

What Causes What: Causal Directionality Moderates Evaluative Conditioning Effects

Social Psychological and
Personality Science
1–9

© The Author(s) 2026

Article reuse guidelines:

sagepub.com/journals-permissions

DOI: 10.1177/19485506261426645

journals.sagepub.com/home/spp



Simone Mattavelli¹ , Marine Rougier² , and Tal Moran³ 

Abstract

Past research showed that Evaluative Conditioning (EC) effects are shaped by the framing of the relationship between conditioned stimuli (CS) and unconditioned stimuli (US). We tested if causal directionality—whether the CS causes the US or the US causes the CS—moderates EC effects. Across two preregistered experiments ($N = 680$), participants viewed food brands (CSs) paired with facial expressions (USs). In Experiment 1, the EC effect was stronger in the CS-causes-US than in the US-causes-CS condition, but this emerged only in self-reports. Experiment 2 extended these findings to a health-related context, with expressions framed as physical comfort/discomfort. Stronger EC effects emerged in the CS-causes-US condition on both self-report and Implicit Association Test (IAT). Aggregated analyses suggest that the superiority of the CS-causes-US (vs. US-causes-CS) condition emerged on both measures. These results refine a propositional view of EC, showing that pairing may function as a symbolic cue conveying more than co-occurrence.

Keywords

Evaluative Conditioning, causality, inference, propositional account

Handling Editor: André Mata

From choosing a cookie brand to forming first impressions, our likes and dislikes shape countless everyday decisions (Allport, 1935). One powerful way we come to hold these evaluations is through Evaluative Conditioning (EC), the phenomenon by which an individual's evaluation of a stimulus (i.e., conditioned stimulus, or CS) changes due to its pairing with another positive or negative stimulus (i.e., unconditioned stimulus, or unconditioned stimuli [US]; De Houwer, 2007; see Hofmann et al., 2010 and Moran et al., 2023 for a review). For instance, when a neutral brand is repeatedly paired with positive (vs. negative) images, people tend to develop a more favorable evaluation of the brand (e.g., Pleyers et al., 2007).

A growing body of research shows that the specific relation between the paired stimuli moderates EC effects (e.g., Moran & Bar-Anan, 2013; Unkelbach & Fiedler, 2016; Zanon et al., 2014). Building on recent studies showing that EC effects are stronger when the US and the CS are told to be causally linked (Hughes et al., 2019), we investigated the role of CS-US causal directionality. Namely, we argue that a relation in which the CS causes the US should not be equivalent in terms of its effect on the CS evaluation to a relation in which the US causes the CS. Because the former framing provides clearer, more diagnostic information about the valence of the CS, it should yield stronger EC effects.

Mental Propositions and EC: The Importance of CS-US Relationship

Early theorizing of EC posited that this effect is mediated by the automatic formation and activation of associative links between the CS and the US: learning their co-occurrence leads the CS to activate the US representation and its evaluation, following the Hebbian principle “cells (here stimuli) that fire together wire together” (Baeyens et al., 1992; Levey & Martin, 1975). Contrasting this view, a propositional account of EC posits that stimulus pairings influence liking through the formation and activation of propositions specifying the relation between the stimuli (e.g., Mitchell et al., 2009). For example, repeated pairing of a cookie brand with the image of a famous actor may increase liking for the brand because individuals activate the proposition “the actor likes/eats/loves this cookie,” forming a meaningful relational link. Naturally,

¹University of Milano-Bicocca, Italy

²Ghent University, Belgium

³The Open University of Israel, Raanana, Israel

Corresponding Author:

Simone Mattavelli, Department of Psychology, University of Milano-Bicocca, I, Piazza dell'Ateneo Nuovo, Milan 20126, Italy.

Email: simone.mattavelli@unimib.it

individuals would interpret co-occurrence as a symbolic cue for a relation of similarity or assimilation between the CS and the US, based on the general heuristic that things appearing together are likely similar (De Houwer & Hughes, 2016). This leads to inferring that the CS shares the US's valence, thereby producing an EC effect (De Houwer, 2018).

However, this default interpretation is flexible. A wide body of EC research has shown that relational information can significantly moderate both the direction and strength of EC effects (e.g., Bading et al., 2025; Fiedler & Unkelbach, 2011; Förderer & Unkelbach, 2012; Hu et al., 2017; Hughes et al., 2019; Kurdi & Banaji, 2019; Moran & Bar-Anan, 2013; Moran et al., 2016; Zanon et al., 2014). For example, Förderer and Unkelbach (2012) found reversal in EC effects when the CSs disliked (vs. liked) the USs. Overall, studies show that EC effects are stronger for assimilative CS-US relations (e.g., the CS and US are friends, the CS loves/causes the US) than for contrastive ones (e.g., enemies, hates/prevents; Fiedler & Unkelbach, 2011; Hu et al., 2017). Such findings highlight that the way individuals construe relationships between stimuli powerfully shapes the resulting evaluative effect.

Causality in CS-US Relationship: Recent Evidence and an Open Question

Notwithstanding the superior EC effects observed when the CS-US pairing supports assimilative inferences, distinct assimilative relationships generate distinct evaluative conclusions. Hughes et al. (2019) demonstrated that EC effects are stronger when the CS is described as *causing* the US than when described as *predicting* the US. This finding suggests that causal relations support more robust evaluative learning, likely because they provide a stronger basis for inference (e.g., “the CS causes something negative, therefore it must be negative itself”). Causal inferences more readily imply dispositional or intrinsic properties of the CS, making it easier to generalize the evaluative implication to the CS itself. Hence, causality appears to facilitate deeper evaluative integration than other assimilative relational qualifiers.

Yet not all causal relations have the same evaluative consequences. A key feature of causality is its directionality: the cause and the effect play different functional roles and are not interchangeable. This asymmetry is fundamental to how people mentally represent causal relations (e.g., Sloman & Lagnado, 2015; Waldmann & Holyoak, 1992). Specifically, the belief that consuming a product (CS) causes negative affect (US) differs from the belief that negative affect (US) causes the product use (CS). The former implies the product has intrinsic negative value; the latter is less clear about intrinsic negative value and may even suggest the opposite—that the product is used to alleviate a negative affect, which could imply positive value. Hence,

there should be a larger chance of inferring the negativity of the CS in the former than in the latter case.

Despite the theoretical importance of this distinction, no prior studies have directly tested whether causal directionality moderates EC effects. If the CS (e.g., a drink) is seen as the cause of a positive US (e.g., a smile), the CS may be evaluated more favorably than when it is merely the consequence of the same US. Demonstrating that causal directionality moderates EC effects would not only deepen our theoretical understanding of the mechanisms underlying evaluative learning, but would also reveal how cause-effect interpretations might boost/prevent changes in evaluation resulting from stimuli pairing that characterize other phenomena in social-cognitive research.

Overview of the Experiments

Using food brands as CSs and facial expressions as USs, we examined how different instruction-based manipulations of CS-US causal directionality affect EC effects. We hypothesized stronger shifts in the evaluation of the CS when it is presented as the cause of the US, as opposed to its effect.

In Experiment 1, participants saw food brands (CSs) paired with facial expressions showing happiness or sadness (USs). They were told either that eating the food caused the emotion (CS-causes-US), that the emotion led to eating the food (US-causes-CS), or simply that the two were causally related (unspecified). The inclusion of this non-directional condition was intended to test whether one of the two directional conditions was the one that participants defaulted to. We found stronger EC effects in self-reported evaluations in the CS-causes-US condition than in the US-causes-CS condition. However, this difference did not emerge in the indirect evaluation measure (Implicit Association Test [IAT]; Greenwald et al., 1998).

Experiment 2 focused on the two directional conditions and increased the diagnosticity of the US for CS judgments. We redefined the USs as physical comfort/discomfort and assessing CS healthiness (rather than general valence) to strengthened the alignment between US meaning and CS evaluation, fostering more meaningful propositional reasoning. As expected, EC effects were significantly stronger in the CS-causes-US than in the US-causes-CS condition across self-report and IAT measures.

Transparency and Openness

The experiments were preregistered on Open Science Framework (OSF; Experiment 1: <https://osf.io/7aqce>; Experiment 2: <https://osf.io/vhspq>). Materials, data, and analysis code are available on the OSF (Experiment 1: <https://osf.io/xu6jh>; Experiment 2: <https://osf.io/ja7wy>). We have not conducted additional experiments on this research question. We report all manipulations and

measures used. The experiments received formal approval from the internal Ethics Committee.

Experiments 1 and 2

Sample Size Determination

Both experiments employed a mixed design, with US valence manipulated within participants and CS-US causal relationship manipulated between participants. Three method variables were manipulated between participants: measures order (self-reports vs. IAT first), IAT block order (EC consistent vs. inconsistent block first) and CS-US assignment (Brand X vs. Brand Y with US_{pos}).

We based our power analyses for both experiments on the expected difference in EC effects across two CS-US causal directionality conditions (i.e., CS-causes-US vs. US-causes-CS), considering both self-reported evaluations and the IAT. For Experiment 1, we aimed to detect effects as small as *Cohen's d* = 0.35, using the “pwr” package in R (Champely, 2017), with $\alpha = .05$ and power $(1 - \beta) = .95$. This analysis suggested a required sample of 107 participants per group, for a total of 321, and we planned to recruit up to 360 participants. A Sequential Bayes Factor (SBF) design was employed (Schönbrodt & Wagenmakers, 2018), with thresholds of $BF > 6$ or < 0.16 , a minimum of 90 participants, and an increment of 90 participants per test, up to a maximum of 360.

For Experiment 2, we targeted a smaller effect size of $d = 0.30$ and applied the same power analysis approach, which indicated 145 participants per group (290 total). We planned to recruit up to 320 participants and adopted the same SBF design, with a minimum of 80 participants, increments of 80 per test, and a maximum of 320 participants. In both experiments, we ultimately collected the maximum planned sample size.

Participants and Procedure

Three hundred sixty participants (190 males, 161 females, nine unspecified, $M_{age} = 40.80$, $SD_{age} = 11.29$) completed Experiment 1 and 220 participants (176 males, 136 females, eight unspecified, $M_{age} = 38.51$, $SD_{age} = 12.07$) completed Experiment 2.¹ Participants in both experiments were native English-speaking recruited via Prolific Academic (<https://www.prolific.com/>) in exchange for a £1.6 monetary reward. After completing demographics and giving their consent to participate, participants were introduced to the EC phase by receiving different instructions (depending on the CS-US causal directionality condition). Next, participants completed a self-report evaluation of the CSs and an IAT (in counterbalanced order). Finally, participants answered two questions assessing memory of and belief in causality instructions and one question assessing memory for contingencies. Participants were thanked and debriefed.

Stimuli

CS Stimuli. Two novel food brands were pretested ($N = 89$) for valence on a response scale of 1 (*negative*) to 9 (*positive*); Ambik: $M = 5.16$, $SD = 1.11$, Safom: $M = 5.13$, $SD = 0.89$, $t(88) = 0.29$, $p = .769$, $d = 0.03$, $BF_{10} = 0.12$, and arousal on a response scale of 1 (*very calm*) to 9 (*very excited*); Ambik: $M = 5.05$, $SD = 1.04$, Safom: $M = 4.98$, $SD = 1.06$, $t(88) = -0.71$, $p = .475$, $d = 0.06$, $BF_{10} = 0.14$.

US Stimuli. Ten facial stimuli from the Karolinska Directed Emotional Faces (KDEF, Lundqvist et al., 1998) were selected after conducting a pretest ($N = 148$). The pretest included 32 distinct facial identities, each with a happy and a sad version. Each participant rated 32 faces: 16 identities were shown in their happy version, and 16 in their sad version, with assignment counterbalanced so that no participant ever saw both expressions of the same individual. Facial stimuli were rated on a 7-point scale from 1 (*very sad*) to 7 (*very happy*). From these, we selected 10 identities for which the happy and sad versions showed clearly distinct ratings (happy: $M = 5.69$, $SD = 0.29$; sad: $M = 2.33$, $SD = 0.35$). In the main studies' conditioning procedure, participants saw all 10 distinct identities: five were presented in their happy version, the other five in their sad version.

Manipulation

Experiment 1. All participants were informed that they would see food brands appearing with some faces displaying emotional expressions. In the CS-causes-US condition, participants were told that faces portrayed “emotional expressions observed in those individuals right after eating the food products belonging to each brand. Thus, the products of the two brands caused the emotional expression.” In the US-causes-CS condition, participants were told faces portrayed “emotional expressions observed in individuals right before they decided eating the products belonging to each brand. Thus, the emotion they experienced caused those individuals to eat the product of the paired brand.” In the control condition, participants were told that “faces and food stimuli were causally linked” without any further specifications.

Experiment 2. Participants were informed that they would see food brands appearing with some faces on the screen and that faces would portray expressions of physical comfort or physical discomfort. In the CS-causes-US condition, participants were told that faces portrayed “expressions of physical comfort or physical discomfort observed on those faces right after eating the food products of each brand. Thus, the products of the two brands caused the physical comfort or discomfort seen in the paired faces.” In the US-causes-CS condition, participants were told that faces portrayed “expressions of physical comfort or physical discomfort observed on those faces before eating the

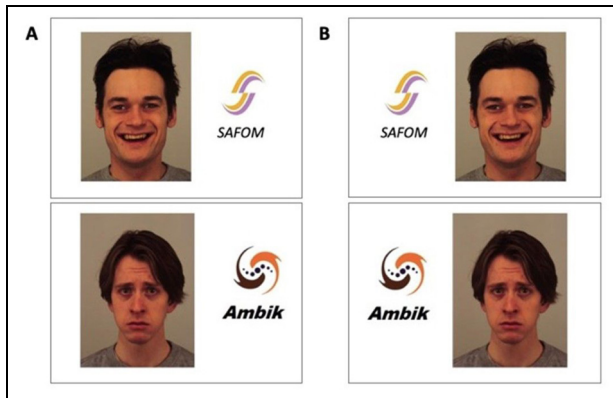


Figure 1. EC procedure in the “US-causes-CS” condition (A panel) and in the “CS-causes-US” condition (B panel). Sample images are AM06HAS and AM22SAS from KDEF (Lunqvist et al., 1998)

food products of each brand. Thus, the physical comfort or discomfort caused those individuals to eat the product of the paired brand.”

EC Procedure. Identical in the two experiments, the EC procedure consisted of two blocks of 20 trials each. Each trial included a pair of food brand (CS) and an emotional face (US) that remained on the screen for 2,500 ms (ITI of 1,000 ms). During each block, one brand was presented twice with each of the five positive faces, whereas the other brand was presented twice with each of the five negative faces. The order of trials was random. The CS-US location on screen varied across conditions, to reinforce the implied causal directionality: in the CS-causes-US condition, the CS appeared on the left side of the screen (and the US on the right side); in the US-causes-CS condition, the CS appeared on the right side of the screen (and the US on the left side, see Figure 1 for trials’ example); in the control condition (Experiment 1), the CS appeared on the right side in half of the trials and on the left in the other half of the trials.

Measures

Self-Reports. Self-reported ratings were assessed using four semantic differential scales. Participants indicated their general impression of the CS stimuli using a scale ranging from -5 to $+5$ with 0 as a neutral point. In Experiment 1, the four end-points of the scales were: *negative–positive*, *unpleasant–pleasant*, *bad–good*, and *I don’t like it–I like it*. In Experiment 2, the four end-points of the scales were: *Spoiled–Fresh*, *Harmful–Beneficial*, *Nonnutritive–Nourishing*, *Weakening–Vitalizing*. A mean evaluative rating was calculated for each CS by averaging scores from these four scales (Cronbach’s $\alpha = .98$ in both experiments).

IAT. We used the seven-block IAT (Nosek et al., 2005). In Experiment 1, images of the two brands served as of

target labels, and the words “Good” and “Bad” attribute labels. Four positive (*Good*, *Pleasant*, *Positive*, *Great*) and four negative (*Bad*, *Unpleasant*, *Negative*, *Awful*) adjectives served as attribute stimuli and images of the two brands served as target stimuli. In Experiment 2, the words “Healthy” and “Unhealthy” served as attribute labels and four adjectives (*Healthy*: *Fresh*, *Beneficial*, *Nourishing*, *Vitalizing*; *Unhealthy*: *Spoiled*, *Harmful*, *Nonnutritive*, *Weakening*) were used as stimuli in the trials. The IAT D score (Greenwald et al., 2003) was calculated such that a positive score indicated a preference for the brand paired with US_{pos} . The internal consistency of the IAT was $\alpha = .78$ in Experiment 1 and $\alpha = .81$ in Experiment 2.

Instruction Memory. Participants indicated what they had been told about the CS-US relationship at the start of the experiment by selecting one of four options: (a) eating the products of the two brands caused the emotion[physical (dis)comfort], (b) the emotion [physical (dis)comfort] caused those individuals to eat the product of the paired brand, (c) the emotion[physical (dis)comfort] and the paired food products are causally linked (but no additional information was provided on the direction of this causality), (d) I do not remember.

Belief in the Instructions. Participants rated how much they believed the initial instruction about the CS-US relationship on a scale from -4 (*I did not believe you at all*) to $+4$ (*I completely believed you*) with 0 (*I somewhat believed you*) as a midpoint.

Contingency Memory. Participants were asked to recall which food brand had been paired with happy[physical comfort] vs. sad[physical discomfort] faces, choosing between three options: (a) Ambik-happy faces[physical comfort] and Safom-sad faces[physical discomfort], (b) Ambik-sad faces[physical discomfort] and Safom-happy faces[physical comfort], and (c) I do not remember.

Data Preparation

In line with the preregistrations, we excluded participants who had more than 10% of fast IAT trials ($RT > 300$ ms) from the IAT analysis (Experiment 1: $N = 16$; Experiment 2: $N = 10$; Greenwald et al., 2003). After exclusions, the final sample sizes allowed us to detect effects of approximately $d = 0.33$ (Experiment 1) and $d = 0.29$ (Experiment 2) for self-report ratings and $d = 0.34$ (Experiment 1) and $d = 0.29$ (Experiment 2) for IAT scores with 95% power at $\alpha = .05$.

We introduced a minor deviation from the preregistered plan for the analyses of self-report ratings: instead of computing difference scores (i.e., EC effect) as the dependent variable in a three-factorial design (preregistered analysis), we analyzed the average ratings assigned to the two CSs in a 3 (CS-US directionality) \times 2 (US valence) mixed-design

analysis of variance (ANOVA). Although the main findings replicate those yielded by the preregistered difference-score analyses, this approach allows for a more direct examination of how the effect of CS-US directionality varies across US valence. The preregistered analyses, along with other preregistered analyses on participant subsets, are fully reported in the Supplementary Materials, including (a) exclusions based on incorrect memory of the instructions on CS-US directionality, (b) exclusions based on incorrect CS-US contingency memory, and (c) analyses examining the role of instruction believability.

Results

Experiment 1

Self-Reports. Descriptives (means and standard deviations) for all the presented analyses are reported in Table 1. We conducted a 2 (US valence: positive vs negative, within-participants) × 3 (CS-US directionality: CS-causes-US vs US-causes-CS vs control, between-participants) mixed ANOVA. There was a significant main effect of US valence, $F(1, 357) = 1,018.56, p < .001, \eta_p^2 = 0.74$, 95% confidence interval (CI) = [0.70, 0.77], $BF_{10} > 10^5$, and a significant main effect of CS-US directionality, $F(2, 357) = 4.46, p = .012, \eta_p^2 = .02$, 95% CI = [0.002, 0.12], $BF_{10} = 0.034$. The interaction term was significant, $F(2, 357) = 7.59, p < .001, \eta_p^2 = .04$, 95% CI = [0.02, 0.16], $BF_{10} = 965.92$, indicating that the effect of US valence varied across the three directionality conditions (see Figure 2, A panel).

Table 1. Means (and Standard Deviations) of Self-Report and IAT Measures in the Different CS-US Causal Directionality Conditions Across Experiments 1 and 2.

Outcome measure	CS-US directionality		
	CS-causes-US M (SD)	US-causes-CS M (SD)	Control M (SD)
Experiment 1			
Self-reports			
CS_{pos}	3.63 (1.71)	3.11 (2.05)	3.54 (1.36)
CS_{neg}	-3.41 (2.18)	-2.09 (2.71)	-2.62 (2.53)
IAT score	0.43 (0.43)	0.35 (0.45)	0.41 (0.43)
Experiment 2			
Self-reports			
CS_{pos}	3.72 (2.04)	2.79 (2.36)	/
CS_{neg}	-3.22 (2.42)	-1.88 (2.79)	/
IAT score	0.43 (0.43)	0.32 (0.45)	/

Supporting our prediction, the effect of US valence (the EC effect) was stronger in the CS-causes-US than in the US-causes-CS, $t(357) = 3.90, p < .001, d = 0.51$, 95% CI = [0.19, 0.82], $BF_{10} = 88.32$. The comparison between CS-causes-US and control conditions was not significant, $t(357) = 1.87, p = .149, d = 0.24$, 95% CI = [-0.07, 0.55], $BF_{10} = 0.92$, nor was the comparison between US-causes-CS and control conditions, $t(357) = 2.06, p = .100, d = 0.27$, 95% CI = [-0.58, 0.04], $BF_{10} = 0.98$.

We also tested separately the effect of CS-US directionality across the two US types. For the positive US, the main effect of CS-US directionality was significant, $F(2,$

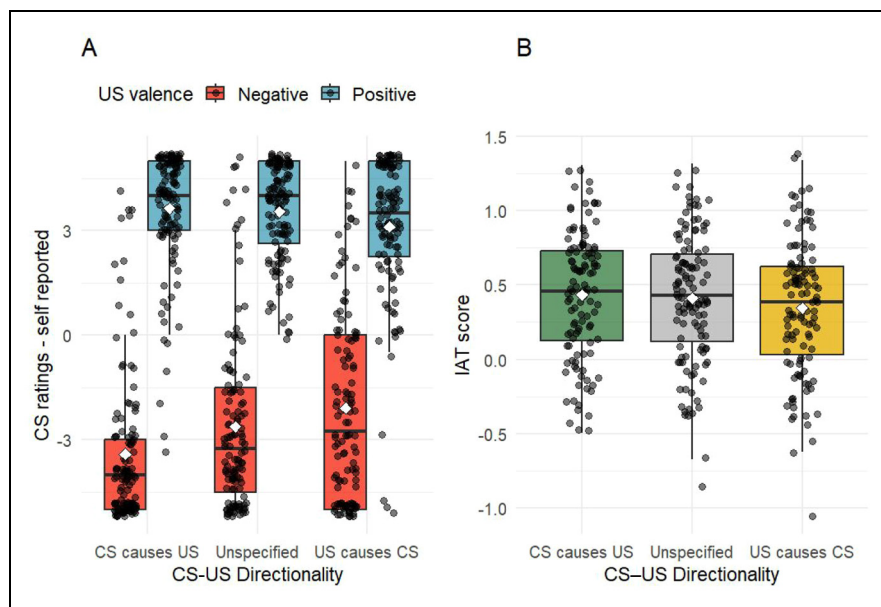


Figure 2. EC Effects Across the Three CS-US Causal Directionality Conditions for Self-Reported Ratings (A Panel) and IAT Scores (B Panel) in Experiment 1
The boxes are the interquartile range, and the middle bars represent the median; the red diamonds represent the mean.

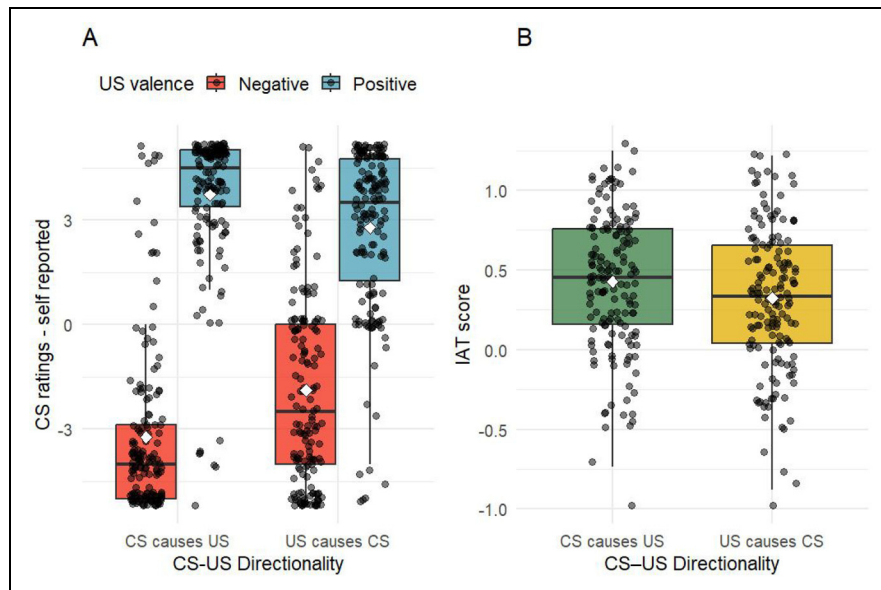


Figure 3. EC Effects Across the Two CS-US Causal Directionality Conditions for Self-Reported Ratings (A Panel) and IAT Scores (B Panel)—Experiment 2. The boxes are the interquartile range, and the middle bars represent the median; the red diamonds represent the mean

357) = 3.14, $p = .044$, $\eta_p^2 = .02$, 95% CI = [0.00, 0.09], $BF_{10} = 0.55$. Pairwise comparisons revealed a significant difference between the CS-causes-US and US-causes-CS conditions, $t(357) = 2.35$, $p = .050$, $d = 0.28$, 95% CI = [0.02, 0.54], $BF_{10} = 1.24$, with CS_{pos} rated as more positive in the former. The comparisons between CS-causes-US and the control condition, and between control and US-causes-CS, were not significant, $t(357) = 0.41$, $p = .901$, $d = 0.06$, 95% CI = [-0.19, 0.31], $BF_{10} = 0.16$ and $t(357) = 1.95$, $p = .127$, $d = 0.25$, 95% CI = [-0.01, 0.51], $BF_{10} = 0.83$, respectively. For the negative US, the main effect of CS-US directionality was significant, $F(2, 357) = 8.48$, $p < .001$, $\eta_p^2 = .045$, 95% CI = [0.02, 0.17], $BF_{10} = 69.33$. Pairwise comparisons revealed a significant difference between the CS-causes-US and US-causes-CS, $t(357) = 4.09$, $p < .001$, $d = 0.54$, 95% CI = [0.28, 0.80], $BF_{10} = 366.69$, with CS_{neg} rated as more negative in the former. In addition, the CS-causes-US group rated the CS_{neg} as more negative than the control group, $t(357) = 2.46$, $p = .038$, $d = 0.33$, 95% CI = [0.08, 0.59], $BF_{10} = 3.18$, whereas comparison between the control and US-causes-CS condition was not significant, $t(357) = 1.67$, $p = .217$, $d = 0.20$, 95% CI = [0.05, 0.46], $BF_{10} = 0.46$.

Implicit Association Test. We conducted a one-way ANOVA, with CS-US directionality as the independent variable. The CS-US causal directionality effect did not emerge, $F(2, 341) = 1.22$, $p = .296$, $\eta_p^2 = .01$, 95% CI = [0.00, 0.03], $BF_{10} = 0.10$ (see Figure 2, B panel). The post hoc test indicated a non-significant difference between CS-causes-US and US-causes-CS conditions, $t(341) = 1.50$, $p = .295$, $d = 0.20$, 95% CI = [-0.12, 0.52], $BF_{10} = 0.41$,

with an only descriptively stronger EC effect in CS-causes-US than those in US-causes-CS (Table 1). The comparison between CS-causes-US and control condition was not significant, $t(341) = 0.37$, $p = .928$, $d = 0.05$, 95% CI = [-0.27, 0.37], $BF_{10} = 0.15$ nor was the comparison between US-causes-CS and control condition, $t(341) = 1.15$, $p = .483$, $d = 0.15$, 95% CI = [-0.47, 0.17], $BF_{10} = 0.26$.

Experiment 2

Self-Reports. The 2 (US valence) \times 2 (CS-US directionality) mixed ANOVA indicated a significant main effect of US valence, $F(1, 318) = 588.26$, $p < .001$, $\eta_p^2 = .65$, 95% CI = [0.59, 0.70], $BF_{10} > 10^5$, and a non-significant main effect of CS-US directionality, $F(1, 318) = 2.58$, $p = .109$, $\eta_p^2 = .01$, 95% CI = [0.00, 0.04], $BF_{10} = 0.11$. The interaction term was significant, $F(1, 318) = 22.61$, $p < .001$, $\eta_p^2 = .07$, 95% CI = [0.02, 0.13], $BF_{10} = 3.08$. Supporting our prediction (see Figure 3, A panel), the effect of US valence (the EC effect) was stronger in the CS-causes-US, $t(158) = 22.43$, $p < .001$, $d = 3.12$, 95% CI = [2.46, 3.78], $BF_{10} > 10^5$, than in the US-causes-CS condition, $t(160) = 12.80$, $p < .001$, $d = 1.81$, 95% CI = [1.36, 2.27], $BF_{10} > 10^5$.

Across both positive and negative US valence, the CS-causes-US condition produced more extreme evaluations (Table 1). For CS_{pos}, scores were significantly more positive in the CS-causes-US than in the US-causes-CS condition, $t(318) = 3.80$, $p < .001$, $d = 0.42$, 95% CI = [0.20, 0.65], $BF_{10} = 110.32$. Similarly, for CS_{neg}, scores were significantly more negative in the CS-causes-US condition compared with the US-causes-CS condition, $t(318) = 4.59$, $p < .001$, $d = 0.51$, 95% CI = [0.29, 0.74], $BF_{10} = 2,333.86$.

Implicit Association Test. We found a significant effect of CS-US causal directionality, $F(1, 308) = 4.34, p = .038, \eta_p^2 = .01, 95\% \text{ CI} = [0.00, 0.03], BF_{10} = 0.99$ (see Figure 3, B panel), with a stronger EC effect in CS-causes-US than in US-causes-CS (Table 1).

Aggregated Analyses (Non-Preregistered). To examine the robustness of the effects, we aggregated data from the two experiments, including only participants assigned to the two directionality conditions ($N = 557$ for self-reports; $N = 534$ for the IAT). For the sake of brevity, we report only the critical effects relevant to our hypotheses here; the full set of analyses is reported in the Supplementary Materials.

Self-reports. A 2 (US valence) $\times 2$ (CS-US directionality) mixed ANOVA revealed a significant interaction between US valence and CS-US directionality, $F(1, 555) = 36.66, p < .001, \eta_p^2 = .06, 95\% \text{ CI} = [0.03, 0.10], BF_{10} > 10^5$. Across US valence, EC effects were stronger when the CS was framed as causing the US than as the effect of the US.

We also tested the difference in the absolute extremity of the CS evaluations across US valence. To place positive-paired and negative-paired CSs on a comparable scale, we reverse-scored the evaluations of CSs paired with negative USs, so that higher values always reflected more extreme (i.e., more strongly valenced) evaluations. This analysis again yielded a significant interaction between US valence and causal directionality, $F(1, 555) = 9.27, p = .002, \eta_p^2 = .02, 95\% \text{ CI} = [0.002, 0.04], BF_{10} = 0.14$. This indicates that causal directionality does not influence evaluative inferences uniformly and provides relevant information concerning the evaluative inferences determined by the interplay between US valence and directionality.

Implicit Association Test. We found a significant effect of CS-US causal directionality, $F(1, 532) = 6.52, p = .011, \eta_p^2 = .01, 95\% \text{ CI} = [0.00, 0.04], BF_{10} = 2.26$, with a stronger EC effect in CS-causes-US than in US-causes-CS.

Discussion

Two experiments examined whether the causal directionality between CS and US influences EC effects. In Experiment 1, where emotional expressions of happiness versus sadness served as USs, stronger EC effects (effects of US valence) emerged in self-reports when the CS was framed as causing (instead of being caused by) the US, but this effect did not generalize to IAT scores. Experiment 2 reduced inferential ambiguity by reframing the USs as physical states and aligning outcomes with health-related judgments. Under these conditions, causal directionality moderated EC effects in both self-report and IAT measures (anecdotal on the latter). Aggregated analyses across experiments confirmed this moderation effect, with

Bayesian analyses revealing that this moderation effect was clear for self-reported ratings while anecdotal for IAT scores. Together, the studies support the idea that causal directionality can shape the strength of EC effects.

We build on a propositional account of EC, which argues that EC arises from inferences about stimulus relations (De Houwer et al., 2020). Our findings refine this view by showing that opposing causal directions between the CS and US moderate EC effects. Moreover, aggregated analyses showed a stronger impact of CS-US directionality for the CSs paired with negative USs, compared to CSs paired with positive USs. We propose that negative inferences about CS_{neg} were facilitated and less ambiguous when the US was the effect of the CS (e.g., “a food that makes someone feel sick must be unhealthy”), whereas for CS_{pos} both causal directions supported similar positive inferences, despite a small but significant effect of directionality.

Moreover, comparing the directional conditions to the control group in Experiment 1 indicated that for positively paired CSs, neither directional condition differed from control, whereas for negatively paired CSs, only the CS-causes-US condition led to more extreme negative ratings. These findings suggest that with positive USs, any interpretation of the unspecified pairing generates inferences (e.g., eating the food caused positive emotions; positive emotions caused eating the food) that lead to similar evaluative conclusions (e.g., the food is good). With negative USs, however, unspecified pairing might produce ambiguous inferences similar to those triggered by the US-causes-CS condition (e.g., negative emotions caused eating the food, thus this food might not be necessarily negative). In contrast, inferences in the CS-causes-US condition should be less equivocal (e.g., eating the food caused negative emotions, thus the food is negative). Future work could directly assess the specific inferences formed under different causal framings to validate this explanation.




Identifying causal directionality as a moderator of EC opens new research avenues. Whereas our effects relied on explicit instructions about CS-US causality, future studies could investigate whether similar moderation effects emerge from subtler cues. For example, causal directionality could be implied through brief scenarios presented before conditioning, or through simple non-verbal signals (e.g., arrows indicating CS→US vs. US→CS) displayed during the task. Likewise, structural properties of CS-US pairings, like temporal proximity (Gast et al., 2016), temporal order, spatial proximity, or contingency patterns may shape spontaneously inferred causality and thus influence EC strength and direction. Employing such indirect manipulations would clarify whether causal directionality modulates evaluative learning without explicit instruction and would also reduce potential concerns about demand characteristics (Corneille & Lush, 2023).

Our findings link EC research with broader work on learning and social cognition. Conditioning principles have been increasingly applied to impression formation and

person perception research (e.g., Hughes et al., 2024; Mattavelli, Fiamberti, et al., 2023; Mattavelli, Masi, & Brambilla, 2023; Rougier et al., 2023). For example, Mattavelli, Fiamberti, et al. (2023) showed that judgments of facial trustworthiness depend on threat-relevant contextual information (background picture), and this effect is moderated by the nature of the emotion displayed by the face in the context: happy targets in threatening contexts were judged less trustworthy than fearful ones, especially when observers inferred the target caused rather than suffered from the threat. Hence, trustworthiness is shaped not only by contextual threat but also by how observers construe causality between them in the situation. Highlighting causal directionality as a shared moderator, our work creates conceptual links between phenomena that have traditionally been studied in relative isolation (De Houwer et al., 2019).

In sum, this research demonstrates that causal directionality influences EC effects. When forming likes and dislikes, individuals might consider not just whether two stimuli co-occur, but how they are causally related. These findings refine propositional EC accounts by showing that inferences about causal relations shape evaluative outcomes, and they underscore the symbolic value of pairings, revealing a conceptual bridge between EC and broader psychological phenomena.

ORCID iDs

Simone Mattavelli  <https://orcid.org/0000-0002-8934-8016>
 Marine Rougier  <https://orcid.org/0000-0002-9467-2726>
 Tal Moran  <https://orcid.org/0000-0002-4681-0725>

Funding

The authors disclosed receipt of the following financial support for the research, authorship, and/or publication of this article: This research was supported by the Israel Science Foundation (ISF) grant #870-23 and the Open University of Israel research grant #41454 to Tal Moran

Declaration of Conflicting Interests

The authors declared no potential conflicts of interest with respect to the research, authorship, and/or publication of this article.

Supplemental Material

Supplemental material is available in the online version of the article.

Note

1. In Experiment 1, after the third round of data collection ($N = 270$), we discovered a programming error in one condition. Consequently, we excluded the 46 participants from that group. After correcting the error, we collected data from 136 additional participants (instead of the planned

90), yielding a final sample of $N = 360$. We also adjusted the assignment probabilities to increase representation (0.45) in the critical group. We reported this deviation in the preregistration protocol.

References

- Allport, G. W. (1935). Attitudes. In C. A. Murchison (Ed.), *A Handbook of Social Psychology* (pp. 798–844). Clark University Press.
- Bading, K. C., Barth, M., & Rothermund, K. (2025). Evidence for an evaluative effect of stimulus co-occurrence may be inflated by evaluative differences between assimilative and contrastive relations. *Cognition and Emotion*, *39*, 1995–2017. <https://doi.org/10.1080/02699931.2025.2460099>
- Baeyens, F., Eelen, P., Crombez, G., & Van den Bergh, O. (1992). Human evaluative conditioning: Acquisition trials, presentation schedule, evaluative style and contingency awareness. *Behaviour Research and Therapy*, *30*(2), 133–142. [https://doi.org/10.1016/0005-7967\(92\)90136-5](https://doi.org/10.1016/0005-7967(92)90136-5)
- Champely, S. (2017). *pwr: Basic functions for power analysis* (Version 1.2.1) [Computer software]. <https://CRAN.R-project.org/package=pwr>
- Corneille, O., & Lush, P. (2023). Sixty years after Orne's American psychologist article: A conceptual framework for subjective experiences elicited by demand characteristics. *Personality and Social Psychology Review*, *27*(1), 83–101. <https://doi.org/10.1177/10888683221104368>
- 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. (2018). Propositional models of evaluative conditioning. *Social Psychological Bulletin*, *13*(3), 1–21. <https://doi.org/10.5964/spb.v13i3.28046>
- De Houwer, J., & Hughes, S. (2016). Evaluative conditioning as a symbolic phenomenon: On the relation between evaluative conditioning, evaluative conditioning via instructions, and persuasion. *Social Cognition*, *34*(5), 480–494. <https://doi.org/10.1521/soco.2016.34.5.480>
- 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>
- Fiedler, K., & Unkelbach, C. (2011). Evaluative conditioning depends on higher order encoding processes. *Cognition and Emotion*, *25*(4), 639–656. <https://doi.org/10.1080/02699931.2010.513497>
- Förderer, S., & Unkelbach, C. (2012). Hating the cute kitten or loving the aggressive pit-bull: EC effects depend on CS-US relations. *Cognition and Emotion*, *26*(3), 534–540. <https://doi.org/10.1080/02699931.2011.588687>
- Gast, A., Langer, S., & Sengewald, M. A. (2016). Evaluative conditioning increases with temporal contiguity. The influence of stimulus order and stimulus interval on evaluative

- conditioning. *Acta Psychologica*, 170, 177–185. <https://doi.org/10.1016/j.actpsy.2016.07.002>
- Greenwald, A. G., McGhee, D. E., & Schwartz, J. L. (1998). Measuring individual differences in implicit cognition: The implicit association test. *Journal of Personality and Social Psychology*, 74(6), 1464–1480. <https://doi.org/10.1037/0022-3514.74.6.1464>
- Greenwald, A. G., Nosek, B. A., & Banaji, M. R. (2003). Understanding and using the implicit association test: I. An improved scoring algorithm. *Journal of Personality and Social Psychology*, 85(2), 197–216. <https://doi.org/10.1037/0022-3514.85.2.197>
- Hofmann, W., De Houwer, J., Perugini, M., Baeyens, F., & Crombez, G. (2010). Evaluative conditioning in humans: A meta-analysis. *Psychological Bulletin*, 136(3), 390–421. <https://doi.org/10.1037/a0018916>
- Hu, X., Gawronski, B., & Balas, R. (2017). Propositional versus dual-process accounts of evaluative conditioning: I. The effects of co-occurrence and relational information on implicit and explicit evaluations. *Personality and Social Psychology Bulletin*, 43(1), 17–32. <https://doi.org/10.1177/0146167216673351>
- Hughes, S., Unkelbach, C., Rougier, M., & De Houwer, J. (2024). The shared feature principle in person perception: Shared features lead to assumptions about other features. *Collabra: Psychology*, 10(1), 117780. <https://doi.org/10.1525/collabra.117780>
- Hughes, S., Ye, Y., Van Dessel, P., & De Houwer, J. (2019). When people co-occur with good or bad events: Graded effects of relational qualifiers on evaluative conditioning. *Personality and Social Psychology Bulletin*, 45(2), 196–208. <https://doi.org/10.1177/0146167218781340>
- Kurdi, B., & Banaji, M. R. (2019). Attitude change via repeated evaluative pairings versus evaluative statements: Shared and unique features. *Journal of Personality and Social Psychology*, 116(5), 681–703. <https://doi.org/10.1037/pspa0000151>
- Levey, A. B., & Martin, I. (1975). Classical conditioning of human “evaluative” responses. *Behaviour Research and Therapy*, 13(4), 221–226. [https://doi.org/10.1016/0005-7967\(75\)90026-1](https://doi.org/10.1016/0005-7967(75)90026-1)
- Lundqvist, D., Flykt, A., & Öhman, A. (1998). *The Karolinska Directed Emotional Faces (KDEF) [CD-ROM]*. Department of Clinical Neuroscience, Psychology Section, Karolinska Institutet.
- Mattavelli, S., Fiamberti, G. C., Masi, M., & Brambilla, M. (2023). The “Happy Face Killer” in the eyes of the beholder: Relational encoding of facial emotions in context influences trustworthiness attributions. *Journal of Experimental Social Psychology*, 109, 104517. <https://doi.org/10.1016/j.jesp.2023.104517>
- Mattavelli, S., Masi, M., & Brambilla, M. (2023). Not just about faces in context: Face–context relation moderates the impact of contextual threat on facial trustworthiness. *Personality and Social Psychology Bulletin*, 49(3), 376–390. <https://doi.org/10.1177/01461672211065933>
- Mitchell, C. J., De Houwer, J., & Lovibond, P. F. (2009). The propositional nature of human associative learning. *Behavioral and Brain Sciences*, 32(2), 183–198. <https://doi.org/10.1017/S0140525X09000855>
- Moran, T., & Bar-Anan, Y. (2013). The effect of object–valence relations on automatic evaluation. *Cognition & Emotion*, 27(4), 743–752. <https://doi.org/10.1080/02699931.2012.732040>
- Moran, T., Bar-Anan, Y., & Nosek, B. A. (2016). The assimilative effect of co-occurrence on evaluation above and beyond the effect of relational qualifiers. *Social Cognition*, 34(5), 435–461. <https://doi.org/10.1521/soco.2016.34.5.435>
- Moran, T., Nudler, Y., & Bar-Anan, Y. (2023). Evaluative conditioning: Past, present, and future. *Annual Review of Psychology*, 74(1), 245–269. <https://doi.org/10.1146/annurev-psych-032420-031815>
- Nosek, B. A., Greenwald, A. G., & Banaji, M. R. (2005). Understanding and using the Implicit Association Test: II. Method variables and construct validity. *Personality and Social Psychology Bulletin*, 31(2), 166–180. <https://doi.org/10.1177/0146167204271418>
- Pleyers, G., Corneille, O., Luminet, O., & Yzerbyt, V. (2007). Aware and (dis) liking: Item-based analyses reveal that valence acquisition via evaluative conditioning emerges only when there is contingency awareness. *Journal of Experimental Psychology: Learning, Memory, and Cognition*, 33(1), 130–144. <https://doi.org/10.1037/0278-7393.33.1.130>
- 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), 84560. <https://doi.org/10.1525/collabra.84560>
- Schönbrodt, F. D., & Wagenmakers, E. J. (2018). Bayes factor design analysis: Planning for compelling evidence. *Psychonomic Bulletin & Review*, 25(1), 128–142. <https://doi.org/10.3758/s13423-017-1230-y>
- Slooman, S. A., & Lagnado, D. (2015). Causality in thought. *Annual Review of Psychology*, 66(1), 223–247. <https://doi.org/10.1146/annurev-psych-010814-015135>
- 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>
- Waldmann, M. R., & Holyoak, K. J. (1992). Predictive and diagnostic learning within causal models: Asymmetries in cue competition. *Journal of Experimental Psychology: General*, 121(2), 222–236. <https://doi.org/10.1037/0096-3445.121.2.222>
- Zanon, R., De Houwer, J., Gast, A., & Smith, C. T. (2014). When does relational information influence evaluative conditioning? *Quarterly Journal of Experimental Psychology*, 67(11), 2105–2122. <https://doi.org/10.1080/17470218.2014.907324>

Author Biographies

Simone Mattavelli is an assistant professor at the University of Milano-Bicocca. His research investigates how beliefs and context shape judgments, preferences, and trust. He focuses on phenomena like evaluative conditioning, face perception, and the illusory truth effect.

Marine Rougier is a postdoctoral researcher at Ghent University. She studies spontaneous preferences and how learning processes shape approach–avoidance behavior. She also explores how beliefs about traits influence learning within the feature transformation effect framework.

Tal Moran is an associate professor at the Open University of Israel. Her research focuses on automatic social behavior and cognition, particularly attitudes and stereotypes. She also investigates emotion regulation, emphasizing the role of construal level in managing different emotions.