. The temporal nature of the data (i.e., the trial-specific structure) was a nuisance needed to probe the impact of the behavioral task, but did not provide any meaningful or interpretable information. Before conducting any statistical inference, all analyses were planned to avoid having to conduct any adjustment for multiple comparisons and to ensure analysis results were not selected to maximize statistical significance. Any comparisons of models were conducted without viewing p-values, confidence intervals or otherwise. Moreover, models were parameterized to provide the comparisons/contrasts of interest and thus no post-hocs were necessary. As such, Naive - No-surprise comparisons were not conducted. To assess whether there were differences in observed rates of head entries across the entire test day between the three groups, we conducted a GLM adjusting for baseline responding.

241 Specifically, we included

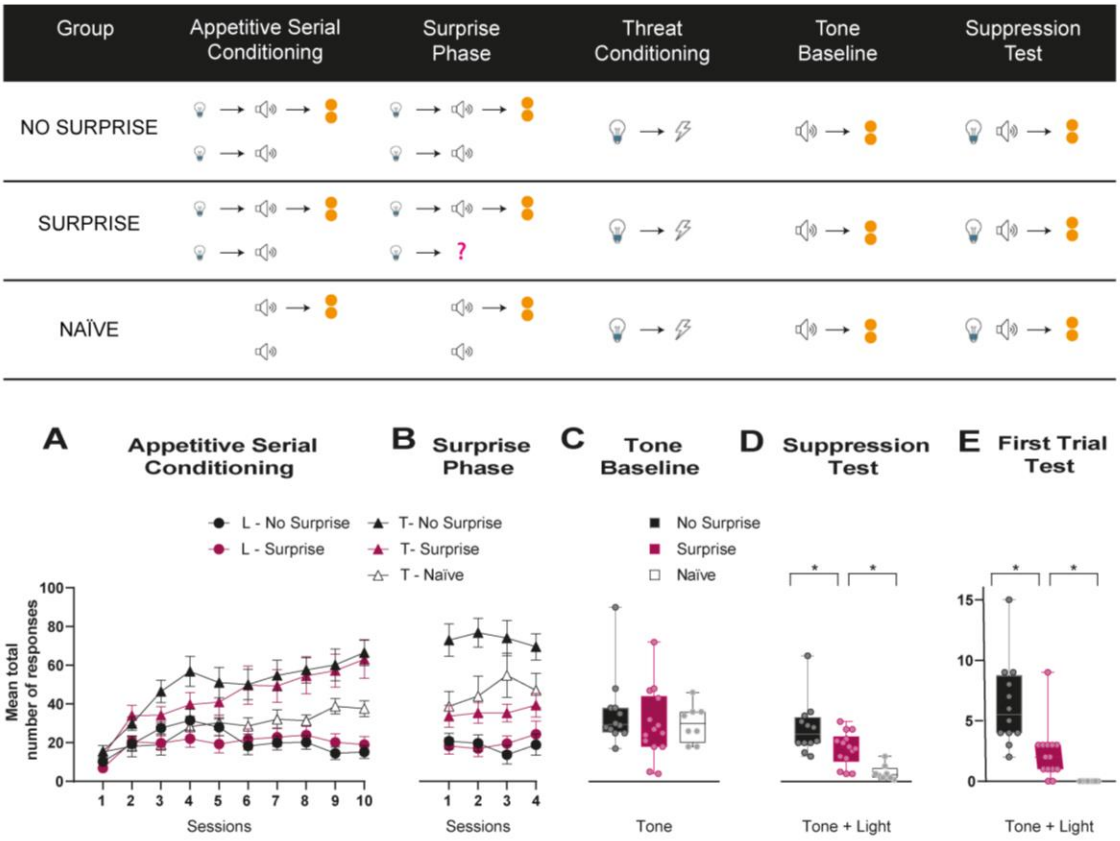

Figure 1. The table at the top shows the experimental design, a modified version of the Wilson et al. (1992) serial prediction task. Two groups of rats, Surprise and No Surprise, received appetitive serial conditioning in which a light was followed by a tone that signaled the delivery of food on a partial-reinforcement basis. As expected given their temporal arrangement, the tone evoked more conditioned responding than the light in both the Surprise and No-Surprise groups **(Panel A)**. A third, Naïve group also received partial reinforcement with the tone, but the latter was never preceded by the light. Following this phase, the No-surprise and Naïve groups continued to receive identical training, but in the Surprise group the tone was unexpectedly omitted on nonreinforced trials—a treatment that has repeatedly been shown to enhance the associability of the light. Performance during this phase is shown in **Panel B**. All groups then received a threat conditioning session in which the light was paired with foot shock. Later that day, all rats were placed back in the conditioning boxes and given reinforced presentations with the tone alone to provide a measure of baseline responding **(Panel C)**. A subsequent Suppression test in which the light was presented simultaneously with the tone in extinction revealed greater suppression (i.e., more threat conditioning to the light) in the Surprise than No-surprise group, indicating that surprise-induced associability enhancements can cross behavior system boundaries. As expected, the greatest level of suppression was observed in Naïve rats, for whom the light was novel during threat conditioning **(Panel D)**. This pattern was apparent on the very first test trial **(Panel E)**.

baseline responding as a covariate in the model so that the interpretation of the parameter

estimates corresponding to group differences were all conditional on fixed levels of baseline responding. A negative binomial with a log link was employed to account for potential overdispersion. During model building, we fit Poisson, negative binomial and quasi-Likelihood approach (quasi-Poisson) models and before examining p-values, compared models based upon the degree to which it accounted for overdispersion. To determine whether the model accounted for potential overdispersion we inspected fitted values vs. squared Pearson residual plots and conducted the appropriate likelihood ratio test using Pearson residuals. The only model that did not reach a statistically significant test for the presence of overdispersion was the negative-binomial model (Tone baseline: p= 0.255; Suppression test: p = 0.0996) and thus we based inference off this model for both phases.

Results

Panel A of Figure 1 depicts the mean total (i.e., summed across trials) number of responses to the light (L) and tone (T) cues across the 10 sessions of the Appetitive serial conditioning phase. As expected, based on the serial arrangement of the cues, asymptotic responding to the tone (i.e., last two sessions) was significantly higher in the Surprise and No Surprise groups than that to the light. Surprise and No-surprise rats responded to the light at a rate that was, respectively, 67.4% (95% CI: [64.1%, 70.4%], $p<0.001$) and 76.5% (95% CI: [73.6%, 79.1%], $p<0.001$) less than they did to the tone, conditional on animal specific random intercepts. There were no statistically significant differences between Surprise and No-surprise animals in their rate of responding to the light ($p=1$) or the tone ($p=1$), conditional on the random effects. Likewise, no significant differences in the rate of responding to the tone were detected between the Surprise and Naïve ($z=1.563$; $p=0.709$), or the No-surprise and Naïve groups, conditional on the random effects ($z=-1.987$; $p=0.282$). The numerically greater rate of responding to the tone in the Surprise and No-surprise groups relative to the Naïve group, however, is likely explained

by the fact that the tone was signalled by the light in the former groups, allowing the animals to prepare for its arrival and respond at the magazine at cue onset.

In the Surprise phase, conditioned responding to the cues proceeded in similar fashion in all groups (Panel B, Figure 1). Inspection of Panel B suggests that Surprise rats responded to the tone much less than in the previous stage, but this is of course an artifact of those animals receiving half of the tone presentations relative to the prior stage and the other two groups. To account for this difference, we included an offset in the model (see Methods). With this adjustment, we found that, as in the previous stage, the Surprise and No-surprise groups responded to the light at a rate that was substantially less than that to the tone (71.4% in Surprise rats, 95% CI: [68.4%, 74.1%]), $p<0.001$, and 77.2% in No-surprise rats, 95% CI: [74.6%, 79.6%], $p<0.001$). There was no significant difference between these groups in their rates of responding to the light ($z=1.096$; $p=1$). Likewise, no significant differences were detected between these groups, or between either of them and the Naïve group, in their rate of responding to the tone (Surprise vs. No-surprise, $z= 0.068$; $p=1$; Surprise vs. Naïve, $z= 1.685$; $p=0.552$; No-surprise vs. Naïve, $z= -1.584$; $p=0.679$). All interpretations for this phase are all conditional on rat-specific random intercepts.

Following the threat conditioning session with the light, reinforced presentations with the tone during the Tone-baseline phase (Panel C, Figure 1) produced no statistically significant differences between the Surprise and No-surprise groups ($z=0.992$; $p=0.321$), or between the Surprise and Naïve and groups ($z=0.076$; $p=0.940$). Crucially, in the subsequent Suppression test, greater suppression of magazine activity during the light/tone compound was observed in the Surprise than the No-surprise group (Panel D, Figure 1). Indeed, adjusting for baseline responding to the tone, the No-surprise group responded 41.8% (95% CI: [7.6%, 86.9%]) more than the Surprise group ($z=2.479$; $p=0.013$) across the entire test session. This difference suggests that the unexpected omission of the tone was effective in increasing the associability

of the light, and that such an increase facilitated threat conditioning with that stimulus. Interestingly, the surprising omission of the tone did not fully restore the light's associability to its original (novelty) levels, as suggested by the even greater suppression of magazine activity observed in the Naïve group in the test. Adjusting for baseline responding to the tone, the Naïve group responded 71.9% less (95% CI: [57.1%, 81.5%]) than the Surprise group (z=-5.886; p<0.001) across the test session. In addition to differences in preexposure history, a potential source for the latter outcome is the pre-existing association of the light with the positive-valence tone in the Surprise group, which may have interfered with the acquisition or expression of threat conditioning at test.

To bolster these findings, we conducted two additional analyses of our test results. The first of these analyses was motivated by the presence of an outlier in the No-surprise group whose responses where unusually high on the test day, as revealed by inspection of Panel D. To rule out the possibility that this outlier may have driven the critical difference between the Surprise and No-Surprise groups, we repeated the above analysis in its absence. This analysis confirmed a significance difference between these groups (z=2.429; p=0.015), with the No-surprise group responding 43.4% (95% CI: [7.2%, 91.8%]) more than the Surprise group.

The second analysis focused exclusively on the first trial of the Suppression test (Panel E, Figure 1). We reasoned that the light's increased associability in the Surprise group would facilitate the acquisition of a light→food association across tone/light→food test trials, thereby weakening the light's enhanced ability to suppress responding to the tone relative to the No-Surprise group. For this reason, we predicted the difference between these two groups on the first test trial to be particularly pronounced. Consistent with this prediction, the No-surprise group responded 145.3% (95% CI: [60.8%, 274.2%]) more than the Surprise group (z=4.164; p < 0.001). Eliminating the No-Surprise group's outlier produced similar results (z=3.804; p <

0.001), with the No-surprise group responding 146.4% (95% CI: [54.8%,292.1%]) more than the Surprise group.

Discussion

Here, we employed a modified version of a serial prediction task (Wilson et al., 1992) to examine the scope of surprise-induced associability changes to predictive cues during learning. Specifically, we violated the expectancies generated by a light serially conditioned with food (feeding system) and tested the associability of this cue by pairing it with foot shock, an aversive exteroceptive stimulus engaging the defense system. To our knowledge, this is the first demonstration that associability enhancements can be expressed across behavior systems. Before proceeding, it is worth noting that the associability of Pavlovian cues is not only modulated by how uncertain the outcome is (the uncertainty principle) but also—and paradoxically—by how well the cue predicts it (the *predictiveness principle*) (Mackintosh, 1975; Luque et al., 2017; for an attempt at resolving this paradox, see Esber & Haselgrove, 2011). According to Mackintosh's theory, the associability of a cue increases to the extent that the cue proves to be a better predictor of the outcome than all other stimuli present (Mackintosh, 1975; Haselgrove et al., 2010). In the serial prediction task employed here, the theory predicts that, if anything, the associability of the light should decline during the Surprise phase as the cue becomes a relatively worse predictor of the tone—a secondary reinforcer. If, as proposed by some (LePelley, 2004; Pearce & Mackintosh, 2010), the uncertainty and predictiveness principles reflect two competing mechanisms of associability modulation under the control of distinct neural circuits (Holland & Maddux, 2010), then it is unclear whether our findings would generalize to predictiveness-driven associability increases.

With that caveat in mind, our findings carry important implications regarding the nature of associability changes. From a behavior system's approach (Timberlake, 1993; 1994), the results may be explained in terms of the close relationship between the feeding and defense systems.

For many animals, foraging for food implies increasing their exposure to predators, making it essential to simultaneously attend to signals for food and threat. Crucially, preys and predators may share common predictive cues (e.g., a glimpse of a moving object, a rustle in the undergrowth), and thus it makes adaptive sense for associability increases to food cues to also benefit learning in connection with threat. It remains to be established, however, whether associability increases occurring within the defensive system can in turn be expressed in the feeding system, or indeed whether associability changes can universally transfer across any arbitrary pair of behavior systems. Such is of course the prediction of attentional theories of associative learning (e.g., Pearce & Hall, 1980; Mackintosh, 1975; Esber & Haselgrove, 2011). From this perspective, associability increments reflect greater attentional processing of the cue rather than enhanced associative plasticity within a particular behavior system. This assumption is consistent with evidence that the same uncertainty conditions that facilitate learning also promote stronger overt attentional responses to the cue (Kaye & Pearce, 1984; Swan & Pearce, 1988; Beesley et al., 2015; Luque et al., 2017; Easdale et al., 2019).

Critically, our assertion that associability enhancements can be expressed across behavior systems hinges on the assumption that the light—the target cue in the current study—was able to gain access to both the feeding and defense systems. One question raised by the low level of magazine approach evoked by the light in the Surprise and No-surprise groups during the first two phases of the study is whether this stimulus was capable of engaging the feeding system at all. That it did so is suggested by the greater level of responding to the tone in these groups relative to the Naïve group, which, as mentioned above, is readily explained if the light alerted the rats of the imminence of the tone and prepared them to respond at the magazine. A related question is whether the greater suppression of magazine activity observed at test in Surprise than No-surprise animals truly reflects a stronger activation of the defense system by the light (e.g., more freezing) or simply its greater proclivity to elicit orienting

responses that compete with magazine approach. While our study does not directly address this possibility, it should be noted that, by the same token, one would expect orienting responses to the light to increase similarly in the Surprise group of a standard serial prediction task, where the associability of the light is tested by directly pairing it with food. If so, those orienting responses should be expected to compete with magazine approach there as well, thus hindering rather than facilitating conditioning, which is of course the opposite result to that typically observed.

A competing explanation for the greater suppression of magazine activity we observed at test in Surprise relative to No-surprise animals is that the omission of the tone in the Surprise phase may have extinguished some of the positive valence of the light rather than enhancing its associability. Compounded with threat conditioning, this reduced positive valence would, according to this account, produce less magazine approach at test relative to No-surprise animals. The issue with this interpretation is that it predicts a drop in responding to the light (i.e., extinction) in Surprise animals during the Surprise phase, which was not observed. Given the low level of responding to the light (a serially conditioned cue), it is possible that such extinction was obscured by a floor effect. We regard it as more plausible, however, that the unaltered relationship between light and food as well as the remaining tone presentations in the Surprise group prevented a loss of positive valence substantial enough to produce a sizeable decrement at test.

Due to various advantages, the standard serial prediction task has been extensively used to characterize the neural substrates of surprise-induced associability enhancements. One such advantage is that it permits decoupling the encoding of associability increases at the time of surprise induction from the expression of those increases at the time of learning. This advantage has permitted the discovery, for instance, that the central nucleus of the amygdala (CeA; Holland & Gallagher, 2006) and the substantia nigra pars compacta (SNc; Lee et al.,

2008) are critical for the encoding, but not the expression of associability increases, although this has only been demonstrated in appetitive conditioning with food. On the other hand, the substantia innominata/nucleus basalis magnocellularis (SI/nBM; Holland & Gallagher, 2006), the secondary visual cortex (V2; Schiffino & Holland, 2016), and the dorsolateral striatum (DLS; Asem et al., 2015) are necessary for the expression, but not the encoding of associability increases. Such associability expression, however, has only been tested within the same behavior system (feeding), and thus it is unclear whether these regions would also be necessary in the current version of the task. Interestingly, the posterior parietal cortex (PPC), which has long been implicated in attention in humans and non-human primates (e.g., Mesulam, 1981; Posner & Petersen, 1990; Desimone & Duncan, 1995; Corbetta & Shulman, 2002) is so far the only region identified as being critical to the encoding, consolidation and expression of associability enhancements (Schiffino et al., 2014), suggesting it might constitute a locus for storing the cue-specific associability memory. Whether this mnemonic representation in PPC is fully detached from the motivational, emotional and behavior system-specific properties of the cue (i.e., whether it provides a neural substrate for the results observed here) remains to be established. Once again, the current procedure should help make this determination.

The present findings may also carry clinical significance. For instance, they suggest the possibility of expediting counterconditioning procedures by coupling them with associability-boosting manipulations (Keller et al., 2020). In behavioral therapy, counterconditioning refers to a collection of procedures that seek to modify maladaptive behaviors by associating their triggering events with an outcome of the opposite affective valence (Keller et al., 2020; Konorski & Szwejkowska, 1956). Recent studies show that counterconditioning therapies have greater efficacy (Engelhard et al., 2014; Kerkhof et al., 2011; Raes & De Raedt, 2012) and resistance to relapse (Kang et al., 2018) than exposure therapies based on extinction procedures. This relative advantage is thought to derive from the greater evaluative learning that takes place when the triggering stimulus is experienced not merely in the absence of its associated outcome

(as in extinction), but in the presence of another of opposite affective sign. On the downside, however, counterconditioning typically requires more training than de novo acquisition or extinction (Scavio & Gormezano, 1980; Peck & Bouton, 1990; Bouton & Peck, 1992), and, to that extent, it could benefit from prior manipulations that enhance the associability of the stimuli and responses involved. The induction of surprise in a manner similar to that used here might provide one such manipulation. While our findings remain to be extended and replicated in humans, they point in a promising direction for future research.

References

Asem, J. S. A., Schiffino, F. L., Holland, P. C. (2015). Dorsolateral striatum is critical for the expression of surprise-induced enhancements in cue associability. *European Journal of Neuroscience*, 42, 2203–2213. https://doi.org/10.1111/ejn.13001

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(11), 2175-2199. https://doi.org/10.1080/17470218.2015.1009919

Bouton, M.E., Peck, C.A. (1992) Spontaneous recovery in cross-motivational transfer (counterconditioning). *Animal Learning & Behavior* 20, 313–321. https://doi.org/10.3758/BF03197954

Bucci, D. J., & MacLeod, J. E. (2007). Changes in neural activity associated with a surprising change in the predictive validity of a conditioned stimulus. *European Journal of Neuroscience, 26*(9), 2669–2676. https://doi.org/10.1111/j.1460-9568.2007.05902.x

Cabrera, F., Jiménez, Á. A., & Covarrubias, P. (2019). Timberlake's behavior systems: A paradigm shift toward an ecological approach. *Behavioural Processes, 167,* Article 103892. https://doi.org/10.1016/j.beproc.2019.103892

Chiba, A. A., Bucci, D. J., Holland, P. C., Gallagher, M. (1995) Basal forebrain cholinergic lesions disrupt increments but not decrements in conditioned stimulus processing. *J Neurosci* 15:7315–7322. https://doi.org/10.1523/JNEUROSCI.15-11-07315.1995

Collins, L., & Pearce, J. M. (1985). Predictive accuracy and the effects of partial reinforcement on serial autoshaping. *Journal of Experimental Psychology: Animal Behavior Processes, 11*(4), 548–564. https://doi.org/10.1037/0097-7403.11.4.548

Corbetta, M., & Shulman, G. L. (2002). Control of goal-directed and stimulus-driven attention in the brain. *Nature reviews. Neuroscience*, 3(3), 201–215. https://doi.org/10.1038/nrn755

Courville, A. C., Daw, N. D., & Touretzky, D. S. (2006). Bayesian theories of conditioning in a changing world. *Trends in Cognitive Sciences*, *10*(7), 294-300. https://doi.org/10.1016/j.tics.2006.05.004

Dayan, P., Kakade, S., & Montague, P. R. (2000). Learning and selective attention. *Nature neuroscience*, *3 Suppl*, 1218–1223. https://doi.org/10.1038/81504

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

Dickinson, A., Hall, G., & Mackintosh, N. J. (1976). Surprise and the attenuation of blocking. *Journal of Experimental Psychology: Animal Behavior Processes, 2*(4), 313322. https://doi.org/10.1037/0097-7403.2.4.313

Domjan, M., & Gutiérrez, G. (2019). The behavior system for sexual learning. *Behavioural processes*, *162*, 184-196. https://doi.org/10.1016/j.beproc.2019.01.013

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 (2006), 72(2), 193–208. https://doi.org/10.1080/17470218.2017.1363257

Engelhard, I. M., Leer, A., Lange, E., Olatunji, B. O. (2014) Shaking That Icky Feeling: Effects of Extinction and Counterconditioning on Disgust-Related Evaluative Learning. *Behavior Therapy*, Vol 45, (5), 708-719. https://doi.org/10.1016/j.beth.2014.04.003.

Esber, G. R., Haselgrove, M. (2011) Reconciling the influence of predictiveness and uncertainty on stimulus salience: a model of attention in associative learning. *Proceedings of the Royal Society B: Biological Sciences* 278(1718):2553-61. https://doi.org/10.1098/rspb.2011.0836

Esber, G. R., Torres-Tristani, K., Holland, P. C. (2015). Amygdalo-striatal interaction in the enhancement of stimulus salience in associative learning. *Behav Neurosci*.;129(2):87-95. http://dx.doi.org/10.1037/bne0000041

Griffiths, O., Johnson, A. M., & Mitchell, C. J. (2011). Negative Transfer in Human Associative Learning. *Psychological Science*, *22*(9), 1198–1204. https://doi.org/10.1177/0956797611419305

Hall, G., & Pearce, J. M. (1982). Restoring the Associability of a Pre-Exposed CS by a Surprising Event. *The Quarterly Journal of Experimental Psychology Section B*, *34*(3b), 127–140. https://doi.org/10.1080/14640748208400881

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*(4), 456–470. https://doi.org/10.1037/a0018528

Hogarth, L., Dickinson, A., Janowski, M., Nikitina, A., & Duka, T. (2008). The role of attentional bias in mediating human drug-seeking behaviour. *Psychopharmacology*, *201*(1), 29–41. https://doi.org/10.1007/s00213-008-1244-2

Holland, P. C. (1984). Unblocking in Pavlovian appetitive conditioning. *Journal of Experimental Psychology: Animal Behavior Processes, 10*(4), 476–497. https://doi.org/10.1037/0097-7403.10.4.476

Holland, P. C., & Gallagher, M. (1993). Amygdala central nucleus lesions disrupt increments, but not decrements, in conditioned stimulus processing. *Behavioral neuroscience*, *107*(2), 246–253. https://doi.org/10.1037//0735-7044.107.2.246

Holland, P. C., & Gallagher, M. (2006). Different roles for amygdala central nucleus and substantia innominata in the surprise-induced enhancement of learning. *The Journal of neuroscience : the official journal of the Society for Neuroscience,* 26(14), 3791–3797. https://doi.org/10.1523/JNEUROSCI.0390-06.2006.

Holland PC, Maddux J-M. (2010). Brain systems of attention in associative learning. In: Mitchell CJ, LePelley ME, editors. *Attention and learning.* Oxford University Press; Oxford; pp. 305–349

Holland, P. C., Schiffino, F. L. (2016) Mini-review: Prediction errors, attention and associative learning. *Neurobiol Learn Mem*.;131:207-15. https://doi.org/10.1016/j.nlm.2016.02.014

Kang, S., Vervliet, B., Engelhard, I. M., van Dis, E., & Hagenaars, M. A. (2018). Reduced return of threat expectancy after counterconditioning verus extinction. *Behaviour research and therapy*, *108*, 78–84. https://doi.org/10.1016/j.brat.2018.06.009

Kaye, H., & Pearce, J. M. (1984). The strength of the orienting response during Pavlovian conditioning. *Journal of Experimental Psychology: Animal Behavior Processes, 10*(1), 90–109. https://doi.org/10.1037/0097-7403.10.1.90

Keller, N. E.; Hennings, A. C.; Dunsmoor, J. E. (2020) Behavioral and neural processes in counterconditioning: Past and future directions. *Behaviour Research and Therapy*, Vol 125, 103532. https://doi.org/10.1016/j.brat.2019.103532.

Kerkhof, I., Vansteenwegen, D., Baeyens, F., Hermans, D. (2011) Counterconditioning: An effective technique for changing conditioned preferences. *Experimental Psychology*, 58, pp. 31-38, https://doi.org/10.1027/1618-3169/a000063

Konorski, J., Szwejkowska, G. (1956) Reciprocal transformations of heterogeneous conditioned reflexes. *Acta Biologica Experimentalis*, 11, 141-165.

Le Pelley, M. E. (2004). The role of associative history in models of associative learning: A selective review and a hybrid model. *The Quarterly Journal of Experimental Psychology B: Comparative and Physiological Psychology, 57B*(3), 193–243. https://doi.org/10.1080/02724990344000141

Lee, H. J., Youn, J. M. O. M. J., Gallagher, M., Holland, P. C. (2008) Temporally-limited role of substantia nigra-central amygdala connections in surprise-induced enhancement of learning. *Eur J Neurosci*.; 27(11): 3043–3049. https://doi.org/10.1111/j.1460-9568.2008.06272.x

Luque, D., Vadillo, M. A., Le Pelley, M. E., & Beesley, T. (2017). Prediction and uncertainty in associative learning: Examining controlled and automatic components of learned attentional biases. *Quarterly Journal of Experimental Psychology*, 70(8), 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*, Vol. 82, No. 4, 276-298. https://doi.org/10.1037/h0076778

Mackintosh, N. J. & Turner, C. (1971) Blocking as a function of novelty of CS and predictability of UCS. *Quarterly Journal of Experimental Psychology*, 23:4, 359-366. https://doi.org/10.1080/14640747108400245

Mesulam M. M. (1981). A cortical network for directed attention and unilateral neglect. *Annals of neurology*, 10(4), 309–325. https://doi.org/10.1002/ana.410100402

Mitchell, C. J., Le Pelley, M. E. (2010) Attention and Learning. Cambridge University Press, Cambridge.

Pearce, J. M., & Hall, G. (1979). Loss of associability by a compound stimulus comprising excitatory and inhibitory elements. *Journal of Experimental Psychology: Animal Behavior Processes, 5*(1), 19–30. https://doi.org/10.1037/0097-7403.5.1.19

Pearce, J. M., & Hall, G. (1980). A model for Pavlovian learning: Variations in the effectiveness of conditioned but not of unconditioned stimuli. *Psychological Review, 87*(6), 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 *Attention and associative learning: from brain to behaviour* (eds Mitchell C. J., Le Pelley M. E.), pp. 11–39 Oxford, UK: Oxford University Press.

Pearce, J. M., Kaye, H., & Hall, G. (1982). Predictive accuracy and stimulus associability: Development of a model for Pavlovian conditioning. In M. L. Commons, R. J. Herrnstein, & A. R. Wagner (Eds.), *Quantitative analyses of behavior: Acquisition, vol 3* (pp. 241–255). Cambridge, MA: Ballinger.

Peck, C. A.; Bouton, M. E. (1990) Context and performance in aversive-to-appetitive and appetitive-to-aversive transfer. *Learning and Motivation*, Vol. 21, (1), 1–31. doi: https://doi.org/10.1016/0023-9690(90)90002-6

Posner, M.I. & Petersen, S.E. (1990). The attention system of the human brain. *Annu. Rev. Neurosci., 13*, 25–42. https://doi.org/10.1146/annurev.ne.13.030190.000325

Raes, A. K., Raedt, R. D. (2012) The Effect of Counterconditioning on Evaluative Responses and Harm Expectancy in a Fear Conditioning Paradigm. *Behavior Therapy*, Vol 43, (4), 757-767, https://doi.org/10.1016/j.beth.2012.03.012.

Roesch, M. R., Esber, G. R., Li, J., Daw, N. D., & Schoenbaum, G. (2012). Surprise! Neural correlates of Pearce-Hall and Rescorla-Wagner coexist within the brain. *The European journal of neuroscience*, *35*(7), 1190–1200. https://doi.org/10.1111/j.1460-9568.2011.07986.x

Russo, S., Burns, N., & Baetu, I. (2019). Mackintosh, pearce-hall and time: An EEG study on Inhibition ofreturn. *Biological psychology*,146, 107731. https://doi.org/10.1016/j.biopsycho.2019.107731

569 Scavio, M. J., & Gormezano, I. (1980). Classical–classical transfer: Effects of prior appetitive
570 conditioning upon aversive conditioning in rabbits. *Animal Learning & Behavior, 8*(2),
571 218–224. https://doi.org/10.3758/BF03199598
572 Schiffino, F. L., & Holland, P. C. (2016). Secondary visual cortex is critical to the expression of
573 surprise-induced enhancements in cue associability in rats. *European Journal of*
574 *Neuroscience*, *44*(2), 1870-1877. https://doi.org/10.1111/ejn.13286
575 Schiffino, F. L., Zhou, V., Holland, P. C. (2014). Posterior parietal cortex is critical for the
576 encoding, consolidation, and retrieval of a memory that guides attention for learning.
577 *European Journal of Neuroscience, 39*, 640–649. https://doi.org/10.1111/ejn.12417
578 Swan, J. A. & Pearce, J. M. (1988). The orienting response as an index of stimulus associability
579 in rats. *J. Exp. Psychol: Anim. Behav. Proc. 4,* 292–301.
580 https://doi.org/10.1037/0097-7403.14.3.292
581 Timberlake W. (1993). Behavior systems and reinforcement: an integrative approach. *Journal of*
582 *the experimental analysis of behavior*, *60*(1), 105–128.
583 https://doi.org/10.1901/jeab.1993.60-105
584 Timberlake, W. (1994) Behavior systems, associationism, and Pavlovian conditioning. *Psychon*
585 *Bull Rev* 1, 405–420. https://doi.org/10.3758/BF03210945
586 Wilson, P. N., Boumphrey, P., & Pearce, J. M. (1992). Restoration of the Orienting Response to
587 a Light by a Change in its Predictive Accuracy. *The Quarterly Journal of Experimental*
588 *Psychology Section B*, *44*(1b), 17–36. https://doi.org/10.1080/02724999208250600