Emotion

Learning Biases to Angry and Happy Faces During Pavlovian Aversive Conditioning

Yoann Stussi, Gilles Pourtois, Andreas Olsson, and David Sander

Online First Publication, March 19, 2020. http://dx.doi.org/10.1037/emo0000733

American Psychological Association

Emotion

ISSN: 1528-3542

http://dx.doi.org/10.1037/emo0000733

Learning Biases to Angry and Happy Faces During Pavlovian Aversive Conditioning

Yoann Stussi
University of Geneva

Gilles Pourtois
Ghent University

Andreas Olsson
Karolinska Institute

David Sander
University of Geneva

Learning biases in Pavlovian aversive conditioning have been found in response to specific categories of threat-relevant stimuli, such as snakes or angry faces. This has been suggested to reflect a selective predisposition to preferentially learn to associate stimuli that provided threats to survival across evolution with aversive outcomes. Here, we contrast with this perspective by highlighting that both threatening (angry faces) and rewarding (happy faces) social stimuli can produce learning biases during Pavlovian aversive conditioning. Using a differential aversive conditioning paradigm, the present study ($N = 107$) showed that the conditioned response to angry and happy faces was more readily acquired and more resistant to extinction than the conditioned response to neutral faces. Strikingly, whereas the effects for angry faces were of moderate size, the conditioned response persistence to happy faces was of relatively small size and influenced by interindividual differences in their affective evaluation, as indexed by a Go/No-Go Association Task. Computational reinforcement learning analyses further suggested that angry faces were associated with a lower inhibitory learning rate than happy faces, thereby inducing a greater decrease in the impact of negative prediction error signals that contributed to weakening extinction learning. Altogether, these findings provide further evidence that the occurrence of learning biases in Pavlovian aversive conditioning is not specific to threat-related stimuli and depends on the stimulus' affective relevance to the organism.

Keywords: Pavlovian conditioning, learning, emotion, happy faces, angry faces

Supplemental materials: http://dx.doi.org/10.1037/emo0000733.supp

Learning to predict and anticipate impending threats in the environment holds a critical survival value to organisms (e.g., LeDoux & Daw, 2018). A basic form of learning whereby this skill is achieved is Pavlovian aversive conditioning (e.g., Delgado, Olsson, & Phelps, 2006; LaBar, Gatenby, Gore, LeDoux, & Phelps, 1998; Phelps & LeDoux, 2005). In this procedure, organisms learn to associate a stimulus from the environment (the conditioned stimulus) with a biologically aversive outcome (the unconditioned stimulus) through single or repeated contingent pairing (Pavlov, 1927; Rescorla, 1988), thereby endowing the conditioned stimulus with a predictive and emotional value eliciting an anticipatory response (the conditioned response). Research on Pavlovian conditioning has generally focused on identifying principles that apply across different types of stimuli irrespective of their nature (Pavlov, 1927; Pearce & Hall, 1980; Rescorla & Wagner, 1972). Certain associations have, however, been revealed to be more easily formed and maintained than others (Garcia & Koelling, 1966; Öhman & Mineka, 2001; Seligman, 1970, 1971).

Yoann Stussi, Swiss Center for Affective Sciences, Campus Biotech and Laboratory for the Study of Emotion Elicitation and Expression, Department of Psychology (FPSE), University of Geneva; Gilles Pourtois, Cognitive and Affective Psychophysiology Laboratory, Department of Experimental Clinical and Health Psychology, Ghent University; Andreas Olsson, Department of Clinical Neuroscience, Division of Psychology, Karolinska Institute; David Sander, Swiss Center for Affective Sciences, Campus Biotech and Laboratory for the Study of Emotion Elicitation and Expression, Department of Psychology (FPSE), University of Geneva.

This research was supported by the National Center of Competence in Research (NCCR) Affective Sciences, financed by the Swiss National Science Foundation (51NF40-104897), and hosted by the University of Geneva, and by a Doc.CH Grant (P0GEP1_159057) and an Early Postdoc. Mobility fellowship (P2GEP1_187911) from the Swiss National Science Foundation to Yoann Stussi. We thank Chloé Da Silva Coelho for her help with data collection, as well as Sylvain Delplanque and Eva R. Pool for their insightful comments on this work. The data reported in the present study and the code used for data analysis are available on the Open Science Framework (https://doi.org/10.17605/OSF.IO/DK2NP).

Correspondence concerning this article should be addressed to Yoann Stussi, who is now at the Department of Psychology, Harvard University, Northwest Lab Building, 52 Oxford Street, Cambridge, MA 02138. E-mail: ystussi@fas.harvard.edu

Surprisingly, mechanisms underlying such learning biases remain yet not well elucidated.

Major theoretical models put forward, such as the preparedness (Seligman, 1970, 1971) and fear module (Öhman & Mineka, 2001) theories, adopt an evolutionary perspective according to which organisms are biologically predisposed by evolution to preferentially associate stimuli that provided threats to the species' survival with aversive events. In agreement with this view, learning biases have been found in response to stimuli from specific animal and social threat-relevant categories, such as snakes, angry faces, or outgroup faces, in that these stimuli are more readily and persistently associated with an aversive outcome than nonthreatening stimuli, such as birds, happy faces, or ingroup faces (e.g., Ho & Lipp, 2014; Öhman & Dimberg, 1978; Öhman, Eriksson, & Olofsson, 1975; Öhman, Fredrikson, Hugdahl, & Rimmö, 1976; Öhman & Mineka, 2001; Olsson, Ebert, Banaji, & Phelps, 2005; but see Åhs et al., 2018; Davey, 1995; Mallan, Lipp, & Cochrane, 2013).

An alternative framework to these accounts derives from appraisal theories of emotion (e.g., Sander, Grafman, & Zalla, 2003; Sander, Grandjean, & Scherer, 2005, 2018), and proposes that the occurrence of learning biases in Pavlovian aversive learning is driven by a mechanism of relevance detection that is not selective to threat (Stussi, Brosch, & Sander, 2015; Stussi, Ferrero, Pourtois, & Sander, 2019; Stussi, Pourtois, & Sander, 2018). This model holds that stimuli detected as relevant to the individual's concerns—such as their goals, needs, or values (Frijda, 1986; Pool, Brosch, Delplanque, & Sander, 2016)—benefit from enhanced Pavlovian conditioning beyond stimulus valence and evolutionary status per se, and that such preferential learning is critically dependent on individual differences in stimulus affective evaluation. Congruent with this hypothesis, initial evidence (Stussi et al., 2018) has shown that, similar to threat-relevant stimuli (angry faces or snakes), positive stimuli with high biological relevance to the organism (baby faces or erotic stimuli) can likewise induce learning biases during Pavlovian aversive conditioning.

Here, we sought to gain further insights into the mechanisms that modulate emotional learning in humans by comparing these two competing models through the investigation of Pavlovian aversive conditioning to threatening (angry faces), rewarding (happy faces), and neutral (neutral faces) social stimuli. On the one hand, extant evidence has documented the existence of learning biases to angry but not to happy faces in Pavlovian aversive conditioning (see, e.g., Bramwell, Mallan, & Lipp, 2014; Dimberg & Öhman, 1996; Esteves, Parra, Dimberg, & Öhman, 1994; Mazurski, Bond, Siddle, & Lovibond, 1996; Öhman & Dimberg, 1978; Öhman & Mineka, 2001; Rowles, Lipp, & Mallan, 2012), thereby mostly supporting the predictions of the preparedness and fear module theories.[1] On the other hand, the relevance detection model predicts that both angry and happy faces should be preferentially learned during Pavlovian conditioning relative to neutral faces because of their higher affective relevance, but that learning biases to happy faces should be smaller than to angry faces and more sensitive to interindividual differences in their affective evaluation. Indeed, happy faces have been suggested to generally have a lower level of relevance to the organism than stimuli with heightened biological relevance, such as angry or baby faces (Brosch, Pourtois, & Sander, 2010; Brosch, Sander, Pourtois, & Scherer, 2008; Pool et al., 2016). Whereas the latter stimuli are likely to be consistently detected as highly relevant across individuals because of their importance for the organism and species' survival, happy faces can carry several meanings (Ambadar, Cohn, & Reed, 2009; Martin, Rychlowska, Wood, & Niedenthal, 2017) and their processing may vary as a function of the situation and individual differences, such as extraversion for instance (Canli, Sivers, Whitfield, Gotlib, & Gabrieli, 2002). Nonetheless, prior research has mainly used small sample sizes (typical n by group ranged between 15 and 25), hence undermining the possibility to detect potentially small learning biases and explore whether learning biases to happy faces can be mapped onto interindividual differences.

In the present study, we implemented a differential Pavlovian aversive conditioning paradigm in a relatively large sample size ($N = 107$) to test the predictions of the relevance detection model. Two angry, happy, and neutral faces were used as conditioned stimuli (CSs). One stimulus (CS+) from each CS category was systematically associated with a mild electric stimulation, whereas the other stimulus (CS−) was never paired with the stimulation. We operationalized the conditioned response (CR) as the differential skin conductance response (SCR) to the CS+ minus CS− from the same CS category, which served as an index of learning (e.g., Olsson et al., 2005; Stussi et al., 2015, 2018, 2019). We also used computational modeling (Li, Schiller, Schoenbaum, Phelps, & Daw, 2011; Lindström, Golkar, & Olsson, 2015; Rescorla & Wagner, 1972; Stussi et al., 2018) to characterize the learning biases associated with angry and happy faces as opposed to neutral faces by extracting and comparing learning parameters for these CS categories. Additionally, we examined interindividual differences in affective evaluation of happy faces in two ways. First, we considered participants' extraversion (see Canli et al., 2002) based on the rationale that individuals high in extraversion should tend to appraise happy faces as more relevant to their concerns than individuals lower in this trait (Sander et al., 2003, 2005). Second, we assessed implicit associations between the face categories and importance (e.g., Critcher & Ferguson, 2016) through a Go/No-Go Association Task (GNAT; Nosek & Banaji, 2001). This task aimed at measuring the strength with which participants associated the face categories with the attribute of importance, thereby serving as a proxy of individuals' affective relevance evaluation of the faces. Specifically, we reasoned that the more individuals appraised the faces as affectively relevant, the more easily and rapidly they should associate these faces with importance (vs. unimportance).

As learning biases are generally reflected by a faster acquisition of a CR and/or an enhanced resistance to extinction of that CR (e.g., Öhman & Mineka, 2001), we predicted that (a) the CR to angry faces would be more readily acquired and more resistant to extinction than the CR to both happy faces and neutral faces across participants, whereas (b) the CR to happy faces would be acquired more readily and more resistant to extinction than the CR to

[1] Of note, Bramwell et al. (2014) reported resistance to extinction to outgroup race happy faces, thereby indicating that happy faces may lead to preferential aversive learning under certain circumstances. This effect was not driven by negative evaluation of outgroup happy faces, which were evaluated as more pleasant than ingroup happy faces at the explicit level, whereas no difference in positive or negative evaluation was found between them at the implicit level. Nevertheless, no resistance to extinction was observed to ingroup happy faces, which suggests that the enhanced persistence of threat conditioned to outgroup happy faces was likely driven by the faces' race category.

neutral faces. Moreover, we hypothesized that (c) participants' extraversion level, as well as the sensitivity and rapidity with which they associated happy faces with the attribute of importance versus unimportance, would predict the CR acquisition readiness and persistence to happy faces.

Method

Participants

There were 117 students from the University of Geneva who participated in the experiment, which was approved by the Faculty of Psychology and Educational Sciences ethics committee at the University of Geneva. They provided informed consent and received partial course credit for their participation. Ten participants were excluded from the analyses because of technical problems ($n = 2$), for displaying virtually no SCR ($n = 2$), for failing to acquire a CR to at least one of the CSs+ ($n = 5$), or for withdrawing from the study early ($n = 1$). These exclusion criteria were determined before data collection (see Olsson et al., 2005; Olsson & Phelps, 2004; Stussi et al., 2015, 2018, 2019). The final sample size consisted of 107 participants (85 women, 22 men), aged between 19 and 34 years old (mean age = 21.85 ± 2.57 years). Two participants were further excluded from the computational modeling analyses because their individual parameters could not be estimated because of a lack of SCR to all the angry face CSs during the experiment (see online supplemental materials). The sample size was established before data collection on the basis of the current heuristic suggesting a sample of at least 100 participants for studies considering interindividual differences (see, e.g., Dubois & Adolphs, 2016). For counterbalancing purposes, we aimed to recruit a minimum sample size of 104 participants exhibiting differential conditioning to at least one of three CS categories. We stopped collecting data at the end of the academic year and ascertained that the established sample size had been reached. A sensitivity power analysis performed with G*Power 3 (Faul, Erdfelder, Lang, & Buchner, 2007) indicated that this sample size allowed for detecting a smallest population effect size of $d_z = 0.242$ with a power of 80% using a one-tailed paired-sample t test.

Apparatus and Stimuli

The experiment took place in a sound-attenuated experimental chamber. The stimuli were presented using MATLAB (The MathWorks Inc., Natick, MA) with the Psychophysics Toolbox extensions (Brainard, 1997; Pelli, 1997) and displayed on a 23-in. LED monitor. Eight angry, eight happy, and eight neutral male face stimuli from the Karolinska Directed Emotional Faces (KDEF; Lundqvist, Flykt, & Öhman, 1998) were used either as targets or as distractors in the GNAT (see online supplemental materials). Four word stimuli related to the attribute of importance (i.e., important words; "important," "relevant," "significant," and "impactful") and four word stimuli related to the attribute of unimportance (i.e., unimportant words; "unimportant," "irrelevant," "insignificant," and "secondary") were also used both as targets and distractors.

In the differential Pavlovian aversive conditioning procedure, the CSs consisted of two male angry (model numbers AM10ANS, AM29ANS), two male happy (AM07HAS, AM22HAS), and two male neutral (AM11NES, AM31NES) faces taken from the KDEF (Lundqvist et al., 1998). These faces were selected based on the correct identification (hit rate range: 89.06–100%) and intensity ratings (mean intensity range: 5.73–7.63) of their respective emotional expression (Goeleven, De Raedt, Leyman, & Verschuere, 2008). Each face served both as a CS+ and as a CS−, counterbalanced across participants. Subjective ratings performed before the conditioning procedure (see online supplemental materials) on a visual analog scale from 0 (*very unpleasant*) to 100 (*very pleasant*) indicated that the angry faces were evaluated as unpleasant ($M = 15.29$, $SD = 15.76$), the happy faces as pleasant ($M = 68.28$, $SD = 20.39$), and the neutral faces as relatively neutral ($M = 43.47$, $SD = 13.07$). The unconditioned stimulus (US) was a mild electric stimulation (200-ms duration) delivered to the participants' right wrist through a unipolar pulse electric stimulator (STM200; BIOPAC Systems Inc., Goleta, CA). The CR was assessed through SCR measured with two Ag-AgCl electrodes (6-mm contact diameter) filled with 0.5% NaCl electrolyte gel. The electrodes were attached to the distal phalanges of the second and third digits of the participants' left hand. SCR was continuously recorded during the conditioning procedure with a sampling rate of 1000 Hz by means of a BIOPAC MP150 system (Santa Barbara, CA). The SCR data were analyzed offline with AcqKnowledge software (Version 4.4; BIOPAC Systems Inc., Goleta, CA).

Procedure

Between 2 to 8 months before their participation in the study, participants completed the French version of the NEO Five-Factor Inventory (NEO-FFI; Costa & McCrae, 1992; Rolland, Parker, & Stumpf, 1998). Upon arrival at the laboratory, they were informed about the general layout of the experiment, provided written informed consent, and performed the GNAT. Participants were next asked to evaluate the to-be-CSs according to various dimensions (see online supplemental materials) before undergoing the differential Pavlovian aversive conditioning procedure. Finally, they were asked again to provide subjective ratings of the CSs after conditioning (see online supplemental materials) and were debriefed.

Differential Pavlovian aversive conditioning. Before conditioning, the electrodes for measuring SCR and delivering the electric stimulation were attached to participants. A work-up procedure was then performed to individually calibrate the electric stimulation intensity ($M = 34.55$ V, $SD = 7.57$, range = 20–50 V) to a level reported as "uncomfortable, but not painful." The differential Pavlovian aversive conditioning procedure (see Figure 1a,b) comprised three contiguous phases. In the initial habituation phase, the six CSs were each presented twice without being reinforced. During the subsequent acquisition phase, each CS was presented seven times. This phase always started with a reinforced CS+ trial. Each CS+ was paired with the US with a partial reinforcement schedule, five of the seven CS+ presentations coterminating with the US delivery, whereas the CS− from each CS category was never associated with the US. The use of a partial reinforcement schedule aimed to potentiate the CR resistance to extinction, hence optimizing the examination of differences between the

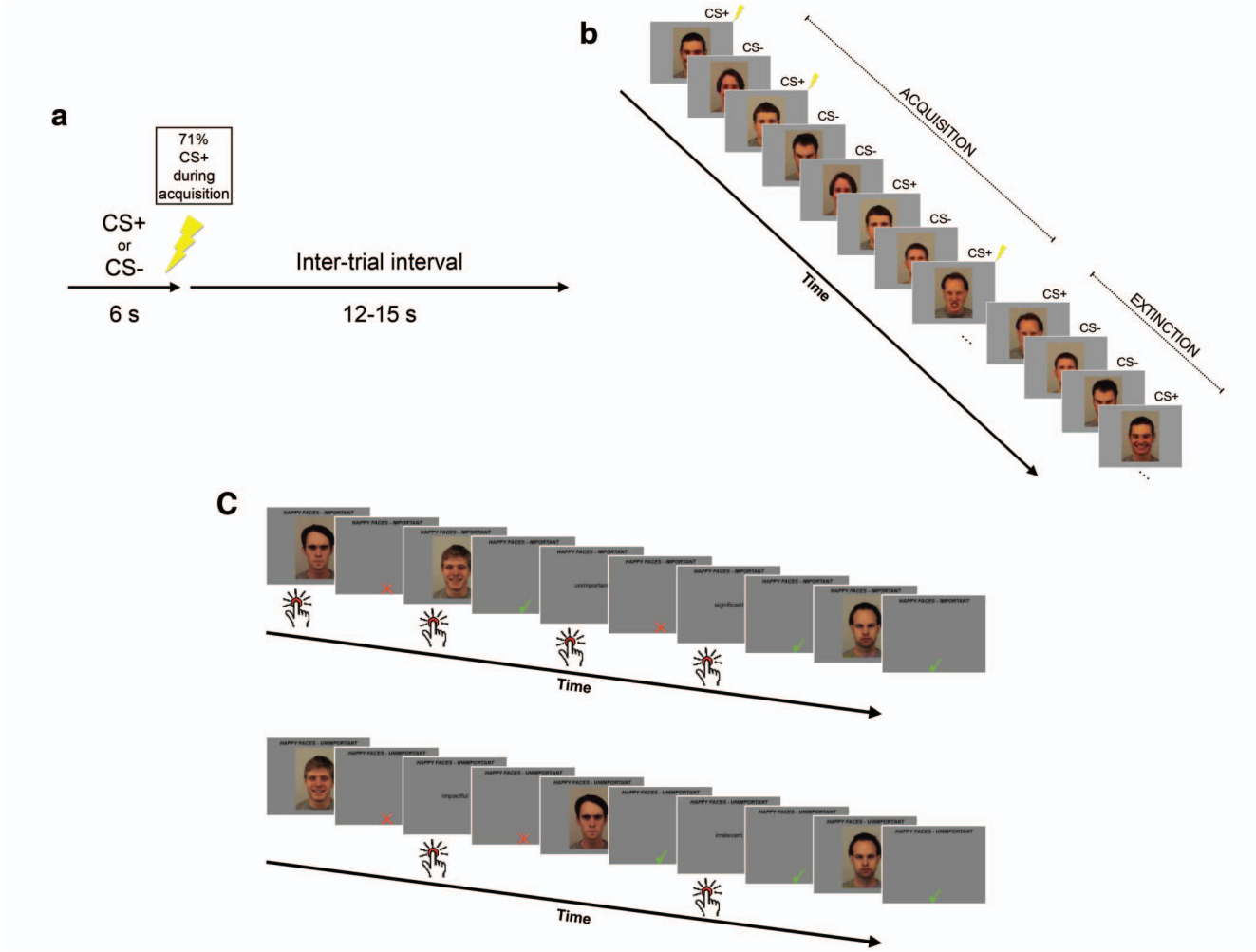

Figure 1. Schematic representation of the experimental procedures. (a) Within-trial structure during the differential Pavlovian aversive conditioning procedure: two angry, happy, and neutral faces were presented as conditioned stimuli (CSs) in a pseudorandom order for 6 s during three contiguous phases (habituation, acquisition, and extinction). Five of the seven CS+ trials (71%) for each face category coterminated with an electric stimulation during acquisition. Trials were separated by an intertrial interval ranging from 12 to 15 s. (b) Illustration of the overall differential Pavlovian aversive conditioning structure during acquisition and extinction. Acquisition consisted of presentations of the six CSs on a partial reinforcement schedule, whereas extinction consisted of presentations of the same CSs while the electric stimulation was no longer delivered. (c) Illustration of the Go/No-Go Association Task: examples of five trials in which participants had to detect whether the faces and the words belonged to the target categories "Happy faces" or "Important words" (upper panel), or to the target categories "Happy faces" or "Unimportant words" (lower panel). If the face or word belonged to one of the two target categories, the correct response was to press 'A' on the keyboard, but to withdraw from responding otherwise. After each response, participants received feedback consisting of either a green check or a red cross for correct and incorrect responses, respectively. The different faces shown (AM02NES, AM07HAS, AM10ANS, AM11NES, AM22HAS, AM23HAS, AM24ANS, AM29ANS, AM31NES) were taken from *The Karolinska Directed Emotional Faces—KDEF*, by D. Lundqvist, A. Flykt, & A. Öhman, 1998, Stockholm, Sweden: Karolinska Institutet, Department of Clinical Neuroscience, Psychology Section. Copyright 1998 by Karolinska Directed Emotional Faces database, which allows their free use for scientific publication (see kdef.se). Reprinted with permission. See the online article for the color version of this figure.

three CS categories used. The final extinction phase consisted of six unreinforced presentations of each CS. During all the conditioning phases, the CSs were presented for 6 s with an intertrial interval varying from 12 to 15 s. The CSs' presentation order was pseudorandomized into eight different orders to counterbalance the associations between the face stimuli and CS type (CS+ vs. CS−) across the three CS categories (angry vs. happy vs. neutral).

NEO Five-Factor Inventory (NEO-FFI). The NEO-FFI is a standard personality inventory measuring the Big Five personality traits consisting of neuroticism, extraversion, openness to experience, agreeableness, and conscientiousness (Costa & McCrae, 1992). It comprises 60 items (12 per trait), each of which is measured on a 5-point Likert scale ranging from 0 (*strongly disagree*) to 4 (*strongly agree*). Given our a priori hypotheses, we focused here on extraversion (M = 28.23, SD = 5.69, range = 10–40, Cronbach's α = .76; see Figure S1 in the online supplemental materials). Exploratory analyses including the other personality traits are reported in the online supplemental materials.

Go/No-Go Association Task. In the GNAT, participants were presented with faces from three emotional categories (angry vs. happy vs. neutral) and words from two categories (important vs. unimportant). In each trial, a face or a word was displayed at the center of the screen. Participants were instructed to press as quickly and accurately as possible on the "A" key if the stimulus was a member of a target category (Go trials), but to withdraw from responding otherwise (No-Go trials). Throughout the task, the labels of the target categories were continuously displayed at the top of the screen as a reminder. After each trial, feedback about participants' response was displayed at the bottom of the screen (i.e., a green check for correct or a red cross for incorrect) during a 150-ms intertrial interval (see Figure 1c).

The GNAT began with a practice session of five blocks in which there was only a single target category (see online supplemental materials). The experimental session ensued and was composed of three parts, each divided into two blocks. Within each part, a specific face category was one of the two target categories with important words being the other target category in Block 1, and unimportant words the other target category in Block 2. The order of the three parts as a function of the face categories was counterbalanced between participants. Each block consisted of 96 trials: 16 training trials and 80 critical trials. Four faces from the target face category and two faces from each distractor face category were presented intermixed with the four important and the four unimportant words in a pseudorandom order. The response deadline was idiosyncratically adapted to the participants' reaction times (RTs) and response accuracy (see, e.g., Coppin et al., 2016; Nosek & Banaji, 2001): When response was correct (for both Go and No-Go trials) and RT faster than the arbitrary response deadline (for Go trials), the response deadline for the next trial was set as 500 ms or as 666 ms if RT was slower than 500 ms but faster than 666 ms (for Go trials); otherwise, it was set as 800 ms.

Participants' RTs and response accuracy were recorded for each trial. All trials with RTs faster than 100 ms were excluded from analysis. Data for all errors and distracter items were removed from the RTs analysis. According to signal detection theory, we calculated a d' score for each block within each part of the GNAT experimental session, considering only critical trials (Nosek & Banaji, 2001). We converted the proportions of hits (correct Go-responses to targets) and false alarms (incorrect Go-responses to distractors) to z scores before computing the difference between them, thereby obtaining d'. Hit and false-alarm rates equal to 0 or 1 were replaced with 1/(2N) and 1 – 1/(2N), respectively, where N is the number of trials (Macmillan & Creelman, 2005). A differential d' index was then calculated by subtracting the d' scores of the second block (Target Face Category + Unimportant Words) from those of the first block (Target Face Category + Important Words; see, e.g., Coppin et al., 2016). Higher values on this index indicated higher accuracy when faces from the target face category and important words were targets in comparison with when faces from the target face category and unimportant words were targets. Additionally, we computed a differential index for RTs by subtracting the mean RTs of the first block to those of the second block, higher values reflecting faster responses when faces from the target face category and important words were targets relative to when faces from the target face category and unimportant words were targets. The differential d' and RTs indices served as indicators of the strength of association between the faces categories and the attribute of importance versus that of unimportance (Nosek & Banaji, 2001). Although d' scores are usually used as the main dependent variable in the GNAT, we measured both indicators because RTs have been suggested to be more reliable than d' scores because of their measurement on a continuous (vs. dichotomic) scale at the trial level (Nosek & Banaji, 2001).

Response Definition

SCR was scored for each trial as the peak-to-peak amplitude difference in skin conductance of the largest response starting in the 0.5–4.5 s temporal window after CS onset. The minimal response criterion was 0.02 μS, and responses below this criterion were scored as zero and remained in the analysis. A low-pass filter (Blackman −92 dB, 1 Hz) was applied on the SCR data before analysis. SCRs were detected automatically with AcqKnowledge software and manually checked for artifacts and response detection. Trials containing artifacts affecting the scoring of event-related SCRs (0.17%) were removed from the subsequent analyses. The raw SCRs were scaled according to each participant's mean unconditioned response (UR), and square-root-transformed to normalize the distributions. The UR was scored as the peak-to-peak amplitude difference in skin conductance of the largest response starting in the 0.5–4.5 s temporal window after the US delivery, and the mean UR was calculated across all USs for each participant. The habituation means comprised the first two presentations of each CS (i.e., Trials 1 and 2). To tease apart effects of faster conditioning from those of larger conditioning, the acquisition means were split into an early (i.e., the first three presentations of each CS following the first pairing between the CS+ of a given CS category and the US; Trials 4 to 6) and a late (i.e., the following three presentations of each CS; Trials 7 to 9) phase (see, e.g., Lonsdorf et al., 2017; Olsson, Carmona, Downey, Bolger, & Ochsner, 2013; Stussi et al., 2015, 2018, 2019). This allowed us to specifically examine the CR acquisition readiness during early acquisition. The first acquisition trial for each CS was removed from the CR analysis because the CSs+ became predictive of the US only after their first association therewith. The extinction means encompassed the last six presentations of each CS (i.e., Trials 10 to 15). The conditioning data analyses were performed on the CR, which was calculated as the SCR to the CS+ minus the SCR to the CS- from the same CS category (e.g., Olsson et al., 2005; Stussi et al., 2015, 2018, 2019). This procedure allows for reducing preexisting differences in emotional salience between the different CS categories (Olsson et al., 2005).

Computational Modeling

Based on previous research (Stussi et al., 2018), we constructed a simple reinforcement learning model to characterize Pavlovian aversive conditioning to angry, happy, and neutral faces (for further details, see online supplemental materials). We adapted the standard version of the Rescorla-Wagner model (Rescorla & Wagner, 1972) by implementing distinct learning rates for positive (i.e., when the outcome is not predicted or more than expected; excitatory learning) and negative (i.e., when the outcome is omitted or less than expected; inhibitory learning) prediction errors instead of a single learning rate (see Niv, Edlund, Dayan, & O'Doherty, 2012; Stussi et al., 2018). Excitatory and inhibitory learning rates exert an influence on associative learning by altering the impact of positive and negative prediction error signals, respectively, on the CS predictive value (see Niv & Schoenbaum, 2008). In the dual-learning-rate Rescorla-Wagner model, the predictive value (or associative strength) V of a given CS j is updated based on the sum of the current predictive value V_j at trial t, and the prediction error between the predictive value V_j and the outcome R at trial t, weighted by different learning rates for positive and negative prediction errors as follows:

$$V_j(t+1) = \begin{cases} V_j(t) + \alpha^+ \cdot (R(t) - V_j(t)) \; if \; R(t) - V_j(t) > 0 \\ V_j(t) + \alpha^- \cdot (R(t) - V_j(t)) \; if \; R(t) - V_j(t) < 0 \end{cases}$$

where the learning rate for positive prediction errors α^+ and the learning rate for negative prediction errors α^- are free parameters within the range [0, 1]. If the US was delivered on the current trial t, $R(t) = 1$, else $R(t) = 0$. This model allows for parsimoniously accounting for how specific stimulus categories can accelerate acquisition (through the excitatory learning rate) and enhance resistance to extinction (through the inhibitory learning rate) of the CR.

The learning-rate parameters were estimated, and the trial-by-trial CS values calculated, by fitting the model to the individual normalized (i.e., scaled and square-root-transformed) SCR data separately for each CS category. Model comparison indicated that the dual-learning-rate Rescorla-Wagner model provided the best fit to the SCR data relative to alternative models (see online supplemental materials). Accordingly, we compared the estimated excitatory and inhibitory learning-rate parameters across the three different CS categories used (angry vs. happy vs. neutral).

Statistical Analyses

The differential d' and the differential RT indices derived from the GNAT were each analyzed with a one-way repeated-measures analysis of variance (ANOVA) with face category (angry vs. happy vs. neutral) as a within-participant factor. Statistically significant main effects were followed up with a multiple comparison procedure using Tukey's HSD tests when applicable.

Following standard practice in the human conditioning literature (e.g., Lonsdorf et al., 2017; Olsson et al., 2005; Stussi et al., 2015, 2018, 2019), the SCR data was analyzed separately for each conditioning phase. The habituation and extinction phases and the estimated learning rates were each analyzed with a one-way repeated-measures ANOVA with CS category (angry vs. happy vs. neutral) as a within-participant factor. The acquisition phase was analyzed with a two-way repeated-measures ANOVA with CS category (angry vs. happy vs. neutral) and time (early vs. late) as within-participant factors. One-sample t tests were additionally performed to test whether differential conditioning occurred for the CS categories across the entire acquisition phase. To specifically test our a priori hypotheses, we conducted planned contrast analyses comparing the CR during early acquisition and during extinction, as well as the estimated learning rates, to (a) angry versus neutral faces, (b) happy versus neutral faces, and (c) angry versus happy faces. As these contrasts were nonorthogonal, we applied a Holm-Bonferroni sequential procedure (Holm, 1979) to correct for multiple comparisons. The alpha level of the contrast with the lowest p value was set as $\alpha = .05/3 = .0167$, the alpha level with the second lowest p value as $\alpha = .05/2 = .025$, and the alpha level with the highest p value as $\alpha = .05$. For each planned contrast, we also calculated the Bayes factor (BF_{10}) quantifying the likelihood of the data under the alternative hypothesis compared with the likelihood of the data under the null hypothesis (e.g., Dienes, 2011; Rouder, Speckman, Sun, Morey, & Iverson, 2009). Because we expected moderate effects for angry faces and relatively small effects for happy faces, we used a noninformative Cauchy prior distribution with a width of 0.5 for the comparisons between angry and happy faces and between angry and neutral faces (see Stussi et al., 2018), and of 0.25 for the comparison between happy and neutral faces. When our theory-driven hypotheses clearly predicted the direction of the expected effects, we performed one-sided testing to test them (one-sample t tests, contrasts a, b, and c).

To assess our a priori hypotheses that extraversion, as well as the sensitivity and the rapidity with which happy faces were associated with the attribute of importance predicted the CR acquisition readiness and persistence to these faces, we conducted multiple linear regression analyses. These analyses tested whether the CR acquisition readiness (i.e., during early acquisition) and persistence (i.e., during extinction), along with the excitatory and inhibitory learning-rate estimates, to happy faces were predicted by participants' (a) extraversion level, (b) differential d' index for happy faces, and (c) differential RT index for happy faces. Further exploratory multiple linear regression analyses carried out on the CR and the learning rates to angry and neutral faces to investigate the specificity of these predictive effects are reported in the online supplemental materials.

All statistical analyses were performed with RStudio (RStudio Team, 2016). Huynh-Feldt adjustments of degrees of freedom were applied for repeated-measures ANOVAs when appropriate. Partial eta squared (η^2) or Hedges' g_{av} (or g_z) and their 90 or 95% confidence interval (CI) were used as estimates of effect sizes (see Lakens, 2013) for the repeated-measures ANOVAs and the planned contrasts analyses (or one-sample t tests), respectively, whereas the coefficient of determination R^2 along with its 90% CI was used for multiple linear regressions.

Results

Pavlovian Aversive Conditioning

Figure 2 depicts the mean SCR to angry, happy, and neutral faces across the habituation, acquisition, and extinction phases of the differential Pavlovian aversive conditioning separately for the CS+ and the CS−. In the habituation phase, no preexisting dif-

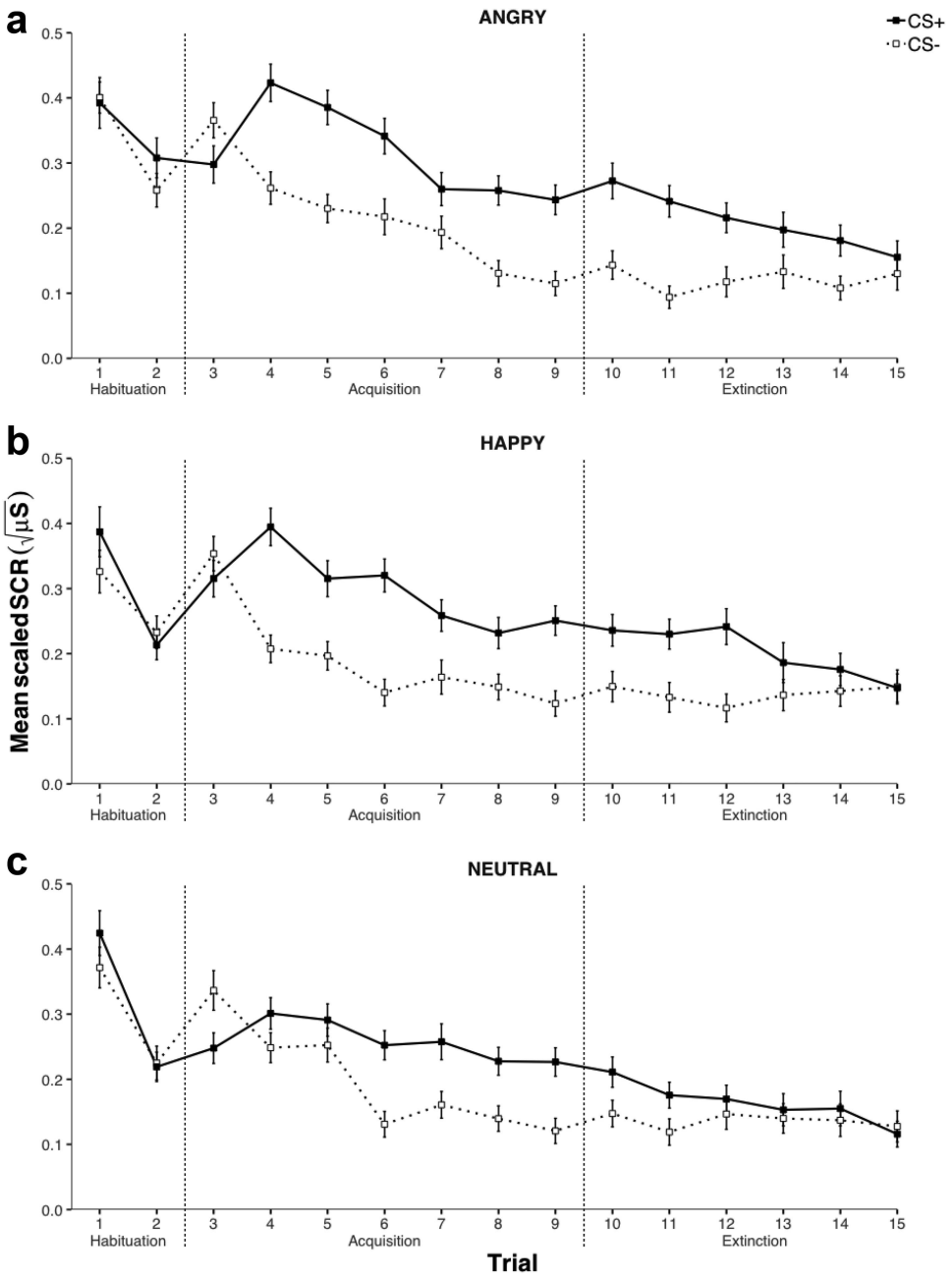

Figure 2. Mean scaled skin conductance response (SCR) to the conditioned stimuli as a function of the conditioned stimulus type (CS+ vs. CS−) across trials. Mean scaled SCR to (a) angry faces, (b) happy faces, and (c) neutral faces. Error bars indicate ±1 *SEM* adjusted for within-participant designs (Morey, 2008).

ference in differential SCR across the CS categories (angry vs. happy vs. neutral) was found, $F(2, 212) = 0.003$, $p = .997$, partial $\eta^2 = .00003$, 90% CI [.000, .0006].

Analysis of the acquisition phase revealed successful differential conditioning to all three CS categories, as reflected by larger SCRs to the CS+ than to the CS− for angry, $t(106) = 7.44$, $p < .001$ (one-tailed), $g_z = 0.714$, 95% CI [0.505, 0.931], happy, $t(106) = 8.10$, $p < .001$ (one-tailed), $g_z = 0.777$, 95% CI [0.564, 0.998], and neutral faces, $t(106) = 5.97$, $p < .001$ (one-tailed), $g_z = 0.573$, 95% CI [0.372, 0.781]. The CS categories, however, differentially influenced the CR acquisition as indicated by a main effect of CS category, $F(2, 212) = 3.27$, $p = .040$, partial $\eta^2 = .030$, 90% CI [.001, .071]. The interaction effect between CS category and time did not yield statistical significance, $F(2, 212) = 2.60$, $p = .076$, partial $\eta^2 = .024$, 90% CI [.000, .062]. Congruent with our a priori hypothesis, a planned contrast analysis showed that the CR to angry faces was more readily acquired than the CR to neutral faces during early acquisition, $t(106) = 2.60$, $p = .005$

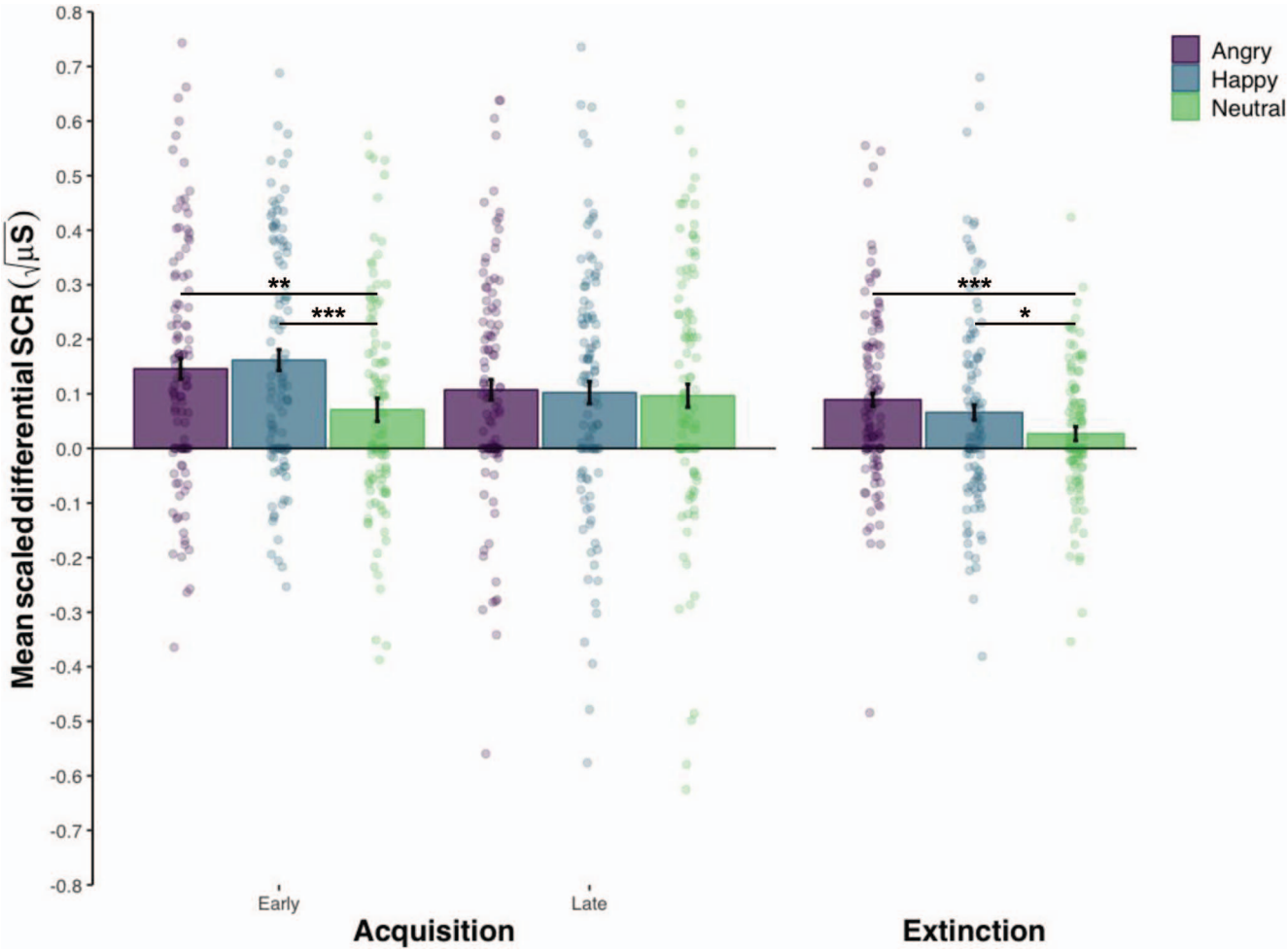

Figure 3. Mean conditioned response (scaled differential skin conductance response [SCR]) as a function of the conditioned stimulus category (angry vs. happy vs. neutral) during (early and late) acquisition and extinction. The dots indicated data for individual participants. Error bars indicated ±1 *SEM* adjusted for within-participant designs (Morey, 2008). Asterisks indicate statistically significant differences between conditions ($^{***}p < .001$, $^{**}p < .01$, $^{*}p < .05$, one-tailed, Holm-Bonferroni corrected). See the online article for the color version of this figure.

(one-tailed), $g_{av} = 0.358$, 95% CI [0.084, 0.636], $BF_{10} = 6.642$ (see Figure 3). The CR to happy faces was likewise more readily acquired than to neutral faces, $t(106) = 3.25$, $p < .001$ (one-tailed), $g_{av} = 0.442$, 95% CI [0.169, 0.720], $BF_{10} = 41.237$, whereas there was no statistical difference in CR acquisition readiness to angry faces compared with happy faces, $t(106) = -0.58$, $p = .717$ (one-tailed), $g_{av} = -0.073$, 95% CI [−0.324, 0.177], $BF_{10} = 0.101$ (see Figure 3). No statistical differences emerged between the three CS categories during late acquisition (all ps $> .92$, $0.02 < g_{av}$s < 0.05, all BFs$_{10} < 0.32$).

Critically, the CR persistence was also modulated by the CS categories during extinction, $F(2, 212) = 5.97$, $p = .003$, partial $\eta^2 = .053$, 90% CI [.011, .104]. As predicted, the CR to angry faces was more resistant to extinction than the CR to neutral faces, $t(106) = 3.69$, $p < .001$ (one-tailed), $g_{av} = 0.432$, 95% CI [0.196, 0.672], $BF_{10} = 133.200$. Similarly, the CR to happy faces was more persistent than to neutral faces, $t(106) = 2.01$, $p = .024$ (one-tailed), $g_{av} = 0.247$, 95% CI [0.003, 0.493], $BF_{10} = 2.777$ (see Figure 3). By comparison, we did not observe an enhanced CR persistence to angry faces relative to happy faces, $t(106) = 1.28$, $p = .102$ (one-tailed), $g_{av} = 0.133$, 95% CI [−0.072, 0.339], $BF_{10} = 0.573$.

Estimated Learning Rates

Analysis of the excitatory learning-rate estimates revealed no statistically significant main effect of CS category, $F(2, 208) = 2.50$, $p = .085$, partial $\eta^2 = .023$, 90% CI [.000, .061]. A more focused planned contrast analysis indicated that happy faces were associated with a higher excitatory learning rate than neutral faces, $t(104) = 2.05$, $p = .022$ (one-tailed), $g_{av} = 0.232$, 95% CI [0.007, 0.460], $BF_{10} = 2.986$ (see Figure 4a), but this difference was not statistically significant when correcting the alpha level for this contrast ($\alpha = .0167$). No statistical differ-

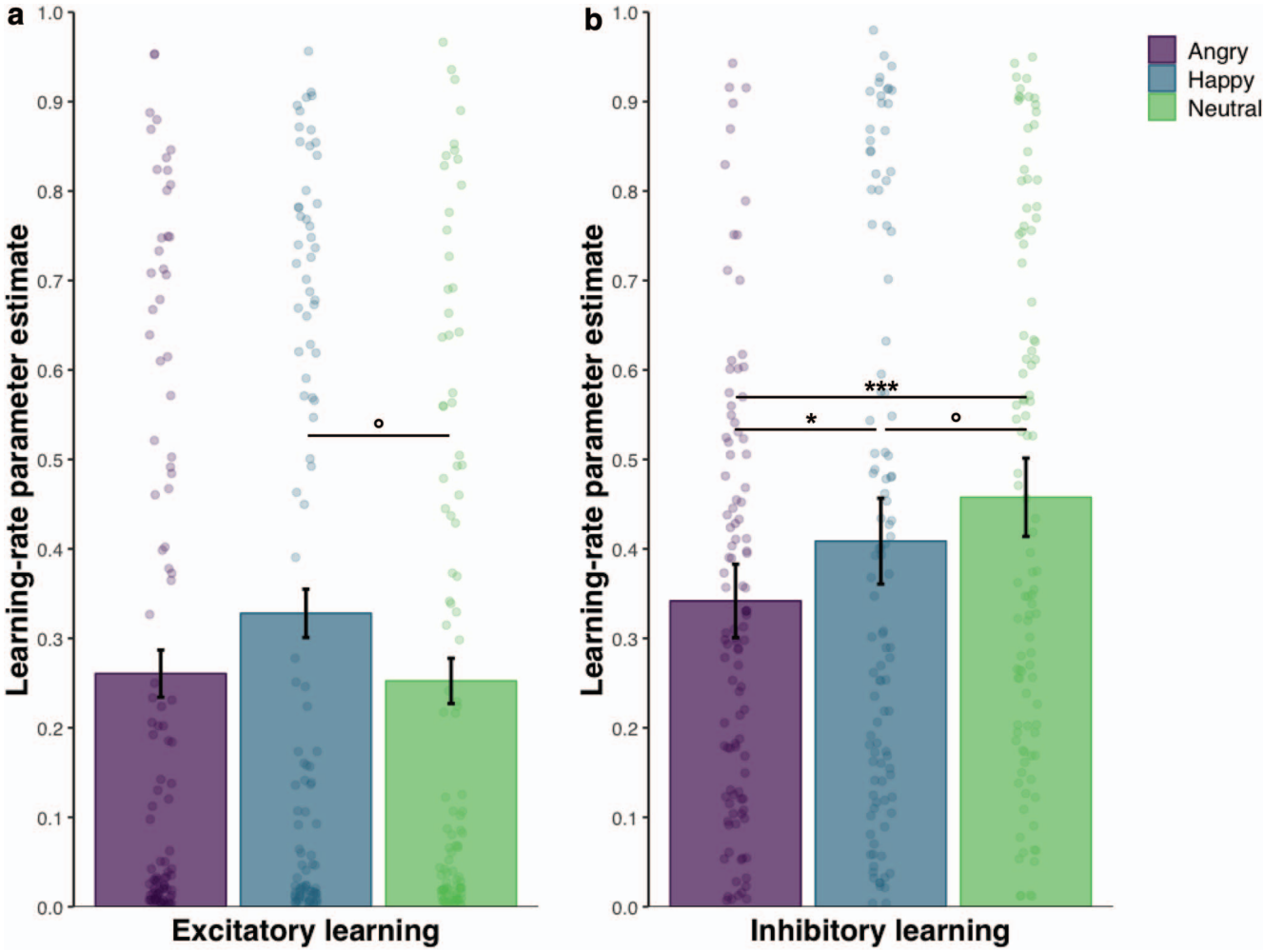

Figure 4. Learning-rate parameter estimates of the Rescorla-Wagner model implementing dual learning rates using the best-fitting parameters for positive predictions errors (excitatory learning) and negative prediction errors (inhibitory learning) as a function of the conditioned stimulus category (angry vs. happy vs. neutral). The dots indicate data for individual participants. Error bars indicate ±1 *SEM* adjusted for within-participant designs (Morey, 2008). Asterisks indicate statistically significant differences between conditions (*** $p < .001$, * $p < .05$, ° $p < .10$, one-tailed, Holm-Bonferroni corrected). See the online article for the color version of this figure.

ence in excitatory learning rate was observed between angry and happy faces, $t(104) = -1.76$, $p = .959$ (one-tailed), $g_{av} = -0.205$, 95% CI [−0.438, 0.026], $BF_{10} = 0.058$, or between angry and neutral faces, $t(104) = 0.23$, $p = .410$ (one-tailed), $g_{av} = 0.027$, 95% CI [−0.203, 0.257], $BF_{10} = 0.181$. By contrast, the CS categories differentially affected the estimated inhibitory learning rates, $F(2, 208) = 5.95$, $p = .003$, partial $\eta^2 = .054$, 90% CI [.011, .106]. These estimates were lower for angry faces than for neutral faces, $t(104) = -3.52$, $p < .001$ (one-tailed), $g_{av} = -0.434$, 95% CI [−0.686, −0.186], $BF_{10} = 78.801$, and happy faces, $t(104) = -2.14$, $p = .017$ (one-tailed), $g_{av} = -0.242$, 95% CI [−0.468, −0.018], $BF_{10} = 2.477$, whereas they were marginally lower for happy faces compared with neutral faces, $t(104) = -1.33$, $p = .093$ (one-tailed), $g_{av} = -0.164$, 95% CI [−0.409, 0.079], $BF_{10} = 1.015$ (see Figure 4b), although the latter difference did not yield statistical significance and the evidence for it remained inconclusive.

Go/No-Go Association Task

The analysis of the differential d' index showed a statistically significant main effect of face category (angry vs. happy vs. neutral), $F(2, 212) = 15.46$, $p < .001$, partial $\eta^2 = .127$, 90% CI [.061, .193]. The differential d' index was higher for happy faces ($M = 0.15$, $SD = 0.55$) than for angry ($M = -0.20$, $SD = 0.46$; $p < .001$, $g_{av} = 0.683$, 95% CI [0.407, 0.965]) and neutral faces ($M = -0.10$, $SD = 0.44$; $p < .001$, $g_{av} = 0.493$, 95% CI [0.222, 0.769]), whereas there was no statistical difference between angry and neutral faces ($p = .273$, $g_{av} = 0.219$, 95% CI [−0.030, 0.469]). These results suggest that participants exhibited a greater sensitivity to the association between the attribute of importance versus unimportance with happy faces than either angry or neutral faces. Conversely, the differential RT index did not differ statistically across the face categories, $F(2, 212) = 2.45$, $p = .089$, partial $\eta^2 = .023$, 90% CI [.000, .059].

Table 1
Results for the Multiple Linear Regression Analyses

	Conditioned response to happy faces during early acquisition (N = 107)					Conditioned response to happy faces during extinction (N = 107)					Estimated excitatory learning rate to happy faces (N = 105)					Estimated inhibitory learning rate to happy faces (N = 105)				
Predictor	b	SE	β	t(103)	p	b	SE	β	t(103)	p	b	SE	β	t(101)	p	b	SE	β	t(101)	p
Intercept	0.073	0.106		0.69	.494	0.027	0.087		0.31	.759	0.069	0.169		0.41	.685	0.446	0.150		2.97	.004**
Extraversion	0.003	0.004	.085	0.87	.388	0.002	0.003	.046	0.50	.621	0.009	0.006	.146	1.49	.140	−0.002	0.005	−.031	−0.32	.750
Differential d' index	−0.005	0.039	−.013	−0.13	.896	−0.028	0.032	−.082	−0.87	.386	0.083	0.062	.133	1.33	.187	0.076	0.055	.137	1.37	.173
Differential reaction time index	−0.001	0.001	−.096	−0.96	.341	0.002	0.0005	.360***	3.83	<.001	−0.000	0.001	−.019	−0.19	.852	−0.002	0.001	−.195	−1.95	.054
R^2			.015					.131					.041					.047		

** $p < .01$. *** $p < .001$.

Regression Analyses

The multiple linear regression analyses on the CR to happy faces (see Table 1) showed that participants' extraversion level, differential d' index for happy faces, and differential RT index for happy faces did not predict the CR to happy faces during early acquisition (all ps $> .34$) where they only explained 1.51% of its variance ($R^2 = .015$, 90% CI [.000, .048], adjusted $R^2 = -.014$, $F(3, 103) = 0.53$, $p = .664$). However, these three predictors explained 13.06% of the variance of the CR to happy faces during extinction ($R^2 = .131$, 90% CI [.031, .224], adjusted $R^2 = .105$, $F(3, 103) = 5.16$, $p = .002$). Whereas extraversion and the differential d' index for happy faces did not predict the CR to happy faces (both ps $> .38$), the CR to happy faces was predicted by the differential RT index for these faces, $b = 0.002$, 95% CI [0.001, 0.003], β = .360, $t(103) = 3.83$, $p < .001$, reflecting that participants who were faster to associate happy faces with the attribute of importance than that of unimportance exhibited a larger CR to happy faces during extinction (see Figure 5). Regarding the excitatory and inhibitory learning rates (see Table 1), participants' extraversion level, differential d' index for happy faces, and differential RT index for happy faces explained 4.09% ($R^2 = .041$, 90% CI [.000, .100], adjusted $R^2 = .012$, $F(3, 101) = 1.44$, $p = .236$) and 4.71% ($R^2 = .047$, 90% CI [.000, .110], adjusted $R^2 = .019$, $F(3, 101) = 1.66$, $p = .180$) of their variance, respectively. No significant relationship emerged between the predictors and the excitatory and inhibitory learning-rate estimates (all ps $> .05$). For angry and neutral faces, no statistically significant relationship was observed between participants' extraversion level, differential d' index, and differential RT index, and the CR during early acquisition and extinction as well as the learning-rate estimates (all ps $> .08$; see online supplemental materials).

Discussion

In this study, we aimed to test the predictions of two competing theoretical approaches of emotional learning. More particularly, we tested the hypothesis deriving from appraisal theories that enhanced emotional learning is driven by a relevance detection mechanism that is not specific to threat, and depends on individual differences in affective relevance appraisal. This hypothesis departs from the preparedness and fear module theories, according to which enhanced emotional learning is selective to threat. To that end, we compared Pavlovian aversive conditioning to threat-related (angry faces), positive (happy faces), and neutral (neutral faces) social stimuli and investigated the influence of interindividual differences in affective evaluation on this process. Altogether, our results showed that both angry and happy faces were preferentially associated with an aversive outcome during Pavlovian conditioning relative to neutral faces, and that the persistence of this association for happy faces was related to interindividual differences in their affective evaluation.

The conditioned response to angry and happy faces was more readily acquired and more persistent than the conditioned response to neutral faces; thus, reflecting learning biases associated with these stimuli. Moreover, the conditioned response to happy faces during extinction was greater in participants who were faster to associate them with the attribute of importance (vs. unimportance) in the GNAT. In comparison, no such relationship was found for angry and neutral faces (see online supplemental materials). Whereas the results obtained for angry faces align with well-established findings in the human conditioning literature (e.g., Öhman & Dimberg, 1978; Rowles et al., 2012; see also Dimberg & Öhman, 1996; Mallan et al., 2013; Öhman & Mineka, 2001), the occurrence of learning biases to happy faces challenges the view that enhanced Pavlovian aversive conditioning is selective to threat-relevant stimuli (Öhman & Mineka, 2001; Seligman, 1971). Conversely, our results indicate that positive stimuli with moderate affective relevance can also be rapidly and persistently associated with an aversive event, with these effects being moderate to small. They further show that individual differences in affective evaluation may affect the emergence of learning biases. In this respect, our findings replicate and expand recent evidence supporting the appraisal-based predictions according to which preferential Pavlovian aversive learning is driven by affective relevance without being bound to a specific valence or inherent threat value, and can be modulated by individual differences in the way the stimulus is appraised in relation to the individual's concerns (Stussi et al., 2018, 2019).

At the computational level, the effects of greater persistence of the conditioned response to angry faces was characterized by a lower inhibitory learning rate. More specifically, the learning rate for negative prediction errors was lower to angry faces than to happy and neutral faces. This lower inhibitory learning altered the impact of negative prediction error signals, which likely contributed to weakening inhibitory learning underlying

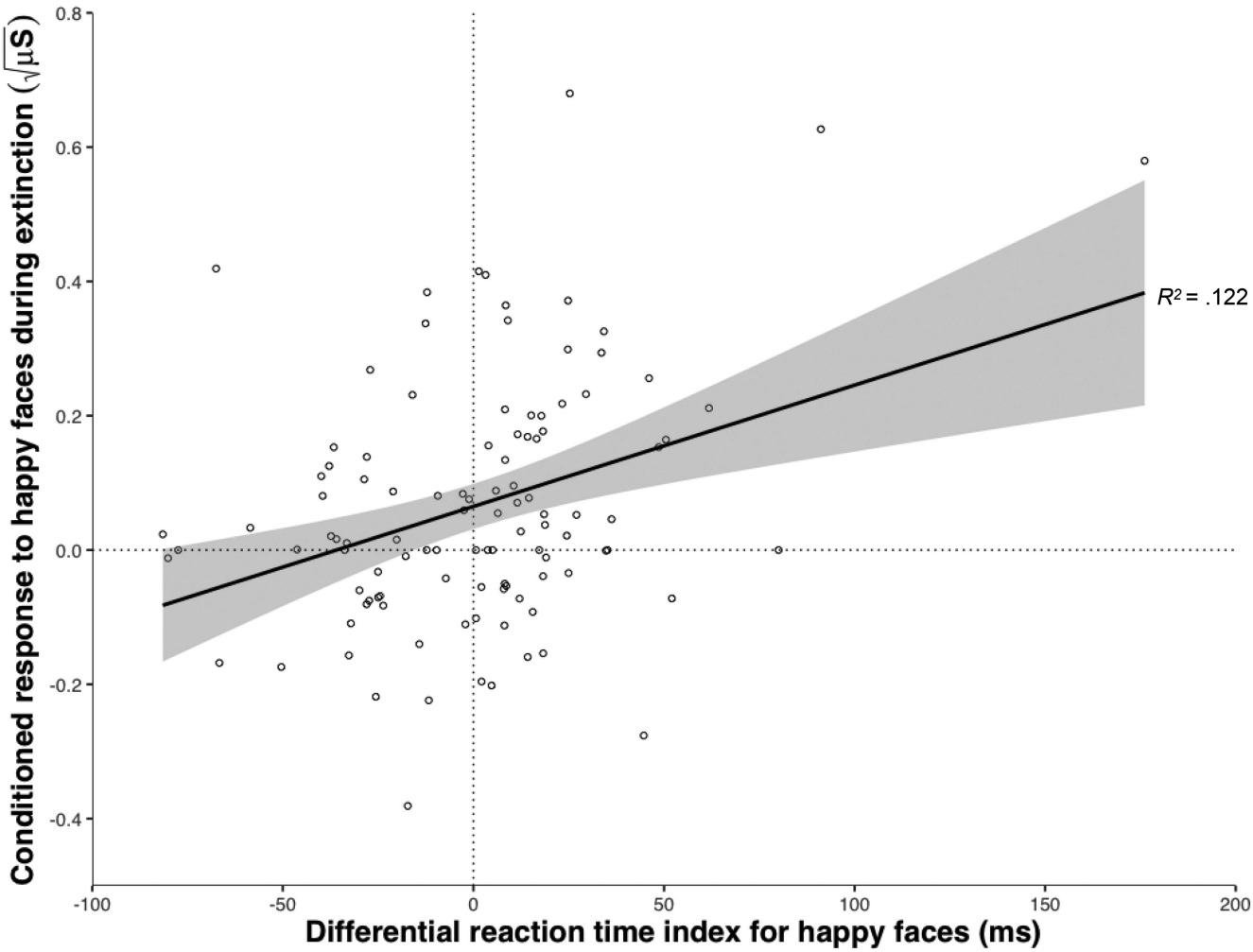

Figure 5. Relationship between the differential reaction time (RT) index for happy faces in the Go/No-Go Association Task (mean RTs in the block where happy faces and the attribute of importance were target categories minus mean RTs in the block where happy faces and the attribute of unimportance were target categories) and the conditioned response to happy faces during extinction. The line represents the fitted regression line using least squares estimation and 95% confidence interval.

extinction (Dunsmoor, Niv, Daw, & Phelps, 2015). The observation that angry faces were associated with a lower inhibitory learning rate than happy faces additionally suggests that angry faces led to more persistent Pavlovian aversive conditioning, even though this difference was not visible when using conventional summary statistics on the conditioned response during extinction. This finding dovetails with the notion that happy faces hold a generally lower level of relevance to the organism than angry faces (Brosch et al., 2008, 2010; Pool et al., 2016), hence entailing smaller learning biases than angry faces. Happy faces were associated with a marginally lower inhibitory learning rate relative to neutral faces, but only inconclusive evidence was observed for this difference. Further evidence is required to determine whether the heightened conditioned response persistence to happy compared with neutral faces could be underlain by a lower inhibitory learning rate. In comparison, we did not find strong evidence that faster acquisition of the conditioned response to angry and happy faces than to neutral faces was driven by a higher excitatory learning rate. These results are partially inconsistent with previous studies using reward learning paradigms (Watanabe & Haruno, 2015; Watanabe, Sakagami, & Haruno, 2013), which reported that threat-related (i.e., fearful) faces not only accelerated learning in comparison with neutral faces, but also increased the associated excitatory learning rate. Tentatively, this discrepancy may be because of habituation effects in the skin conductance response in the present case, which could have biased the estimation of the excitatory learning rates and mitigated the emergence of robust differences between the face categories.

The fact that happy faces led to a relatively small learning bias during extinction could potentially account for failures to report a resistance-to-extinction effect for this specific emotional category in prior research (see, e.g., Bramwell et al., 2014; Esteves, et al., 1994; Mazurski et al., 1996; Öhman & Dimberg, 1978; Rowles et

al., 2012; see also Dimberg & Öhman, 1996; Öhman & Mineka, 2001). Indeed, past studies have generally used between-participants designs (but see Bramwell et al., 2014) that are less sensitive than within-participant designs (see, e.g., Ho & Lipp, 2014) and, importantly, often with modest sample sizes, typically varying from 15 to 25 participants by group. These two methodological factors likely contributed to hindering the possibility to reveal the existence of learning biases to happy faces given that, as our results suggest here with the use of a larger sample and stringent within-participant design, this bias has a small effect size.[2] It is, therefore, highly desirable in future research to set up adequately powered experiments when the goal is to explore differences in Pavlovian aversive learning to happy compared with neutral or angry faces.

Although our study shows that interindividual differences in stimulus affective evaluation can exert an effect on learning biases in Pavlovian conditioning, we only found a clear relationship between the conditioned response to happy faces during extinction and the differential RT index, but not with the differential d' index—this dissociation likely stemming from the putative lower reliability of this latter index (Nosek & Banaji, 2001)—or during early acquisition. In addition, we found no evidence that interindividual differences in extraversion affected the conditioned response to happy faces during either early acquisition or extinction, which is at odds with our predictions. Speculatively, this null result might arise from a relative lack of heterogeneity in the current sample's extraversion scores (see online supplemental materials Figure S1; see Rolland et al., 1998, for a comparison with normative data from a similar student population). For these reasons, caution is warranted in the interpretation of the specific dimensions that underlain the impact of individual differences in affective evaluation on the conditioned response to happy faces during extinction, and these findings await replication in future studies before stronger conclusions might be drawn.

Another caveat pertains to the GNAT that we used in the sense that it probably did not provide a direct and pure measure of the affective relevance or importance value of the face categories. Results of this task showed that participants more easily associated happy faces with importance (vs. unimportance) than they did for angry and neutral faces. This suggests that the GNAT rather captured the stimuli's valence and may have reflected participants' preferences or liking toward the face categories (Nosek & Banaji, 2001). Accordingly, it is possible that differential preferences toward happy faces actually drove the conditioned response persistence to these faces in the present study.

As angry and happy faces are usually considered as more arousing than neutral faces, it could be argued that these faces induced enhanced Pavlovian aversive conditioning because of their higher arousal value rather than, or in addition to, their affective relevance. Appraisal theories (e.g., Sander et al., 2003, 2005, 2018) suggest that stimuli appraised as relevant to the organism's concerns often trigger a physiological state of arousal that can be felt consciously as a consequence of the elicitation of a motivational state (see Montagrin & Sander, 2016; Pool et al., 2016), hence rendering it difficult to disentangle the specific contributions of affective relevance and arousal from one another. Although we cannot rule out that arousal contributed to our findings, it seems unlikely that they were solely determined by felt and/or physiological arousal (see Stussi et al., 2018, for a related discussion). In fact, previous studies (Hamm, Greenwald, Bradley, & Lang, 1993; Hamm & Stark, 1993; Hamm & Vaitl, 1996) have reported that highly arousing negative and positive stimuli, without taking into account their affective relevance to the organism, did not produce preferential Pavlovian aversive conditioning relative to less arousing stimuli. Moreover, supplementary analysis of the habituation phase revealed that (a) angry faces elicited larger skin conductance responses than happy faces before conditioning, whereas no difference emerged between angry and neutral faces, and between happy and neutral faces, and (b) the skin conductance responses to the various face categories during habituation did not correlate with the conditioned response to these stimuli during early acquisition and extinction.[3] These considerations suggest that an explanation in terms of arousal alone does not satisfactorily account for the occurrence of differential learning biases to both angry and happy faces.

Alternatively, our results could also be interpreted as reflecting the involvement of two different mechanisms instead of a single relevance detection mechanism: a specialized mechanism selectively acting on threat-related stimuli that is consistently engaged across individuals, and a more general one acting on affectively relevant stimuli that is more sensitive to individual differences. Future research is needed to disentangle these two competing explanations, for instance by investigating at the neurobiological level whether learning biases in Pavlovian aversive conditioning occurring in response to threat-relevant stimuli are underpinned by a threat-specific mechanism that is functionally distinct from a mechanism of relevance detection.

In conclusion, the present study highlights that positive stimuli with a relatively moderate level of relevance can be readily and persistently associated with an aversive outcome as is the case for threat-relevant stimuli; thus, replicating and extending recent work showing that learning biases in Pavlovian aversive conditioning are not specific to threat-related stimuli, but can likewise occur for positive emotional stimuli (Stussi et al., 2018). Our results further-

[2] Additional post hoc power analyses corroborated this assumption in indicating that achieved power to detect a small effect as reported in the present study ($g_{av} = 0.247$) using a one-tailed t test and an alpha level of .05 with a sample size ranging from 15 to 25 participants per group would vary between 23.14 and 32.83% for a within-participant design, and between 16.24 and 21.66% for a between-participant design.

[3] A repeated-measures ANOVA with CS type (CS+ vs. CS−) and CS category (angry vs. happy vs. neutral) as within-participant factors performed on the skin conductance response data during habituation revealed a main effect of CS category, $F(2, 212) = 4.20$, $p = .016$, partial $\eta^2 = .038$, 90% CI [.004, .083]. Further post hoc comparisons using Tukey's HSD tests indicated that angry faces elicited larger skin conductance responses than happy faces ($p = .012$, $g_{av} = 0.215$, 95% CI [0.064, 0.369]), whereas no statistically significant difference was found between angry and neutral faces ($p = .190$, $g_{av} = 0.129$, 95% CI [−0.019, 0.279]) or between happy and neutral faces ($p = .497$, $g_{av} = -0.088$, 95% CI [−0.239, 0.062]). Pearson's correlation analyses moreover showed no statistically significant relationship between the skin conductance responses to the different faces during habituation and the conditioned response to these faces during the early acquisition phase ($-.129 <$ all $rs(105) < .100$, all $ps > .18$) or during the extinction phase ($.001 <$ all $rs(105) < .129$, all $ps > .18$). Of note, computational learning models incorporating a Pavlovian bias to account for possible differences in inherent responding to the various CS categories did not provide a better fit to the normalized SCR data than the modified Rescorla-Wagner model implementing dual learning rates (see online supplemental materials).

more suggest that interindividual differences may play a key role in the development of these learning biases (Stussi et al., 2019; see also Lonsdorf & Merz, 2017). In this context, our study suggests that the determinants of Pavlovian aversive conditioning are more flexible than previously thought and may adaptively rely on the interaction between the stimulus at play and the individuals' current concerns. These findings thereby contribute to further advancing and refining our understanding of the basic mechanisms underlying emotional learning in humans, and could ultimately provide insights into impairments in this process that are typically associated with specific emotional disorders, including anxiety, phobia, or addictions.

References

Åhs, F., Rosén, J., Kastrati, G., Fredrikson, M., Agren, T., & Lundström, J. N. (2018). Biological preparedness and resistance to extinction of skin conductance responses conditioned to fear relevant animal pictures: A systematic review. *Neuroscience and Biobehavioral Reviews, 95,* 430–437. http://dx.doi.org/10.1016/j.neubiorev.2018.10.017

Ambadar, Z., Cohn, J. F., & Reed, L. I. (2009). All smiles are not created equal: Morphology and timing of smiles perceived as amused, polite, and embarrassed/nervous. *Journal of Nonverbal Behavior, 33,* 17–34. http://dx.doi.org/10.1007/s10919-008-0059-5

Benjamini, Y., & Hochberg, Y. (1995). Controlling the false discovery rate: A practical and powerful approach to multiple testing. *Journal of the Royal Statistical Society Series B. Methodological, 57,* 289–300. http://dx.doi.org/10.1111/j.2517-6161.1995.tb02031.x

Boll, S., Gamer, M., Gluth, S., Finsterbusch, J., & Büchel, C. (2013). Separate amygdala subregions signal surprise and predictiveness during associative fear learning in humans. *European Journal of Neuroscience, 37,* 758–767. http://dx.doi.org/10.1111/ejn.12094

Brainard, D. H. (1997). The Psychophysics Toolbox. *Spatial Vision, 10,* 433–436. http://dx.doi.org/10.1163/156856897X00357

Bramwell, S., Mallan, K. M., & Lipp, O. V. (2014). Are two threats worse than one? The effects of face race and emotional expression on fear conditioning. *Psychophysiology, 51,* 152–158. http://dx.doi.org/10.1111/psyp.12155

Brosch, T., Pourtois, G., & Sander, D. (2010). The perception and categorisation of emotional stimuli: A review. *Cognition and Emotion, 24,* 377–400. http://dx.doi.org/10.1080/02699930902975754

Brosch, T., Sander, D., Pourtois, G., & Scherer, K. R. (2008). Beyond fear: Rapid spatial orienting toward positive emotional stimuli. *Psychological Science, 19,* 362–370. http://dx.doi.org/10.1111/j.1467-9280.2008.02094.x

Canli, T., Sivers, H., Whitfield, S. L., Gotlib, I. H., & Gabrieli, J. D. E. (2002). Amygdala response to happy faces as a function of extraversion. *Science, 296,* 2191. http://dx.doi.org/10.1126/science.1068749

Coppin, G., Pool, E., Delplanque, S., Oud, B., Margot, C., Sander, D., & Van Bavel, J. J. (2016). Swiss identity smells like chocolate: Social identity shapes olfactory judgments. *Scientific Reports, 6,* 34979. http://dx.doi.org/10.1038/srep34979

Costa, P. T., Jr., & McCrae, R. R. (1992). *Revised NEO Personality Inventory (NEO-PI-R) and NEO Five-Factor Inventory (NEO-FFI) professional manual*. Odessa, FL: Psychological Assessment Resources.

Critcher, C. R., & Ferguson, M. J. (2016). "Whether I like it or not, it's important": Implicit importance of means predicts self-regulatory persistence and success. *Journal of Personality and Social Psychology, 110,* 818–839. http://dx.doi.org/10.1037/pspa0000053

Davey, G. C. L. (1995). Preparedness and phobias: Specific evolved associations or a generalized expectancy bias? *Behavioral and Brain Sciences, 18,* 289–297. http://dx.doi.org/10.1017/S0140525X00038498

de Berker, A. O., Tirole, M., Rutledge, R. B., Cross, G. F., Dolan, R. J., & Bestmann, S. (2016). Acute stress selectively impairs learning to act. *Scientific Reports, 6,* 29816. http://dx.doi.org/10.1038/srep29816

Delgado, M. R., Olsson, A., & Phelps, E. A. (2006). Extending animal models of fear conditioning to humans. *Biological Psychology, 73,* 39–48. http://dx.doi.org/10.1016/j.biopsycho.2006.01.006

Dienes, Z. (2011). Bayesian versus orthodox statistics: Which side are you on? *Perspectives on Psychological Science, 6,* 274–290. http://dx.doi.org/10.1177/1745691611406920

Dimberg, U., & Öhman, A. (1996). Behold the wrath: Psychophysiological responses to facial stimuli. *Motivation and Emotion, 20,* 149–182. http://dx.doi.org/10.1007/BF02253869

Dubois, J., & Adolphs, R. (2016). Building a science of individual differences from fMRI. *Trends in Cognitive Sciences, 20,* 425–443. http://dx.doi.org/10.1016/j.tics.2016.03.014

Dunsmoor, J. E., Niv, Y., Daw, N., & Phelps, E. A. (2015). Rethinking Extinction. *Neuron, 88,* 47–63. http://dx.doi.org/10.1016/j.neuron.2015.09.028

Esteves, F., Parra, C., Dimberg, U., & Öhman, A. (1994). Nonconscious associative learning: Pavlovian conditioning of skin conductance responses to masked fear-relevant facial stimuli. *Psychophysiology, 31,* 375–385. http://dx.doi.org/10.1111/j.1469-8986.1994.tb02446.x

Faul, F., Erdfelder, E., Lang, A.-G., & Buchner, A. (2007). G*Power 3: A flexible statistical power analysis program for the social, behavioral, and biomedical sciences. *Behavior Research Methods, 39,* 175–191. http://dx.doi.org/10.3758/BF03193146

Frijda, N. H. (1986). *The emotions*. London: Cambridge University Press.

Garcia, J., & Koelling, R. A. (1966). Relation of cue to consequence in avoidance learning. *Psychonomic Science, 4,* 123–124. http://dx.doi.org/10.3758/BF03342209

Gershman, S. J. (2016). Empirical priors for reinforcement learning models. *Journal of Mathematical Psychology, 71,* 1–6. http://dx.doi.org/10.1016/j.jmp.2016.01.006

Goeleven, E., De Raedt, R., Leyman, L., & Verschuere, B. (2008). The Karolinska Directed Emotional Faces: A validation study. *Cognition and Emotion, 22,* 1094–1118. http://dx.doi.org/10.1080/02699930701626582

Guitart-Masip, M., Huys, Q. J. M., Fuentemilla, L., Dayan, P., Duzel, E., & Dolan, R. J. (2012). Go and no-go learning in reward and punishment: Interactions between affect and effect. *NeuroImage, 62,* 154–166. http://dx.doi.org/10.1016/j.neuroimage.2012.04.024

Hamm, A. O., Greenwald, M. K., Bradley, M. M., & Lang, P. J. (1993). Emotional learning, hedonic change, and the startle probe. *Journal of Abnormal Psychology, 102,* 453–465. http://dx.doi.org/10.1037/0021-843X.102.3.453

Hamm, A. O., & Stark, R. (1993). Sensitization and aversive conditioning: Effects on the startle reflex and electrodermal responding. *Integrative Physiological and Behavioral Science, 28,* 171–176. http://dx.doi.org/10.1007/BF02691223

Hamm, A. O., & Vaitl, D. (1996). Affective learning: Awareness and aversion. *Psychophysiology, 33,* 698–710. http://dx.doi.org/10.1111/j.1469-8986.1996.tb02366.x

Ho, Y., & Lipp, O. V. (2014). Faster acquisition of conditioned fear to fear-relevant than to nonfear-relevant conditional stimuli. *Psychophysiology, 51,* 810–813. http://dx.doi.org/10.1111/psyp.12223

Holm, S. (1979). A simple sequentially rejective multiple test procedure. *Scandinavian Journal of Statistics, 6,* 65–70.

LaBar, K. S., Gatenby, J. C., Gore, J. C., LeDoux, J. E., & Phelps, E. A. (1998). Human amygdala activation during conditioned fear acquisition and extinction: A mixed-trial fMRI study. *Neuron, 20,* 937–945. http://dx.doi.org/10.1016/S0896-6273(00)80475-4

Lakens, D. (2013). Calculating and reporting effect sizes to facilitate cumulative science: A practical primer for *t*-tests and ANOVAs. *Frontiers in Psychology, 4,* 863. http://dx.doi.org/10.3389/fpsyg.2013.00863

LeDoux, J., & Daw, N. D. (2018). Surviving threats: Neural circuit and computational implications of a new taxonomy of defensive behaviour. *Nature Reviews Neuroscience, 19,* 269–282. http://dx.doi.org/10.1038/nrn.2018.22

Li, J., Schiller, D., Schoenbaum, G., Phelps, E. A., & Daw, N. D. (2011). Differential roles of human striatum and amygdala in associative learning. *Nature Neuroscience, 14,* 1250–1252. http://dx.doi.org/10.1038/nn.2904

Lindström, B., Golkar, A., & Olsson, A. (2015). A clash of values: Fear-relevant stimuli can enhance or corrupt adaptive behavior through competition between Pavlovian and instrumental valuation systems. *Emotion, 15,* 668–676. http://dx.doi.org/10.1037/emo0000075

Lonsdorf, T. B., Menz, M. M., Andreatta, M., Fullana, M. A., Golkar, A., Haaker, J., . . . Merz, C. J. (2017). Don't fear 'fear conditioning': Methodological considerations for the design and analysis of studies on human fear acquisition, extinction, and return of fear. *Neuroscience and Biobehavioral Reviews, 77,* 247–285. http://dx.doi.org/10.1016/j.neubiorev.2017.02.026

Lonsdorf, T. B., & Merz, C. J. (2017). More than just noise: Inter-individual differences in fear acquisition, extinction and return of fear in humans - Biological, experiential, temperamental factors, and methodological pitfalls. *Neuroscience and Biobehavioral Reviews, 80,* 703–728. http://dx.doi.org/10.1016/j.neubiorev.2017.07.007

Lundqvist, D., Flykt, A., & Öhman, A. (1998). *The Karolinska Directed Emotional Faces—KDEF*. Stockholm, Sweden: Karolinska Institutet, Department of Clinical Neuroscience, Psychology Section.

Macmillan, N. A., & Creelman, C. D. (2005). *Detection theory: A user's guide*. New York, NY: Psychology Press.

Mallan, K. M., Lipp, O. V., & Cochrane, B. (2013). Slithering snakes, angry men and out-group members: What and whom are we evolved to fear? *Cognition and Emotion, 27,* 1168–1180. http://dx.doi.org/10.1080/02699931.2013.778195

Martin, J., Rychlowska, M., Wood, A., & Niedenthal, P. (2017). Smiles as multipurpose social signals. *Trends in Cognitive Sciences, 21,* 864–877. http://dx.doi.org/10.1016/j.tics.2017.08.007

Mazurski, E. J., Bond, N. W., Siddle, D. A. T., & Lovibond, P. F. (1996). Conditioning with facial expressions of emotion: Effects of CS sex and age. *Psychophysiology, 33,* 416–425. http://dx.doi.org/10.1111/j.1469-8986.1996.tb01067.x

Montagrin, A., & Sander, D. (2016). Emotional memory: From affective relevance to arousal. *Behavioral and Brain Sciences, 39,* e216. http://dx.doi.org/10.1017/S0140525X15001879

Morey, R. D. (2008). Confidence intervals from normalized data: A correction to Cousineau (2005). *Tutorials in Quantitative Methods for Psychology, 4,* 61–64. http://dx.doi.org/10.20982/tqmp.04.2.p061

Niv, Y., Edlund, J. A., Dayan, P., & O'Doherty, J. P. (2012). Neural prediction errors reveal a risk-sensitive reinforcement-learning process in the human brain. *The Journal of Neuroscience, 32,* 551–562. http://dx.doi.org/10.1523/JNEUROSCI.5498-10.2012

Niv, Y., & Schoenbaum, G. (2008). Dialogues on prediction errors. *Trends in Cognitive Sciences, 12,* 265–272. http://dx.doi.org/10.1016/j.tics.2008.03.006

Nosek, B. A., & Banaji, M. R. (2001). The Go/No-go Association Task. *Social Cognition, 19,* 625–666. http://dx.doi.org/10.1521/soco.19.6.625.20886

Öhman, A., & Dimberg, U. (1978). Facial expressions as conditioned stimuli for electrodermal responses: A case of "preparedness"? *Journal of Personality and Social Psychology, 36,* 1251–1258. http://dx.doi.org/10.1037/0022-3514.36.11.1251

Öhman, A., Eriksson, A., & Olofsson, C. (1975). One-trial learning and superior resistance to extinction of autonomic responses conditioned to potentially phobic stimuli. *Journal of Comparative and Physiological Psychology, 88,* 619–627. http://dx.doi.org/10.1037/h0078388

Öhman, A., Fredrikson, M., Hugdahl, K., & Rimmö, P.-A. (1976). The premise of equipotentiality in human classical conditioning: Conditioned electrodermal responses to potentially phobic stimuli. *Journal of Experimental Psychology: General, 105,* 313–337. http://dx.doi.org/10.1037/0096-3445.105.4.313

Öhman, A., & Mineka, S. (2001). Fears, phobias, and preparedness: Toward an evolved module of fear and fear learning. *Psychological Review, 108,* 483–522. http://dx.doi.org/10.1037/0033-295X.108.3.483

Olsson, A., Carmona, S., Downey, G., Bolger, N., & Ochsner, K. N. (2013). Learning biases underlying individual differences in sensitivity to social rejection. *Emotion, 13,* 616–621. http://dx.doi.org/10.1037/a0033150

Olsson, A., Ebert, J. P., Banaji, M. R., & Phelps, E. A. (2005). The role of social groups in the persistence of learned fear. *Science, 309,* 785–787. http://dx.doi.org/10.1126/science.1113551

Olsson, A., & Phelps, E. A. (2004). Learned fear of "unseen" faces after Pavlovian, observational, and instructed fear. *Psychological Science, 15,* 822–828. http://dx.doi.org/10.1111/j.0956-7976.2004.00762.x

Pauli, W. M., Larsen, T., Collette, S., Tyszka, J. M., Seymour, B., & O'Doherty, J. P. (2015). Distinct contributions of ventromedial and dorsolateral subregions of the human substantia nigra to appetitive and aversive learning. *The Journal of Neuroscience, 35,* 14220–14233. http://dx.doi.org/10.1523/JNEUROSCI.2277-15.2015

Pavlov, I. P. (1927). *Conditioned reflexes*. London, UK: Oxford University Press.

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

Pelli, D. G. (1997). The VideoToolbox software for visual psychophysics: Transforming numbers into movies. *Spatial Vision, 10,* 437–442. http://dx.doi.org/10.1163/156856897X00366

Phelps, E. A., & LeDoux, J. E. (2005). Contributions of the amygdala to emotion processing: From animal models to human behavior. *Neuron, 48,* 175–187. http://dx.doi.org/10.1016/j.neuron.2005.09.025

Pool, E., Brosch, T., Delplanque, S., & Sander, D. (2016). Attentional bias for positive emotional stimuli: A meta-analytic investigation. *Psychological Bulletin, 142,* 79–106. http://dx.doi.org/10.1037/bul0000026

Prévost, C., McNamee, D., Jessup, R. K., Bossaerts, P., & O'Doherty, J. P. (2013). Evidence for model-based computations in the human amygdala during Pavlovian conditioning. *PLoS Computational Biology, 9,* e1002918. http://dx.doi.org/10.1371/journal.pcbi.1002918

Rescorla, R. A. (1988). Pavlovian conditioning. It's not what you think it is. *American Psychologist, 43,* 151–160. http://dx.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. Prosky (Eds.), *Classical conditioning II: Current research and theory* (pp. 64–99). New York, NY: Appleton-Century-Crofts.

Rolland, J. P., Parker, W. D., & Stumpf, H. (1998). A psychometric examination of the French translations of the NEO-PI-R and NEO-FFI. *Journal of Personality Assessment, 71,* 269–291. http://dx.doi.org/10.1207/s15327752jpa7102_13

Rouder, J. N., Speckman, P. L., Sun, D., Morey, R. D., & Iverson, G. (2009). Bayesian t tests for accepting and rejecting the null hypothesis. *Psychonomic Bulletin & Review, 16,* 225–237. http://dx.doi.org/10.3758/PBR.16.2.225

Rowles, M. E., Lipp, O. V., & Mallan, K. M. (2012). On the resistance to extinction of fear conditioned to angry faces. *Psychophysiology, 49,* 375–380. http://dx.doi.org/10.1111/j.1469-8986.2011.01308.x

RStudio Team. (2016). *RStudio: Integrated development for R*. Boston, MA: RStudio, Inc. Retrieved from https://www.rstudio.com/

Sander, D., Grafman, J., & Zalla, T. (2003). The human amygdala: An evolved system for relevance detection. *Reviews in the Neurosciences, 14,* 303–316. http://dx.doi.org/10.1515/REVNEURO.2003.14.4.303

Sander, D., Grandjean, D., & Scherer, K. R. (2005). A systems approach to appraisal mechanisms in emotion. *Neural Networks, 18,* 317–352. http://dx.doi.org/10.1016/j.neunet.2005.03.001

Sander, D., Grandjean, D., & Scherer, K. R. (2018). An appraisal-driven componential approach to the emotional brain. *Emotion Review, 10,* 219–231. http://dx.doi.org/10.1177/1754073918765653

Schwarz, G. (1978). Estimating the dimension of a model. *Annals of Statistics, 6,* 461–464. http://dx.doi.org/10.1214/aos/1176344136

Seligman, M. E. P. (1970). On the generality of the laws of learning. *Psychological Review, 77,* 406–418. http://dx.doi.org/10.1037/h0029790

Seligman, M. E. P. (1971). Phobias and preparedness. *Behavior Therapy, 2,* 307–320. http://dx.doi.org/10.1016/S0005-7894(71)80064-3

Stussi, Y., Brosch, T., & Sander, D. (2015). Learning to fear depends on emotion and gaze interaction: The role of self-relevance in fear learning. *Biological Psychology, 109,* 232–238. http://dx.doi.org/10.1016/j.biopsycho.2015.06.008

Stussi, Y., Ferrero, A., Pourtois, G., & Sander, D. (2019). Achievement motivation modulates Pavlovian aversive conditioning to goal-relevant stimuli. *npj Science of Learning, 4,* 4. http://dx.doi.org/10.1038/s41539-019-0043-3

Stussi, Y., Pourtois, G., & Sander, D. (2018). Enhanced Pavlovian aversive conditioning to positive emotional stimuli. *Journal of Experimental Psychology: General, 147,* 905–923. http://dx.doi.org/10.1037/xge0000424

Watanabe, N., & Haruno, M. (2015). Effects of subconscious and conscious emotions on human cue-reward association learning. *Scientific Reports, 5,* 8478. http://dx.doi.org/10.1038/srep08478

Watanabe, N., Sakagami, M., & Haruno, M. (2013). Reward prediction error signal enhanced by striatum-amygdala interaction explains the acceleration of probabilistic reward learning by emotion. *The Journal of Neuroscience, 33,* 4487–4493. http://dx.doi.org/10.1523/JNEUROSCI.3400-12.2013

Zhang, S., Mano, H., Ganesh, G., Robbins, T., & Seymour, B. (2016). Dissociable learning processes underlie human pain conditioning. *Current Biology, 26,* 52–58. http://dx.doi.org/10.1016/j.cub.2015.10.066

Received August 14, 2019
Revision received December 18, 2019
Accepted January 13, 2020 ■