Unconstraining Evaluative Conditioning Research by Using the Reverse Correlation Task

Marine Rougier and Jan De Houwer

Department of Experimental Clinical and Health Psychology, Ghent University, Belgium

Author Note

Marine Rougier is working at Ghent University as a postdoctoral researcher under the supervision of Jan De Houwer. Jan De Houwer is a professor at Ghent University since 2002, where he heads the Learning and Implicit Processes Laboratory.

This research was made possible by the Methusalem grant *BOF22/MET_V/002* of Ghent University to Jan De Houwer. Correspondence should be addressed to Marine Rougier, Department of Experimental-Clinical and Health Psychology, Ghent University, H.Dunantlaan 2, 9000 Ghent, Belgium. Email: Marine.Rougier@UGent.be.

The pre-registration files, materials, data, and analytic (R) scripts are made publicly available at https://osf.io/autbm/?view_only=caf1c58a239146089657f2c9ed5cb7f7. Authors report having no conflict of interest in publishing this work. Studies received approval (number 2021/39) from the ethical committee of the Faculty of Psychology and Educational Sciences at Ghent University.

Abstract

In the evaluative conditioning effect, pairing neutral stimuli (conditioned stimuli) with valenced stimuli (unconditioned stimuli) changes the evaluation of the former. We examined this effect with a reverse correlation task that assesses how participants visually remember the conditioned stimuli. Importantly, this measure 1) does not require participants to evaluate stimuli and 2) allows to capture multiple trait attributions. In a pre-registered experiment with US prolific academic users, we observed an evaluative conditioning effect in both an evaluation task and a reverse correlation task. Moreover, the effect in the reverse correlation task went beyond mere changes in valence. Our work opens new empirical and theoretical challenges for future conditioning research.

Keywords: Evaluative conditioning; Reverse correlation; Impression formation; Social judgment.

Unconstraining Evaluative Conditioning Research by Using the Reverse Correlation Task

Evaluative conditioning (EC) research showed that the evaluation of an initially neutral stimulus (Conditioned Stimulus or "CS") changes due to pairings with a positive or a negative stimulus (Unconditioned Stimulus or "US"; De Houwer, 2007; for a review, see Moran et al., 2023). To assess changes in the perception of the CS, virtually all past studies relied on (direct or indirect) measures that require participants to evaluate stimuli. For instance, participants could be asked to directly evaluate the CSs on a rating scale (e.g., "how positive vs. negative is the stimulus?"). In more indirect tasks, participants could be asked to classify CSs in the same way as other "good" or "bad" stimuli (Implicit Association Test [IAT]) or to categorize target words as "good" vs. "bad" after they were preceded by CSs (evaluative priming task). In the current work, we extended EC research by testing whether it can be observed in a visual memory measure. Specifically, we tested whether neutral faces are remembered in a visually biased fashion (i.e., as more positive- vs. negative-looking) after being paired with positive vs. negative images.

Prior research showed that minimal information about individuals (e.g., trustworthy behavior) can bias how people remember their faces (e.g., how trustworthy a face looks; Dotsch et al., 2013). Even participants' own actions can bias face memory: after having approached one group of neutral faces and avoided another one, participants reported a more positive facial representation of the approached (vs. avoided) group (Rougier et al., 2021). In these studies, visual representations [1] were assessed using the reverse correlation task (Dotsch & Todorov, 2012; Mangini & Biederman, 2004). In this task, participants compare noisy faces (neutral face mixed with different with random noise) to select the best match for a target category (e.g., the approached group). Crucially, the random noise slightly alters the

[1] Following the reverse correlation literature, we use the term "visual representations" when referring to the *outcome* of the task for the measured category and not to the cognitive (mental) representation itself.

face so that it can (mis)match – by chance only – with the way participants visually remember the target category. After many trials, averaged noise patterns selected by a (sample of) participant(s) form the "Classification Images" (CI). Typically, CIs are subsequently evaluated by independent judges to quantify the visual bias (e.g., to what extent CIs look trustworthy).

The questions we address in the current work are whether the mere co-occurrence of neutral faces (CSs) with positive and negative stimuli (USs) can bias the way participants remember these faces and, if so, which facial attributes are influenced. The reverse correlation task differs in several interesting ways from the evaluative measures typically used in EC research. First, because it is a measure of face *memory* (i.e., participants' task is to *'select the face that looks the most like'* the target category), it does not require participants to directly evaluate stimuli as good or bad. Although the EC effect has already been demonstrated using indirect evaluation tasks (IAT, evaluative priming) that do not involve a direct evaluation of the CS, these tasks still require participant to evaluative stimuli. We know of only two EC studies with measurement tasks that do not require stimulus evaluation (De Houwer et al., 1998; Spruyt et al., 2004). In those studies, participants' task was to name as quickly as possible a valenced target stimulus (e.g., word 'happiness') that was preceded by a CS. Faster reaction times were expected when the two shared the same valence (e.g., CS paired with a positive US followed by the target 'happiness'). Results, however, were mixed as the predicted effect was observed by Spruyt et al. (2004) but not by De Houwer et al. (1998). Using the reversed correlation task in an EC procedure can shed further light on whether EC can emerge in a task not requiring stimulus evaluation. [2]

[2] Prior research showed that direct stimulus evaluation *during* CS-US pairings strengthens EC (Gast & Rothermund, 2011). Note that we examine the need for direct stimulus evaluation *after* the CS-US pairings. Also note that a distinction needs to be made between whether a task requires direct stimulus evaluation and whether participants adopt the goal to evaluate stimuli; even in tasks without direct stimulus evaluation, people can still adopt the goal to evaluate stimuli. Hence, our studies cannot determine whether EC requires an evaluative goal.

Second, whereas previous measures focused solely on changes in valence, the reverse correlation task can capture also other changes in the CSs' impression, including changes that may be ineffable to the participants themselves (Mangini & Biederman, 2004). Indeed, when performing the task, participants can spontaneously choose to use any criteria of interest to select the face (e.g., they can select the face that looks the most intelligent or the most attractive). This allows for an almost limitless range of visual outcome variations, potentially resulting in complex combinations of facial biases (e.g., a face appearing both incompetent yet trustworthy). In the present context, pairing a CS with positive/negative USs could thus alter the perception of various positive/negative CS features like trustworthiness or incompetence. As De Houwer et al. (2019) noted, it is more of a historical coincidence that such variety of changes have not been assessed in EC research. In line with this idea, Rougier et al. (2023) demonstrated that pairing faces high vs. low on attractiveness (US) with other medium attractive faces (CS) influenced how the latter were perceived on personality traits (e.g., sociability). Of note, this previous research departs from our work in that we manipulated USs valence (instead of another feature, such as attractiveness).

Interestingly, Rougier et al. observed that this effect stems from the conceptual links between attractiveness and the personality traits (e.g., the effect was larger for sociability than for intelligence because sociability is more conceptually related to attractiveness; see Kim et al., 1996, and Förderer & Unkelbach, 2015, for related research). In our case, a question is whether the observed changes reflect a valence effect – so that the visual representations merely vary on positive/negative features (i.e., merely positive- vs. negative-looking) – or whether this effect goes beyond the manipulated feature of valence – so that the effect emerges more strongly for some categories of traits even when controlling for the valence of the traits.

To examine these issues, we paired two faces named ‘John’ vs. ‘Andy’ (CSs) with positive vs. negative pictures (USs). Then, we asked participants to recognize John vs. Andy’s face in a reverse correlation task and to complete evaluative self-reports. In a second part, we tested whether the visual representation of the CS face paired with positive (vs. negative) USs would be evaluated more positively by independent judges. We also explored whether differences between those visual representations varied not only on valence (i.e., more/less positive) but also in terms of other features (i.e., larger difference on socially-relevant features). To test that, we asked the judges to evaluate the visual representations on various personality and physical features.

To increase the chances of finding effects that go beyond valence, we tested features that varied not only on valence (i.e., how positive/negative is the feature) but also on social relevance. Socially-relevant traits refer to the dimension of social judgment that includes warmth- and communion-related traits (Abele et al., 2021). Traits of this dimension are also typically high in ‘other-relevance’ because they carry unconditional positive/negative consequences for individuals interacting with the trait holder (Peeters, 1983) and are strongly related to approach/avoidance tendencies (e.g., Rougier et al., 2021; Wentura et al., 2000). Socially-irrelevant traits refer to the dimension of social judgment that includes competence- and agency-related traits. Traits of this dimension are typically high in ‘self-relevance’ because they carry unconditional positive/negative consequences for the trait holder. Because traits of the warmth/communion dimension and other-relevant are more socially significant, we explored whether participants’ visual representations of CSs were specifically biased toward those traits. To determine if this effect was merely a secondary consequence of changes in valence (i.e., socially-relevant traits are also more valenced) or had an independent effect, we statistically controlled for the traits’ valence in our analyses.

Method

Transparency and Openness

Pre-registrations on the OSF include estimation of sample sizes, criteria for data exclusions, all manipulations, and all measures. Any deviation from the pre-registrations is signaled in the main text, following JARS (Kazak, 2018). Pre-registration files, materials, data, and analytic (R) scripts are made publicly available at https://osf.io/autbm/?view_only=caf1c58a239146089657f2c9ed5cb7f7. Data were analyzed using RStudio, version 1.4.1106 (RStudio Team, 2021) with packages and versions detailed in the Results section.

Power Analysis and Sample Size

This experiment required two independent power analyses: One for the "face producers" who underwent the EC procedure and the outcome measures, and one for the "judges" who rated the CIs obtained in the first part. Likelihood to detect effects in the reverse correlation (i.e., differences between CIs) depends on adequate sampling of both groups, whereas power in self-reports depends solely on the face producers sample.

For the face producers, we relied on similar work testing the effect of an approach/avoidance training on the reverse correlation task (i.e., $N = 110$ in Rougier et al., 2021; Exp. 1). To secure for a smaller effect, we recruited 200 participants, providing 80% power to detect a minimum EC effect of $dz = 0.18$ (5% false-positive rate in a two-tailed t-test with paired samples). For the judges, a sample of 100 participants (similar to Rougier et al., $N = 101$ in Experiment 3B) provided 80% power to detect a $dz = 0.25$ difference in CIs ratings (5% false-positive rate in a two-tailed t-test with paired samples).

Part 1: Creation of Classification Images Resulting from the EC Procedure

Participants and Design

One hundred and ninety-eight Prolific Academic users ($M_{age} = 41.02$, $SD_{age} = 14.09$; 96 men, 97 women, and 5 reporting "other" – self-categorizing as neither a man or a woman)

took part in exchange for a monetary compensation (£7.00/hour). Eligibility criteria included being an American English speaker, no prior participation in our lab experiments, computer access, and ≥98% approval rate. Following our pre-registration, we removed participants with >30% response time under 200 ms in the reverse correlation task (N = 15), reporting being named John or Andy (N = 2), and failing the attention check (N = 3), resulting in 178 participants with valid data. This experiment followed a within-participants design focusing on 2 US valence conditions (positive vs. negative).

Procedure

We programmed the experiment with jsPsych (de Leeuw, 2015) and administered it online. Participants provided their consent before starting the experiment.

Pairing Procedure. Participants were informed that they would be presented with pictures of John and Andy displayed on the right (left) of the screen together with other pictures on the left (right; random assignment). We asked them to look carefully at all pictures and try to remember which ones were presented together. CS pictures comprised two neutral male faces selected from the NimStim database (Tottenham et al., 2009); slightly blurred and presented in black and white. The names 'Andy' and 'John' were always presented below the faces (random name-faces assignment). USs consisted in 8 positive and 8 negative pictures selected from the IAPS database (Lang et al., 2005) not containing any human face. Positive and negative pictures did not differ in terms of valence extremity and arousal (all ps > .35). The face-name combination (CS) was randomly assigned to a pairing with positive vs. negative USs. Pairs were displayed during 2500 ms with a 1000 ms inter-trial interval. Each 16 CS-US pair was randomly presented twice, totaling 32 trials.

Reverse Correlation. Participants then engaged in a reverse correlation task in which their task was to "recognize the face that is the most similar to the face of Andy or John". In

one block (400 trials), they chose the face resembling Andy (John) and, in a second block (400 trials), the face resembling John (Andy; random order). [3]

Each trial featured two noisy faces (512 x 512 grayscale pixels) which consisted of a base image (blurred morph of John and Andy's faces, converted in gray scale) overlaid with unique random noise (see Figure 1). We generated 400 pairs, each with distinct noise pattern (R package rcicr version 0.3.4.1 with default settings; Dotsch, 2015). For each pair, the oriented noise created the "oriented" image, and the corresponding inverse noise created the "inverted" image (see Figure 1). Noisy face pairs were consistent across blocks and participants, but were randomly assigned to trial and position (left vs. right).

Figure 1

Base image and associated examples of stimuli (pair with images having opposite patterns of noise) for a given random noise in the reverse correlation task

Self-Reports. Then, participants indicated to what extent they agreed with the following statements (from 1 = *totally disagree* to 7 = *totally agree*): "I like Andy", "I like

[3] To reduce the number of trials, one can rely on an alternative RC procedure known as the "Brief-RC" (Schmitz et al., 2021).

John", "I think that Andy is a nice person", and "I think that John is a nice person" (always in the stated order). In one item, participants were asked to select the number 4 (attention check).

Memory of the US Valence, Demographics, and Exploratory Questions. Participants then reported the valence of the USs paired with each CS (response options: "pleasant pictures", "unpleasant pictures", and "I do not remember"). After answering demographic and exploratory questions (see Supplementary Materials), participants reported whether their first name was Andy or John ("Is your name Andy or John?", response options: "yes, my name is John or Andy", "no"). Participants were then thanked and debriefed.

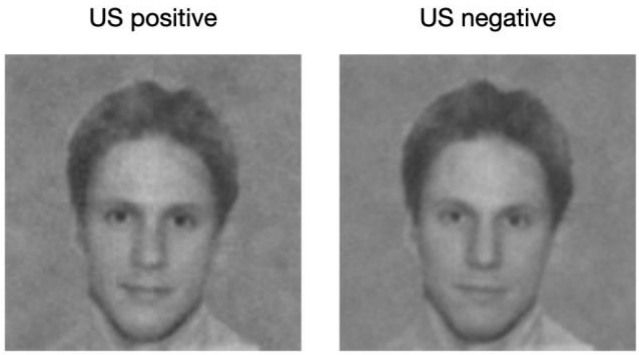
Construction of the Classification Images. We constructed CIs at condition and subgroup levels based on US valence. Condition-level CIs combined all participants' responses within a condition (US positive vs. US negative), whereas subgroup-level CIs used randomly selected individual responses within a condition. We adopted this strategy based on recent work suggesting that only relying on condition-level CIs is problematic (Cone et al., 2021). Indeed, when aggregating face producers' responses, the variability stemming from them is ignored (e.g., that some participants might produce a CI less pleasant than others in the US positive condition), ultimately increasing Type I error. Subgroup-level CIs, however, more likely retain part of this variability (see Rougier et al., 2021). Hence, we relied on both condition- and subgroup-level CIs.

For condition-level CIs, we averaged all the selected noises for all participants within each conditions and superimposed this average noise to the base image, resulting in two condition-level CIs (see Figure 2; constant scaling factor used in the *rcicr* package: 0.004). For subgroup-level CIs, we built each CI using the data of five participants within a US valence condition. Data of the same five participants were selected to compute a subgroup CI in the US positive and in the US negative conditions. This procedure was repeated until we

generated 356 subgroup CIs (random selection with replacement), with half (178) in the US positive condition and half (178) in the US negative condition (constant scaling factor: 0.012).

Figure 2

Condition-level classification images as a function of the US valence (positive vs. negative)



Part 2: Classification Image Ratings by Independent Judges

Participants

One hundred and one Prolific Academic users (M_{age} = 41.17, SD_{age} = 13.61; 56 men, 43 women, and 2 reporting "other") took part in exchange for a monetary compensation (£7.00/hour; same pre-screening criteria as in Part 1). We excluded 4 participants having less than 5% variance in their ratings.

Material

We selected 54 traits from literature on the attractiveness halo effect (Bassili, 1981; Eagly et al., 1991; e.g., traits related to vanity, concerns for others, adjustment), social judgment (Scheider et al., *under review*; traits related to sociability, integrity, intellectual competences, potency), and face perception (Oosterhof & Todorov, 2008; traits related to physical threat and physical attractiveness; see Supplementary Materials).

In a pilot study (N = 39, M_{age} = 36.92, SD_{age} = 14.88, 16 men, 20 women, 3 reporting "other"), we collected ratings on traits valence (i.e., from -3 = *extremely negative* to +3 =

extremely positive), self-relevance (i.e., to what extent a trait is consequential for the trait holder), other-relevance (i.e., to what extent a trait is consequential for the individuals living nearby the trait holder; both scales from 1 [*low consequences*] to 7 [*high consequences*]), and face-readability (i.e., to what extent is it easy to infer a personality trait on the basis of someone's face; from 1 [*not easy at all*] to 7 [*extremely easy*]). We computed trait valence, self-relevance, other relevance, and face-readability scores by averaging ratings for each trait. We collected traits face-readability to control for its potential effect. Indeed, we wanted to exclude the possibility that the observed differences were due to the fact that some traits (e.g., socially-relevant) were more strongly linked to facial features than other (socially-irrelevant). For self-other relevance, we computed a score of difference between self- and other-relevance values (i.e., we subtracted the other-relevance value to the self-relevance value for each trait so that the lower the score, the more other-relevant was the trait).

Procedure

We programmed the online study via JsPsych. Participants evaluated 12 faces on a series of traits. We informed them that faces were blurred to make the task more challenging and encouraged them to answer as honestly and as spontaneously as possible. The study had two parts. In a first part, participants evaluated the CIs on the 54 personality traits (from 0 = *not [trait] at all* to 5 = *totally [trait]*). Before the rating, we briefly displayed the CIs during one second each (automatic pace, random presentation). Then, participants rated the faces one by one in a random order and the items were presented adjacent to each other (random order). In a second part, they evaluated the CIs on liking (i.e., "I like this person") and positivity (i.e., "I think that this is a nice person", both from 1 = *not at all* to 7 = *very much*), always in this order. We also used an attention check ("Please answer 4 to this statement").

Each part had participants evaluate two condition-level CIs in one block and a sample of 10 subgroup-level CIs (5 in the US positive and 5 in the US negative condition) in another

block. Subgroup CIs were randomly selected from the whole pool of 356 images (that resulted from the 178 participants providing two CIs in the first part). Demographics were collected as in Part 1.

Results

First, we tested the traditional EC effect with face producers' self-reports of liking and positivity (Part 1) and whether it was moderated by the US valence memory, as typically observed in EC literature (e.g., Stahl & Unkelbach, 2009). Second, we tested the EC effect when considering the evaluative and personality ratings (as a positivity score) of the CIs obtained in Part 2. Third, we tested whether the observed differences on CIs' personality ratings could go beyond mere valence. When possible, we relied on mixed-models that allow considering more than one unit of analysis (e.g., participants, traits, and subgroup CIs; Judd et al., 2012; Westfall et al., 2014) as well as continuous within-participant variables (e.g., traits valence), which cannot be handled with traditional OLS regressions. [4]

We report frequentist and Bayesian analyses when performing OLS regression. [5] We reported the BF_{10} (evidence in favor of H_1) when the frequentist analysis reached significance ($p < .05$), and the BF_{01} (evidence in favor of H_0) when it did not. We carried out the analyses with the JZS default Bayes factor (ttestBF function) of the *BayesFactor* R package (version 0.9.12-4.2.; Morey et al., 2015).

EC Effect in Self-Reports (OLS Regression)

We considered US valence (positive: +0.5; negative: -0.5) and US valence memory (correct: +0.5; incorrect: -0.5) as contrast-coded predictors and liking and positivity ratings from face producers as outcome measures. Liking was higher for the CS paired with positive USs than with negative USs, $t(176) = 3.47$, $p < .001$, $dz = 0.26$, 95% CI [0.11; 0.41], $BF_{10} >$

[4] We pre-registered mixed-model analyses for subgroup-level CIs but not for condition-level CIs. The use of OLS regression vs. mixed-model does not impact the significance of the presented results.
[5] Given that there does not exist well-established metrics for the effect size and Bayes factor in mixed-models, we did not report them.

100 (see Table 1 for all descriptives), as well as positivity, $t(176) = 3.57$, $p < .001$, $dz = 0.27$, 95% CI [0.12; 0.42], $BF_{10} > 100$. The memory of the pairings moderated both effects when considering liking, $t(176) = 4.88$, $p < .001$, $d = 0.73$, 95% CI [0.43; 1.04], $BF_{10} > 100$, and positivity, $t(176) = 4.60$, $p < .001$, $d = 0.69$, 95% CI [0.39; 0.99], $BF_{10} > 100$, so that US valence effects were larger for participants having a correct ($N = 156$) than an incorrect memory ($N = 22$).

Table 1

Mean (and standard deviation) values of face producers' ratings of the CSs (liking, positivity) and judges' ratings of the CIs (liking, positivity, and personality) per US valence condition and level of CI

		US valence	
Measure	**Level of CI**	Positive	Negative
Face producers' liking	*NA*	4.03 (1.35)	2.18 (1.52)
Face producers' positivity rating	*NA*	4.19 (1.35)	2.33 (1.54)
Judges' liking	Condition	3.37 (1.13)	2.71 (1.27)
	Subgroup	2.76 (1.22)	2.33 (1.31)
Judges' positivity rating	Condition	3.44 (1.19)	2.60 (1.32)
	Subgroup	2.78 (1.22)	2.30 (1.33)
Judges' personality ratings (average positivity score)	Condition	3.07 (1.19)	2.71 (1.20)
	Subgroup	2.85 (1.26)	2.63 (1.31)
Judges' personality ratings for warmth- and communion-related traits	Condition	2.91 (1.23)	2.54 (1.27)
	Subgroup	2.86 (1.24)	2.52 (1.29)
Judges' personality ratings for competence- and agency-related traits	Condition	3.02 (1.17)	2.95 (1.21)
	Subgroup	2.99 (1.18)	2.92 (1.23)

Note. Standard deviation values are presented in parentheses. The symbol *NA* means "Non Applicable". The US valence effect is expected to be represented by a higher rating in the US positive than in the US negative condition.

EC Effect in the Reverse Correlation Task

Liking and Positivity Ratings

We considered the US valence as predictor and liking and positivity ratings from the judges as outcome measures. For condition-level CIs, we used OLS regressions whereas for subgroup-level CIs we used mixed-model analyses (judges and subgroup CIs being random factors). A main effect of US valence emerged in the expected direction on liking for both condition-level CIs, $t(96) = 5.55$, $p < .001$, $d = 0.57$, 95% CI [0.35; 0.78], $BF_{10} > 100$, and subgroup-level CIs, $t(118.36) = 5.10$, $p < .001$ (see Table 1). This effect was also significant on positivity ratings for both condition-level CIs, $t(96) = 5.82$, $p < .001$, $d = 0.59$, 95% CI [0.38; 0.81], $BF_{10} > 100$, and subgroup-level CIs, $t(113.97) = 5.14$, $p < .001$.

Personality Ratings

We then tested the US valence effect on personality ratings by reversing personality ratings (and corresponding valence scores) for negatively valenced traits. This way, the higher the personality rating, the more positive the perception of a CI. We considered US valence, traits valence (centered on zero), and their interaction as predictors (fixed effects) and personality ratings as the outcome measure. For condition-level CIs, we estimated a mixed-model having judges and traits as random factors. For subgroup-level CIs, we additionally considered subgroup CIs as a random factor. All effects remained significant when controlling for the face readability of the traits.

A main effect of US valence emerged for both condition-level CIs, $t(112.51) = 7.18$, $p < .001$, and subgroup-level CIs, $t(228.54) = 2.49$, $p = .013$ (see Table 1). The effect of US valence was significantly moderated by the traits valence extremity for both condition-level CIs, $t(83.24) = 5.08$, $p < .001$, and subgroup-level CIs, $t(104.19) = 3.69$, $p < .001$, so that the

difference observed between positive and negative US conditions increased when traits valence extremity increased.

Is There a Change Beyond Valence?

We categorized traits as belonging to the warmth/communion dimension (sociability- and integrity-related traits) vs. competence/agency dimension of judgment (intellectual competences- and potency-related traits) and considered their continuous score of self-other relevance. We then conducted mixed-models using the same random factors as previously. Note that the effects reported below remained significant when controlling for traits valence extremity and face readability.

First, we considered the US valence, the type of trait dimension (warmth/communion: +0.5; competence/agency: -0.5), and their interaction as fixed effects and personality ratings as the outcome measure. The US valence effect was larger for traits related to the warmth/communion dimension than to the competence/agency dimension for both condition-level CIs, $t(76.08) = 4.19$, $p < .001$, and subgroup-level CIs, $t(85.43) = 5.12$, $p < .001$. This implies that CIs in the positive USs condition were rated higher than CIs paired with negative USs on traits like warm or moral, but this difference was smaller (but still significant) for traits like intelligent or strong (see Table 1).

Second, we tested similar models but instead of the type of trait dimension we used the traits self-other relevance score (centered on zero). The moderation by the traits self-other relevance score on the US valence effect was not significant for the condition-level CIs, $t(61.13) = 1.97$, $p = .053$ [6], but it was significant for subgroup-level CIs, $t(65.98) = 2.04$, $p = .045$, so that the US valence effect increased when traits' other-relevance increased.

Discussion

[6] This effect reached significance when controlling for the traits valence, $t(68.28) = 2.39$, $p = .02$.

In this work, we investigated whether pairing faces with positive/negative stimuli can bias visual memory of those faces. We relied on a reverse correlation task, that is, a measure that does not require participants to directly evaluate stimuli and that allows exploring changes on multiple facial features.

We found an EC effect in self-reports of face producers and in the reverse correlation task. In the latter task, judges' personality ratings also indicated that the CSs' visual representations differed in terms of their social-relevance: CIs in the US positive vs. negative condition differed more on socially-relevant traits (e.g., warmth) than irrelevant traits (e.g., cleverness). Importantly, these differences were not solely due to variations in valence perception, as we controlled for trait valence.

These findings empirically contribute to EC research in several ways. First, they confirm that EC effects can occur when participants are not directly asked to evaluate stimuli. Interestingly, the effect size in the reverse correlation task was even twice as large (dz = 0.55-0.57) as for the self-reports (dz = 0.26-0.27), showing that reliable EC effects can be achieved in this context.

Second, pairing a CS with positive/negative stimuli influenced how the CS was perceived on other features than the manipulated one (US valence). This aligns with prior research showing that the manipulated feature of the US (e.g., attractiveness) can influence how the CS is perceived on other, conceptually related, features (e.g., sociability; Rougier et al., 2023). However, our work provides evidence for the effect of US valence across stimulus features beyond valence. Indeed, the observed changes on the visual representations were a function of the US valence but the facial features that were influenced were not the most positive/negative ones; they were the most socially-relevant ones (i.e., warmth/communion traits). Hence, although Rougier et al. (2023) already observed an impact of a manipulated feature across other stimulus features, our study is different in that (1) the manipulated feature

differed (i.e., valence rather than attractiveness) and (2) we observed effects that go beyond the manipulated feature (i.e., we observed effects on other traits even after controlling for the valence of the traits whereas in Rougier et al. this effect depended on how the other traits were conceptually related to attractiveness).

The fact that pairing CSs with positive/negative USs has an influence beyond valence poses important theoretical challenges. EC effects are typically explained through either associative or propositional processes. The associative perspective suggests that EC effects result from the formation and activation of an association between CS and US representations (e.g., Rescorla & Wagner, 1972; see also Gast & Rothermund, 2011). In contrast, the propositional perspective posits that EC arises from the formation and activation of propositions about relations in the environment (e.g., "the CS and US go together"; De Houwer, 2009, 2018; see De Houwer et al., 2020, 2021, for reviews). In our case, the USs did not contain any socially-relevant material but non-human positive and negative pictures (e.g., cute kittens, garbage). From an associative perspective, it is hard to see how the CS can be influenced on dimensions (e.g., warmth) that were not manipulated on the USs (i.e., USs should not vary systematically on warmth) and that cannot be accounted for by valence alone. This result is more consistent with propositional processes, suggesting that individuals activate propositions about how the person on the picture relates to the US (e.g., "this person has kittens") and infer personality attributes accordingly (e.g., "he is warm"). In this scenario, the US itself may not possess socially-relevant attributes, but how it is assumed to relate to the CS could lead to these types of attribution.

In future studies, one could investigate if attributions resulting from pairings vary based on the type of US and CS. When faces serve as CSs, socially-related attributions are likely to emerge, as these traits are pivotal in social interactions (e.g., Abele & Wojciszke, 2007; Rougier et al., 2021; Wentura et al., 2000). Conversely, for products, other dimensions

like tastiness or healthiness may be more central. This would support the notion that pairing effects depend on the specific inferences people make when CSs co-occur with positive/negative stimuli.

Moreover, the reverse correlation has potential in addressing current debates in EC research. For instance, some EC studies have reported contrast effects where a more positive evaluation of the CS paired with a negative US emerged, as compared to the CS paired with the positive US (Alves & Imhoff, 2023; Unkelbach & Fiedler, 2016). Whereas it is assumed that a typical EC effect reflects actual changes in stimuli perception, there is a debate on whether contrast effects could rather reflect an anchoring effect (e.g., change in how participants use the scale based on a context stimulus or an anchor; Frederick & Mochon, 2012). Because the reverse correlation does not involve any anchors, it can directly address this debate.

Conclusion

Relying on the reverse correlation task, we tested for the first time whether pairing faces with positive vs. negative stimuli could bias face memory. Our findings suggest that an EC effect can emerge when participants are not asked to evaluate stimuli (i.e., recognition instruction), but also that the face memory bias extends beyond mere changes in valence. This work strengthens and extends empirical evidence for the EC effect, while also challenging some associative theoretical approaches.

References

Abele, A. E., Ellemers, N., Fiske, S. T., Koch, A., & Yzerbyt, V. (2021). Navigating the social world: Toward an integrated framework for evaluating self, individuals, and groups. *Psychological Review, 128*(2), 290–314. https://doi.org/10.1037/rev0000262

Abele, A. E., & Wojciszke, B. (2007). Agency and communion from the perspective of self versus others. *Journal of Personality and Social Psychology, 93*(5), 751–763. https://doi.org/10.1037/0022-3514.93.5.751

Alves, H., & Imhoff, R. (2023). Evaluative context and conditioning effects among same and different objects. *Journal of Personality and Social Psychology, 124*(4), 735–753. https://doi.org/10.1037/pspa0000323

Bassili, J. N. (1981). The attractiveness stereotype: Goodness or glamour? *Basic and Applied Social Psychology*, *2*(4), 235–252. https://doi.org/10.1207/s1 5324834basp0204_1

Cone, J., Brown-Iannuzzi, J. L., Lei, R., & Dotsch, R. (2021). Type I Error Is Inflated in the Two-Phase Reverse Correlation Procedure. *Social Psychological and Personality Science*, *12*(5), 760–768. https://doi.org/10.1177/1948550620938616

De Houwer, J. (2007). A conceptual and theoretical analysis of evaluative conditioning. *The Spanish journal of psychology, 10*(2), 230–241. https://doi.org/10.1017/s1138741600006491

De Houwer, J. (2009). The propositional approach to associative learning as an alternative for association formation models. *Learning & Behavior, 37*(1), 1□–20. https://doi.org/10.3758/lb.37.1.1

De Houwer, J. (2018). Propositional models of evaluative conditioning. *Social Psychological Bulletin, 13*(3), 1–21. https://doi.org/10.5964/spb.v13i3.28046

De Houwer, J., Hermans, D., & Eelen, P. J. (1998). Affective and identity priming with episodically associated stimuli. *Cognition & Emotion*, *12*(2), 145-169. https://doi.org/10.1080/026999398379691

De Houwer, J., Richetin, J., Hughes, S., & Perugini, M. (2019). On the assumptions that we make about the world around us: A conceptual framework for feature transformation effects. *Collabra: Psychology, 5*(1), 43. https://doi.org/10.1525/collabra.229

De Houwer, J., Van Dessel, P., & Moran, T. (2020). Attitudes beyond associations: On the role of propositional representations in stimulus evaluation. In *Advances in experimental social psychology* (Vol. 61, pp. 127-183). Academic Press. https://doi.org/10.1016/bs.aesp.2019.09.004

De Houwer, J., Van Dessel, P., & Moran, T. (2021). Attitudes as propositional representations. *Trends in Cognitive Sciences*, *25*(10), 870-882. https://doi.org/10.1016/j.tics.2021.07.003

de Leeuw, J. R. (2015). jsPsych: A JavaScript library for creating behavioral experiments in a web browser. *Behavior Research Methods*, 47(1), 1–12. https://doi.org/10.3758/s13428-014-0458-y

Dotsch, R. (2015). rcicr: Reverse-correlation image-classification toolbox (R Package Version 0.3.0) [Computer software]. https://rdrr.io/cran/ rcicr/man/rcicr-package.html

Dotsch, R., & Todorov, A. (2012). Reverse correlating social face perception. *Social Psychological and Personality Science, 3*, 562–571. https://doi.org/10.1177/1948550611430272

Dotsch, R., Wigboldus, D. H. J., & van Knippenberg, A. (2013). Behavioral information biases the expected facial appearance of members of novel groups. *European Journal of Social Psychology, 43*(1), 116–125. https://doi.org/10.1002/ejsp.1928

Eagly, A. H., Ashmore, R. D., Makhijani, M. G., & Longo, L. C. (1991). What is beautiful is good, but...: A meta-analytic review of research on the physical attractiveness stereotype. *Psychological Bulletin*, *110*(1), 109–128. https://doi.org/10.1037/0033-2909.1 10.1.109

Förderer, S., & Unkelbach, C. (2015). Attribute conditioning: Changing attribute-assessments through mere pairings. *Quarterly Journal of Experimental Psychology, 68*(1), 144–164. https://doi.org/10.1080/17470218.2014.939667

Frederick, S. W., & Mochon, D. (2012). A scale distortion theory of anchoring. *Journal of Experimental Psychology: General, 141*(1), 124–133. https://doi.org/10.1037/a0024006

Gast, A., & Rothermund, K. (2011). What you see is what will change: Evaluative conditioning effects depend on a focus on valence. *Cognition and Emotion, 25*(1), 89–110. https://doi.org/10.1080/02699931003696380

Judd, C. M., Westfall, J., & Kenny, D. A. (2012). Treating stimuli as a random factor in social psychology: A new and comprehensive solution to a pervasive but largely ignored problem. *Journal of Personality and Social Psychology, 103*(1), 54–69. https://doi.org/10.1037/a0028347

Kazak, A. E. (2018). Editorial: Journal article reporting standards. *American Psychologist, 73*(1), 1–2. https://doi.org/10.1037/amp0000263

Kim, J., Allen, C. T., & Kardes, F. R. (1996). An Investigation of the Mediational Mechanisms Underlying Attitudinal Conditioning. *Journal of Marketing Research*, *33*(3), 318–328. https://doi.org/10.1177/002224379603300306

Lang, P. J., Bradley, M. M., & Cuthbert, B. N. (2005). *International Affective Picture System (IAPS)* [Database record]. APA PsycTests. https://doi.org/10.1037/t66667-000

Mangini, M., & Biederman, I. (2004). Making the ineffable explicit: Estimating the information employed for face classifications. *Cognitive Science, 28*(2), 209–226. https://doi.org/10.1207/s15516709cog2802_4

Moran, T., Nudler, Y., & Bar-Anan, Y. (2023). Evaluative conditioning: Past, present, and future. *Annual Review of Psychology, 74*, 245-269. https://doi.org/10.1146/annurev-psych-032420-031815

Morey, R. D., Rouder, J. N., Jamil, T., & Morey, M. R. D. (2015). Package 'bayesfactor'. *URLh http://cran/r-projectorg/web/packages/BayesFactor/BayesFactor pdf i (accessed 1006 15)*.

Oosterhof, N. N., & Todorov, A. (2008). The functional basis of face evaluation. *Proceedings of the National Academy of Sciences of the United States of America, 105*(32), 11087–11092. https://doi.org/10.1073/pnas .0805664105

Peeters, G. (1983). Relational and informational pattern in social cognition. In W. Doise & S. Moscovici (Eds.), *Current issues in European social psychology* (pp. 201–237). Cambridge University Press.

Rescorla, R. A., & Wagner, A. R. (1972). A theory of Pavlovian conditioning: Variations in the effectiveness of reinforcement and nonreinforcement. *Classical conditioning II: Current research and theory*, 2, 64-99.

Rougier, M., De Houwer, J., Richetin, J., Hughes, S., & Perugini, M. (2023). From Halo to Conditioning and Back Again: Exploring the Links Between Impression Formation and Learning. *Collabra: Psychology, 9*(1). https://doi.org/10.1525/collabra.84560

Rougier, M., Schmitz, M., & Yzerbyt, V. (2021). When my actions shape your looks: Experience-based properties of approach/avoidance bias the visual representation of others. *Journal of Personality and Social Psychology, 120*(5), 1146–1174. https://doi.org/10.1037/pspa0000268

RStudio Team. (2021). *RStudio: Integrated Development Environment for R*. RStudio. http://www.rstudio.com/

Scheider, J., Barbedor, J., Yzerbyt, V. & Abele, A. (*under review*). A Novel Approach to the Evaluation of Groups: Type of Group and Facet of Evaluation matter.

Schmitz, M., Rougier, M., & Yzerbyt, V. (2021, March 24). Introducing the Brief Reverse Correlation. https://doi.org/10.31234/osf.io/xg693

Spruyt, A., Hermans, D., De Houwer, J., & Eelen, P. (2004). Automatic non-associative semantic priming: Episodic affective priming of naming responses. *Acta Psychologica*, *116*(1), 39-54. https://doi.org/10.1016/j.actpsy.2003.12.012

Stahl, C., & Unkelbach, C. (2009). Evaluative learning with single versus multiple unconditioned stimuli: The role of contingency awareness. *Journal of Experimental Psychology: Animal Behavior Processes*, *35*(2), 286–291. https://doi.org/10.1037/a0013255

Tottenham, N., Tanaka, J. W., Leon, A. C., McCarry, T., Nurse, M., Hare, T. A., ... & Nelson, C. (2009). The NimStim set of facial expressions: Judgments from untrained research participants. *Psychiatry research*, *168*(3), 242-249. https://doi.org/10.1016/j.psychres.2008.05.006

Unkelbach, C., & Fiedler, K. (2016). Contrastive CS-US relations reverse evaluative conditioning effects. *Social Cognition*, *34*(5), 413-434. https://doi.org/10.1521/soco.2016.34.5.413

Wentura, D., Rothermund, K., & Bak, P. (2000). Automatic vigilance: The attention-grabbing power of approach- and avoidance-related social in- formation. *Journal of Personality and Social Psychology, 78*(6), 1024–1037. https://doi.org/10.1037/0022-3514.78.6.1024

Westfall, J., Kenny, D. A., & Judd, C. M. (2014). Statistical power and optimal design in experiments in which samples of participants respond to samples of stimuli. *Journal of Experimental Psychology: General, 143*(5), 2020–2045. https://doi.org/10.1037/xge0000014

Supplementary Material Section

Exploratory Questions

Demand Awareness. Open-ended question: "What do you think the researchers were trying to achieve in this study?".

Influence Awareness. "Do you think that, in TASK 1, the pleasantness of the pictures presented on the right [left] (i.e., together with John and Andy) influenced your responses in TASK 2 when you had to select the face that looked the most like John or Andy?", response options: "Yes", "No", "I do not know".

Demand Compliance. "When we asked you to complete TASK 2, did you respond truthfully? Or did you try to fake your response (i.e., tried to tell us what you thought we wanted to hear)? Please be honest here (it will not affect payment in any way).", response options: "Yes - I faked my response based on what I thought the researchers wanted to find", "No - my responses were based on how I genuinely felt", "I do not know".

List of Personality Traits Used in Part 2 of the Experiment

Hereafter is the complete list of the 54 personality traits and personality outcomes as a function of the underlying personality dimension. Sociability: sociable, fun-loving, likeable, popular, friendly, funny, agreeable, warm. Intellectual competence: intelligent, skilful, rational, scientific, ambitious, hard-working, likely to achieve career success. Concerns for others: sensitive, empathic, compassionate, generous, modest, egoistic (R). Integrity: trustworthy, honest, faithful, sneaky (R), moral, sincere. Adjustment: well-adjusted, satisfied, happy, confident, likely to have a positive self-regard, healthy. Potency: strong, self-assertive, dominant, likely to act as a leader. Vanity: vain, elitist, snobbish, shallow, humble (R), materialistic, pompous, prudish (R), boastful. Physical threat: violent, aggressive, hateful, hurtful. Physical attractiveness: physically attractive, cute, disgusting (R), beautiful.

Note. (R) indicates that the trait is reversed.