Beyond valence transfer in an evaluative conditioning paradigm: On the nature of the phenomenon and its relation to personality

Florina Gabriela Huzoaica*[1], Jan De Houwer[2], Marco Perugini[3], Andrei Rusu[1], and Florin Alin Sava[1]

[1] Department of Psychology, West University of Timișoara, Romania

[2] Department of Experimental Clinical and Health Psychology, Ghent University, Belgium

[3] Department of Psychology, University of Milano-Bicocca, Italy

Author Note

Florina Gabriela Huzoaica https://orcid.org/0009-0005-1977-4895

Jan De Houwer https://orcid.org/0000-0003-0488-5224

Marco Perugini https://orcid.org/0000-0002-4864-6623

Andrei Rusu https://orcid.org/0000-0003-0036-4149

Florin Alin Sava https://orcid.org/0000-0001-8898-1306

*Correspondence should be addressed to Huzoaica Florina-Gabriela, Department of Psychology, West University of Timișoara, 4 Vasile Pârvan Boulevard, Timișoara 300223. Email: huzoaica.florina@e-uvt.ro.

Beyond valence transfer in an evaluative conditioning paradigm: On the nature of the phenomenon and its relation to personality

Date of submission: 28.03.2025

Word Count: 15.570

Abstract

Inspired by previous work on the relation between evaluative conditioning (EC) and impression formation (e.g., halo effects), we investigated whether the pairing of a neutral conditioned stimulus (CS) and a valenced unconditioned stimulus (US) changes not only the valence of the CS but also judgments about other CS features (e.g., *trustworthiness*). Additionally, we assessed whether individual differences in neuroticism and agreeableness moderate these effects. In two pre-registered experiments (N = 289 for Experiment 1 and N = 211 for Experiment 2), we found an impact of CS-US pairings not only on ratings of the valence of CSs but also on ratings of other CS features. The evidence for the moderating role of personality at the dimension levels (e.g., neuroticism) was mixed. Theoretical and practical implications of the results, alongside future research directions, are discussed.

Keywords: evaluative conditioning, halo effect, neuroticism, agreeableness, impression formation

We frequently lack information regarding the objects we encounter in our environment. For example, when meeting a new colleague, we often have little solid ground for forming an accurate initial impression, making it difficult to determine our expectations, thoughts, and behaviour. Nonetheless, people often make wide-ranging inferences about the properties of objects in the environment that go well beyond the available social cues (Kunda & Thagard, 1996; Chua & Freeman, 2022; Uleman & Saribay, 2012), inferring features from the limited information they have on a particular object. These inferences can be based on stereotypes (Chen et al., 2022; Hamilton et al., 1994), specific heuristic beliefs (e.g., “beautiful is good”; Lorenzo et al., 2010), or a dichotomous view expressed in terms of *like* vs. *dislike* (i.e., “if you like kittens, you dislike pit-bulls”; Förderer & Unkelbach, 2012).

There are a plethora of empirical studies suggesting that people tend to generalise the initial information towards other aspects such as physical features (e.g., attractiveness, Batres & Shiramizu, 2022; Lorenzo et al., 2010; Han et al., 2018), social features (e.g., competence and warmth halo, Brannon et al., 2017; Ashton-James et al., 2019), personality traits (e.g., extraversion, Srivastava et al., 2010) or behaviours (e.g., body rocking, Mor & Berkson, 2003). A particularly interesting phenomenon that encompasses this tendency is the halo effect (e.g., Zanna & Hamilton, 1977). Relevant to impression formation and attitude change, the halo effect was studied in the context of several topics, including impression formation (Forgas & Laham, 2017), nutrition (Richetin et al., 2021; Demartini et al., 2018), health (Burton et al., 2015), and tourism (Nicolau et al., 2020). For example, one classical study (Dion et al., 1972) examined the attractiveness halo effect, or as they called it, the "what is beautiful is good" effect. They found that physically attractive individuals were consistently rated higher on positive personality traits such as intelligence, kindness, and honesty compared to less attractive individuals. This line of

research showed that the halo effect involves biased inferences based on a specific feature of the object that changes the overall impression of that particular object (Nisbett & Wilson, 1977).

Another phenomenon that is relevant in this context is evaluative conditioning (EC). It typically involves a transfer of valence from an unconditioned stimulus (US) that already possesses a specific valence to a neutral conditioned stimulus (CS) solely as a result of the repeated pairing of the two stimuli (De Houwer, 2007). EC shows that people make assumptions about the evaluative properties of the CS merely based on other stimuli it co-occurs with, being a robust learning effect that has been studied extensively (Baeyens et al., 1998; Kurdi & Banaji, 2017; De Houwer et al., 2001; Sava et al., 2020; Walther et al., 2011, see Hofmann et al., 2010, and Moran et al., 2023, for reviews). Attribute conditioning (AC) is also relevant in this context. Like EC, it involves the pairing of a CS and a US, but it focuses on changes in specific traits rather than general valence. For instance, when a novel person (CS) is paired with the picture of an athlete (US), the novel person is perceived as more athletic (e.g., Förderer & Unkelbach, 2015).

In this paper, we examine whether a simple pairing of two social stimuli will result not only in a feature transfer (i.e., the valence of the CS) but also in feature transformation, that is, changes in the judgments about other features of the CS (i.e., whether the CS is trustworthy). We also examine whether this phenomenon is moderated by personality traits (e.g., agreeableness). As such, our research can shed new light on the way people form impressions about other individuals based on limited information.

We first introduce the feature transformation framework (De Houwer et al., 2019) and discuss why this framework is relevant for impression formation and evaluative conditioning studies. Finally, we explain how the current studies can add to the literature on evaluative

conditioning and impression formation, as well as why personality traits might be relevant in this context.

The Feature Transformation Framework in Person Perception

The feature transformation framework (De Houwer et al., 2019) includes, as working concepts, source and target objects, source and target features, and feature transformation. Feature transformation occurs when a feature of a source object influences the assumptions made about a feature of a target object. These concepts can be applied to various impression formation phenomena, including the halo effect, EC, and AC.

In terms of the feature transformation framework, in studies on the halo effect, the source and target objects are the same (i.e., the same person), while the source feature (e.g., attractiveness) and target feature (e.g., intelligence) differ. In this context, feature transformation occurs when people make assumptions about the target feature based on the source feature (e.g., attractiveness results in higher ratings for intelligence; Batres & Shiramizu, 2022; Han et al., 2018).

The EC and AC effects, on the other hand, entail a different source object (e.g., a well-known and liked person called Bob) and a target object (e.g., a novel and hence neutral person called Joe) that are presented in the same space and time (e.g., Bob and Joe often hang together). These effects occur when the features of the source object (e.g., the liking of Bob) give rise to assumptions about the same feature of the target object (e.g., the liking of Joe).

In a recent paper, Hughes and collaborators (2020) differentiated between feature transfer and feature transformation. Feature transfer implies that the source and target object features are the same (e.g., valence) and that the feature of the target object changes in line with the value of the feature of the source object (i.e., the target object becomes positive if the source object is

positive). Feature transformation, on the other hand, entails that the source and target object features are different (e.g., inferring competence from attractiveness).[1] Based on this terminology, the standard EC and AC effects would be instances of feature transfer, and the halo effect would involve feature transformation.

Interestingly, there are no a priori reasons why EC or AC studies should be limited to assessing feature transfer. In principle, just like in halo studies, researchers could assess multiple target features in EC and AC studies whereas until now, they assessed only the target feature that was identical to the source feature (i.e., valence in EC and a specific attribute such as athleticism in AC). If researchers would simply measure multiple target features, they might observe an impact of one source feature (e.g., valence or athleticism) on other target features. Note that the feature transformation framework only highlights this caveat in the literature but does not provide a theory for predicting whether or explaining how other target features would be influenced. Nevertheless, the framework has already inspired novel research.

First, Rougier et al. (2023) examined a new phenomenon called the pairing-based halo effect, placed at the crossroads of conditioning and halo effects. They examined it in the context of person perception (e.g., attractiveness halo) and nutrition (e.g., health halo). In one study, for instance, they used attractiveness as a source feature and examined if the pairing of a target person with a (un)attractive source person influenced the ratings of the target person on a variety of traits. As in the halo effect, they indeed observed changes in multiple traits, in line with the attractiveness stereotype.

[1] Feature transformation also entails situations in which source and target features are identical but the change in the target feature is not in the direction of the value of the source feature (i.e., the target object becomes negative if the source object is positive). Because this case is not relevant for the present paper, we will only consider feature transformation that involves different source and target features.

Second, Rougier and De Houwer (2023) found similar effects using valence as the source feature in a reverse correlation task. The first group of participants underwent a typical EC procedure in which neutral CSs (faces of unknown persons) were paired with positive or negative USs (visual scenes other than faces). Then, the participants saw pairs of pictures, each showing one of the CS pictures, but random visual noise was added to the picture. Their task was to select from each pair the picture that most resembled the CS picture presented during the EC procedure. Based on these responses, a classification image was constructed for each CS by calculating the visual average of the selected pictures. These classification images looked like blurry versions of the CS pictures. A second group of participants then rated the classification images not only on valence but also on several personality traits. Rougier and De Houwer (2023) observed a strong feature transfer effect: the classification image of the CS that was paired with positive USs was liked more than the classification image of the CS that was paired with negative USs. However, they also observed clear feature transformation effects: the classification image of the CS that was paired with positive USs was rated higher compared to the other CS on warmth-related personality traits.

Importantly, Rougier and De Houwer (2023) also examined a possible explanation for feature transformation. More specifically, they tested whether the changes in personality ratings went beyond changes in valence. In principle, changes in personality traits might be instances of the so-called “global impression effect” (Nisbett & Wilson, 1977; also see Förderer & Unkelbach, 2011, for a discussion of this effect in the context of AC). For instance, participants might rate a CS that was paired with a positive US not only as more positive but also as having a warmer personality because the increase in positivity of the CS results in a global positive impression which in turn results in higher ratings on all positive traits, including warmth.

Rougier and De Houwer (2023) found evidence for this explanation but also found two indications that feature transformation went beyond a global impression effect: feature transformation effects (1) did not occur for all positive traits to an equal degree and (2) were still present after statistically controlling for the valence of the traits.

The Role of Personality in Person Perception

It is widely accepted that a perceiver is not a blank canvas (Uleman & Saribay, 2012). Impressions are formed by people who all have their own personalities. It is, therefore, conceivable that, as for many other things, individual differences in personality dimensions can have an impact on impression formation.

In selecting personality traits that could moderate EC and feature transformation more generally, we relied on the existing literature linking EC and impression formation to personality (see De Houwer et al., 2023). For instance, in a study of EC, Vogel et al. (2019) concluded that neuroticism and agreeableness were moderators of the EC effect. Subsequent studies provided more evidence for the correlation with agreeableness (Ingendahl & Vogel, 2023) and revealed a more complex picture with regard to the correlation with neuroticism (e.g., Bunghez et al., 2023; Casini et al., 2023). In the impression formation literature, agreeableness is one of the most evaluative personality traits within the Big Five taxonomy (Rau et al., 2021) that organises information about others and is considered a central dimension of interpersonal judgment (Ames & Bianchi, 2008). Likewise, neuroticism was associated with a negative shift in evaluations in the context of impression formation (Uziel, 2015), and it also presents a strong link to psychopathology (Krueger, 1999; Lakdawalla & Hankin, 2008; Kercher et al., 2009; Kotov et al., 2010).

The Current Research

Building on previous research, we conducted two experiments that studied feature transformation in a standard EC procedure, starting from a pilot study. Although the results of Rougier and colleagues (Rougier et al., 2023; Rougier & De Houwer, 2023) revealed feature transformation effects, they used procedures that differed from standard EC paradigms. On the one hand, Rougier et al. (2023) manipulated attractiveness rather than valence. On the other hand, Rougier and De Houwer (2023) assessed the likeability and the personality traits of the CSs indirectly (i.e., using a reversed correlation task) rather than directly (i.e., ratings given by the participants who experienced the EC procedure). In contrast, we used a standard EC procedure in which we manipulated US valence and measured both CS liking and CS personality traits using ratings by the participants who saw the CS-US pairings. In line with Rougier and De Houwer (2023), we tested whether possible feature transformation effects are due to changes in valence and whether they go beyond the effect of mere valence.

As our first hypothesis (H1a), we expected that CS-US pairings would influence not only the likeability ratings (i.e., feature transfer) but also the ratings of other (personality and social) features of the CS (i.e., feature transformation). We also expected that the likeability ratings would mediate the ratings of the other features (H1b). We also looked for indications of whether feature transformation involved more than just changes in likeability.

Our studies also go beyond previous research by examining the role of the participant's personality. Based on the results previously obtained (Vogel et al., 2019; Bunghez et al., 2023; Casini et al., 2023; Ingendahl & Vogel, 2022, 2023), we examined whether neuroticism and agreeableness moderate the effect of CS-US pairings on liking (i.e., feature transfer) (H2a), the effect of CS-US pairings on other features (i.e., feature transformation) (H2b), and the extent to

which the effect of CS-US pairings on other features is mediated by the effect of pairings on liking (H2c).

Ethics and Open Science

All studies were approved by the Scientific Council of Research and Creation at the West University of Timisoara regarding compliance with ethical aspects in scientific research (No. 57994/11.11.2020, No. 65585/22.11.2021, respectively No. 84392/09.11.2023). The studies were pre-registered on the Open Science Framework (OSF) website and are available at: https://osf.io/nk68c/?view_only=afbe0991a9df46db8acc32a1ea93769a (pilot study), https://osf.io/tp8ha/?view_only=b5337b26c800452f8408ba612f298590 (experiment 1), https://osf.io/bfah8/?view_only=6f47583b96994b868679254688e39287 (experiment 2). While the experiments adhered to the preregistered plan, we acknowledge minor deviations in the planned sample size and statistical analysis. All deviations are thoroughly documented on the OSF platform within the Pre-registration deviations files for each experiment.

Experiment 1 [2]

Method

Sample Size Strategy

We targeted a good statistical power ($1 - \beta = 0.80$) at an alpha level of 0.05 to detect a mediation effect. Starting from the correlation coefficients of Condition, Likeability ratings and overall Features score, obtained in the pilot study, the Monte Carlo Power Analysis for Indirect Effects (Schoemann et al., 2017) for the first hypothesis (H1b) indicated that we required 90

[2] Initially, we conducted a pilot study to explore whether the transfer of valence (positive vs. negative) predicts the subsequent evaluations on specific features and to examine if neuroticism or agreeableness moderates this transfer. Because the study was underpowered and the results, therefore, provided little information regarding the role of personality, we decided to explore these relations further in the next two experiments. More information about the pilot study can be found in the first section of the Supplemental Materials.

participants. Note that the within-subject manipulation of the US valence in our study introduces a dependency in the data, which is not accounted for by this approach, as it assumes independent observations. However, Montoya (2023) has shown that in most scenarios, given the same parameters, a within-subject mediation requires half the sample size of the equivalent between-subjects mediation. Consequently, we can be confident that the experiment was sufficiently powered to detect the mediation effect, given that the sample size for the analyses was 237 participants (see below).

For the moderation analyses (H2), we conducted an a priori power analysis for linear multiple regression with G*Power (Faul et al., 2009), which indicated that we needed 263 participants to detect a small to medium effect size (Cohen's $f^2 = 0.05$) with sufficient statistical power ($1 - \beta = 0.80$). Therefore, considering the results of the power analyses and the established exclusion criteria, we set our target sample size at 285 participants. We recognize, however, that interactions with personality traits may involve smaller effect sizes, particularly for ordinal interactions where the primary effect is attenuated rather than reversed (Sommet et al., 2023). To further explore this issue, we conducted power analysis simulations, which underlined that we might not have had sufficient power to detect very small effects (see Section 12 of Supplemental Materials).

Participants

Participants ($N = 289$; 99 female, 190 male, $M_{age} = 40$ years, $SD_{age} = 10$) were recruited through Daedalus (https://ro.daedalusonline.eu/), an online recruiting platform.

Based on the preregistered analysis plan, we excluded from the analyses data collected from participants who did not complete all the experiment phases and had more than 50%

incorrect answers (more than 2 wrong answers out of 4 answers) at the valence awareness task. Thus, we ended up with 237 participants (84 female, 153 male, M_{age} = 39.39 years, SD_{age} = 9.74).

Materials

Personality assessment. We selected 12 items from the Big Five Inventory-2 (BFI-2; Soto & John, 2017) to assess neuroticism (the Negative Emotionality scale) and another 12 items to measure agreeableness (the Agreeableness scale). Negative Emotionality and Agreeableness were measured on a 5-point Likert scale (1 = *strongly disagree* to 5 = *strongly agree*). Scores were computed for Negative Emotionality (α =.83) and its three underlying facets (Anxiety, Depression, and Emotional Volatility), and for Agreeableness (α =.72) and its three related facets (Compassion, Respectfulness, and Trust). Overall, at the level of dimensions, both scales had a good internal consistency.

To capture neuroticism and agreeableness from a psychobiological perspective, we also selected two additional scales from the Zuckerman-Kuhlman Personality Questionnaire (ZKPQ; Zuckerman, 2002): Neuroticism-Anxiety scale (19 items) and the Aggression-Hostility scale (17 items). The response options were presented in a dichotomic style (*True/False*). These scales also presented a good internal consistency (α = .90 and α = .71, respectively). Section 4 of Supplemental Materials provides the means, standard deviations, and reliability scores for each scale and its factors.

Stimuli. In the EC procedure, we used 4 neutral faces (all female) retrieved from The Karolinska Directed Emotional Faces (Goeleven et al., 2008) that served as CSs. As for USs, we used 20 color images (10 positive and 10 negative) extracted from the International Affective Picture System IAPS (Bradley & Lang, 2007). More details regarding the stimuli are presented in Section 2 of Supplemental Materials.

Procedure

In the first part of the experiment, participants were briefed on the study overview, signed the informed consent, and provided demographic data (age, gender, education level). Following this, they completed the personality questionnaires.

The second part of the experiment started with a pre-rating phase of CSs, followed by the classical EC procedure where the CS-US pairs were presented in trials. During the pre-rating phase, each participant evaluated 4 images, which afterwards served as CSs, alongside 8 filler images, using a scale from -3 to 3, where -3 means *Very unpleasant*, and 3 means *Very pleasant*.

In the acquisition phase, participants received two blocks of 20 trials each, resulting in 40 trials. In between blocks, we asked the participant to pay attention to the task. In both blocks, the same two CSs were always been paired with different USs of positive valence, and the other CSs were always paired with different USs of negative valence. Each trial contained two images presented simultaneously onscreen: the neutral face (CS) on the left side of the screen, and the valenced image (positive or negative US) appeared on the right side of the screen. Each CS-US pair was presented for 2500ms, with 1000ms intertrial interval, and appeared randomly 5 times in the block. Across trials, a CS was paired with different USs of the same valence. On each trial, the assignment of a CS to a US of a particular valence was determined randomly.

In the last part of the experiment, the participants were asked to rate the four CSs in terms of likeability and five specific features (*friendly*, *trustworthy*, *strong*, *calm,* and *humble*). These features were selected to reflect aspects of social judgement (Scheider et al., 2022; Warmth – *friendliness*; Morality - *trustworthiness*) and personality traits (Romano et al., 2023; Emotionality - *strength*; Agreeableness – *calm*; Honest-Humility – *humbleness*). The CSs appeared onscreen in a randomized order. First, the participants received the following

instruction: “In the next phase, on the computer's screen, you will see the previously presented pictures of human faces. Your task will be to assess each picture that appears onscreen by selecting the value that best expresses your opinion.” The CSs’likeability was evaluated by using a -3 to +3 rating scale with *Pleasant* vs. *Unpleasant* as labels (-3 = *Very unpleasant* to 3 = *Very pleasant*). Secondly, to evaluate the other features, we gave participants the following instructions: “In the next phase, on the computer's screen, you will see the same pictures of human faces. Your task will be to assess each picture that appears onscreen by indicating the impression that each individual left, based on the following set of characteristics.” As for the likeability ratings, a scale from -3 to 3 was used for all other features, using the following labels: Unfriendly – Friendly; Untrustworthy – Trustworthy; Strong – Vulnerable; Tense – Calm; Arrogant – Humble. Each feature pair was displayed individually onscreen. Both CS stimuli and features were randomised.

Further, the participants responded to exploratory questions (valence awareness, demand compliance, study objective). The four CSs were presented again onscreen to assess the participant's valence awareness in the evaluative learning sequence. Under each image, a question was displayed as follows: “With what kind of pictures was this image paired during the first task?” with three possible answers (Negatively valenced pictures, Positively valenced pictures, I don't know).

Next, each participant received the following question to measure demand compliance: “Previously, when you evaluated each individual, on what did you base your answers?” with four possible answers (“I answered based on what I thought the researcher expected from me”; “I answered based on what I learned about the stimuli in the previous task”; “I answered based on what I felt towards the stimuli”; “I don’t know why I answered like that”).

Finally, the participants were asked about the study's objective, debriefed and thanked for their participation.

Design. This experiment had a within-subjects design with *US valence* (positive vs. negative) as a factor. Other method factors varied between participants: *stimulus assignment* (CS1/CS2/CS3/CS4 identity assigned to the positive/negative images), *US identity* (Set 1 vs. Set 2)[3] and *block assignment* (6 block combinations). The EC trials were randomly presented in each block.

Analytic strategy. As the primary dependent variables, we used self-reported evaluations of likeability (how pleasant or unpleasant the target is) provided by the participants and the explicit ratings of the five evaluated features of the CSs (*friendly, trustworthy, strong, calm,* and *humble).* As the moderator variables, we introduced in the analyses the scores obtained by the participants on each personality scale: Negative Emotionality and Agreeableness (BFI-2; Soto & John, 2017), Neuroticism-Anxiety, and Aggression-Hostility (ZKPQ; Zuckerman, 2002).

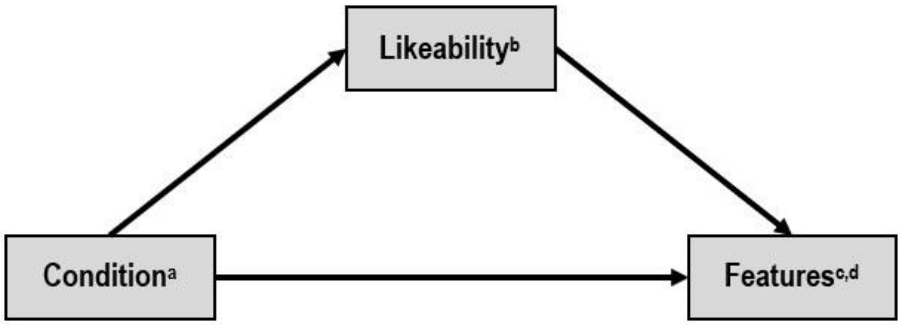
We first conducted a paired-sample *t*-test to check if the CS-US pairings had an effect on the likeability ratings of the CSs (feature transfer). We also conducted a mixed 2 x 5 ANOVA to test whether the impact of pairings was significant for each of the other features (feature transformation). This ANOVA was followed up with mediation analyses that allowed us to test whether the CS-US pairings influenced the ratings for the other features through the changes in liking. More specifically, we employed mixed models using the *lme4* package version 1.1-35.4 (Bates et al., 2015) and we used the *lmerTest* package version 3.1-3 (Kuznetsova et al., 2017) to provide the p-values. To visualise the interaction effects for the outcome variables, we used the *interactions* package (Bauer & Curran, 2005). The Causal Mediation approach developed by

[3] We created two sets of images used as USs, so that each CS would always be paired with a different US. In both sets of images, we had 5 images rated as negative and 5 images rated as positive.

Tingley et al. (2014) was used to test the mediation effect of the Condition (USpos for CSs paired with positive USs and USneg for CSs paired with negative USs) through general changes in liking (EC) on each outcome (overall CS features, and each specific CS features). Following the specifications for this type of analysis, we built two sets of multilevel models to test the mediating effect of changes in liking. In the first set of models, we used condition as a fixed effect and a random intercept for subjects to predict the mediator. In the second set of models, the mediator was also included in the analysis as a fixed effect predictor, alongside the condition. Finally, we performed the mediation analysis using the mediate function from the *mediation* package version 4.5.0 (Tingley et al., 2014) testing the Average Causal Mediation Effect (ACME), Average Direct Effect (ADE), total effect, and the proportion mediated by the mediator. Additionally, we used package *lmerTest* version 3.1-3 to obtain p values for the fixed effects (Kuznetsova et al., 2017). A visual depiction of the tested mediation models can be found in Figure 1.

Figure 1

Visual representation of the mediation model

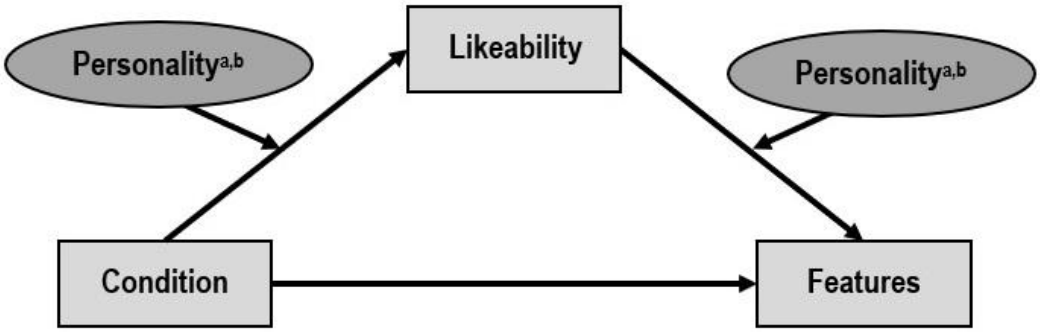


Note. This model was used for both experiments. The dependent variables (Features) were mostly similar, with some additional features in the second experiment. The model was also tested with an overall CS features score.
a = CSs paired with positive or negative USs; b = CS likeability ratings; c = for Experiment 1, the following features were used: friendly, trustworthy, strong, calm, and humble; d = for Experiment 2, the following features were used: warm, sincere, competent, emotionally stable, friendly, trustworthy, strong, calm, and humble.

To examine the role of personality, we explored whether the magnitude of the mediation depends on personality-feature links (see Figure 2). To do so, we built mixed-models that allowed us to address the moderating role of personality in (1) the effects of condition on liking, (2) the effects of liking on each of the other five features, and (3) the simultaneous effects of both links in which liking mediates the effect of condition on each of the other five features for each personality dimension (see Figure 2).

Figure 2

Visual representation of the moderated mediation model



Note. This overall model was used for both experiments.
a = the following personality scales were used for experiment 1: Neuroticism-Anxiety, Aggression-Hostility (ZKPQ; Zuckerman, 2002) and Negative Emotionality, Agreeableness (BFI-2; Soto & John, 2017); b = the following personality scales were used for experiment 2: Emotionality, Honest-Humility, Agreeableness, Conscientiousness, Extraversion, Openness (HEXACO–60; Ashton & Lee, 2009).

The models were also tested with an overall CS features score. In this way, we could test separately if personality moderates the first (condition to liking) and/or the second (liking to other feature) term of the indirect effect model, as well as both of them simultaneously within the context of a mediation model, for each of the personality dimensions and for each of the five features. First, a linear mixed-effects model was fitted to assess the relationship between condition and likeability ratings, moderated by personality (neuroticism or agreeableness). This

model included condition as a fixed effect and a random intercept for subjects to account for repeated measures. Next, another linear mixed-effects model was fitted to examine the effect of likeability ratings on feature evaluations, with the condition and the interaction between likeability ratings and personality dimensions as predictors. This model also included a random intercept for subjects. To test the moderated mediation hypothesis, the mediate function from the *mediation* package was used. This function integrated the results of the two fitted models and performs simulation-based inference to estimate the indirect effect of condition on feature score through likeability ratings, moderated by personality. The *sjPlot* package version 2.8.16 was used to visualise the interaction effects for the outcome variables (Ludecke, 2014). Additionally, we investigated the role of demand compliance and beliefs about the study's objective. More information about data preparation can be found in Section 3 of the Supplemental Materials.

Additionally, at the reviewers' suggestion, we conducted the analyses on the dataset without excluding the participants based on the established criteria. The mixed ANOVA and mediation results were largely consistent, whereas most moderation effects and trends disappeared. Complete results can be found in Section 5 of the Supplemental Materials.

Results

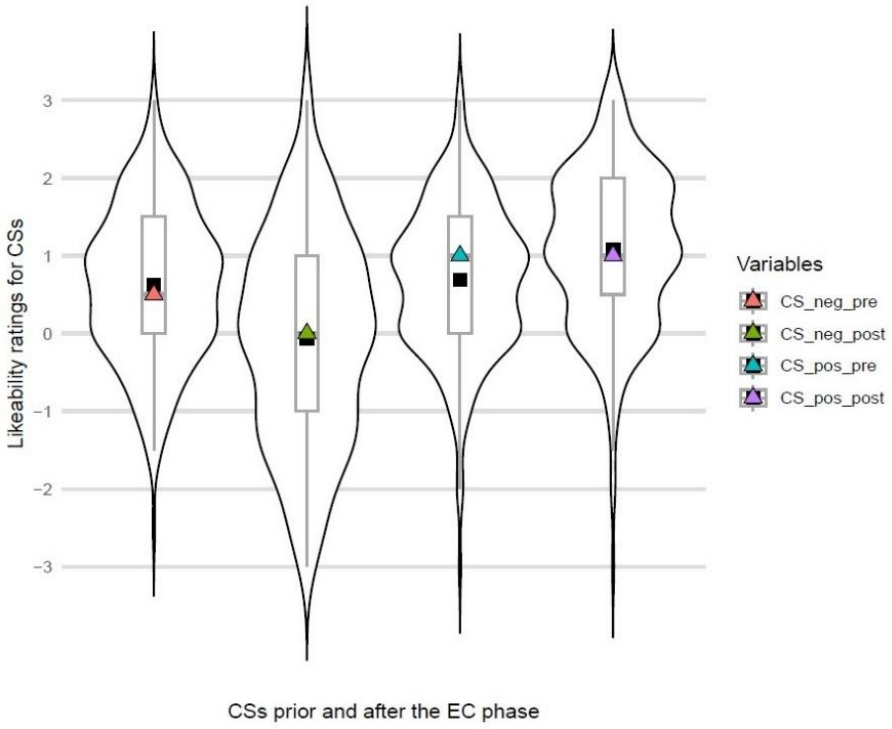
Overall effects

Before conducting the analyses, we averaged the pre-acquisition likeability ratings and the post-acquisition likeability ratings. To determine the change in liking, we subtracted the score of the pre-acquisition evaluation from the post-acquisition evaluation and obtained a score for CSneg (CS paired with negative USs) and one for CSpos (CS paired with positive USs). We first conducted a paired-sample *t*-test that compared the change scores for CSneg and CSpos. The results indicated a significant difference post-acquisition, $t(236) = 10.46$, $p < .001$, with a

large effect size (d = 0.97, 95% CI [0.78 to 1.17]), revealing that the EC effect was present, and thus that a transfer of valence occurred. A visual representation of the likeability ratings before and after the acquisition phase can be found in Figure 3.

Figure 3

The statistics of the likeability ratings for CSs in the two conditions prior-acquisition and post-acquisition



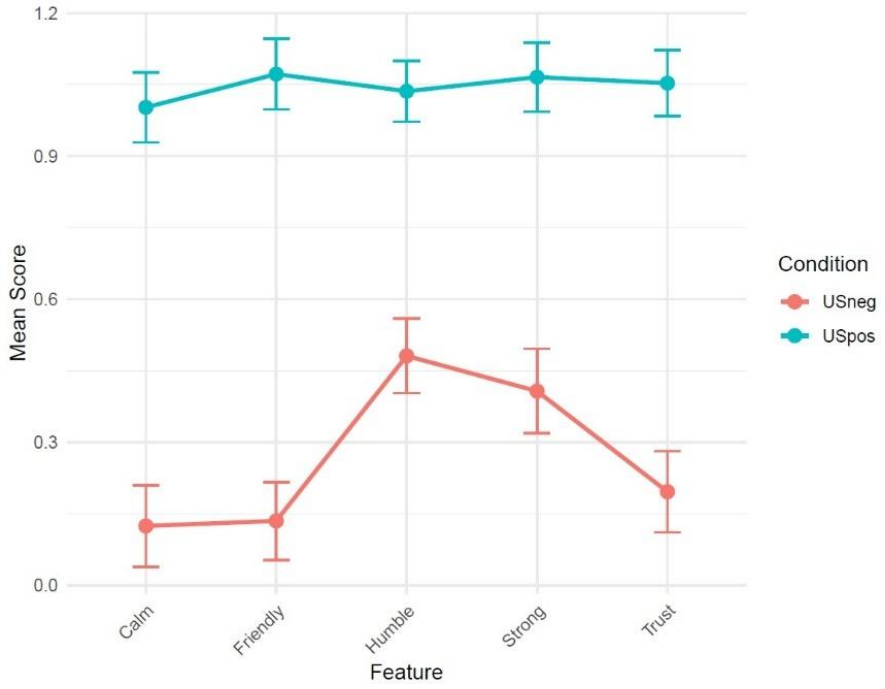
Note. The violin plots present the summary statistics and the density of the evaluations for the CSs paired with positive and negative USs, as can be seen in the description of the variables. A boxplot for the response distribution is represented for each plot. The black square represents the mean of the likeability ratings for the two conditions, while the triangle represents the median of the likeability ratings for the two conditions.

Further, we conducted a mixed 2 x 5 ANOVA, where we introduced Condition (USpos for CSs paired with positive USs and USneg for CSs paired with negative USs) and the five levels of Feature (*friendly, trustworthy, strong, calm,* and *humble*), with a corrected Greenhouse-

Geisser estimate (GGe) given a violation of the sphericity assumption for Features (Mauchly's W = 0.76, p < .001). The main effect of Condition was significant, indicating differences between the two levels, $F(1, 236) = 82.08$, $p < .001$, with a medium effect size $\eta_G^2 = 0.0961$. The main effect of Feature was also significant, indicating differences in evaluation between the different features, $F(3.49, 823.55) = 5.50$, $p < .001$, $\eta_G^2 = 0.0004$. The interaction between Feature and Condition was also significant, with a small effect size, $F(3.54, 834.50) = 5.91$, $p < .001$, $\eta_G^2 = 0.0037$. A visual representation of the interaction between Feature and Condition is depicted in Figure 4.

Figure 4

Interaction plot between the Feature evaluations as a function of Condition



Note. Error bars represent the 95% confidence intervals.

The post-hoc tests indicated significant differences between CSpos and CSneg, suggesting that the condition significantly affected the ratings across all features. Larger effects were observed for the features *calm*, *friendly*, and *trustworthy*, compared to the features *humble* and *strong*, which revealed medium-size effects.

Mediation analyses

Firstly, we examined whether the condition would predict the subsequent ratings of the additional features through the change in liking. As expected, the mixed effects analyses for the first set of models indicated that the mediator (change in liking) is significantly predicted by Condition $b = 1.08$, $SE = 0.08$, $t(710) = 14.05$, $p < .001$.

In the second set of models, we introduced an overall CS features score composed of the five specific features that had a high level of internal consistency ($\alpha = .88$). The results underlined that the change in liking was a significant predictor for the overall CS features score $b = 0.27$, $SE = 0.02$, $t(937.38) = 9.97$, $p < .001$, and also for all the specific CS features: *friendliness* ($b = 0.29$, $p < .001$), *trustworthiness* ($b = 0.27$, $p < .001$), *strength* ($b = 0.24$, $p < .001$), *calm* ($b = 0.29$, $p < .001$) and *humbleness* ($b = 0.27$, $p < .001$).

The results for the mediation analyses indicated that the relationship between Condition and Features score is partially mediated by the change in liking. More specifically, the change in liking mediated 38% of the effect of the Condition on the overall CS features score, which implies that the initial inferred valence is only partly responsible for the further evaluations of the other features. Estimated mediation effects, confidence intervals and p-values for all the models are presented in Table 1.

Table 1

Mediation effects of the condition through EC effects on the overall score of features and on specific features

		Condition (USneg – Uspos)		
		Estimate	**95% CI [LL, UL]**	***p***
Features	ACME	0.30	[0.23, 0.37]	<.001
	ADE	0.48	[0.34, 0.62]	<.001
	Total effect	0.77	[0.64, 0.91]	<.001
	Proportion mediated	0.38	[0.29, 0.50]	<.001
Friendly	ACME	0.31	[0.23, 0.40]	<.001
	ADE	0.63	[0.45, 0.80]	<.001
	Total effect	0.94	[0.77, 1.11]	<.001
	Proportion mediated	0.33	[0.24, 0.44]	<.001
Trustworthy	ACME	0.29	[0.22, 0.38]	<.001
	ADE	0.57	[0.40, 0.74]	<.001
	Total effect	0.86	[0.71, 1.01]	<.001
	Proportion mediated	0.34	[0.25, 0.46]	<.001
Strong	ACME	0.26	[0.17, 0.35]	<.001
	ADE	0.40	[0.21, 0.59]	<.001
	Total effect	0.66	[0.48, 0.83]	<.001
	Proportion mediated	0.39	[0.25, 0.58]	<.001
Calm	ACME	0.31	[0.23, 0.41]	<.001
	ADE	0.56	[0.36, 0.75]	<.001
	Total effect	0.87	[0.68, 1.04]	<.001
	Proportion mediated	0.36	[0.25, 0.50]	<.001
Humble	ACME	0.29	[0.21, 0.37]	<.001
	ADE	0.27	[0.10, 0.44]	<.001
	Total effect	0.56	[0.40, 0.71]	<.001
	Proportion mediated	0.51	[0.35, 0.76]	<.001

Note. ACME = average causal mediation effect, the indirect effect through the mediator; ADE = average direct effect, the effect of the predictor on the outcome when subtracting the effect of the mediator; Total effect = the effect of the predictor on the outcome without taking into account the mediator; Proportion mediated = the proportion of the effect of the predictor mediated by the mediator.

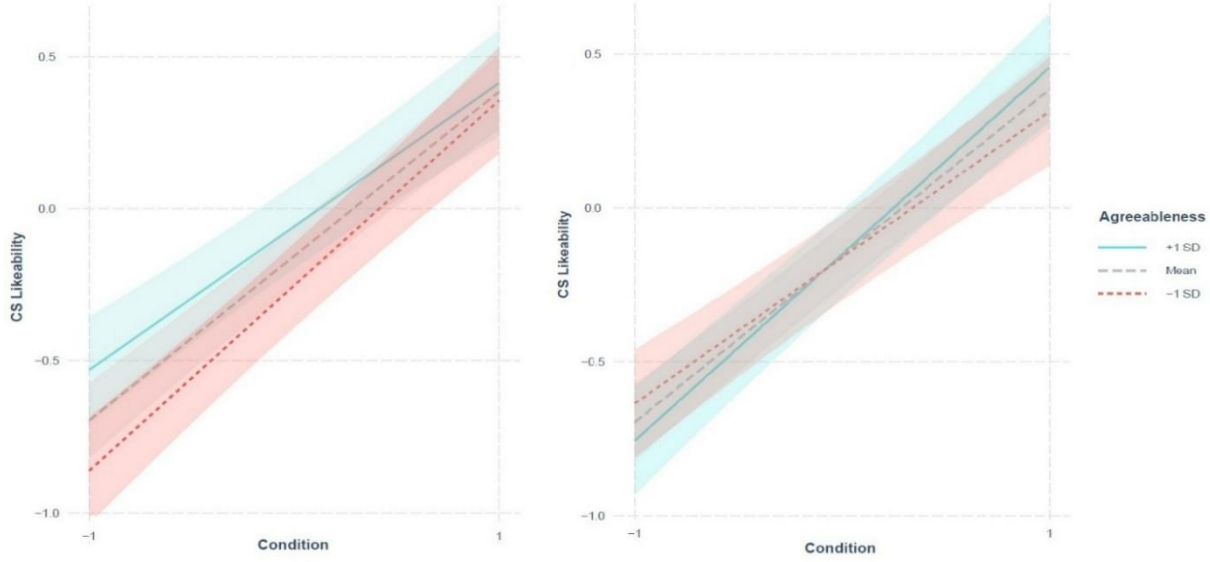
Moderation analyses

Firstly, a linear mixed-effects model was fitted to examine the interaction of personality with Condition on the likeability ratings of the CSs (model 1). When agreeableness was included in the model (the BFI scale), the results indicated that the main effect of Condition (b = 0.54, *SE*

= 0.04, $t(709^4) = 14.07$, $p = .002$) was statistically significant, but the main effect of agreeableness was not ($b = 0.001$, $SE = 0.009$, $t(235) = 0.11$, $p = .916$). However, agreeableness and Condition had an interaction effect close to significance, $b = 0.02$, $SE = 0.007$, $t(709) = 1.71$, $p = .088$. The decomposition of the effect showed that the manipulation (Condition) had a larger effect on the likeability evaluations (that is, there was a larger EC effect) for increasing levels of agreeableness (+1SD), $b = 0.61$, $SE = 0.05$, $t(709) = 11.15$, $p < .001$, compared to lower levels of agreeableness (-1SD), $b = 0.47$, $SE = 0.05$, $t(709) = 8.73$, $p < .001$ (see Figure 5, right panel). Note, however, that this effect disappeared when we conducted the analysis on the dataset with all participants, as the reviewers suggested (see Section 5 of Supplemental Materials).

Figure 5

The interaction effect between agreeableness and Condition (CS-US pairing manipulation) in predicting the CS likeability ratings



Note. The slopes of CS likeability ratings for each level of agreeableness. Condition was contrast coded with -1 for the pairing of the CSs with the negative USs, and with 1 for the pairing of the CSs with positive USs. On the left side is presented the slope for agreeableness as measured by ZKPQ (Aggression-Hostility scale), and on the right side is presented the slope for agreeableness as measured by BFI (Agreeableness scale).

[4] The degrees of freedom were estimated using the Satterthwaite degrees of freedom approximation (Kuznetsova et al., 2017), which takes into account the data structure and model complexity.

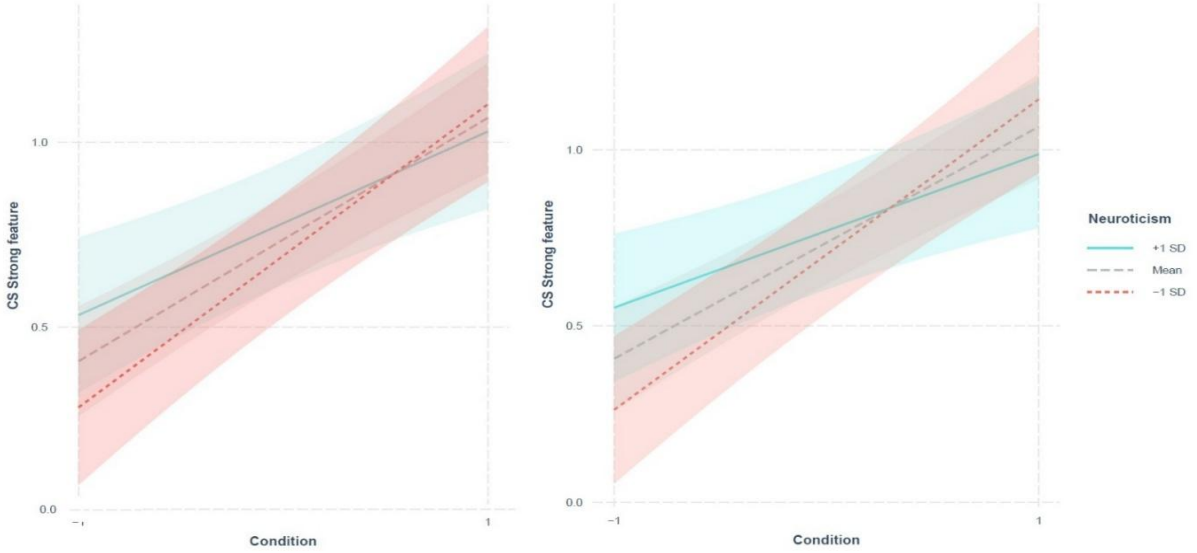
The same trend was observed for agreeableness as measured by ZKPQ (Aggression-Hostility scale). In this model, Condition significantly predicted the likeability ratings $b = 0.54$, $SE = 0.04$, $t(709) = 14.06$, $p = .002$, while agreeableness was close to significance $b = 0.03$, $SE = 0.02$, $t(235) = 1.94$, $p = .053$. The interaction between Condition and agreeableness was also close to significance ($b = -0.02$, $SE = 0.01$, $t(709) = -1.79$, $p = .074$). Also this trend disappeared when we conducted the analysis on the dataset with all participants (see Section 5 of Supplemental Materials).The slope analysis further illustrated this interaction, as the participants with a low score on the Aggression-Hostility (-1SD), $b = 0.61$, $SE = 0.05$, $t(709) = 11.21$, $p < .001$ scale reported a stronger effect on the likeability ratings in both conditions, compared to the high scores on the scale (+1SD), $b = 0.47$, $SE = 0.05$, $t(709) = 8.68$, $p < .001$, where the effect seemed to decrease (see Figure 5, left panel).

Secondly, when exploring the interaction effects of personality with Condition and the impact on specific feature evaluations (model 2), it was observed that the feature *strong* was significantly predicted by the interaction between Condition and neuroticism. In the case of neuroticism measured with the BFI scale, there was a significant main effect of Condition, $b = 0.33$, $SE = 0.05$, $t(709) = 7.13$, $p < .001$, but not of neuroticism, $b = 0.005$, $SE = 0.009$, $t(235) = 0.56$, $p = .575$. The interaction between neuroticism and Condition was also significant, $b = -0.02$, $SE = 0.006$, $t(709) = -2.41$, $p = .016$, indicating that as neuroticism increases, the effect of the Condition on the feature *strong* decreases. The decomposition of the effect indicated that at low levels of neuroticism (-1SD), $b = 0.44$, $SE = 0.07$, $t(709) = 6.74$, $p < .001$, the effect of Condition is stronger, compared to high levels of neuroticism (+1SD), $b = 0.22$, $SE = 0.07$, $t(709) = 3.33$, $p < .001$ (see Figure 6, right panel). The results of the model with neuroticism measured by the ZKPQ scale, were similar. The main effect of Condition was significant, $b =$

0.33, *SE* = 0.05, *t*(709) = 7.11, *p* < .001, while the effect of neuroticism was not, *b* = 0.009, *SE* = 0.01, *t*(235) = 0.73, *p* = .462. Nevertheless, the interaction between Condition and neuroticism was close to significant, *b* = -0.02, *SE* = 0.009, *t*(709) = -1.75, *p* = .081. This interaction suggested that when the neuroticism level was lower (-1SD), *b* = 0.41, *SE* = 0.07, *t*(709) = 6.26, *p* < .001, the difference between positive (paired with US+) and negative (paired with US-) CSs was larger, compared to the difference at higher neuroticism levels (+1SD), *b* = 0.24, *SE* = 0.07, *t*(709) = 3.79, *p* < .001 (see Figure 6, left panel). However, this trend disappeared when we conducted the analysis on the dataset with all participants (see Section 5 of Supplemental Materials).

Figure 6

The interaction effect between neuroticism and Condition (CS-US pairing manipulation) in predicting the CS strength ratings

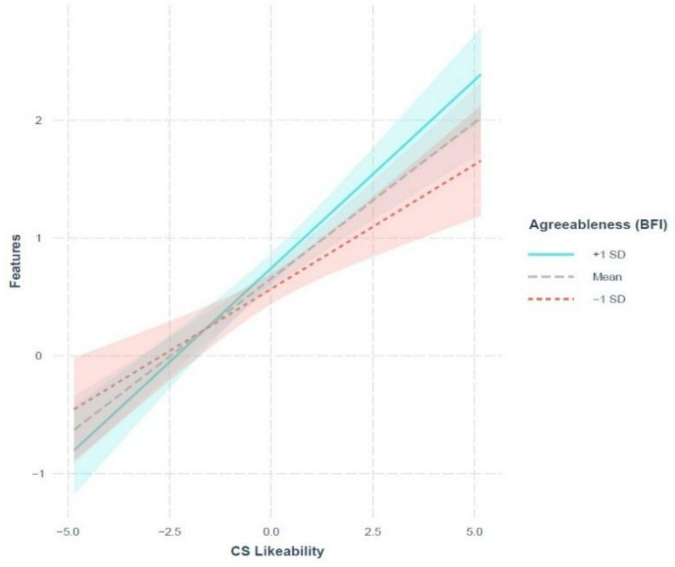


Note. The slopes of CS likeability ratings for level of neuroticism. Condition was contrast coded with -1 for the pairing of the CSs with the negative USs, and with 1 for the pairing of the CSs with positive USs. On the left side is presented the slope for neuroticism as measured by ZKPQ, and on the right side is presented the slope for neuroticism as measured by BFI.

Finally, while testing the moderation effect of personality on the mediation model (model 3), the results underlined that agreeableness, as measured by BFI, might be a potential moderator of these effects. The analyses showed that agreeableness, $b = 0.02$, $SE = 0.009$, $t(234.10) = 1.92$, $p = .056$, and Condition, $b = 0.24$, $SE = 0.04$, $t(762.48) = 6.90$, $p < .001$, predicted the overall CS features score, as well as the likeability rating, $b = 0.27$, $SE = 0.03$, $t(938.80) = 9.47$, $p < .001$. Further, it was observed that agreeableness had a close to significant interaction effect with the likeability rating, $b = 0.02$, $SE = 0.006$, $t(931.59) = 1.83$, $p = .067$, in predicting the overall CS features score. More information is presented in Figure 7.

Figure 7

The interaction effect between agreeableness and likeability ratings in predicting the CS overall features ratings



Note. The slopes for the overall features score for each level of agreeableness. As agreeableness increases, the slope of the relationship between CS likeability ratings and overall CS features score becomes steeper, indicating a stronger positive association at higher levels of agreeableness.

Figure 7 indicates that for individuals with high levels of agreeableness, (+1SD), b = 0.32, $p < .001$, the slope of the effect is slightly more positive, compared to lower levels of agreeableness (-1SD), b = 0.21, $p < .001$, suggesting that they rate overall CS features more positively when they already like the CS, and more negatively when they disliked the CS. Once more, however, this trend disappeared when we conducted the analysis on the dataset with all participants (see Section 5 of Supplemental Materials).

The mediation analysis for this model underlined a partial mediation as both the average causal mediation effect (ACME), b = 0.15, 95% CI [0.11,0.18], $p < .001$, and the average direct effect (ADE), b = 0.24, 95% CI [0.17,0.31], $p < .001$ were significant.

No other results approached significance, except for some exploratory analyses. Notably, it was observed that neuroticism, as measured by BFI, interacts with likeability to influence *strength* ratings, while agreeableness, as measured by BFI, interacts with likeability to affect *calm* ratings. More details about the exploratory analyses can be found in Section 6 of the Supplemental Materials.

Discussion

The first aim of this experiment was to examine whether the mere pairing of a CS with a valenced US may influence judgements about other features and whether such effects would go beyond effects on likeability. Overall analyses (*t*-test and mixed ANOVA) showed the effects of CS-US pairings not only on liking (feature transfer) but also on other features (feature transformation). The mediation effect for the overall CS features score suggests that higher likeability ratings lead to higher features scores, indicating that individuals who perceive the target more positively are likely to infer more positive attributes also for the other features of the target (e.g., more friendly, trustworthy, strong, calm, and humble). Specifically, the effects of

CS-US pairings on all other features (*friendly, trustworthy, strong, calm,* and *humble*) were partially mediated by the change in liking, with a proportion of mediation ranging from 33% to 46%. The obtained partial mediation might suggest that the likeability ratings are only partly responsible for the additional features ratings, pointing out other potential pathways beyond valence that might be relevant in this particular context.

With regards to the role of the personality of the participants, we obtained some results that provided directional but not statistically significant evidence. The results for the interaction between personality and the effect of pairing on liking indicated that agreeableness, as measured by both the Big Five Inventory (BFI) and the Zuckerman-Kuhlman Personality Questionnaire (ZKPQ), had a close to significant effect on likeability evaluations. These findings might suggest that individuals with higher levels of agreeableness may have slightly higher likeability ratings for the CSs, although the effect is only a tendency that does not reach statistical significance. Additionally, we explored the interaction effects of personality with pairings on the other features evaluation. The feature *strong* was significantly predicted by neuroticism, as measured by the BFI, suggesting that individuals with higher levels of neuroticism rated the CS stimuli as more vulnerable. For ZKPQ, this effect was close to significance, supporting a similar trend. These results align with previous research indicating that neuroticism is often associated with a tendency to perceive stimuli more negatively (Lahey, 2009).

Lastly, the study investigated the moderation effect of personality on the mediation model. The results showed that agreeableness, as measured by BFI, had a close to significant interaction effect with likeability ratings. The mediation analysis revealed partial mediation. This underlined that while agreeableness might influence the relationship between the condition and likeability, the mediation by likeability is only partial, with a significant portion of the effect

being direct. These findings suggest the importance of considering individual differences in personality in the context of impression formation, contributing to valuable insights into the fields of social psychology and personality research.

Experiment 1 had some limitations. Firstly, some of the effects observed in this study, particularly those involving personality dimensions such as agreeableness and neuroticism, were close to but did not reach conventional levels of significance. Moreover, all the trends that involved personality dimensions disappeared when we conducted the analysis on the dataset with all participants (see Section 5 of Supplemental Materials). A more highly powered study could try to probe again the same traits as well as to extend the attention to other main personality traits as a way to provide a more comprehensive test of the impact of personality.

Secondly, the specific evaluated features (e.g., *strength*, *trustworthiness*) were chosen based on theoretical considerations, but they are still only a limited sample of possible features. Including a more comprehensive set of features could provide a richer understanding of how social judgements are formed. We used an EC procedure with only four CSs that were paired with different USs of either positive or negative valence. It remains to be seen whether our results generalize to other more commonly used EC procedures such as procedures with more CSs and fixed CS-US pairs.

Experiment 2

In our second experiment, we adjusted the procedure to address the main limitations identified in the previous study. Specifically, to address the first limitation, we expanded our approach by measuring all six main personality dimensions of the HEXACO model (Honesty-Humility, Emotionality, Extraversion, Agreeableness, Conscientiousness, and Openness to Experience; Ashton & Lee, 2009) to shed light on possible moderators that were not previously

considered. To address the second limitation, we extended the list used to evaluate CSs to nine features. To further enhance the robustness of our findings, we also increased the number of conditioned stimuli (CSs) from two to three in both positive and negative conditions, while keeping the CS-US pairs constant across all trials. This adjustment aimed to generalize our findings with another type of EC procedure.

Method

Sample size strategy

Due to resource constraints and participant availability, we set a target sample size of 200 participants for Experiment 2 after accounting for potential exclusions, therefore adjusting sample size expectations based on practical considerations such as resources and feasibility (Lakens, 2022). While we aimed for a statistical power of $1 - \beta = 0.80$ at an alpha level of 0.05, we recognize that, whereas the sample size was more than adequate to detect the mediation effect with sufficient power, it was smaller than ideal for detecting very small effects for the moderation hypothesis. We conducted sensitivity analyses to determine the effects that could be reliably detected under the given conditions for both the mediation and moderation hypotheses (see Section 12 of Supplemental Materials). We discuss power considerations in more detail later in the paper.

Participants

For this study, we recruited Romanian-speaking students (N = 211; 177 female, 31 male, 3 other, M_{age} = 21.37 years, SD_{age} = 6.01). In return, participants who completed all phases of the study received course credits. The study took approximately 20 minutes to finish.

Based on the preregistered exclusion criteria, we eliminated from the dataset the participants who did not complete all the experiment phases, had more than three wrong answers

(from a total of 6) at the valence awareness task, and responded incorrectly to most of the attention questions (2 or 3 out of 3). In total, we excluded 64 participants ending up with a sample of 157 participants (140 female, 23 male, 3 other, M_{age} = 21.28 years, SD_{age} = 5.86).

Materials

Personality assessment. To assess the participant's personality, we used HEXACO–60 (Ashton & Lee, 2009), the short version of the HEXACO Personality Inventory. All 6 dimensions of the personality structure used in this model presented a good internal consistency: Emotionality (α = .74), Honest-Humility (α = .72), Agreeableness (α = .76), Conscientiousness (α = .77), Extraversion (α = .84), Openness (α = .70). All details about the psychometric properties of the scale can be found in Section 9 of the Supplemental Materials.

Stimuli. As the CSs we retrieved six neutral faces from The Karolinska Directed Emotional Faces (Goeleven et al., 2008). The neutral faces were selected based on the highest-rated emotion (in this case, *Neutral*) and the mean intensity score.

As the USs, we used 6 color images (3 positive and 3 negative) extracted from the International Affective Picture System IAPS (Bradley & Lang, 2007). These images were selected to be socially relevant, capturing real-life experiences. Additionally, we used 16 filler images in the pre-rating phase. More details can be found in Section 7 of the Supplemental Materials.

Procedure

In the first part of the experiment, participants received a briefing on the study and were invited to sign the informed consent.

Similar to the previous study, the experiment started with a pre-rating phase, in which each participant evaluated the 6 neutral faces (which further served as CSs), alongside 6 images that served as USs, and 16 filler images, using a scale from -3 to 3.

In the acquisition phase, the randomly assigned block was repeated, resulting in 48 trials in total. The participant was asked to pay attention to the task between blocks. Each trial consisted of 2 images presented simultaneously onscreen: the neutral face (CS) on the left side of the screen and the valenced image (positive or negative US) on the right side of the screen. The exposure time for each CS-US pair was extended from 2500ms to 3000ms, as in the pilot study (see Footnote 2), with 800ms intertrial space. Each CS-US pair appeared randomly 4 times in the block. In between trials, we used a fixation cross so that the participant would focus on the center of the screen.

In the second part of the experiment, the participants were asked to evaluate the 6 CS in terms of likeability and in terms of nine features, of which four (*warm*, *sincere*, *competent*, and *emotionally stable*) were added to the five used in the previous study (*friendly*, *trustworthy*, *strong*, *calm*, and *humble*). The participants received instructions similar to those from the previous experiment. The CSs'likeability was evaluated by using one dichotomous feature: *Pleasant* vs. *Unpleasant* rated on a 7-point scale from -3 to 3.

The same type of scale was used for all specific features (traits): Friendly – Unfriendly, Warm – Cold; Sincere – Insincere; Trustworthy – Untrustworthy; Competent – Incompetent; Emotionally stable – Hypersensitive; Strong – Vulnerable; Calm – Tense; Humble – Arrogant. The features were grouped into two blocks: friendliness, warmth, trustworthiness, sincerity and competence (block 1), and strength, emotional stability, calm, humbleness (block 2). Both the

CSs and the groupings of features were randomized, but the features within each page always appeared in the same fixed order.

To assess the participant's valence awareness in the evaluative learning sequence, the six CSs (randomly chosen), were presented again on the computer's screen. Additionally, each participant was asked about explicit demand compliance. The instructions were similar to Experiment 1.

In the last part of the study, the participants were asked to complete a personality questionnaire (HEXACO–60; Ashton & Lee, 2009). Alongside the personality items, we included three attention checks to verify whether a participant paid attention to the questions (e.g., *For this item, please choose number four - agree*). Lastly, the participants were asked to provide demographic information about themselves (gender, age, education level, student code), were debriefed and thanked for participating in the study.

Design. For this study, we had a within-subjects design with *US valence* (positive vs. negative) as a factor. Other method factors varied between participants: *stimulus assignment* as the images that served as CSs (CS1/CS2/CS3/CS4/CS5/CS6) were randomly paired with one of the positive or the negative images that served as USs (US1, US2, US3, US4, US5, US6). Before starting the experiment, participants were randomly assigned to one of the task combinations. All trials appeared onscreen in a random order.

Analytic strategy. In the analyses, we included self-reported data resulting from the experimental task: likeability evaluations of stimuli, pre-acquisition and post-acquisition, and ratings on the additional 9 features (*friendliness, warmth, sincerity, trustworthiness, competence, emotional stability, strength, calm,* and *humbleness*) post-acquisition.

As the possible moderators of the feature transfer effect, we included in the analyses the six personality scales from HEXACO–60 (Ashton & Lee, 2009), with a focus on two dimensions: Emotionality and Agreeableness.

Similar to the first experiment, we conducted a paired-sample *t*-test to check the effect of the condition on likeability ratings (feature transfer) and a mixed 2x9 ANOVA to test whether the impact of pairings was significant for the other features (feature transformation).

To test the role of changes in liking in the effect of CS-US pairings on other features and the role of personality in the effect of CS-US pairings on liking and on other features, we used the same logic as in Experiment 1, conducting a mediation analysis and a moderation analysis (the previous models, model 1, model 2, and model 3, were tested). Figures 1 and 2 show the visual representation of the main tested models.

We also explored the role of demand compliance and beliefs regarding the study's objective. Section 8 of the Supplemental Materials provides more information on data preparation, exclusion criteria, score transformation and processing, and testing for normal distributions.

In response to the reviewers' suggestion, we again conducted analyses on the full dataset without excluding participants based on the established criteria. The results for the mixed ANOVA and the mediation analyses remained similar in significance. However, one moderation effect disappeared (see below for details). The complete results are presented in Section 10 of the Supplemental Materials.

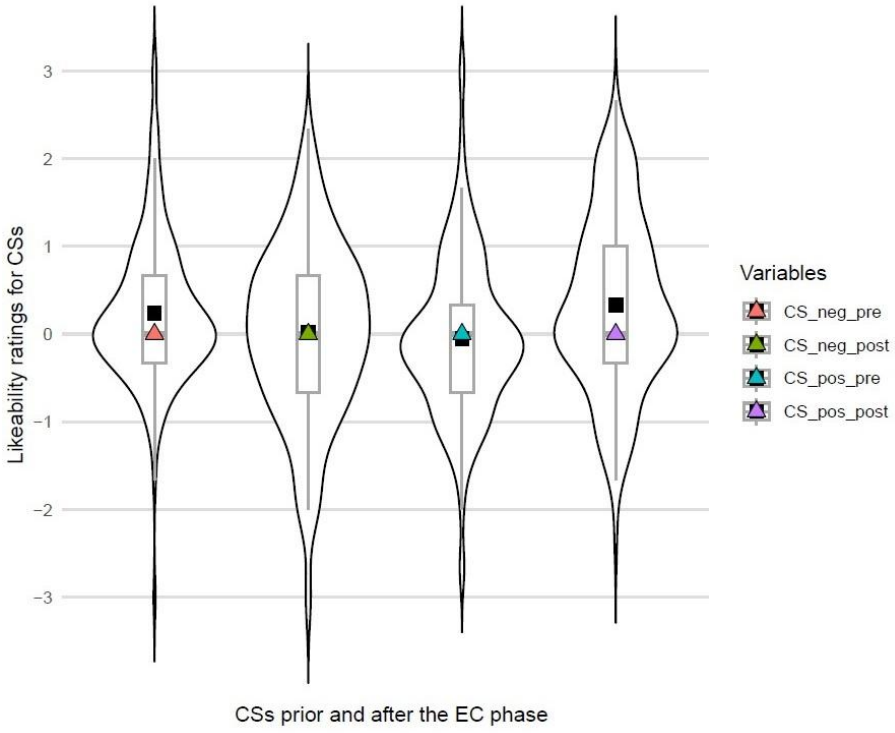
Results

Overall effects

The paired-sample *t*-test results showed that the change scores of CSneg and CSpos indicated a significant difference post-acquisition, $t(156) = 5.64$, $p < .001$, with a medium effect size ($d = 0.54$) and a 95% confidence interval ranging from 0.31 to 0.76. Figure 8 shows the pre and post-acquisition phase likeability ratings.

Figure 8

The statistics of the likeability ratings for CSs in the two conditions prior-acquisition and post-acquisition

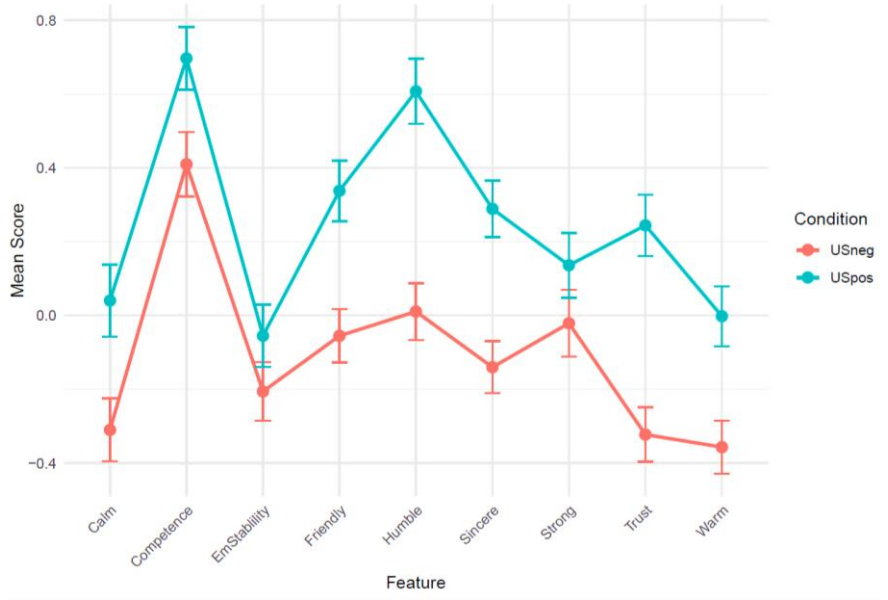


Note. The violin plots present the summary statistics and the density of the evaluations for the CSs paired with positive and negative USs, as can be seen in the description of the variables. For each plot, we represented a boxplot for the response distribution. The black square represents the mean of the likeability ratings for the two conditions, while the triangle represents the median of the likeability ratings for the two conditions.

A 2 x 9 mixed-design ANOVA was conducted to examine the impact of the nine levels of Feature (friendliness, warmth, sincerity, trustworthiness, competence, emotional stability, strength, calm, and humbleness) and two levels of Condition (USpos for CSs paired with positive USs and USneg for CSs paired with negative USs) on the participant's ratings. We used the Greenhouse-Geisser estimate (GGe) to adjust the degrees of freedom, as we were confronted with a violation of the sphericity assumption for Features (Mauchly's W = 0.18, $p < .001$) Condition had a significant main effect, $F(1, 156) = 17.12$, $p < 0.001$, $\eta_G^2 = 0.03$. The main effect of Feature was also significant, indicating differences in evaluation between the different features, $F(5.69, 884.13) = 27.15$, $p < .001$, $\eta_G^2 = 0.05$. The interaction between Feature and Condition (Figure 9) was also significant, with a small effect size, $F(5.31, 831.74) = 4.01$, $p < .001$, $\eta_G^2 = 0.005$.

Figure 9

Interaction plot between the Feature evaluations as a function of Condition



Note. Error bars represent the 95% confidence intervals.

Post-hoc pairwise comparisons using Bonferroni correction were conducted to explore the differences between USneg and USpos conditions within each feature. The results showed that except for *emotionally stable* and *strong*, ratings on all features differed significantly as a function of Condition, with *humble* and *trustworthy* showing the largest effects.

Mediation analyses

First, we tested whether Condition would predict the subsequent evaluations of the additional features through the change in liking. Results showed that the mixed effects for the first set of models indicated that the mediator (change in liking) is significantly predicted by Condition $b = 0.61$, $SE = 0.09$, $t(784) = 6.47$, $p < .001$.

The second set of models underlined that the change in liking predicted the overall CS features score (composite score of the nine features, $\alpha = .91$), $b = 0.37$, $SE = 0.02$, $t(931.53) = 19.23$, $p < .001$, and the specific CS features, *friendly* ($b = 0.51$, $p < .001$), *warmth* ($b = 0.47$, $p < .001$), *sincerity* ($b = 0.39$, $p < .001$), *trustworthiness* ($b = 0.51$, $p < .001$), *competence* ($b = 0.35$, $p < .001$), *emotional stability* ($b = 0.25$, $p < .001$), *strength* ($b = 0.21$, $p < .001$), *calm* ($b = 0.27$, $p < .001$) and *humbleness* ($b = 0.29$, $p < .001$).

The mediation model underlined that the relationship between Condition and features score is partially mediated by the change in liking. More specifically, the change in liking mediated 61% of the effect of the condition on the overall CS features score.

For some specific features, such as *friendliness*, *warmth*, *competence*, and *calm*, the condition effect was fully mediated by the change in liking. Estimated mediation effects, confidence intervals, and p-values are presented in Table 2.

Table 2

Mediation effects of the condition through EC effects on the overall score of features and on specific features

		Condition (USneg – Uspos)		
		Estimate	**95% CI [LL, UL]**	***p***
Features	ACME	0.22	[0.15, 0.29]	<.001
	ADE	0.14	[0.03, 0.26]	<.01
	Total effect	0.37	[0.24, 0.50]	<.001
	Proportion mediated	0.61	[0.43, 0.88]	<.001
Friendliness	ACME	0.31	[0.21, 0.41]	<.001
	ADE	0.08	[-0.09, 0.25]	0.33
	Total effect	0.39	[0.21, 0.57]	<.001
	Proportion mediated	0.79	[0.54, 1.37]	<.001
Warmth	ACME	0.29	[0.21, 0.37]	<.001
	ADE	0.08	[-0.08, 0.24]	0.42
	Total effect	0.36	[0.18, 0.54]	<.001
	Proportion mediated	0.80	[0.52, 1.37]	<.001
Sincerity	ACME	0.23	[0.15, 0.31]	<.001
	ADE	0.20	[0.04, 0.36]	<.01
	Total effect	0.43	[0.25, 0.60]	<.001
	Proportion mediated	0.54	[0.36, 0.88]	<.001
Trustworthiness	ACME	0.31	[0.21, 0.41]	<.001
	ADE	0.27	[0.10, 0.41]	<.001
	Total effect	0.57	[0.38, 0.75]	<.001
	Proportion mediated	0.53	[0.38, 0.75]	<.001
Competence	ACME	0.21	[0.14, 0.28]	<.001
	ADE	0.08	[-0.09, 0.24]	0.43
	Total effect	0.29	[0.12, 0.47]	<.01
	Proportion mediated	0.73	[0.44, 1.69]	<.01
Emotionally stable	ACME	0.15	[0.10, 0.22]	<.001
	ADE	-0.01	[-0.24, 0.21]	0.99
	Total effect	0.15	[-0.09, 0.36]	0.19
	Proportion mediated	0.87	[-6.66, 9.62]	0.19
Strong	ACME	0.13	[0.07, 0.19]	<.001
	ADE	0.04	[-0.19, 0.27]	0.73
	Total effect	0.16	[-0.06, 0.39]	0.15
	Proportion mediated	0.70	[-4.82, 7.00]	0.15
Calm	ACME	0.16	[0.11, 0.24]	<.001
	ADE	0.19	[-0.05, 0.42]	0.11
	Total effect	0.35	[0.12, 0.58]	<.001
	Proportion mediated	0.47	[0.25, 1.41]	<.001
Humble	ACME	0.18	[0.11, 0.25]	<.001
	ADE	0.42	[0.22, 0.63]	<.001
	Total effect	0.60	[0.40, 0.80]	<.001
	Proportion mediated	0.29	[0.18, 0.47]	<.001

Note. ACME = average causal mediation effect, the indirect effect through the mediator; ADE = average direct effect, the effect of the predictor on the outcome when subtracting the effect of the mediator; Total effect = the effect of the predictor on the outcome without taking into account the mediator; Proportion mediated = the proportion of the effect of the predictor mediated by the mediator.

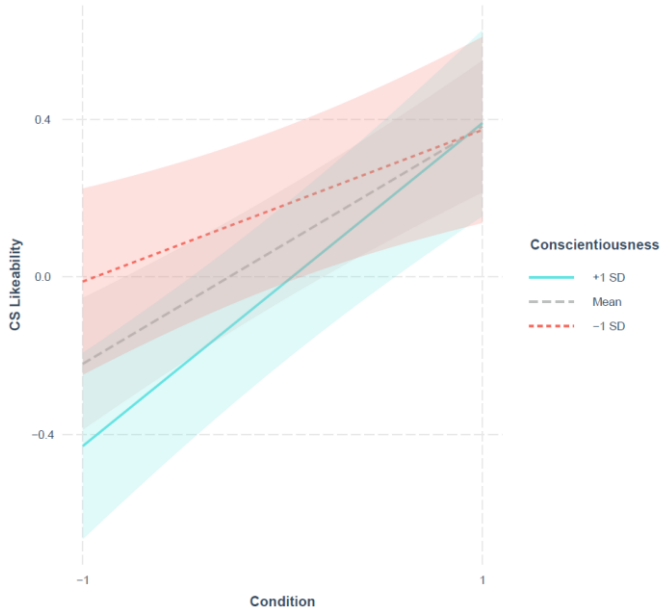
Moderation analyses

First, a linear mixed-effects model was fitted to examine the interaction of personality with Condition in predicting the likeability ratings of the CSs (model 1). The results did not support the second hypothesis, as neuroticism b = -0.01, SE = 0.01, $t(783) = -0.17$, $p = .866$ and agreeableness b = -0.01, SE = 0.01, $t(783) = -0.19$, $p = .849$ did not interact significantly with Condition in predicting the likeability ratings.

However, when inspecting the other four personality factors, it was observed that when we included conscientiousness in the model, it presented an interaction with Condition. Specifically, the main effect of Condition was significant in predicting the likeability ratings, b = 0.30, SE = 0.04, $t(783) = 6.48$, $p < 0.001$, while the effect of conscientiousness, b = -0.02, SE = 0.01, $t(155) = -1.40$, $p = .163$ was not significant. Further, its interaction with Condition was significant, b = 0.02, SE = 0.007, $t(783) = 2.34$, $p = .019$. The slope analysis indicated that at high levels of conscientiousness, (+1SD), b = 0.41, SE = 0.07, $t(783) = 6.24$, $p < .001$, the EC effect was larger (i.e., Condition had a stronger effect on CS likeability) compared to individuals with a low level of conscientiousness, (-1SD), b = 0.19, SE = 0.13, $t(783) = 2.93$, $p < .001$ (Figure 10). However, this interaction disappeared when we conducted the analysis on the dataset with all participants (see Section 10 of Supplemental Materials).

Figure 10

The interaction effect between conscientiousness and Condition (CS-US pairing manipulation) in predicting the CS likeability ratings



Note. The slopes of CS likeability ratings for each level of conscientiousness. Condition was contrast coded with -1 for the pairing of the CSs with the negative USs, and 1 for the pairing of the CSs with positive USs.

Secondly, the results did not indicate any significant effects when it was tested whether the interaction effect of personality with the Condition impacts the other specific feature evaluations of the CSs (model 2).

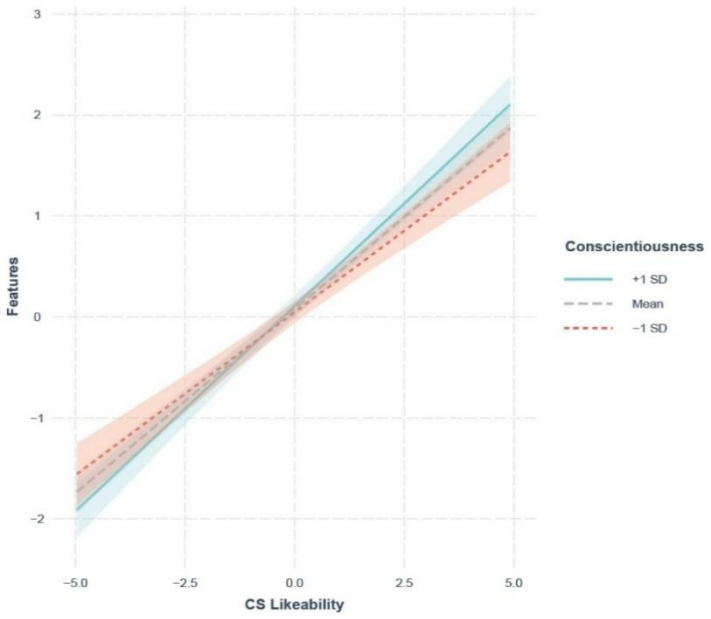
Lastly, while exploring the moderation effect of personality on the mediation model (model 3), the results underlined that neuroticism and agreeableness do not have a significant role in this context, neither on their own (b = 0.001, p = .593 for neuroticism and b = 0.01, p = .891 for agreeableness), and neither in interaction with likeability ratings (b = 0.02, p = .397 for neuroticism and b = 0.01, p = .397 for agreeableness). But, again, conscientiousness seems to have an interaction with the CS likeability ratings.

The main effect of conscientiousness (b = 0.005, SE = 0.006, t(151) = 0.78, p = .436) was not significant in predicting the overall CS features score, but likeability (b = 0.36, SE = 0.19, t(928.80) = 18.96, p < 001) and Condition reached statistical significance b = 0.07, SE = 0.03,

t(790.30) = 2.40, *p* = .017. Conscientiousness had a significant interaction effect with the likeability ratings, *b* = 0.06, *SE* = 0.003, *t*(936.50) = 2.19, *p* = .029, in predicting the overall CS features score. Figure 11 shows that the likeability ratings have a different relation with the overall CS features score based on levels of conscientiousness. Specifically, at higher levels of conscientiousness, (+1SD), *b* = 0.41, *p* < .001, the CS likeability ratings had a stronger effect on the overall CS features compared to individuals with a lower level of conscientiousness, (-1SD), *b* = 0.19, *p* < .001. Additionally, the mediation analysis underlined a partial mediation as both the average causal mediation effect (ACME) *b* = 0.11, 95% CI [0.08,0.15], *p* <. 001, and the average direct effect (ADE) *b* = 0.07, 95% CI [0.02,0.12], *p* = .016 were significant.

Figure 11

The interaction effect between conscientiousness and likeability ratings in predicting the CS overall features ratings



Note. The slopes for the overall features score for each level of conscientiousness. As conscientiousness increases, the slope of the relationship between CS likeability ratings and overall CS features score becomes steeper, indicating a stronger positive association at higher levels of conscientiousness.

We also explored the interaction effects at the level of specific features. It was observed that some features (e.g., *friendliness*, *warmth*), were significantly impacted by personality (e.g., neuroticism). For more details, see Section 11 of the Supplemental Materials.

Discussion

Experiment 2 aimed to replicate and extend the findings of Experiment 1. The results provided strong evidence that CS-US pairings influenced not only liking ratings but also the ratings of other features of the CSs. We also found evidence that the impact of the CS-US pairings on the ratings of the other features was partially mediated by the impact of the manipulation on liking ratings.

The mediation analyses underlined that likeability is a crucial aspect through which individuals attach more meaning to the world, inferring other features to a specific object. However, even after controlling for the likability ratings, significant effects were observed when averaging across all other features and also when analyzing some of the specific CS features separately (e.g., *sincerity*, *trustworthiness*, and *humbleness*). Hence, for these particular cases, the feature transformation effects do not seem merely an instance of a global impression effect.

As in the previous study, we initially focused on the two main personality dimensions of neuroticism and agreeableness and their potential role in moderating the effects of CS-US pairings on likeability ratings, feature evaluations, and the moderated mediation effect. However, the results did not support any of these possibilities. Additional analyses revealed that conscientiousness was the only personality dimension that significantly interacted with the effect of pairing on liking. However, this effect disappeared when we conducted the analysis on the dataset with all participants (see Section 10 of Supplemental Materials). Conscientiousness also influenced the extent to which CS-US pairings affect liking and other features. This finding

suggests that individuals with higher conscientiousness may be more responsive to the conditioning effects on likeability. However, given that the effects were weak and found amongst many other nonsignificant effects, we recommend caution in interpreting them.

General Discussion

In two pre-registered experiments, we sought to investigate the impact of CS-US pairings on both general liking and specific features of the CS. This research extends beyond previous research with traditional EC paradigms by (1) examining not only changes in the liking of CSs but also changes in the ratings of other features of CSs and (2) considering the role of personality as a potential moderator of these effects.

The results were consistent across the two studies, providing strong evidence that the EC pairing procedures do not only result in feature transfer but also in feature transformation effects. Moreover, feature transformation, to a large extent, depended on changes in liking. Finally, we did not observe a clear impact of personality on feature transfer, feature transformation, or the moderation of feature transformation by changes in liking. Below we discuss each of these findings in more detail.

Beyond valence transfer

The fact that the CS-US pairing influenced not only liking but also ratings of other CS features is in line with Rougier et al. (2023) who reported similar effects using different methodological approaches. However, our study contributes to the literature by directly assessing ratings of the specific features through a traditional EC procedure rather than using a reversed correlation method, thereby offering complementary evidence for these effects. Our findings support the conclusion that pairing a neutral stimulus with a liked or disliked stimulus not only changes the liking of the originally neutral stimulus but also the evaluation concerning a broad

range of other features of the neutral stimulus. As such it further highlights the importance of evaluative pairings for impression formation.

It is important to point out that these conditioned changes for features other than valence were unlikely to be instances of AC, that is, a transfer of a specific US feature (e.g., trustworthiness) to a CS as the result of CS-US pairings. We can be quite sure of this conclusion given that we used a wide variety of USs (e.g., a crying child, a car in flames, or an astronaut) for which many of the features would not apply (e.g., it makes sense to say that a car in flames is negative but not that it is sincere or insincere). Therefore, we can confidently conclude that the feature transformation effects were not merely instances of feature transfer effects.

We can also conclude that the observed feature transformation effects were at least partly mediated by conditioned changes in valence (i.e., EC). Hence, the feature transformation effects could, at least partly, be explained in terms of the formation of a global evaluative impression (Nisbett & Wilson, 1977) which subsequently influences judgements on other traits. This idea is in line with cognitive consistency theories (Gawronski & Brannon, 2019), which propose that individuals strive for coherence in their judgments. When a general positive (or negative) impression is formed as the result of CS-US pairings, individuals may start aligning their evaluations of specific traits (such as *friendliness, humbleness,* or *competence*) to fit this overall positive (or negative) view, creating a more consistent and stable overall impression.

We also observed that the percentage of the change in liking that mediated feature transformation was much lower in Experiment 1 (38%) compared to Experiment 2 (61%). This could be due to the fact that only five specific features had to be rated in Experiment 1 whereas nine features were rated in Experiment 2. It is indeed possible that the more features participants

are asked to rate, the more likely they are to adopt the heuristic of rating a feature of a CS on the basis of their general evaluative impression of the CS.

That said, we often observed only partial mediation, suggesting that feature transformation effects in EC procedures are not merely based on a transfer of valence. Moreover, the strength of mediation seemed to depend on the nature of the feature. For example, in both experiments, only relatively weak, partial mediation effects were observed for the features *trustworthy* and *humble*. On the other hand, we observed that feature transformation for features such as *friendly* and *calm* were fully mediated by changes in liking. Given that these features seem to be comparable in valence (and thus in terms of being susceptible to global impression effects), and also in the impact of CS-US pairings, such differences also suggest that the observed feature transformation effects were not entirely due to the formation and application of a global evaluative impression.

At present, we can only speculate about what other processes might be at play. We know from impression formation research that trait inferences are a fundamental cognitive process for efficiently evaluating others (Carlston & Skowronski, 2005; Uleman et al., 2008). Perhaps CS-US pairings sometimes give rise not only to a global evaluative impression but also to more specific beliefs about the CSs (e.g., a person paired with an expensive car might be seen as wealthy; see Van Dessel et al., 2018, for a related proposal, and Rougier et al., 2023, for evidence supporting this idea), which in turn could give rise to inferences about specific other traits (e.g., wealthy person is thought to be intelligent). This idea can be examined in future research by manipulating and probing the beliefs that participants form about the CSs.

The role of the perceiver's personality in impression formation

Previous research on impression formation focused mainly on assessing others' personality traits (Ames & Bianchi, 2008; Rau et al., 2021) but did not consider the perceiver's personality. Our secondary goal was to investigate if the effects of pairings on the rating for liking, of pairings on the ratings for other features, and the mediation by effects of liking on the effect of pairings on the other features, are moderated by the personality of the participant, in particular, neuroticism and agreeableness (Vogel et al., 2019; Casini et al., 2023; Ingendahl & Vogel, 2023).

While some previous studies suggested the relevance of neuroticism (Bunghez et al., 2023; Casini et al., 2023) and agreeableness (Ingendahl & Vogel, 2023) for the EC effect, we did not find clear evidence on this issue, with some tendential effects in the first study that were not confirmed in the second study. The lack of significant moderating effects for neuroticism and agreeableness may indicate that these traits do not substantially influence evaluative judgments, at least in this context and warrants further discussion.

First, the measures of neuroticism and agreeableness used in these studies might not capture the nuanced, facet-level variations that might be more directly linked to specific aspects of the individual differences relevant to the evaluative processes. Likewise, the absence of these moderation effects in our experiments, might also be explained by the situational specificity in trait-performance or the "trait-relevance of the task." According to trait activation theory (Tett & Burnett, 2003), personality traits are more likely to influence behaviour when the situation provides cues that make those traits salient. Hence, the influence of personality on EC effects might be diminished if the stimuli lack clear interpersonal or social content that could engage agreeable individuals' tendency to perceive others as friendly or trustworthy, or if the stimuli are

not strongly valenced or emotionally evocative enough to prompt individuals high in neuroticism, to perceive others as vulnerable or tense.

When it comes to agreeableness, Kammrath and Scholer (2011) argued that highly agreeable individuals judge prosocial behaviours more favourably and antisocial behaviours more unfavourably compared to individuals with low scores of agreeableness. In other words, these individuals demonstrate a more extreme perception of affective stimuli in general, and not just a more positive perception of stimuli. In line with this idea, EC research showed that individuals with higher levels of agreeableness tend to exhibit stronger EC effects compared to individuals with lower levels (Vogel et al., 2019). Nevertheless, it is possible that the type of stimuli in our study lacked sufficient emotional salience or did not align closely with the types of evaluative responses that typically engage agreeableness as a moderating factor. Future research could explore these boundary conditions more explicitly, perhaps by employing a broader range of stimuli or by examining specific stimulus characteristics that interact with individual differences in agreeableness.

As for neuroticism, at a theoretical level, the lack of a significant moderation effect might be explained by the participant's subjective perception of the USs. Previous studies have proposed that highly neurotic individuals experience the USs as more extreme, compared to participants low in neuroticism (Vogel et al., 2019; Casini et al., 2023; Ingendahl & Vogel, 2022; Bunghez et al., 2024). However, this underlying mechanism may rely on specific experimental contexts, such as counterconditioning paradigms (Bunghez et al., 2024), stimuli with high levels of arousal (Casini et al., 2023) or controlling for US evaluations (Ingendahl & Vogel, 2023), which may not align with the conditions in the present study.

Furthermore, when the moderation models were tested on the full dataset, including participants who demonstrated low valence awareness, all previously significant and marginally significant moderation effects disappeared. A possible reason for this observation is that participants who demonstrated low valence awareness were generally less attentive, which could introduce noise in the data (Luck & Vogel, 2013). To further address this, we conducted an ANOVA analysis that revealed a significant interaction between the EC effect and valence awareness (see Sections 5 and 10 of Supplemental Materials). As our analysis suggests, this is not merely a statistical issue, as it highlights the conditions under which the EC effect can be reliably observed. When participants have low awareness, the EC effect becomes less stable, making it more difficult to detect moderation effects, such as those linked to personality traits. Regardless of the merits of this interpretation, the fact that moderation effects were weak and present only after applying exclusion criteria calls for caution in interpreting these effects (Gawronski & Corneille, 2024).

Limitations and future directions

The present study is subject to a series of limitations. Firstly, the design was restricted to a specific set of stimuli because, as CSs, we used only images of women. Hence, one could argue that the findings cannot be generalised. Note, however, that in the pilot study (see Footnote 2), we did include both male and female faces and did not observe a significant difference in how they were evaluated. Secondly, we conducted all our analyses on self-reported data from the participants and did not include implicit measures such as the Implicit Association Test. Hence, we do not know if our results generalize to these other types of measures. Given that similar research sometimes did (e.g., Hughes et al., 2020), but sometimes did not manage to replicate the

explicit effects on the implicit measures (e.g., Kasran et al., 2023), it would be interesting to examine this further.

Thirdly, as indicated above, it would be interesting to study more systematically whether feature transformation and the mediation role of EC in these effects depends on the nature of the feature. We did find some indications that effects were different for different features, but could not draw strong conclusions about the reasons for these differences.

Another limitation is that we used different measures of personality in our studies (Big-Five framework in Experiment 1 and a HEXACO model in Experiment 2), an aspect that makes it more difficult to compare some of the findings while, at the same time, broadens the theoretical perspective used in assessing personality.

A further limitation of this study is the use of a single-item measure for the likeability ratings. While this approach facilitated efficient data collection, single-item measures are inherently less reliable compared to multi-item scales, potentially leading to attenuated effects in the mediation model. To address this issue, future research could increase the reliability of key variables, allowing for more precise tests of the proposed mediational effects and providing stronger evidence for the hypothesized causal relationships.

Lastly, the relatively limited sample sizes, especially in the second experiment, may impact the dependability of some of the findings. While the study was powered to detect small-to-medium interaction effects for the interaction with personality, it may lack sufficient power to detect smaller effects. This limitation likely applies, in general, to research on personality differences in EC, where interactions with personality traits often involve small effect sizes (Sommet et al., 2023). Future research could benefit from designs that accommodate smaller effect sizes, as highlighted in Sommet et al. (2023), and from larger sample sizes.

Conclusion

In two experiments, we addressed a significant gap in the EC literature by investigating how the pairing of two social stimuli in an EC paradigm results not only in feature transfer (i.e., changes in the degree to which a person is liked) but can also lead to feature transformations (i.e., changes in the degree to which a person is considered to be strong). We established that feature transformation effects can occur as the result of EC procedures and that these effects are partially (but not fully) mediated by valence transfer (i.e., changes in liking due to pairings). We also examined the role of the participant's personality in these effects and the mediation of these effects but did not find strong evidence for moderation. Our findings highlight the importance of stimulus pairings in influencing impression formation on a broad range of features.

Contributions

Contributed to the conception and design: Author 1, Author 2, Author 3, Author 4, and Author 5.

Contributed to the acquisition of data: Author 1, and Author 5.

Contributed to analysis and interpretation of data: Author 1, Author 3, and Author 4.

Drafted and/or revised the article: Author 1, Author 2, Author 3, Author 4, and Author 5.

Approved the submitted version for publication: Author 1, Author 2, Author 3, Author 4, and Author 5.

Funding information

This work has received funding from the European Union's Horizon 2020 research and innovation programme [grant details omitted for review] and from Exploratory Research Projects - 2020 Call [grant details omitted for review]. Author 2 is supported by [grant details omitted for review] from [Institution].

Competing interests

The authors have stated that they do not have any conflicts or competing interests to declare.

Supplemental Materials

The supplemental materials are available on the Open Science Framework (OSF) website at the following link: https://osf.io/jganc?view_only=017a1e9e06484081937c5027264774c4.

Data accessibility

The pre-registration files, used materials (e.g., Inquisit code, codebooks), datasets, and R scripts for data preparation and analyses for all the experiments are made publicly available at OSF (https://osf.io/p9wzy/?view_only=017a1e9e06484081937c5027264774c4) in the Files Tab, separately, for each experiment. Data processing and all analyses were conducted in R version 4.4.0 (R Core Team, 2024).

References

Ames, D. R., & Bianchi, E. C. (2008). The agreeableness asymmetry in first impressions: Perceivers' impulse to (mis) judge agreeableness and how it is moderated by power. *Personality and Social Psychology Bulletin, 34*(12), 1719-1736. https://doi.org/10.1177/0146167208323932

Ashton, M. C., & Lee, K. (2009). The HEXACO-60: A short measure of the major dimensions of personality. *Journal of Personality Assessment, 91*(4), 340–345. https://doi.org/10.1080/00223890902935878

Ashton-James, C. E., Tybur, J. M., Grießer, V., & Costa, D. (2019). Stereotypes about surgeon warmth and competence: the role of surgeon gender. *PLoS One, 14*(2), e0211890. https://doi.org/10.1371/journal.pone.0211890

Baeyens, F., De Houwer, J., Vansteenwegen, D., & Eelen, P. (1998). Evaluative conditioning is a form of associative learning: On the artifactual nature of Field and Davey's (1997) artifactual account of evaluative learning. *Learning and motivation*, *29*(4), 461-474. https://doi.org/10.1006/lmot.1998.1007

Bates, D., Mächler, M., Bolker, B. M., & Walker, S. C. (2015). Fitting Linear Mixed-Effects models using lme4. *Journal of Statistical Software*, 67(1), 1-48. https://doi.org/10.18637/jss.v067.i01

Batres, C., & Shiramizu, V. (2022). Examining the “attractiveness halo effect” across cultures. *Current Psychology*, 1-5. https://doi.org/10.1007/s12144-022-03575-0

Bauer, D. J., & Curran, P. J. (2005). Probing Interactions in Fixed and Multilevel Regression: Inferential and Graphical Techniques. *Multivariate behavioral research, 40*(3), 373–400. https://doi.org/10.1207/s15327906mbr4003_5

Bradley, M. M., & Lang, P. J. (2007). The International Affective Picture System (IAPS) in the study of emotion and attention. In J. A. Coan & J. J. B. Allen (Eds.), Handbook of emotion elicitation and assessment (pp. 29–46). Oxford University Press.

Brannon, S. M., Sacchi, D. L.M., & Gawronski, B. (2017). (In)consistency in the eye of the beholder: The roles of warmth, competence, and valence in lay perceptions of inconsistency. *Journal of Experimental Social Psychology, 70,* 80–94. https://doi.org/10.1016/j.jesp.2016.12.011

Bunghez, C., Houwer, J. D., Rusu, A., & Sava, F. A. (2024). Exploring the association between neuroticism and negativity bias in evaluative counterconditioning. *Learning and Motivation, 88,* 102039. https://doi.org/10.1016/j.lmot.2024.102039

Bunghez, C., Rusu, A., De Houwer, J., Perugini, M., Boddez, Y., & Sava, F. A. (2023). The Moderating Role of Neuroticism on Evaluative Conditioning: Evidence From Ambiguous Learning Situations. *Social Psychological and Personality Science, 0*(0). https://doi.org/10.1177/19485506231191861

Burton, S., Cook, L. A., Howlett, E., & Newman, C. L. (2015). Broken halos and shattered horns: overcoming the biasing effects of prior expectations through objective information disclosure. *Journal of the Academy of Marketing Science, 43*(2), 240-256. https://doi.org/10.1007/s11747-014-0378-5

Carlston, D. E., & Skowronski, J. J. (2005). Linking versus thinking: Evidence for the different associative and attributional bases of spontaneous trait transference and spontaneous trait inference. *Journal of Personality and Social Psychology, 89*(6), 884–898. https://doi.org/10.1037/0022-3514.89.6.884

Casini, E., Richetin, J., Sava, F. A., & Perugini, M. (2023). The Moderating Role of Neuroticism on Evaluative Conditioning: New Insights on the Processes Underlying This Relationship. *Collabra: Psychology, 9*(1). https://doi.org/10.1525/collabra.74820

Chen, J. M., Quinn, K. A., & Maddox, K. B. (2022). Bridging the gap between spontaneous behavior-and stereotype-based impressions. In *The Handbook of Impression Formation* (pp. 93-115). Routledge.

Chua, K.-W., & Freeman, J. B. (2022). Learning to judge a book by its cover: Rapid acquisition of facial stereotypes. *Journal of Experimental Social Psychology*, *98,* Article 104225. https://doi.org/10.1016/j.jesp.2021.104225

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), e28046. 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., Perugini, M., Boddez, Y., & Sava, F. (2023). A Roadmap for Future Interactions Between Research on Personality and Learning. *Collabra: Psychology, 9*(1). https://doi.org/10.1525/collabra.88334

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), Article 43. https://doi.org/10.1525/collabra.229

De Houwer, J., Thomas, S., & Baeyens, F. (2001). Association learning of likes and dislikes: A review of 25 years of research on human evaluative conditioning. *Psychological bulletin*, *127*(6), 853. https://doi.org/10.1037//0033-2909.127.6.853

Demartini, E., Ricci, E. C., Mattavelli, S., Stranieri, S., Gaviglio, A., Banterle, A., Richetin, J. & Perugini, M. (2018). Exploring consumer biased evaluations: Halos effects of local food and of related attributes. *International Journal on Food System Dynamics*, *9*(4), 375-389. https://doi.org/10.18461/ijfsd.v9i4.947

Faul, F., Erdfelder, E., Buchner, A., & Lang, A. G. (2009). Statistical power analyses using G* Power 3.1: Tests for correlation and regression analyses. *Behavior research methods*, *41*(4), 1149-1160. https://doi.org/10.3758/BRM.41.4.1149

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

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

Forgas, J. P., & Laham, S. M. (2017). Halo effects. In R. F. Pohl (Ed.), *Cognitive illusions: Intriguing phenomena in thinking, judgment and memory* (pp. 276–290). Routledge/Taylor & Francis Group.

Gawronski, B., & Brannon, S. M. (2019). What is cognitive consistency, and why does it matter? In E. Harmon-Jones (Ed.), *Cognitive dissonance: Reexamining a pivotal theory in psychology* (2nd ed., pp. 91–116). American Psychological Association. https://doi.org/10.1037/0000135-005

Gawronski, B., & Corneille, O. (2024). Unawareness of attitudes, their environmental causes, and their behavioral effects. *Annual Review of Psychology, 76,* 359-384. https://doi.org/10.1146/annurev-psych-051324-031037

Goeleven, E., De Raedt, R., Leyman, L., & Verschuere, B. (2008). The Karolinska Directed Emotional Faces: A validation study. *Cognition and Emotion, 22*(6), 1094–1118. https://doi.org/10.1080/02699930701626582

Goodwin, G. P., Piazza, J., & Rozin, P. (2014). Moral character predominates in person perception and evaluation. *Journal of Personality and Social Psychology, 106*(1), 148-168. https://doi.org/10.1037/a0034726

Green, P., & MacLeod, C. J. (2016). SIMR: An R package for power analysis of generalized linear mixed models by simulation. *Methods in Ecology and Evolution, 7*(4), 493-498. https://doi.org/10.1111/2041-210X.12504

Hamilton, D. L., Stroessner, S. J., & Driscoll, D. M. (1994). Social cognition and the study of stereotyping. In P. G. Devine, D. L. Hamilton, & T. M. Ostrom (Eds.), *Social cognition: Impact on social psychology* (pp. 291–321). Academic Press.

Han, S., Li, Y., Liu, S., Xu, Q., Tan, Q., & Zhang, L. (2018). Beauty is in the eye of the beholder: The halo effect and generalization effect in the facial attractiveness evaluation. *Acta Psychologica Sinica, 50*(4), 363-376. https://doi.org/10.3724/SP.J.1041.2018.00363

Hofmann, W., De Houwer, J., Perugini, M., Baeyens, F., & Crombez, G. (2010). Evaluative conditioning in humans: a meta-analysis. *Psychological bulletin, 136*(3), 390. https://doi.org/10.1037/a0018916

Högden, F., & Unkelbach, C. (2020). The Role of Relational Qualifiers in Attribute Conditioning: Does Disliking an Athletic Person Make You Unathletic? *Personality and Social Psychology Bulletin*. https://doi.org/10.1177/0146167220945538

Hughes, S., De Houwer, J., Mattavelli, S., & Hussey, I. (2020). The shared features principle: If two objects share a feature, people assume those objects also share other features. *Journal of Experimental Psychology: General, 149*(12), 2264–2288. https://doi.org/10.1037/xge0000777

Ingendahl, M., & Vogel, T. (2022). Stimulus Evaluation in the Eye of the Beholder: Big Five Personality Traits Explain Variance in Normed Picture Sets. *Personality Science, 3*. https://doi.org/10.5964/ps.7951

Ingendahl, M., & Vogel, T. (2023). (Why) do big five personality traits moderate evaluative conditioning? The role of US extremity and pairing memory. *Collabra: Psychology*, *9*(1), 74812. https://doi.org/10.1525/collabra.74812

Kammrath, L. K., & Scholer, A. A. (2011). The Pollyanna Myth: How Highly Agreeable People Judge Positive and Negative Relational Acts. *Personality and Social Psychology Bulletin, 37*(9), 1172-1184. https://doi.org/10.1177/0146167211407641

Kasran, S., Hughes, S., & De Houwer, J. (2022). Learning via instructions about observations: Exploring similarities and differences with learning via actual observations. *Royal Society Open Science, 9(3),* 220059. https://doi.org/10.1098/rsos.220059

Kercher, A. J., Rapee, R. M., & Schniering, C. A. (2009). Neuroticism, life events and negative thoughts in the development of depression in adolescent girls. *Journal of Abnormal Child Psychology*, *37*(7), 903–915. https://doi.org/10.1007/s10802-009-9325-1

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

Kotov, R., Gamez, W., Schmidt, F., & Watson, D. (2010). Linking "big" personality traits to anxiety, depressive, and substance use disorders: A meta-analysis. *Psychological Bulletin, 136*(5), 768–821. https://doi.org/10.1037/a0020327

Krueger, R. F. (1999). The structure of common mental disorders. *Archives of General Psychiatry, 56*(10), 921–926. https://doi.org/10.1001/archpsyc.56.10.921

Kunda, Z., & Thagard, P. (1996). Forming impressions from stereotypes, traits, and behaviors: A parallel-constraint-satisfaction theory. *Psychological Review, 103*(2), 284–308. https://doi.org/10.1037/0033-295X.103.2.284

Kurdi, B., & Banaji, M. R. (2017). Repeated evaluative pairings and evaluative statements: How effectively do they shift implicit attitudes? *Journal of Experimental Psychology: General, 146*(2), 194–213. https://doi.org/10.1037/xge0000239

Kuznetsova A, Brockhoff PB, Christensen RHB (2017). lmerTest Package: Tests in Linear Mixed Effects Models. *Journal of Statistical Software*, 82(13), 1-26. https://doi.org/10.18637/jss.v082.i13

Lahey, B. B. (2009). Public Health Significance of Neuroticism. *The American Psychologist, 64*(4), 241. https://doi.org/10.1037/a0015309

Lakens, D. (2022). Sample Size Justification. *Collabra: Psychology*, *8*(1). https://doi.org/10.1525/collabra.33267

Lorenzo, G. L., Biesanz, J. C., & Human, L. J. (2010). What is beautiful is good and more accurately understood: Physical attractiveness and accuracy in first impressions of

personality. *Psychological Science, 21*(12), 1777-1782. https://doi.org/10.1177/0956797610388048

Luck, S. J., & Vogel, E. K. (2013). Visual working memory capacity: from psychophysics and neurobiology to individual differences. *Trends in cognitive sciences, 17*(8), 391-400. https://doi.org/10.1016/j.tics.2013.06.006

Lüdecke D (2024). sjPlot: Data Visualization for Statistics in Social Science. R package version 2.8.16. https://CRAN.R-project.org/package=sjPlot

Montoya A. K. (2023). Selecting a Within- or Between-Subject Design for Mediation: Validity, Causality, and Statistical Power. *Multivariate behavioral research, 58*(3), 616–636. https://doi.org/10.1080/00273171.2022.2077287

Mor, N., & Berkson, G. (2003). Attitudes toward stereotyped behaviors. *Journal of Developmental and Physical Disabilities*, *15*, 351-365. https://doi.org/10.1023/A:1026362200139

Moran, T., Hughes, S., Van Dessel, P., & De Houwer, J. (2022). The Role of Trait Inferences in Evaluative Conditioning. *Collabra: Psychology*, *8*(1). https://doi.org/10.1525/collabra.31738

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

Nicolau, J. L., Mellinas, J. P., & Martín-Fuentes, E. (2020). The halo effect: A longitudinal approach. *Annals of Tourism Research*, *83*, 102938. https://doi.org/10.1016/j.ijhm.2020.102497

Nisbett, R. E., & Wilson, T. D. (1977). The halo effect: Evidence for unconscious alteration of judgments. *Journal of Personality and Social Psychology, 35*(4), 250–256. https://doi.org/10.1037/0022-3514.35.4.250

R Core Team (2024). R: A Language and Environment for Statistical Computing. R Foundation for Statistical Computing. https://www.R-project.org

Rau, R., Carlson, E. N., Back, M. D., Barranti, M., Gebauer, J. E., Human, L. J., Leising, D., & Nestler, S. (2021). What is the structure of perceiver effects? On the importance of global positivity and trait-specificity across personality domains and judgment contexts. *Journal of Personality and Social Psychology*, *120*(3), 745-764. https://doi.org/10.1037/pspp0000278

Richetin, J., Demartini, E., Gaviglio, A., Ricci, E. C., Stranieri, S., Banterle, A., & Perugini, M. (2021). The biasing effect of evocative attributes at the implicit and explicit level: The tradition halo and the industrial horn in food products evaluations. *Journal of Retailing and Consumer Services*, *61*, 101890. https://doi.org/10.1016/j.jretconser.2019.101890

Romano, D., Costantini, G., Richetin, J., & Perugini, M. (2023). The HEXACO Adjective Scales and Its Psychometric Properties. *Assessment*, *0*(0). https://doi.org/10.1177/10731911231153833

Rougier, M., & De Houwer, J. (2023). Unconstraining Evaluative Conditioning Research by Using the Reverse Correlation Task. *Social Psychological and Personality Science*. https://doi.org/10.1177/19485506231217526

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.84560d

Sava, F. A., Payne, B. K., Măgurean, S., Iancu, D. E., & Rusu, A. (2020). Beyond contingency awareness: the role of influence awareness in resisting conditioned attitudes. *Cognition and Emotion, 34*(1), 156-169. https://doi.org/10.1080/02699931.2019.1652146

Scheider, J., Barbedor, J., Yzerbyt, V. & Abele, A. (2022). The facets of different kinds of social groups: A study in three languages.

Schoemann, A. M., Boulton, A. J., & Short, S. D. (2017). Determining power and sample size for simple and complex mediation models. *Social Psychological and Personality Science, 8*(4), 379-386. https://doi.org/10.1177/1948550617715068

Sommet, N., Weissman, D. L., Cheutin, N., & Elliot, A. J. (2023). How Many Participants Do I Need to Test an Interaction? Conducting an Appropriate Power Analysis and Achieving Sufficient Power to Detect an Interaction. *Advances in Methods and Practices in Psychological Science*., *6*(3). https://doi.org/10.1177/25152459231178728

Soto, C. J., & John, O. P. (2017). The next Big Five Inventory (BFI-2): Developing and assessing a hierarchical model with 15 facets to enhance bandwidth, fidelity, and predictive power. *Journal of Personality and Social Psychology, 113*(1), 117–143. https://doi.org/10.1037/pspp0000096

Srivastava, S., Guglielmo, S., & Beer, J. S. (2010). Perceiving others' personalities: Examining the dimensionality, assumed similarity to the self, and stability of perceiver effects. *Journal of Personality and Social Psychology, 98*(3), 520–534. https://doi.org/10.1037/a0017057

Tett, R. P., & Burnett, D. D. (2003). A personality trait-based interactionist model of job performance. *Journal of Applied Psychology, 88*(3), 500-517. https://doi.org/10.1037/0021-9010.88.3.500

Tingley, D., Yamamoto, T., Hirose, K., Keele, L., & Imai, K. (2014). mediation: R Package for Causal Mediation Analysis. *Journal of Statistical Software*, 59(5), 1-38. https://doi.org/10.18637/jss.v059.i05

Uleman, J. S., & Saribay, S. A. (2012). Initial impressions of others. In K. Deaux & M. Snyder (Eds.), *The Oxford handbook of personality and social psychology* (pp. 337–366). Oxford University Press. https://doi.org/10.1093/oxfordhb/9780195398991.013.0014

Uleman, J. S., Saribay, S. A., & Gonzalez, C. M. (2008). Spontaneous inferences, implicit impressions, and implicit theories. *Annual Review of Psychology, 59*(1), 329–360. https://doi.org/10.1146/annurev.psych.59.103006.093707

Uziel, L. (2015). Life seems different with you around: Differential shifts in cognitive appraisal in the mere presence of others for neuroticism and impression management. *Personality and Individual Differences, 73,* 39–43. https://doi.org/10.1016/j.paid.2014.09.023

Van Dessel, P., Hughes, S., & De Houwer, J. (2018). How Do Actions Influence Attitudes? An Inferential Account of the Impact of Action Performance on Stimulus Evaluation. *Personality and Social Psychology Review*. https://doi.org/10.1177/1088868318795730

Vogel, T., Hütter, M., & Gebauer, J. E. (2019). Is evaluative conditioning moderated by Big Five personality traits? *Social Psychological and Personality Science, 10*(1), 94-102. https://doi.org/10.1177/1948550617740193

Walther, E., Weil, R., & Düsing, J. (2011). The role of evaluative conditioning in attitude formation. *Current Directions in Psychological Science, 20*(3), 192-196. https://doi.org/10.1177/0963721411408771

Zanna, M. P., & Hamilton, D. L. (1977). Further evidence for meaning change in impression formation. *Journal of Experimental Social Psychology*, *13*(3), 224-238. https://doi.org/10.1016/0022-1031(77)90045-2

Zuckerman, M. (2002). Zuckerman-Kuhlman personality questionnaire (ZKPQ): An alternative five-factorial model. In B. de Raad & M. Perugini (Eds.), *Big five assessment* (pp. 376–392). Hogrefe & Huber Publishers.