Supplemental Materials for:

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

Table of Contents

Supplemental Materials Section 1 - Pilot study

Method

Participants and design

In this study, we had a sample of 82 participants (60 women, M_{age} = 29.27 years, SD_{age} = 10.14). In the first experimental we had 40 participants (28 women), and in the second experimental condition participated 42 individuals (32 women). No power analysis was performed prior to data collection.

We used a 2x2 mixed design with *Valence* (CSs paired with positive USs vs. CSs paired with negative USs), and *Condition (CS gender)* as a between-subject variable. The condition was created by combining the gender of CSs (2 neutral faces of women vs. 2 neutral faces of men) and the valence of the USs (2 positive vs. 2 negative valenced pictures). Consequently, in the first condition, the participants saw the 2 neutral faces of women paired with positive valenced pictures and the 2 neutral faces of men paired with negative valenced pictures. In the second condition, the participants saw the 2 neutral faces of women paired with negative valenced pictures and the 2 neutral faces of men paired with positive valenced pictures. The Condition factor was counterbalanced among participants. Likewise, the order of the trials and CS-US pairings were randomized between participants to control for their potential effect.

As the moderator variables, we used the scores obtained by the participants on each personality scale: *Neuroticism* and *Agreeableness* (NEO PI-R; Costa & McCrae, 1992), *Emotionality*, *Agreeableness*, and *Honest-Humility* (HEXACO-100; Lee & Ashton, 2018). For measuring EC effects, and other feature transfer effects (hallo/horn effects) we used self-reported evaluations provided by the participants. Data on both likeability (how pleasant or unpleasant is

the individual) and feature transfer (*friendliness*, *trustworthiness*, *invulnerability*, *calm*) was collected in order to distinguish between these 2 effects (De Houwer et al., 2019).

The study design, objectives, and analysis plan were pre-registered on the Open Science Framework website (https://osf.io/nk68c/?view_only=afbe0991a9df46db8acc32a1ea93769a). Prior to conducting this study, we received approval from the Scientific Council of University Research and Creation of the West University of Timișoara.

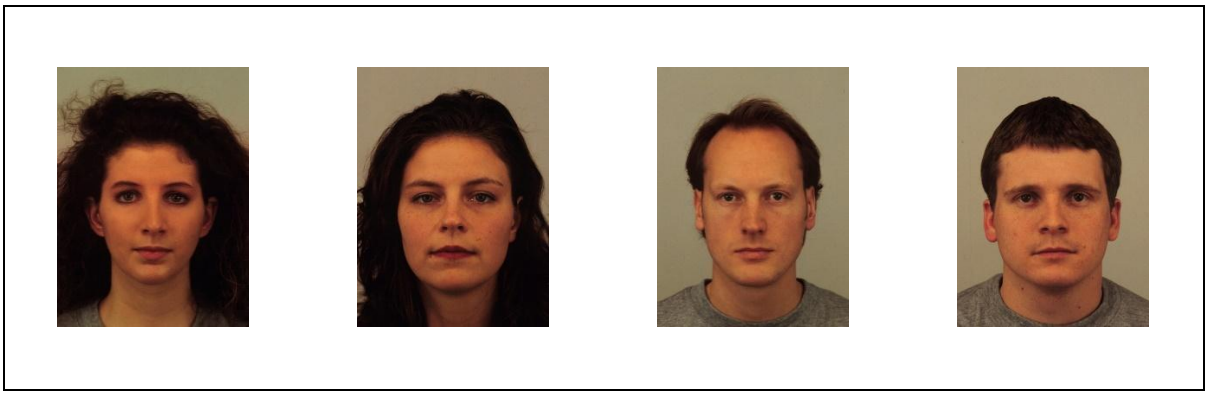
Stimuli

As CSs, four pictures (two men and two female) depicting neutral faces were retrieved from The Karolinska Directed Emotional Faces (Goeleven et al., 2008). The neutral faces (Figure S1.1) were selected from a poll of 20 pictures, based on the highest-rated emotion (in this case, *Neutral*) and the mean intensity score. The hit rate ranged from 96.88 – 89.06, with a mean intensity higher than 5.41.

As USs, four color images (2 positive and 2 negative) were extracted from the International Affective Picture System (IAPS) (Bradley & Lang, 2007). The images (Figure S1.2) were selected to be socially relevant, capturing real-life experiences, with an arousal/intensity value over 4.9, and valence values above 7.5 (for the positive valence), and below 2.5 (for the negative valence).

Figure S1.1

Female and male faces extracted from Karolinska Database that were used as conditioned stimuli in the evaluative conditioning procedure



Note. From left to right: F06NE, F13NE, M31NE, M10NE.

Figure S1.2

Negative and positive images extracted from IAPS used as unconditioned stimuli in the evaluative conditioning procedure

Note. From left to right: Distressed Fem 6311, Black Eye 2345.1, Bride 2209, Father 2160.

Personality assessment

Neuroticism was assessed with 48 items retrieved from *NEO-Personality Inventory Revised* (NEO PI-R; Costa & McCrae, 1992). The subscales had a good reliability ranging from $\alpha = 0.84$ to $\alpha = 0.75$, except for *impulsiveness* which had a questionable internal consistency ($\alpha = 0.57$).

Emotionality was measured with 16 items selected from HEXACO *Personality Inventory-Revised* (HEXACO-100; Lee & Ashton, 2018). Generally, the scales had an acceptable internal consistency (between $\alpha = 0.60$ and $\alpha = 0.77$), but at the scale level, the reliability was poor ($\alpha = 0.52$).

In the case of *Agreeableness*, we used 48 items from NEO PI-R and 16 items from HEXACO. Additionally, we included the dimension *Honest-Humility* from HEXACO. *Agreeableness* measured by NEO PI-R had acceptable reliability at the domain level ($\alpha = 0.74$). Regarding the scales from HEXACO, *Agreeableness* presented an acceptable internal consistency ($\alpha = 0.77$), while *Honest-Humility* had poor reliability ($\alpha = 0.53$).

Procedure

The experiment was programmed in Inquisit 5.0 (Millisecond Software, 2016), and it consisted of three phases. To go through the experiment, the participants were asked to access Inquisit Web.

In the first phase, participants were informed about the study overview and were invited to sign the informed consent. Afterward, they completed the personality questionnaires.

The second part of the experiment started with a *Pre-rating* sequence. During this process, each participant was asked to evaluate the 4 neutral faces on a scale from -3 to 3 in terms of likeability. After this sequence, the *Acquisition phase* began. In this part, each CS-US

pair was presented 5 times on the computer's screen for 3000ms, with 1000ms intertrial space, resulting in 20 trials per condition. The participants were randomly assigned to one of the 2 conditions.

In the last part of the experiment, the participants were asked to evaluate each CS in terms of likeability (*Pleasant* vs. *Unpleasant*) and feature transfer. Each feature (*friendly*, *trustworthy*, *invulnerable,* and *calm)* was rated on a 9-point scale (1 = *Very little* to 9 = *Very much*).

We also assessed the participant's valence awareness of the source-target contingency of the CS-US pairs. To further explore the reasoning behind the participant's answers, and in line with Hughes et al. (2020), we asked 2 additional questions, addressing the reasoning behind their answers and the objective of the study. Lastly, we collected demographic data.

Results

Preliminary analyses

A paired-sample t-test was performed to test if the positive and the negative valence were transferred from the source object (the unconditioned stimuli) to the target object (the conditioned stimuli). A visual representation is depicted in Figure S1.3.

The results displayed in Table S1.1, showed that the exposure to the acquisition phase resulted in a statistically significant change in valence for both the CSs paired with negative USs, $t(81) = 5.01$, $p < .001$), and the CSs paired with positive USs, $t(81) = -7.17$, $p < .001$. On average, the negative CSs were evaluated more positively prior to the acquisition phase ($M = 0.96$, $SD = 0.21$), compared to their evaluation after the acquisition ($M = -0.46$, $SD = 0.26$). In the case of positive CSs, the evaluation was more negative before the acquisition phase ($M = 0.68$, $SD = 0.24$), with a significant increase in likeability after the acquisition phase ($M = 2.21$,

SD = 0.19).

Regarding the change scores for the negative CSs and the positive CSs (M = -2.96, SD = 3.35), the results indicate a significant difference pre and post-acquisition, $t(81) = -7.98$, $p < .001$, $d = 0.88$, with a large effect size and a 95% confidence interval ranging from -3.67 to -2.22.

Analyses showed that the factor Condition had no significant interaction with the Valence factor, therefore, we dropped this variable from subsequent analyses, because it did not impact the EC effect.

Table S1.1

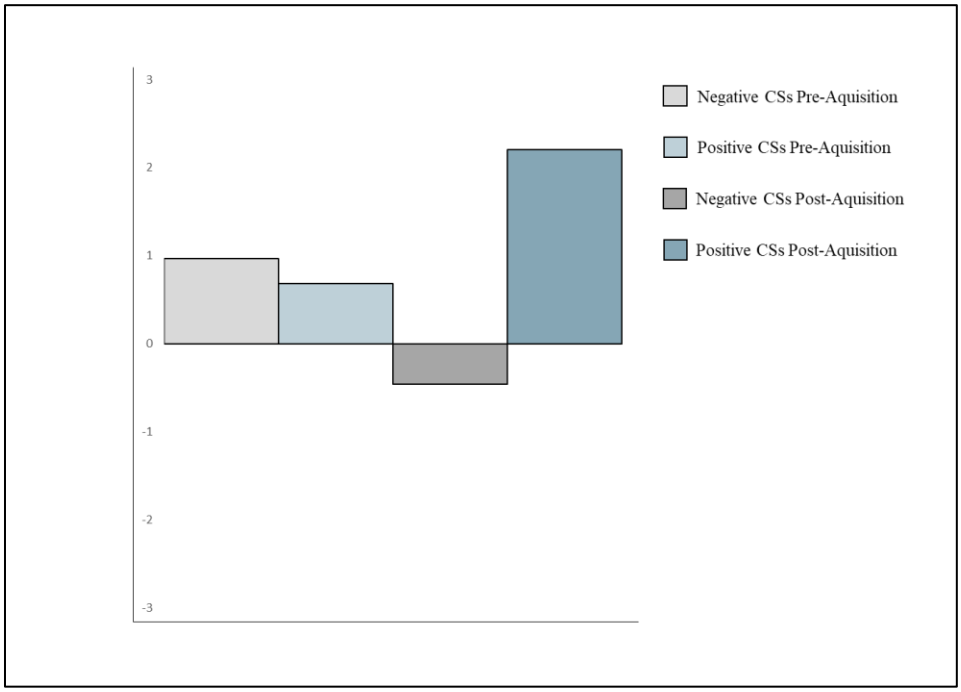
Means, Standard Deviations, and Paired Sample t-tests

	Paired sample t test					
Paired variables	**M**	**SD**	***t***	**df**	**Sig. (two-tailed)**	**Cohen's *d***
NegCS Pre – NegCS Post	1.43	2.58	5.01	81	.000	0.55
PosCS Pre – PosCS Post	-1.52	1.93	-7.17	81	.000	0.78
NegCS – PosCS	-2.96	3.35	- 7.98	81	.000	0.88
Features NegCS – Features PosCS	-2.85	5.62	-4.59	81	.000	0.51

Note. N = 82, M = mean, SD = standard deviation.

Figure S1.3

Scores on the likeability of the CSs pre-acquisition and post-acquisition



Hypothesis testing

We diverted from the initial analytic strategy and used simple linear regressions and two-step hierarchical regressions (Dardas & Ahmad, 2015), which were more suited for this data format and for exploring whether neuroticism or agreeableness has a moderation effect on the link between the valence transfer (EC effect) and the subsequent feature evaluations (hallo/horn effects).

First, it was hypothesized that an overall feature transfer effect will occur through an assimilative feature transformation effect, transferred from the positive, respectively the negative features of the CSs to other related features, such as *friendliness*, *trustworthiness*, *invulnerability*, and *calm* (H1). To test this hypothesis, a simple linear regression was performed to predict the overall features effect, based on the EC effect. For H1a and H2b we tested separately the

predictions from the negative and positive CSs to the negative and positive features.

Results (presented in Table S1.2) showed that 25% of the variance in overall features, can be accounted for by the emergence of the evaluative conditioning effect, $F(1,80) = 26.94$, $p < .001$, 95% CI [0.26, 0.58], $\Delta R = .25$. Furthermore, the likeability ratings for the negative CSs predicted the negative features, $F(1,80) = 4.13$, $p < .05$, 95% CI [0.04, 0.43], $\Delta R = .049$. And the likeability ratings for the positive CSs also predicted the positive features, $F(1,80) = 4.99$, $p < .05$, 95% CI [0.03, 0.46], $\Delta R = .047$.

To test if individual differences moderate the transformation effect in specific directions, we performed two-step hierarchical regressions. The results were not statistically significant, in consequence, we could not reject the null hypothesis.

Table S1.2

Linear regression to predict the overall features evaluations

Predictor	B	95% CI for B		SE	β	R^2
		LL	UL			
(Constant)	.301	-1.15	1.75			
EC effect	.86	5.24	1.18	.16	.51***	.26
(Constant)	1.70	.36	3.04			
Negative CS	-.80	-1.26	-.33	.23	-.36**	.05
(Constant)	1.06	-.39	2.51			
Positive CS	1.18	.58	1.77	.29	.41***	.05

Note. B = unstandardized regression coefficient, CI = confidence interval; LL = lower limit; UL = upper limit, SE = standard error, β = standardized regression coefficient.
*p < .05. **p < .01. ***p < .001

Exploratory analyses

During the analyses for the first objective, we observed that treated separately, both negative CSs, $F(1,80) = 11.69$, $p < .01$, 95% CI [-3.17, -0.85], $\Delta R^2 = .12$, and positive CSs, $F(1,80) = 15.56$, $p < .001$, 95% CI [1.12, 3.41], $\Delta R^2 = .15$, also predicted the overall overall features effect.

While testing for the second objective, we examined if personality dimensions can be valid moderators for the overall features effect. The results have shown that the overall features effect had a moderation interaction effect with Neuroticism (as measured by NEO-PI-R), respectively with Honest-humility (as measured by HEXACO).

Regarding Neuroticism, both models introduced in the multilinear regression were significant. In the first step, the 2 predictors collectively had a medium effect size ($\Delta R^2 = .25$), $F(1,80) = 14.76$, $p < .001$. In the second step, the interaction effect between the EC effect and Neuroticism, indicated that 28% of the variance in overall features, could be accounted for by the interaction between the 2 predictors, $F(1,80) = 11.31$, $p < .001$.

The overall model showed a significant interaction between the introduced variables, N from NEO, EC, GEN, $F(3, 78) = 11.23$, $p < .001$, $R^2 = .30$ and the interactions between neuroticism from NEOPIR, and the EC effect, we obtained $\beta = .21$, $t(81) = 2.11$, $p < .05$, 95% CI [0.01, 0.40]

The overall model showed a significant interaction between the introduced variables, N from HEXACO, EC, GEN, $F(3, 78) = 10.46$, $p < .001$, $R^2 = .28$. The and the interactions between neuroticism from HEXACO and EC was marginally significant: $\beta = .19$, $t(81) = 1.91$, $p = .06$, 95% CI [-0.008, 0.387]

The overall model showed a significant interaction between the introduced variables, HH

from HEXACO, EC, GEN, $F(3, 78) = 11.30$, $p < .001$, $R^2 = .31$. The and the interactions between HH from HEXACO and EC was marginally significant: $\beta = -.20$, $t(81) = -2.05$, $p < .05$, 95% CI [-0.39, -0.006]

In the case of Honest-Humility, the two models introduced in the multilinear regression analysis were both significant. In the first step, the 2 predictors collectively had a medium effect size ($\Delta R^2 = .25$), $F(1,80) = 14.76$, $p < .001$. In the second step, the interaction effect between the EC effect and Honest-Humility was also significant, $F(1,80) = 12.73$, $p < .001$, indicating that 30% of the variance in overall features, can be accounted for by the 2 predictors, collectively.

As a subsequent exploratory analysis, we used the same type of statistical analysis to examine if the moderation effect for overall features exists at the level of specific personality facets. As it is displayed in Table S1.3, the results yielded significant effects for the models that included Anxiety (as both measured by NEO PI-R, and HEXACO), Depression (as measured by NEO PI-R), Fairness, and Modesty (as measured by HEXACO).

Table S1.3

Multilinear regression analysis with personality facets as moderators

	Model	B	95% CI for B		SE	β	R^2	ΔR^2
			LL	UL				
Step 1	(Constant)	-.44	-.70	-.18	.13			
	EC effect	.15	.09	.21	.03	.49***	.27	.25
	Anxiety[a]	.10	-.10	.29	.10	.10		
Step 2	(Constant)	-.04	-.23	.14	.09			
	EC effect	.44	.25	.63	.10	.44***		
	Anxiety[a]	.08	-.10	.27	.09	.08		
	Interaction 1	.29	.08	.50	.11	.26*	.33	.30
Step 1	(Constant)	-.45	-.70	-.20	.13			
	EC effect	.15	.09	.21	.03	.50	.28	.27
	Anxiety[b]	.17	-.02	.36	.10	.17		
Step 2	(Constant)	-.01	-.19	.17	.09			
	EC effect	.47	.29	.65	.09	.47***		
	Anxiety[b]	.15	-.03	.33	.09	.15		
	Interaction 2	.28	.10	.46	.09	.29**	.37	.34
Step 1	(Constant)	-.44	-.70	-.19	.13			
	EC effect	.15	.09	.21	.03	.49***	.28	.26
	Depression[c]	.14	-.05	.33	.10	.14		
Step 2	(Constant)	-.02	.25	.63	.10			
	EC effect	.44	-.04	.33	.09	.44***		
	Depression[c]	.14	.03	.52	.12	.14		
	Interaction 3	.28	.25	.63	.10	.22*	.32	.29
Step 1	(Constant)	-.45	-.71	-.19	.13			
	EC effect	.15	.09	.21	.03	.50***	.27	.25
	Fairness[d]	-.17	-.31	.08	.10	-.12		
Step 2	(Constant)	-.01	-.19	.17	.09			
	EC effect	.49	.30	.67	.09	.49***		
	Fairness[d]	-.14	-.33	.04	.09	-.14		
	Interaction 4	-.30	-.50	-.09	.10	-.27**	.34	.31
Step 1	(Constant)	-.44	-.70	-.18	.13			
	EC effect	.15	.09	.21	.03	.50***	.26	.24
	Modesty[e]	-.07	-.26	.13	.10	-.07		
Step 2	(Constant)	-.06	-.24	.12	.09			
	EC effect	.45	.26	.63	.09	.45***		
	Modesty[e]	-.05	-.23	.13	.09	-.05		
	Interaction 5	-.36	-.55	-.17	.09	-.35***	.38	.35

Note. B = unstandardized regression coefficient, CI = confidence interval; LL = lower limit; UL = upper limit, SE = standard error, β = standardized regression coefficient; Interaction 1 = Anxiety x EC effect; Interaction 2 = Anxiety x EC effect; Interaction 3 = Depression x EC effect; Interaction 4 = Fairness x EC effect; Interaction 5 = Modesty x EC effect; *p < .05. **p < .01. ***p < .001; a_0 = Anxiety as measured by NEO PI-R; b_0 = Anxiety as measured by HEXACO; c_0 = Depression as measured by NEO PI-R; d_0 = Fairness as measured by HEXACO; e_0 = Modesty as measured by HEXACO.

Supplemental Materials Section 2 – Experiment 1: Method and Procedure

Stimuli

In Table S2.1, we present the stimuli selected as CS, and in Table S2.2, the stimuli selected as USs in the evaluative conditioning trials are presented.

Table S2.1

The codes of the images that were used as conditioned stimuli in the evaluative conditioning procedure

		KDEF code	Hit rate (%)	Mean (SD) intensity
Conditioned stimuli	CS1	AF06NES	96.88	5.60 (2.60)
	CS2	AF19NES	92.19	5.48 (2.09)
	CS3	AF07NES	87.50	5.36 (2.04)
	CS4	AF34NES	84.38	5.06 (2.16)
Check images	check1	AF01NES	85.94	5.20 (2.33)
	check2	BM10NES	89.06	5.41 (2.42)
Filler images	filler1	BM08NES	84.38	4.89 (2.17)
	filler3	BM06NES	84.38	4.94 (2.17)
	filler5	AM31NES	89.06	5.73 (2.15)
	filler7	AM14NES	-	-

Note. We retrieved 4 neutral faces (all female) 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. The hit rate ranged from 96.88 – 84.38, with a mean intensity higher than 5.06.

Table S2.2

The codes of the images that were used as unconditioned stimuli in the evaluative conditioning procedure

			IAPS Number	**Valence** (SD)	**Arousal** (SD)
Positive US	Set 1	US6	2160	7.58 (1.69)	5.16 (2.18)
		US7	1340	7.13 (1.57)	4.75 (2.31)
		US8	2900.2	6.62 (1.97)	4.52 (1.92)
		US9	5470	7.35 (1.62)	6.02 (2.26)
		US10	1463	7.45 (1.76)	4.79 (2.19)
	Set 2	US06	2209	7.64 (1.46)	5.59 (2.37)
		US07	8540	7.48 (1.51)	5.16 (2.37)
		US08	2035	7.52 (1.33)	3.69 (2.11)
		US09	8185	7.57 (1.52)	7.27 (2.08)
		US010	1710	8.34 (1.12)	5.41 (2.34)
Negative US	Set 1	US1	6311	2.58 (1.58)	4.95 (2.27)
		US2	2345.1	2.26 (1.46)	5.50 (2.34)
		US3	2375.1	2.20 (1.31)	4.88 (2.21)
		US4	1525	3.09 (1.72)	6.51 (2.25)
		US5	9909	2.78 (1.45)	5.98 (2.04)
	Set 2	US01	9332	2.25 (1.33)	5.34 (2.00)
		US02	2900.1	2.56 (1.41)	4.61 (2.07)
		US03	3230	2.02 (1.30)	5.41 (2.21)
		US04	1274	3.17 (1.53)	5.39 (2.39)
		US05	9230	3.89 (1.58)	5.77 (2.36)
Check images		check01	5726	6.23 (1.60)	2.84 (2.04)
		check02	7140	5.50 (1.42)	2.92 (2.38)
		check03	1441	7.97 (1.28)	3.94 (2.38)
		check04	2880	5.18 (1.44)	2.96 (1.94)
Filler images		filler2	2205	1.95 (1.58)	4.53 (2.23)
		filler4	2360	7.70 (1.76)	3.66 (2.32)
		filler6	6190	3.57 (1.84)	5.64 (2.03)
		filler8	2655	6.88 (2.09)	4.57 (2.19)

Note. The 20 color images (10 positive and 10 negative) were extracted from the International Affective Picture System IAPS (Bradley & Lang, 2007). These images were selected to be socially relevant, capturing real-life experiences, with an arousal value above 3.6, valence values over 6.6 (for positive), and below 4.5 (for negative).

Supplemental Materials Section 3 – Experiment 1: Data preparation

Mean scores for each CS valence (positive, negative) were computed according to the assigned block. This transformation was performed for the pre-acquisition ratings and post-acquisition ratings. Next, a change score for each valence (positive, negative) was computed by subtracting the pre-acquisition evaluations from the post-acquisition evaluations. Hence, two explicit mean rating scores resulted: negative CS and positive CS.

The EC effect was reflected by an explicit score resulting from subtracting the negative CS score from the positive CS score.

We averaged a score for the features (*Friendly, Trustworthy, Strong, Calm, Humble*) for each CS (CS1, CS2, CS3, CS4). Next, in accordance with the US's valence, we computed a mean score for the negative direction CS and one for the positive CS. Hence, we obtained a score for the negative features and one for the positive features. Mean scores were also computed for each specific feature (*Friendly, Trustworthy, Strong, Calm, Humble)* by subtracting the negative score from the positive score. For the feature transfer overall effect, we computed a differential score between the positive direction of the features and the negative direction of the features.

Scores for the personality scales were computed according to the instrument's specifications. All variables included as predictors in the analyses were mean-centered.

Additionally, we screened for data collected from participants who provided more than 2 wrong answers for the valence awareness task (from 4). From the total sample, 82% of the participants were aware of the source-target pairings (2, 3, or 4 correct answers). After we screened for data collected from participants who provided less than 2 correct answers for the valence awareness task, we excluded 52 participants (17.99%) from the final sample.

Regarding the demand compliance variable, the majority of the participants (75.11%) declared that they responded based on what they felt towards the stimuli. Another small percentage of the participants responded that they answered based on what I learned about the stimuli in the previous task (13.08%). And 5.91% of the participants declared that they did not know why they answered like that, and 14 said that they answered based on what they thought the researcher expected from them (5.91%). We further explored the last category of participants in order to determine if they were blind or not to the study objective. Based on their answers, we concluded that they did not know what the experimenter's objectives were, hence we did not exclude them from the analyses.

Supplemental Materials Section 4 – Experiment 1: Descriptive statistics

Prior to analysing the data, we explored the distributions of each variable. In the case of likeability ratings, feature ratings, and personality subscales, we tested the normality assumption by inspecting the visual representations, the symmetry (*Skewness*), and the *pointiness (Kurtosis)* of the data. The values for these parameters are reported in Table S4.1. We considered values greater than -1.96 and lesser than 1.96 to be normally distributed. Table S4.2 provides details about the psychometric proprieties of the personality questionnaires, while Table S4.3 presents the correlation values between the ratings of features and personality scales.

Table S4.1

Means, Standard Deviations, Skewness and Kurtosis for the likeability, features ratings and personality scales

Variable	M (SD)	Skewness	Kurtosis
EC effect	1.08 (1.59)	0.96	0.50
Features	0.78 (1.32)	0.96	1.24
Friendly	0.94 (1.66)	0.59	0.20
Trust	0.86 (1.54)	0.67	0.30
Strong	0.66 (1.69)	0.52	0.6
Calm	0.88 (1.66)	0.48	-0.04
Humble	0.55 (1.47)	0.46	0.56
Negative Emotionality – BFI-2	32.50 (6.89)	0.20	-0.13
Anxiety	11.73 (2.46)	0.05	0.38
Depression	9.95 (2.79)	0.2	-0.44
Emotional Volatility	10.82 (2.75)	0.3	0.13
Agreeableness – BFI-2	43.91 (5.07)	-0.25	0.19
Compassion	14.49 (2.2)	-0.18	-0.26
Respectfulness	15.17 (2.12)	-0.24	0.25
Trust	14.14 (2.02)	-0.22	0.03
Neuroticism-Anxiety – ZKPQ	23.32 (4.77)	0.88	-0.32
Aggression-Hostility – ZKPQ	21.37 (3.1)	0.56	-0.23

Note. N = 237; M = mean; SD = standard deviation; Positive = CSs paired with positive USs; Negative – CSs paired with negative USs.

Table S4.2

Psychometric Proprieties for BFI-2 and ZKPQ scales and subscales

Measure	M	SD	α
Negative Emotionality – BFI-2	32.50	6.89	.83
Anxiety	11.73	2.46	.51
Depression	9.95	2.79	.68
Emotional Volatility	10.82	2.75	.69
Agreeableness – BFI-2	43.91	5.13	.72
Compassion	14.49	2.23	.44
Respectfulness	15.17	2.16	.54
Trust	14.14	2.02	.39
Neuroticism-Anxiety – ZKPQ	23.32	4.77	.90
Aggression-Hostility – ZKPQ	21.37	3.10	.71

Note. N = 237; M = mean; SD = standard deviation; α = Alpha Cronbach.

Table S4.3

Correlations between the ratings of features and personality scales

	1	2	3	4	5	6	7	8	9	10
Likeability	1									
Friendly	.45**	1								
Trustworthy	.42**	.74**	1							
Strong	.38**	.48	.57	1						
Calm	.44**	.68	.64	.51	1					
Humble	.43**	.67	.68	.38	.58	1				
Neuroticism (BFI)	-.02	.03	-.01	-.11	-.04	.05	1			
Agreeableness (BFI)	.05	-.02	-.01	.08	.03	-.04	-.41	1		
Neuroticism-Anxiety (ZKPQ)	.03	.09	.03	-.07	-.02	.07	.73	-.22	1	
Aggression-Hostility (ZKPQ)	-.05	-.04	-.02	-.07	.01	.04	.31	-.41	.23	1

Note. N = 237; **. Correlation is significant at the 0.01 level (2-tailed); *. Correlation is significant at the 0.05 level (2-tailed); We considered *generalization* a factor because the features we introduced were highly correlated, and the overall score on features presented a high internal consistency (α = .88).

Supplemental Materials Section 5 – Experiment 1: Analyses with all participants

Mixed ANOVA

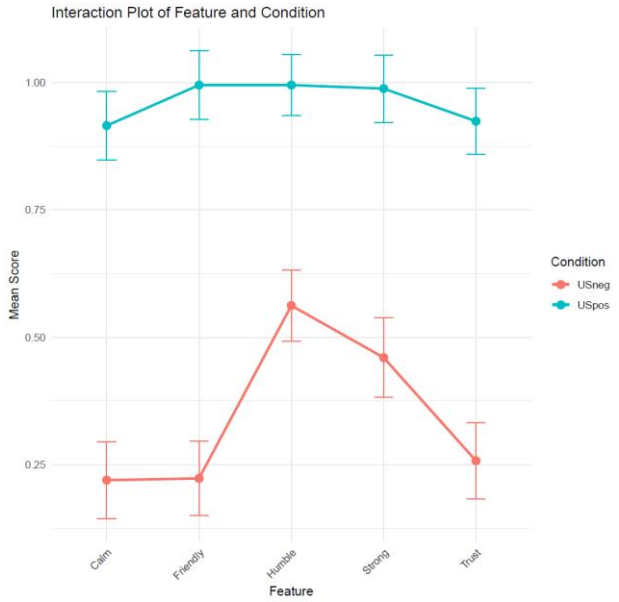
As a suggestion from the reviewers, we conducted the mixed 2 x 5 ANOVA on all dataset, without excluding participants. Just like previously, we introduced Condition (USpos for CSs paired with positive USs and USneg for CSs paired with negative USs) and the five levels of Feature (*friendly, trustworthy, strong, calm,* and *humble*), with a corrected Greenhouse-Geisser estimate (GGe) given a violation of the sphericity assumption for Features (Mauchly's W = 0.79, $p < .001$). The main effect of Condition was significant, indicating differences between the two levels, $F(1, 288) = 64.77$, $p = .001$, with a medium effect size $\eta_G^2 = 0.0637$. The main effect of Feature was also significant, indicating differences in evaluation between the different features, $F(4, 1152) = 7.99$, $p < .001$, $\eta_G^2 = 0.0048$. The interaction between Feature and Condition was also significant, with a small effect size, $F(4, 1152) = 5.17$, $p < .001$, $\eta_G^2 = 0.0026$. The posthoc tests indicated significant differences between CSpos and CSneg, suggesting that the condition significantly affected the ratings across all features. The results were similar to those conducted on the dataset after applying the exclusion criteria. Figure S5.1 depicts a visual representation of the interaction between Feature and Condition.

We also conducted a repeated-measures ANOVA to examine the effect of Condition (EC+ vs. EC−) on participants' response on the Likeability ratings, while including Valence Awareness as a covariate. As expected, the analysis revealed a significant main effect of Condition, $F(1, 2880) = 208.08$, $p < .001$, $\eta^2 = 0.07$, indicating a robust EC effect. The main effect of Valence Awareness was not significant, $F(4, 2880) = 1.49$, $p = .203$, suggesting that Valence Awareness alone did not predict participants' response. However, the interaction between Condition and Valence Awareness was significant, $F(4, 2880) = 48.57$, $p < .001$, $\eta^2 =$

0.06, indicating that the magnitude of the EC effect varied depending on participants' level of Valence.

Figure S5.1

Interaction plot between the Feature evaluations as a function of Condition



Note. Error bars represent the 95% confidence intervals.

Mediation analyses

Similarly, we conducted the mediation analyses on the dataset without excluding the participants based on the valence awareness criteria. Firstly, we examined whether the condition would predict the subsequent ratings of the additional features through the change in liking. The mediator (change in liking) was significantly predicted by Condition $b = 0.95$, $SE = 0.07$, $t(863.29) = 14.02$, $p < .001$. 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 $b = 0.27$, $SE = 0.03$, $t(1137) = 10.49$, $p < .001$, and also for all the specific CS

features: *friendliness* ($b = 0.28$, $p < .001$), *trustworthiness* ($b = 0.27$, $p < .001$), *strength* ($b = 0.29$, $p < .001$), *calm* ($b = 0.28$, $p < .001$) and *humbleness* ($b = 0.26$, $p < .001$). The results for the mediation analyses indicated that the change in liking mediated 41% of the effect of the Condition on the overall CS features score. These results were similar to those conducted on the dataset after applying the exclusion criteria. Estimated mediation effects, confidence intervals and p-values for all the models are presented in Table S5.1.

Table S5.1

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	0.25	[0.20, 0.31]	<.001
	ADE	0.37	[0.25, 0.48]	<.001
	Total effect	0.62	[0.51, 0.74]	<.001
	Proportion mediated	0.41	[0.31, 0.52]	<.001
Friendly	ACME	0.27	[0.20, 0.34]	<.001
	ADE	0.51	[0.36, 0.67]	<.001
	Total effect	0.77	[0.63, 0.92]	<.001
	Proportion mediated	0.34	[0.25, 0.46]	<.001
Trustworthy	ACME	0.26	[0.20, 0.33]	<.001
	ADE	0.41	[0.25, 0.55]	<.001
	Total effect	0.67	[0.52, 0.80]	<.001
	Proportion mediated	0.39	[0.28, 0.55]	<.001
Strong	ACME	0.22	[0.15, 0.29]	<.001
	ADE	0.31	[0.14, 0.49]	<.001
	Total effect	0.53	[0.37, 0.69]	<.001
	Proportion mediated	0.41	[0.26, 0.64]	<.001
Calm	ACME	0.27	[0.19, 0.34]	<.001
	ADE	0.43	[0.26, 0.59]	<.001
	Total effect	0.69	[0.53, 0.85]	<.001
	Proportion mediated	0.38	[0.27, 0.53]	<.001
Humble	ACME	0.25	[0.18, 0.32]	<.001
	ADE	0.19	[0.03, 0.35]	<.001
	Total effect	0.43	[0.29, 0.58]	<.001
	Proportion mediated	0.56	[0.38, 0.86]	<.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

Similarly, we also conducted the moderation analyses on the dataset with all participants, without using the valence awareness exclusion criteria. 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 effects of Condition and agreeableness were not statistically significant (b = 0.48, SE = 0.034, $t(865) = 13.97$, $p < .001$, and b = -0.0003, SE = 0.008, $t(287) = 0.41$, p = .967, respectively). And the interaction between agreeableness and Condition was not significant, b = 0.007, SE = 0.006, $t(865) = 1.07$, p = .284. Similar results were observed for agreeableness as measured by ZKPQ (Aggression-Hostility scale). In this model, Condition significantly predicted the likeability ratings b = 0.48, SE = 0.03, $t(865) = 13.98$, $p < .001$, while agreeableness was not significant b = 0.02, SE = 0.01, $t(287) = 1.61$, p = .108. Lastly, the interaction between Condition and agreeableness was not significant b = -0.01, SE = 0.01, $t(865) = -1.12$, p = .264.

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 = 0.26, SE = 0.04, $t(865) = 6.47$, $p < .001$, but not of neuroticism, b = =0.01, SE = 0.005, $t(286) = 0.28$, p = .781. The interaction between neuroticism and Condition was also significant, b = -0.01, SE = 0.005, $t(865) = -2.21$, p = .027, 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), $b = 0.22$, $SE = 0.07$, $t(709) = 3.33$, $p < .001$. The results of the model with neuroticism measured by the ZKPQ scale did not underline significance. The main effect of Condition was significant, $b = 0.26$, $SE = 0.04$, $t(865) = 6.45$, $p < .001$, while the effect of neuroticism was not, $b = 0.01$, $SE = 0.01$, $t(287) = 1.02$, $p = .309$. The interaction between Condition and neuroticism was not significant $b = -0.01$, $SE = 0.008$, $t(865) = -1.43$, $p = .153$.

Finally, while testing the moderation effect of personality on the mediation model (model 3), the results underlined that Condition, $b = 0.18$, $SE = 0.03$, $t(917.80) = 5.94$, $p < .001$, and or the likeability rating, $b = 0.26$, $SE = 0.03$, $t(1139) = 10.05$, $p < .001$, predicted the overall CS features score, but not agreeableness, $b = 0.01$, $SE = 0.008$, $t(286.10) = 1.26$, $p = .210$. However, it was observed that agreeableness did not have a significant interaction effect with the likeability rating, $b = 0.007$, $SE = 0.005$, $t(1126) = 1.43$, $p = .152$, in predicting the overall CS features score.

In conclusion, most of the effects, and close to significant effects, disappeared after testing the models on the whole dataset without applying the valence awareness criteria.

Importance of valence awareness

Additionally, we also check whether the effects of agreeableness are driven by the participants' better memory of stimulus pairings. To do so, we used the valence awareness (VA) variable coded in an ordinal manner (from 0 to 4) and Agreeableness coded as a scale to conduct a Spearman correlation. The results showed that the correlation was not significant.

Supplemental Materials Section 6 – Experiment 1: Exploratory analyses

After testing the main hypotheses, we explored whether specific feature ratings are moderated by personality.

Neuroticism, as measured by BFI, had a significant interaction effect with likeability ratings b = -1.29, SE = 4.8, t(924.80) = -2.64, p < .001, on the feature *strength* (as opposed to vulnerability). The causal mediation analysis underlined a partial mediation as both the average causal mediation effect (ACME) b = 0.25, 95% CI [0.16,0.35], p <. 001, and the average direct effect (ADE) b = 0.41, 95% CI [0.21,0.61], p <. 001 were significant. These results suggest that the effect of likeability on feature *strength* varies depending on the level of neuroticism. Specifically, as neuroticism increases, the positive relationship between likeability ratings and *strong* ratings weakens.

Another personality dimension, agreeableness, had a significant interaction effect with likeability ratings on the feature *calm*. While likeability ratings did not independently predict feature *calm* ratings, the interaction between likeability and agreeableness, as measured by BFI, was significant, b = .02, SE = .007, t(942.41) = 2.33, p < .001, suggesting that the impact of likeability on *calm* ratings varies with the levels of agreeableness. The causal mediation analysis further revealed that likeability ratings partially mediated the relationship between condition and calm ratings, with 35.3% of the total effect being mediated, as both the average causal mediation effect (ACME) b = 0.31, 95% CI [0.22,0.40], p <. 001, and the average direct effect (ADE) b = 0.57, 95% CI [0.38,0.76], p <. 001 were significant.

Supplemental Materials Section 7 – Experiment 2: Method and Procedure

Six neutral faces (all female) were retrieved 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. The hit rate ranged from 96.88 - 84.38 for CSs, and with a mean intensity higher than 5.06. Additionally, we selected other six images that served as fillers in the pre-rating phase. The pictures used as CS and as US stimuli are presented in Table S7.1.

Table S7.1

The codes of the images that were used as conditioned stimuli in the evaluative conditioning procedure

		KDEF code	Hit rate (%)	Mean (SD) intensity
Conditioned stimuli	CS1	AF06NES	96.88	5.60 (2.60)
	CS2	AF19NES	92.19	5.48 (2.09)
	CS3	AF07NES	87.50	5.36 (2.04)
	CS4	AF34NES	84.38	5.06 (2.16)
	CS5	AF01NES	85.94	5.20 (2.33)
	CS6	AF05NES	89.06	5.33 (2.44)
Filler images	F1	AF06NE	96.88	5.67 (2.24)
	F2	BM10NE	89.06	5.41 (2.42)
	F3	BM21NE	16.00	-
	F4	BM14NE	73.00	-
	F5	AF09NE	55.00	-
	F6	AM31NES	89.06	5.73 (2.15)

Note. KDEF = Karolinska Directed Emotional Faces Database; Hit rate refers to how often participants correctly recognize and label the emotional expression (e.g., neutral in this case).

Six color images (3 positive and 3 negative) were extracted from the International Affective Picture System IAPS (Bradley & Lang, 2007). The images were selected to be socially relevant, capturing real-life experiences, with an arousal value above 3.7, valence values over 7.5 (for positive), and below 2.6 (for negative). Additionally, 10 other images were selected to serve as fillers in the pre-rating phase (Table S7.2).

Table S7.2

The codes of the images that were used as unconditioned stimuli in the evaluative conditioning procedure

		IAPS Number	**Valence** (SD)	**Arousal** (SD)
Negative US	US1	6311	2.58 (1.58)	4.95 (2.27)
	US2	2345.1	2.26 (1.46)	5.50 (2.34)
	US3	2900.1	2.56 (1.41)	4.61 (2.07)
Positive US				
	US4	2209	7.64 (1.46)	5.59 (2.37)
	US5	2035	7.52 (1.33)	3.69 (2.11)
	US6	8540	7.48 (1.51)	5.16 (2.37)
Filler images	F7	1340	7.13 (1.57)	4.75 (2.31)
	F8	1710	8.34 (1.12)	5.41 (2.34)
	F9	5470	7.35 (1.62)	6.02 (2.26)
	F10	9909	2.78 (1.45)	5.98 (2.04)
	F11	6311	2.58 (1.58)	4.95 (2.27)
	F12	1525	3.09 (1.72)	6.51 (2.25)
	F13	1274	3.17 (1.53)	5.39 (2.39)
	F14	3230	2.02 (1.30)	5.41 (2.21)
	F15	2160	7.58 (1.69)	5.16 (2.18)
	F16	1463	7.45 (1.76)	4.79 (2.19)

Supplemental Materials Section 8 – Experiment 2: Data preparation

Mean scores for each CS valence (positive, negative) were computed according to the assigned block. This transformation was performed for the pre-acquisition ratings and post-acquisition ratings. Next, a change score for each valence (positive, negative) was computed by subtracting the pre-acquisition evaluations from the post-acquisition evaluations. Hence, two explicit mean rating scores resulted: negative CS and positive CS. The EC effect was reflected by an explicit score resulting from subtracting the negative CS score from the positive CS score.

We averaged a score for the features (*Friendly, Trustworthy, Strong, Calm, Humble*) for each CS (CS1, CS2, CS3, CS4, CS5, CS6). Next, in accordance with the US's valence, we computed a mean score for the negative direction CS and one for the positive CS. Hence, we obtained a score for the negative features and one for the positive features. Mean scores were also computed for each specific feature (*Friendly, Trustworthy, Strong, Calm, Humble)* by subtracting the negative score from the positive score. For the overall feature transfer effect, we computed a differential score between the positive direction of the features and the negative direction of the features. Scores for the personality scales were computed accordingly to the instrument's specifications.

Concerning the variable of demand compliance, the majority of the participants (78.34%) stated that they responded based on their personal feelings towards the stimuli. A smaller percentage of participants (14.01%) responded that they answered based on what they had learned about the stimuli in the previous task. Additionally, 5.10% of the participants stated that they did not know why they responded in a particular way, while 2.55% said that they answered based on what they thought the researcher expected from them.

We further investigated the last category of four participants to determine if they were aware of the study objective. Based on their responses, we concluded that they were not aware of the experimenter's objectives, and therefore we did not exclude them from the analyses.

Supplemental Materials Section 9 – Experiment 2: Descriptive statistics

Prior to analysing the data, we explored the distributions of each variable. In the case of likeability ratings, feature ratings, and the personality subscales we tested the normality assumption by inspecting the visual representations, the symmetry (*Skewness*), and the *pointiness (Kurtosis)* of the data. We considered values greater than -1.96 and lesser than 1.96 to be normally distributed (Table S9.1).

Table S9.1

Descriptive statistics for features ratings

Variable	M (SD)	Skewness	Kurtosis
Likeability	.60 (1.34)	.55	-.02
Friendly	.39 (1.40)	.43	.21
Warm	.35 (1.38)	.51	.44
Sincere	.43 (1.33)	.52	.79
Trustworthy	.57 (1.38)	.81	.65
Competent	.29 (1.33)	.36	.30
Emotionally stable	.15 (1.56)	.55	.64
Strong	.15 (1.56)	.55	.64
Calm	.35 (1.57)	.62	.11
Humble	.60 (1.47)	.33	1.03

Note. N = 157; M = mean; SD = standard deviation; Positive = CSs paired with positive USs; Negative – CSs paired with negative USs.

Table S9.2 provides details about the psychometric proprieties of the personality questionnaire, while Table S9.3 presents the correlation values between the ratings of features and personality scales.

Table S9.2

Descriptive statistics for HEXACO scales and subscales

Measure	M	SD	α	Skewness	Kurtosis
Honest Humility	34.19	6.45	.72	-.15	-.23
Sincerity	11.08	3.05	.74		
Fairness	10.73	2.91	.57		
Greed avoidance	5.11	1.79	.52		
Modesty	7.26	1.78	.55		
Agreeableness	31.17	6.49	.76	-.40	.09
Forgiveness	6.07	2.08	.57		
Gentleness	9.45	2.75	.67		
Flexibility	8.97	2.34	.59		
Patience	6.67	2.01	.69		
Emotionality	34.86	6.53	.74	-.40	-.49
Fearfulness	10.16	2.63	.45		
Anxiety	7.83	2.09	.74		
Sentimentality	6.10	2	.68		
Dependence	10.77	2.51	.54		
Extraversion	30.46	7.56	.84	-.13	-.61
Social self-esteem	9.05	2.86	.67		
Social boldness	9.04	2.83	.73		
Sociability	6.19	2	.60		
Liveliness	6.18	1.77	.47		
Consciousness	34.62	6.48	.77	-.09	-.46
Organization	6.55	2.26	.74		
Diligence	7.22	1.57	.45		
Perfectionism	11.80	2.33	.60		
Prudence	9.06	2.74	.67		
Openness	39.07	5.55	.70	-.46	-.10
Aesthetic appreciation	7.90	1.81	.55		
Inquisitiveness	6.99	1.82	.28		
Creativity	12.33	2.42	.66		
Unconventionality	11.85	2.03	.42		

Note. N = 157; M = mean; SD = standard deviation; α = Alpha Cronbach

Table S9.3

Correlations between the ratings of features and personality scales

Variable	1	2	3	4	5	6	7	8	9	10	11	12	13	14	15	16
Likeability	1															
Friendly	**.65**	1														
Warm	**.61**	**.70**	1													
Sincere	**.61**	**.64**	**.66**	1												
Trustworthy	**.74**	**.68**	**.72**	**.83**	1											
Competent	**.60**	**.58**	**.50**	**.49**	**.59**	1										
Emotionally stable	**.52**	**.46**	**.45**	**.48**	**.57**	**.67**	1									
Strong	**.50**	**.46**	**.45**	**.48**	**.57**	**.67**	**1**	1								
Calm	**.44**	**.50**	**54**	**.48**	**.59**	**.62**	**.67**	**.67***	1							
Humble	.07	**.47**	**.47**	**.55**	**.57**	**.44**	**.35**	**.35**	**.55**	1						
Emotionality[a]	-.01	.03	-.01	.11	.05	-.01	-.07	-.07	.01	-.06	1					
Agreeableness[a]	-.01	.04	.06	.06	.09	-.01	.04	.04	-.08	.11	**-.18**	1				
Honest-Humility[a]	-.11	.03	**.14**	-.04	-.09	-.01	-.05	-.05	-.12	-.04	-.12	**.36**	1			
Extraversion[a]	.06	-.01	-.01	.01	.04	.04	.02	.02	-.10	.03	**-.41**	**.18**	.01	1		
Consciousness[a]	**.16**	.10	.07	.06	**.16**	**.22**	**.21**	**.21**	**.18**	.02	**-.22**	.05	**.15**	.13	1	
Openness[a]	.07	.13	.11	.02	.07	.03	.11	.11	.05	.06	**-.18**	.13	.04	.08	**.23**	1

Note. N = 157; a = HEXACO personality dimensions; Values in bold suggest that the correlation is significant at the 0.01 level (2-tailed); Values in bold and italic suggest that the correlation is significant at the 0.05 level (2-tailed).

Supplemental Materials Section 10 – Experiment 2: Analyses with all participants

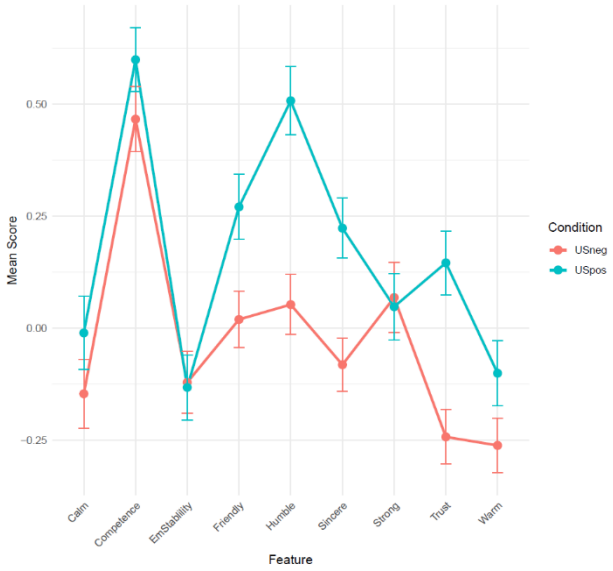
Mixed ANOVA

Just like for Experiment 1, we conducted the analyses on the dataset without excluding the participants based on the valence awareness criteria. Hence, 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 and USneg) on the participant's ratings. Condition had a significant main effect, $F(1, 210) = 7.54$, $p < .001$, $\eta_G^2 = 0.009$. The main effect of Feature was also significant, indicating differences in evaluation between the different features, $F(8, 1680 = 33.14$, $p < .001$, $\eta_G^2 = 0.04$. The interaction between Feature and Condition (Figure S10.1) was also significant, with a small effect size, $F(8, 1680) = 5.91$, $p < 001$, $\eta_G^2 = 0.006$. The results were similar to those conducted on the dataset after applying the exclusion criteria. A visual representation of the interaction between Feature and Condition is depicted in Figure S10.1.

Similarly, as we did for the first experiment, we conducted a repeated-measures ANOVA to examine whether the magnitude of the EC effect varied depending on the participants' level of Valence Awareness. The results were similar but with some differences. Condition had a main significant effect, $F(1, 3784) = 36.89$, $p < .001$, $\eta^2 = 0.01$, while the main effect of Valence Awareness was also significant, $F(6, 3784) = 6.70$, $p < .001$, $\eta^2 = 0.01$. Likewise, the interaction between Condition and Valence Awareness was significant, $F(6, 3784) = 34.31$, $p < .001$, $\eta^2 = 0.05$. This might further support the idea that participants with lower valence awareness experience a weaker EC effect, which in turn reduces the likelihood of detecting moderating relationships with other factors such as personality traits.

Figure S10.1

Interaction plot between the Feature evaluations as a function of Condition



Note. Error bars represent the 95% confidence intervals.

Mediation analyses

Before testing the mediation models, we first checked whether the 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 $b = 0.44$, $SE = 0.12$, $t(4.00) = 3.57$, $p = .023$.

The second set of models underlined that the change in liking predicted the overall CS features score (composite score of the nine features, $\alpha = .91$), $b = 0.37$, $SE = 0.02$, $t(1261) = 22.19$, $p < .001$, and the specific CS features, *friendly* ($b = 0.52$, $p < .001$), *warmth* ($b = 0.46$, $p < .001$), *sincerity* ($b = 0.41$, $p < .001$), *trustworthiness* ($b = 0.51$, $p < .001$), *competence* ($b = 0.32$, $p < .001$), *emotional stability* ($b = 0.25$, $p < .001$), *strength* ($b = 0.21$, $p < .001$), *calm* ($b = 0.29$, $p < .001$) and *humbleness* ($b = 0.30$, $p < .001$). The change in liking mediated 79% of the effect of the condition on the overall CS features score. The results were similar to those conducted on the dataset after applying the exclusion criteria. More information is presented in Table S10.2.

Table S10.2

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	0.16	[0.15, 0.29]	<.001
	ADE	0.04	[-0.05, 0.13]	0.42
	Total effect	0.20	[0.09, 0.31]	<.001
	Proportion mediated	0.79	[0.53, 1.48]	<.001
Friendliness	ACME	0.22	[0.14, 0.31]	<.001
	ADE	0.02	[-0.11, 0.16]	0.33
	Total effect	0.25	[0.08, 0.41]	<.01
	Proportion mediated	0.91	[0.56, 2.30]	<.01
Warmth	ACME	0.21	[0.13, 0.28]	<.001
	ADE	-0.05	[-0.19, 0.09]	0.50
	Total effect	0.16	[0.01, 0.31]	<.05
	Proportion mediated	1.25	[0.57, 7.09]	<.05
Sincerity	ACME	0.18	[0.11, 0.25]	<.001
	ADE	0.13	[0.01, 0.26]	<.05
	Total effect	0.30	[0.17, 0.46]	<.001
	Proportion mediated	0.58	[0.38, 1.00]	<.001
Trustworthiness	ACME	0.22	[0.15, 0.30]	<.001
	ADE	0.16	[0.04, 0.29]	<.01
	Total effect	0.39	[0.24, 0.54]	<.001
	Proportion mediated	0.57	[0.40, 0.86]	<.001
Competence	ACME	0.14	[0.09, 0.20]	<.001
	ADE	-0.05	[-0.15, 0.15]	0.93
	Total effect	0.14	[-0.02, 0.29]	.082
	Proportion mediated	0.96	[-2.76, 6.01]	.082
Emotionally stable	ACME	0.11	[0.07, 0.16]	<.001
	ADE	-0.11	[-0.28, 0.07]	0.21
	Total effect	-0.01	[-0.18, 0.17]	0.91
	Proportion mediated	-0.57	[-21.51, 21.57]	0.91
Strong	ACME	0.09	[0.05, 0.14]	<.001
	ADE	-0.11	[-0.29, 0.07]	0.22
	Total effect	-0.02	[-0.21, 0.17]	0.77
	Proportion mediated	-0.62	[-11.16, 12.41]	0.77
Calm	ACME	0.13	[0.08, 0.19]	<.001
	ADE	0.05	[-0.17, 0.19]	0.96
	Total effect	0.13	[-0.05, 0.32]	0.15
	Proportion mediated	0.84	[-6.84, 11.01]	0.15
Humble	ACME	0.13	[0.08, 0.19]	<.001
	ADE	0.32	[0.16, 0.50]	<.001
	Total effect	0.46	[0.28, 0.63]	<.001
	Proportion mediated	0.29	[0.17, 0.49]	<.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

Similarly, we conducted the moderation models on the dataset with all participants. 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 showed that when we included conscientiousness in the model, it presented an interaction with Condition.

The main effect of Condition was significant in predicting the likeability ratings, b = 0.22, SE = 0.39, $t(1053) = 5.63$, $p < 0.001$, while the effect of conscientiousness was close to significance, b = -0.02, SE = 0.01, $t(209) = -1.79$, $p = .075$. However, its interaction with Condition was not significant $b = 0.008$, $SE = 0.006$, $t(1053) = 1.30$, $p = .193$.

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).

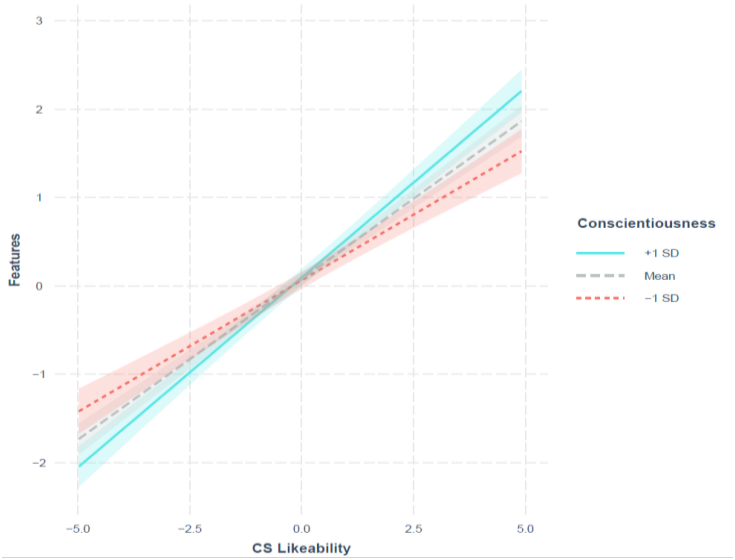
Lastly, while exploring the moderation effect of personality on the mediation model (model 3), the results underlined that neuroticism and agreeableness do not have a significant role in this context, neither on their own and neither in interaction with likeability ratings. But, again, conscientiousness seems to have an interaction with the CS likeability ratings. The main effects of conscientiousness ($b = 0.002$, $SE = 0.005$, $t(207.30) = 0.46$, $p = .642$) and Condition ($b = 0.02$, $SE = 0.02$, $t(1058) = 0.68$, $p = .493$), but likeability reached statistical significance $b = 0.36$, $SE = 0.02$, $t(1260) = 21.95$, $p < .001$. Conscientiousness had a significant interaction effect with the likeability ratings, $b = 0.01$, $SE = 0.001$, $t(1259) = 4.14$, $p < .001$, in predicting the overall CS features score. Figure S11.1 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), $b = 0.54$, $p < .001$, the CS likeability ratings had a stronger

effect on the overall CS features compared to individuals with a lower level of conscientiousness, (-1SD), $b = 0.34$, $p < .001$.

In conclusion, most effects, including those approaching significance, disappeared when the models were tested on the entire dataset without applying the valence awareness criteria.

Figure S11.1

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.

Importance of valence awareness

We also checked whether the effects of agreeableness were driven by the participants' better memory of stimulus pairings, and the results indicated that the correlation was not significant.

Supplemental Materials Section 11 – Experiment 2: Exploratory analyses

After testing the second main hypothesis, we explored whether personality had a moderation role in the mediation model when considering specific features as the dependent variable. Similar analyses were conducted, but in this case, the features were grouped per dimension as follows: *Warmth* – friendliness, warmth; *Morality* - trustworthiness, sincerity; *Ability* – competence; *Emotionality* - emotional stability, strength; *Agreeableness* – calm; *Honest-Humility* – humbleness).

Even though personality had no significant interaction effect with Condition while predicting likeability ratings, it was observed that there is an interaction effect with likeability ratings when predicting specific features. For example, the evaluations for the dimension *Warmth* were impacted by neuroticism in interaction with likeability ratings, $b = .01$, $SE = .003$, $t(869.43) = 2.41$, $p < .05$. In this specific case, there was also found a full mediation, as only the average causal mediation effect (ACME) $b = 0.29$, 95% CI [0.19, 0.39], p<.001, was significant.

Another feature dimension, *Emotionality*, was moderated by personality in interaction with the likeability ratings. More specifically, agreeableness, $b = .02$, $SE = .004$, $t(862.16) = 2.40$, $p < .05$ consciousness, $b = .02$, $SE = .005$, $t(891.19) = 2.36$, $p < .05$ and openness, $b = .02$, $SE = .006$, $t(877.92) = 2.03$, $p < .05$ presented significant effects. Regarding the moderated mediation, we found that only the average causal mediation effect (ACME) was significant in all three models: $b = 0.16$, 95% CI [0.12,0.22], p<.001 for agreeableness, $b = 0.15$, 95% CI [0.11, 0.21], $p < .001$ for consciousness and $b = 0.15$, 95% CI [0.09, 0.21], $p < .001$ for openness, but not the total effect.

A significant moderation effect was also observed in the case of consciousness in interaction with likeability ratings while predicting the *Agreeableness* feature dimension: $b = .02$,

SE = .005, *t*(929.27) = 2.65, *p* < .01. The average causal mediation effect (ACME) was significant (*b* = 0.18, 95% CI [0.12, 0.26], *p* < 0.001), indicating that likeability ratings partially mediate the relationship between condition and the agreeableness feature dimension. This suggests that changes in likeability due to the condition contribute to changes in this dimension. The average direct effect (ADE) was not significant (*b* = 0.17, 95% CI [-0.05, 0.37], *p* = 0.13), indicating the importance of likeability as a mediator in this relationship. The proportion of the total effect mediated by likeability ratings was significant (*b* = 0.52, 95% CI [0.28, 1.34], *p* < 0.01), indicating that approximately 52% of the condition's total effect on the agreeableness feature dimension is mediated by changes in likeability.

Supplemental Materials Section 12 – Power simulations

Mediation analyses in Experiment 1 and Experiment 2

For the key effects of the mediation models power analysis was calculated based on the Monte Carlo Power Analysis for Indirect Effects (Schoemann et al., 2017).

For the first experiment, we used the correlation coefficients obtained in the pilot study: *r*(Condition-Likeability) = 0.55; *r*(Condition-Features) = 0.37; *r*(Likeability - Features) = 0.43 (with an overall average of *r* = 0.45). Fixing power at 0.80, the required number of participants was 88.

For the second experiment, we used the correlation values of the first experiment: *r*(Condition-Likeability) = 0.39; *r*(Condition-Features) = 0.33; *r*(Likeability - Features) = 0.38, with an overall average of *r* = 0.37. In this case, the simulation indicated that we needed an N of 99 participants. Both values are well below the actual sample sizes of the two experiments. Note that the approach of Schoemann et al. (2017) has been developed for a between-subjects independent variable, whereas in our experiments the independent variable was within-subjects.

However, a recent contribution (Montoya, 2023) underlined that within-subject designs tend to have higher empirical power than between-subject designs for mediation, and generally, they require about half of the participants to achieve the same level of power, everything else being equal. Therefore, we can conclude that our experiments were sufficiently powered to detect the hypothesized mediation effects.

Moderation analyses

The sensitivity analyses were conducted based on the R code and information in the Supplementary Materials of Casini et al. (2023). Hence, we assessed the sensitivity of our study

to detect the key interaction effects by using a simulation-based approach, mean-centering the predictors included in the models.

Experiment 1

For **Model 1: Neuroticism** (using the BFI scale of neuroticism; BFI-2; Soto & John, 2017), we fixed all other parameters of the multilevel model, based on the sample size of N = 237, and simulated the smallest detectable effect size with 80% power at $\alpha = .05$. For the interaction effect (rating_Like × BFI_Neuroticism), the sensitivity analysis indicated that the study was sufficiently powered to detect an effect of B = -0.0101. Notably, while the observed interaction effect was in the expected direction, it did not reach statistical significance (B = -0.0037).

For **Model 2: Agreeableness** (using the BFI scale of agreeableness; BFI-2; Soto & John, 2017), the sensitivity analysis indicated that the study was sufficiently powered to detect an effect of B = 0.0166. In our study, the interaction effect was not significant (B = 0.0106) but it was in the expected direction.

Model 1: Neuroticism

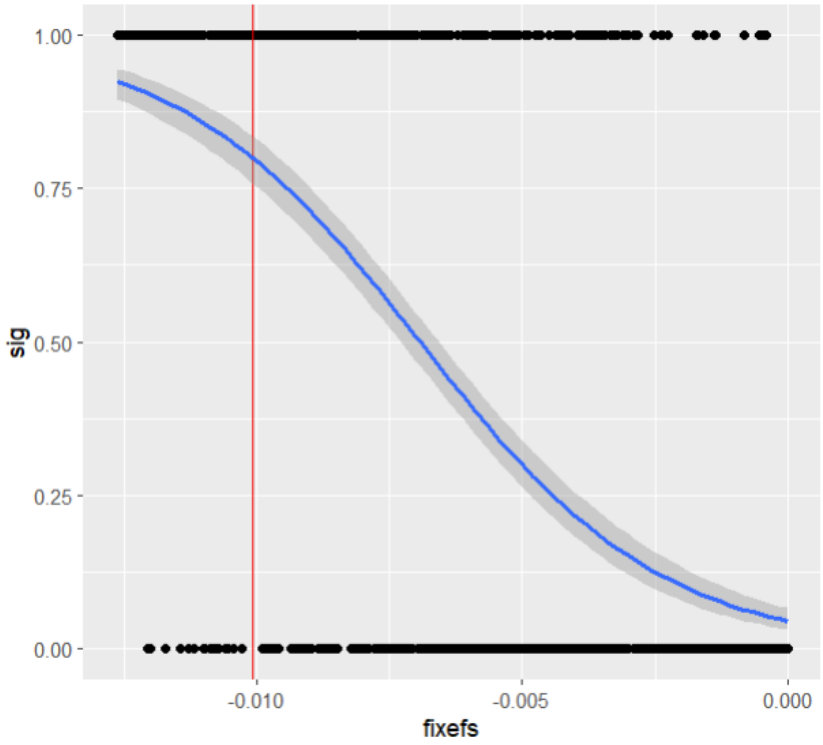
Reproduce results for BFI Neuroticism

```
fit.model <- lmer(Features ~ rating_Like_c * BFI_N_c + Condition + (1 | ID), data =
db.long_clean)
smm <- summary(fit.model)
ci <- confint(fit.model)

smm2 <- smm$coefficients %>%
  round(4) %>%
  data.frame %>%
  select(B = Estimate, SE = Std..Error, t = t.value)

ci2 <- ci %>%
  round(4) %>%
  data.frame %>%
  .[-c(1:2),]
cbind(smm2, ci2)
```

```
                          B     SE       t  X2.5.. X97.5..
(Intercept)           0.6589 0.0463 14.2184  0.5682  0.7496
rating_Like_c         0.2717 0.0278  9.7607  0.2170  0.3262
BFI_N_c              -0.0055 0.0067 -0.8170 -0.0187  0.0077
Condition             0.2405 0.0351  6.8505  0.1718  0.3094
rating_Like_c:BFI_N_c -0.0037 0.0035 -1.0499 -0.0105  0.0032
```

Power analysis with Simpower

```
set.seed(123)
sensitivity <- Simpower(fit.model, effect = "rating_Like_c:BFI_N_c", power = .80, B =
B, logplot = TRUE)
sensitivity$es
## [1] -0.01007473
sensitivity$plot
```



Check the resulted power with the specific effect size value, using simr

```
modPow <- fit.model
fixef(modPow)["rating_Like_c:BFI_N_c"] <- sensitivity$es
set.seed(123)
ps <- powerSim(modPow,
               test = fixed("rating_Like_c:BFI_N_c", method = "t"),
               nsim = B, progress = FALSE)
ps
```

```
Power for predictor 'rating_Like_c:BFI_N_c', (95% confidence interval):
      83.20% (80.74, 85.47)

Test: t-test with Satterthwaite degrees of freedom (package lmerTest)
      Effect size for rating_Like_c:BFI_N_c is -0.010

Based on 1000 simulations, (0 warnings, 0 errors)
alpha = 0.05, nrow = 948

Time elapsed: 0 h 3 m 10 s
```

Model 2: Agreeableness

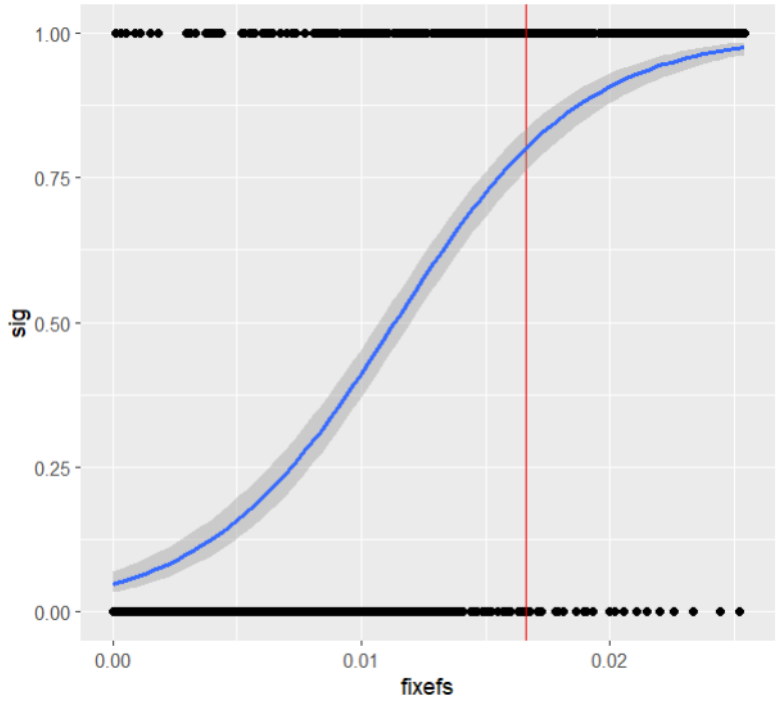
Reproduce results for BFI Agreeableness

```
fit.model2 <- lmer(Features ~ rating_Like_c * BFI_A_c + Condition + (1 | ID), data =
db.long_clean)
smm <- summary(fit.model2)
ci <- confint(fit.model2)

smm2 <- smm$coefficients %>%
  round(4) %>%
  data.frame %>%
  select(B = Estimate, SE = Std..Error, t = t.value)

ci2 <- ci %>%
  round(4) %>%
  data.frame %>%
  .[-c(1:2),]

cbind(smm2, ci2)
```

```
                          B     SE       t  X2.5.. X97.5..
(Intercept)           0.6569 0.0459 14.3123  0.5670  0.7467
rating_Like_c         0.2649 0.0280  9.4711  0.2099  0.3196
BFI_A_c               0.0172 0.0090  1.9153 -0.0004  0.0347
Condition             0.2418 0.0351  6.8946  0.1732  0.3106
rating_Like_c:BFI_A_c 0.0106 0.0058  1.8322 -0.0007  0.0219
```

Power analysis with Simpower

```
set.seed(123)
sensitivity <- Simpower(fit.model, effect = "rating_Like_c:BFI_A_c", power = .80, B =
B, logplot = TRUE)
sensitivity$es
## [1] 0.01660559

sensitivity$plot
```



Check the resulted power with the specific effect size value, using simr

```
modPow <- fit.model2
fixef(modPow)["rating_Like_c:BFI_A_c"] <- sensitivity$es
set.seed(123)
ps <- powerSim(modPow,
               test = fixed("rating_Like_c:BFI_A_c", method = "t"),
               nsim = B, progress = FALSE)
ps
```

```
Power for predictor 'rating_Like_c:BFI_A_c', (95% confidence interval):
      82.10% (79.58, 84.43)

Test: t-test with Satterthwaite degrees of freedom (package lmerTest)
      Effect size for rating_Like_c:BFI_A_c is 0.017

Based on 1000 simulations, (0 warnings, 0 errors)
alpha = 0.05, nrow = 948

Time elapsed: 0 h 3 m 12 s
```

Experiment 2

For **Model 3: Neuroticism** (using the HEXACO scale for neuroticism; HEXACO–60; Ashton & Lee, 2009), we fixed all other parameters of the multilevel model, based on the sample size of N = 157, and simulated the smallest detectable effect size with 80% power at α = .05. For the interaction effect (rating_Like × Emotionality), the sensitivity analysis indicated that the study was sufficiently powered to detect an effect of B = 0.0081. In this study, the observed interaction effect was in the expected direction, but it did not reach statistical significance (B = -0.002).

For **Model 4: Agreeableness** (using the HEXACO scale for agreeableness; HEXACO–60; Ashton & Lee, 2009), the sensitivity analysis indicated that the study was sufficiently powered to detect an effect of B = 0.008. In our study, the interaction effect was not significant (B = 0.0024), but, again, it was in the expected direction.

Model 3: Neuroticism

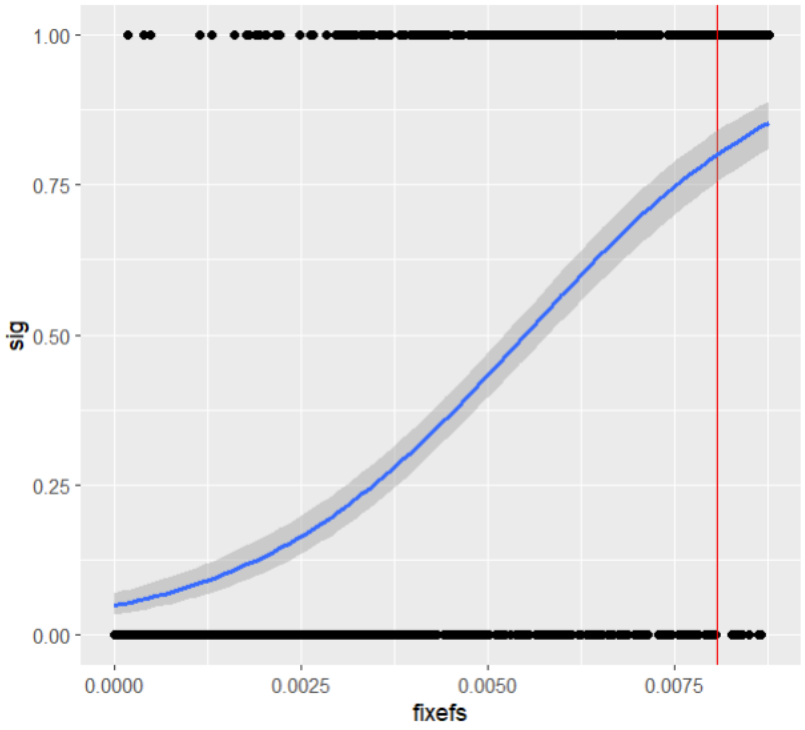
Reproduce results for HEXACO Emotionality

```
fit.model <- lmer(Features ~ rating_Like_c * Emotionality_c + Condition + (1|subject),
data = db.long_clean)
smm <- summary(fit.model)
ci <- confint(fit.model)

smm2 <- smm$coefficients %>%
  round(4) %>%
  data.frame %>%
  select(B = Estimate, SE = Std..Error, t = t.value)

ci2 <- ci %>%
  round(4) %>%
  data.frame %>%
  .[-c(1:2),]

cbind(smm2, ci2)
```

```
                                B     SE       t  X2.5.. X97.5..
(Intercept)                 0.0731 0.0385  1.8997 -0.0022  0.1484
rating_Like_c               0.3671 0.0192 19.1677  0.3286  0.4049
Emotionality_c             -0.0018 0.0059 -0.3044 -0.0134  0.0098
Condition                   0.0719 0.0285  2.5242  0.0162  0.1279
rating_Like_c:Emotionality_c 0.0015 0.0028  0.5339 -0.0040  0.0070
```

Power analysis with Simpower

```
set.seed(123)
sensitivity <- Simpower(fit.model, effect = "rating_Like_c:Emotionality_c", power =
.80, B = B, logplot = TRUE)
sensitivity$es
## [1] 0.008056877
sensitivity$plot
```



Check the resulted power with the specific effect size value, using simr

```
modPow <- fit.model
fixef(modPow)["rating_Like_c:Emotionality_c"] <- sensitivity$es
set.seed(123)
ps <- powerSim(modPow,
               test = fixed("rating_Like_c:Emotionality_c", method = "t"),
               nsim = B, progress = FALSE)
ps
```

```
Power for predictor 'rating_Like_c:Emotionality_c', (95% confidence interval):
      77.00% (74.26, 79.58)

Test: t-test with Satterthwaite degrees of freedom (package lmerTest)
      Effect size for rating_Like_c:Emotionality_c is 0.0078

Based on 1000 simulations, (0 warnings, 0 errors)
alpha = 0.05, nrow = 942

Time elapsed: 0 h 3 m 20 s
```

Model 4: Agreeableness

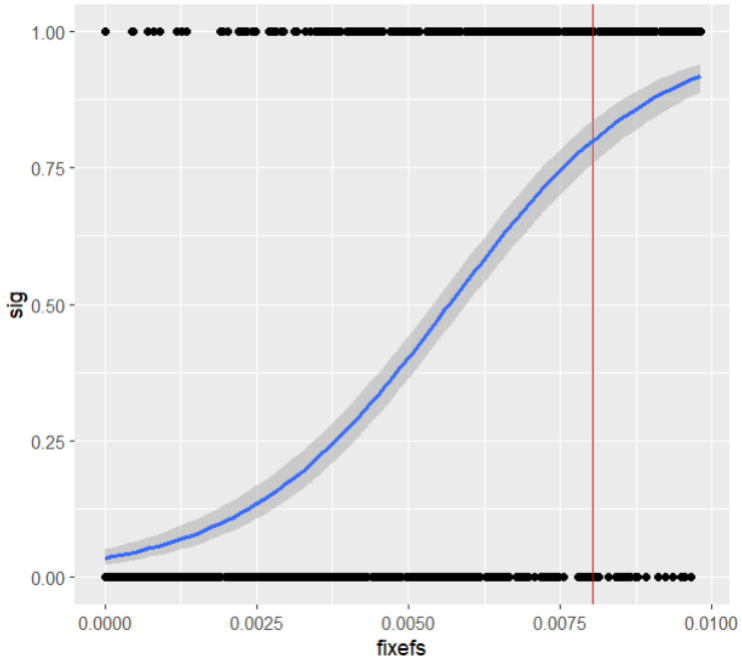
Reproduce results for HEXACO Agreeableness

```
fit.model2 <- lmer(Features ~ rating_Like_c * Agreeableness_c + Condition +
(1|subject), data = db.long_clean)
smm <- summary(fit.model2)
ci <- confint(fit.model2)

smm2 <- smm$coefficients %>%
  round(4) %>%
  data.frame %>%
  select(B = Estimate, SE = Std..Error, t = t.value)

ci2 <- ci %>%
  round(4) %>%
  data.frame %>%
  .[-c(1:2),]

cbind(smm2, ci2)
```

```
                                B     SE       t  X2.5.. X97.5..
(Intercept)                 0.0722 0.0384  1.8784 -0.0030  0.1474
rating_Like_c               0.3683 0.0191 19.2368  0.3298  0.4062
Agreeableness_c             0.0010 0.0059  0.1707 -0.0106  0.0127
Condition                   0.0716 0.0285  2.5137  0.0159  0.1275
rating_Like_c:Agreeableness_c 0.0024 0.0029  0.8467 -0.0032  0.0080
```

Power analysis with Simpower

```
set.seed(123)
sensitivity <- Simpower(fit.model, effect = "rating_Like_c:Agreeableness_c", power =
.80, B = B, logplot = TRUE)
sensitivity$es
## [1] 0.008042314
sensitivity$plot
```



Check the resulted power with the specific effect size value, using simr

```
modPow <- fit.model2
fixef(modPow)["rating_Like_c:Agreeableness_c"] <- sensitivity$es
set.seed(123)
ps <- powerSim(modPow,
               test = fixed("rating_Like_c:Agreeableness_c", method = "t"),
               nsim = B, progress = FALSE)
ps
```

```
Power for predictor 'rating_Like_c:Agreeableness_c', (95% confidence interval):
      81.10% (78.53, 83.48)

Test: t-test with Satterthwaite degrees of freedom (package lmerTest)
      Effect size for rating_Like_c:Agreeableness_c is 0.0080

Based on 1000 simulations, (0 warnings, 0 errors)
alpha = 0.05, nrow = 942

Time elapsed: 0 h 3 m 18
```

References

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

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.

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. Supplementary material [DOCX file]. https://collabra.scholasticahq.com/article/74820-the-moderating-role-of-neuroticism-on-evaluativeconditioning-new-insights-on-the-processes-underlying-this-relationship/attachment/157726.docx?auth_token=soRboJaYvByhXbKH_Gex

Costa, P. T., & McCrae, R. R. (1992). Normal personality assessment in clinical practice: The NEO Personality Inventory. *Psychological Assessment, 4*(1), 5–13. doi: 10.1037/1040-3590.4.1.5

Dardas, L. A., & Ahmad, M. M. (2015). For fathers raising children with autism, do coping strategies mediate or moderate the relationship between parenting stress and quality of life?. *Research in developmental disabilities*, *36*, 620-629. doi: https://doi.org/10.1016/j.ridd.2014.10.047

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

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. doi: 10.1080/02699930701626582

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

Lee, K., & Ashton, M. C. (2018). Psychometric properties of the HEXACO-100. *Assessment, 25*(5), 543–556. doi: https://doi.org/10.1177/1073191116659134

Millisecond Software (2015). Inquisit 5 Iowa Gambling Task [Computer software]. Retrieved from https://www.millisecond.com.

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

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