Can (instructions about) stimulus pairings influence automatic and self-reported evaluations in the presence of more diagnostic evaluative information?

Tal Moran[a], Pieter Van Dessel[a], Colin Tucker Smith[b], & Jan De Houwer[a]

[a]*Ghent University, Belgium*

[b]*University of Florida, USA*

In press: Personality and Social Psychology Bulletin

Evaluative Conditioning (EC) and persuasion are important pathways for shaping evaluations. However, little is known about how these pathways interact. Two preregistered experiments (total N=1,510) examined effects of EC procedures (i.e., stimulus pairings) and EC instructions (i.e., instructions about stimulus pairings) on automatic and self-reported evaluations of social groups in the presence of more diagnostic information about the evaluative traits of those groups. Interestingly, both EC procedures and EC instructions still influenced automatic and self-reported evaluations when participants had read more diagnostic persuasive information. In line with predictions of propositional accounts of evaluation, EC instruction effects on automatic evaluations were not mediated by corresponding changes in self-reported evaluations. These results have theoretical implications and also highlight the important role that (instructions about) stimulus pairings have in social learning.

Keywords: Evaluative Conditioning; Persuasion; Instructions; Automatic Evaluation; Diagnosticity

One of the leading ideas in psychology is that evaluations are important determinants of behavior (e.g., Allport, 1935). People's likes and dislikes seemingly exert a strong influence on bhavior in many contexts such as racial prejudiceand discrimination (e.g., Fiske, 1998), consumption (e.g., Berger & Mitchell, 1989) and political behavior (e.g., Gawronski et al., 2015). Given that evaluations are presumed to guide behavior, a long tradition of research exists in examining how evaluations are acquired and how they can be changed.

Evaluative Conditioning and Persuasion

One of the most studied pathways in this regard is Evaluative Conditioning (EC; De Houwer et al., 2001), which, as a procedure, refers to the repeated pairing of stimuli. Extensive research has found that when a neutral conditioned stimulus (CS) repeatedly occurs with either a positive or negative unconditioned stimulus (US), the evaluation of the CS changes, typically toward the valence of the US (i.e., an EC effect; Hofmann et al., 2010). Indeed, research suggests that EC effects may play an important role in many social and personal phenomena such as racial bias (e.g., Olson & Fazio, 2006), disgust sensitivity (e.g., Schienle et al., 2001) and consumption behavior (e.g., Gibson, 2008). EC explanations may be proposed for a variety of real-world social phenomena such as how features like source attractiveness or credibility influence the perceived positivity or negativity of persuasive messages (e.g., Petty & Cacioppo, 1984), why individuals who are perceived to be in a social relationship with stigmatized persons (i.e., obese people) are derogated (Hebl & Mannix, 2003), or why positive evaluation of the self may transfer to objects belonging to the self (Beggan, 1992). In applied contexts, EC procedures are often used as a method for changing problematic behaviors related to alcohol use (e.g., Houben et al., 2010), unhealthy food consumption (e.g., Shaw et al., 2016), and racism (e.g., Olson & Fazio, 2006).

In research on the formation of evaluations, stimulus pairings are often contrasted with a second learning

TM, PVD, and JDH, Department of Experimental-Clinical and Health Psychology, Ghent University, Dunantlaan 2, 9000 Ghent, Belgium. CTS, Department of Psychology, University of Florida, PO Box 112250, Gainesville, FL 32611, USA. Correspondence concerning this article should be addressed to Tal.MoranYorovich@UGent.be. This manuscript is supported by Ghent University grant BOF16/MET_V/002 to JDH and by grant FWO19/PDS/041 of the Scientific Research Foundation, Flanders to PVD. The authors declare that they have no competing interests. All materials, data and analyses scripts are available at osf.io/athds/.

pathway that involves the presentation of persuasive information (Petty & Cacioppo, 1986). Historically, effects of EC procedures were often seen as fundamentally different from persuasion via arguments (for a discussion, see De Houwer & Hughes, 2016). Indeed, there are important differences between these two pathways at the procedural level: whereas persuasion uses verbal information, EC involves stimulus pairings. Important differences were also assumed at the mental level. For instance, it has been argued that learning via persuasion is deliberative and effortful and involves the formation of propositional representations whereas learning via EC procedures constitutes a more primitive pathway that involves an automatic formation of associations in memory and therefore requires little thought and cognitive effort (Briñol et al., 2009). Some dual-process models of evaluation (e.g., Gawronski & Bodenhausen, 2006; Rydell & McConnell, 2006; Smith & DeCoster, 2000) also map this division onto the division between self-reported (explicit) and automatic (implicit) evaluations such that persuasive messages are assumed to influence self-reported evaluation based on propositional information whereas EC procedures would influence automatic evaluation via the formation of mental associations.

In recent years, propositional theories (De Houwer, 2014; see De Houwer et al., 2020, for a review) have provided a different view on EC and persuasion effects. According to this perspective, both effects are mediated by the formation and activation of propositional representations in memory. In contrast to associations, propositions are units of information that specify how events are related. For example, they can represent that ‘the US is positive’ rather than simply link representations of the concept ‘US’ and the concept ‘positive’. In addition, they have a truth value in that they can be evaluated as being true or false. According to the propositional perspective, both EC procedures and persuasive messages provide information that leads to the formation of propositions. Once these propositions have been formed, they can influence both self-reported and automatic evaluations (De Houwer, 2014). Automatic and self-reported evaluation might, however, differ if they reflect activation of different propositions under different measurement conditions (De Houwer et al., 2020). For example, self-report measures typically provide ample opportunity to disregard beliefs that are considered to be irrelevant or invalid. Measures designed to capture automatic evaluation, on the other hand, typically require fast responding, thus making it more difficult to control which beliefs have an impact on responding.

Propositional theories also highlight that the information provided by stimulus pairings could in principle also be conveyed in a verbal manner (i.e., via instructions about CS-US pairings). Such EC instructions can result in similar propositions and thus should have a similar impact on evaluations as the actual experience of CS-US pairings. Indeed, recent research has demonstrated that automatic evaluations can form and change as a result of EC instructions even when CS-US pairings are never directly experienced (De Houwer, 2006; Gast & De Houwer, 2013; Smith et al., 2020). Intriguingly, evidence suggests that actual pairings do not produce stronger changes in automatic evaluations than EC instructions and that actual pairings do not add to effects of EC instructions (Kurdi & Banaji, 2017). These observations of potent effects of EC instructions on automatic evaluation pose a challenge to models that assume that automatic evaluation only reflects representations that are formed as the result of a slow-learning process that requires the repeated pairing of stimuli (Rydell & McConnell, 2006; Smith & DeCoster, 2000).

Although propositional theories highlight the similarities between EC and persuasion, they do allow for important differences between both learning pathways. Specifically, information that is considered highly diagnostic for inferring stimulus valence should lead to stronger effects on evaluation (Van Dessel, Cone, et al., 2020). In the context of social stimuli, diagnosticity has been defined as the extent to which information is revealing of a person's true nature or character (Cone & Ferguson, 2015). This might often be higher for persuasive information such as trait instructions (e.g., information that a politician is truthful) than for stimulus pairings (e.g., pairing a politician with cuddly babies in an advertisement). These ideas lead to interesting predictions regarding interactions between persuasion and EC (via instructions). Before describing these ideas in more detail, we explain why it is important to examine those interactions and briefly review the limited literature on this topic.

Most often, effects of persuasion and EC (instructions) on evaluation are investigated in isolation (e.g., when only trait information or only information about stimulus pairings is presented). For instance, previous research has shown that, when presented alone, diagnostic trait information (Cone & Ferguson, 2015), EC procedures (Gibson, 2008) and EC instructions (De Houwer, 2006) all influence both self-reported and automatic evaluation. However, outside of laboratory settings, stimulus pairings and persuasive arguments are often encountered together when confronted with new people or objects. For instance, consider pre-election times during which people might be exposed to both pairings of a politician with positive stimuli (e.g., ads in which the politician holds cuddly babies) as well as to information about the traits of this politician (e.g., through magazine articles). Hence, it is important to also understand whether and how these different learning pathways interact.

Only a few papers deal with both the effects of stimulus pairings and the effects of verbal messages akin to those used in persuasion research. Rydell et al. (2006) tested the effects of (subliminal) stimulus pairings and behavioral statements (e.g., “Bob helps old ladies cross the street”) in a single study and found that stimulus pairings influenced only automatic evaluation whereas the behavioral statements influenced only self-reported evaluation. However, this effect recently failed to replicate (Heycke at el., 2018). Whitfield and Jordan (2009) compared the effects of stimulus pairings and behavioral statements in a single study and found that stimulus pairings had a direct effect on automatic evaluation and indirect effect on self-reported evaluation, whereas the opposite pattern was found for the behavioral statements. Finally, Mann et al. (2019) tested the effect of positive diagnostic trait information on evaluation of a target-person that was provided after pairings of the target-person with negative stimuli and found that automatic evaluation was updated in line with the diagnostic information (i.e., became more positive). The latter study thus suggests that the effects of stimulus pairings on automatic evaluation can be counteracted as the result of subsequent trait information.

Because of the limited and mixed nature of the available evidence about the interactions between EC procedures and trait information, we set out to examine in a more systematic manner the effects of stimulus pairings and persuasion-like information about verbal traits. In doing so, we not only examined the effects of actual experienced stimulus pairings (EC procedures) but also the effects of instructions about stimulus pairings (EC instructions). More specifically, we examined an important yet overlooked question: whether (instructions about) stimulus pairings still influence people’s evaluation of social groups even after being exposed to more diagnostic trait information about the groups. This study is of interest because it elucidates whether conditioning processes might have an important impact on our everyday evaluations in a world where people are often exposed to more diagnostic evaluative information. Our studies not only promise to shed new light on how different pathways of forming evaluations interact but, as we explain in the next section, could also constrains ideas about the processes that underlie the formation of evaluations.

EC Effects in the Context of Prior Trait Instructions

Two popular models of evaluation make different predictions about effects of EC procedures and EC instructions in the context of prior trait instructions. First, recent propositional accounts (De Houwer, 2014; Van Dessel, Cone, et al., 2020) argue that both persuasive messages and (instructions about) EC procedures provide information that can give rise to evaluative inferences. As mentioned above, one crucial distinction is that persuasive messages (like information about traits) are typically more diagnostic than mere pairing information (like co-occurrence with babies). Hence, when given ample time and opportunity to consider whether an object (e.g., a person or social group) is good or bad, people might lend more weight to trait information than to information about stimulus pairings. On the other hand, when processing conditions are suboptimal, diagnosticity might have less effect. Given the plausible assumption that conditions for effortful processing are more optimal when completing self-report measures than when responding in tasks designed to capture automatic evaluation (e.g., the IAT; Greenwald et al., 1998), (instructions about) EC procedures should have less influence on self-reported evaluation than on automatic evaluation in the presence of more diagnostic evaluative (trait) information.

Whereas it should not matter for propositional models whether stimulus pairings are experienced or merely instructed, it should matter according to the associative-propositional evaluation (APE) model (Gawronski & Bodenhausen, 2006, 2011). This model, which has been a major force in research on evaluation, postulates that both pairings and verbal information can result in the formation of mental associations. From this perspective, EC instruction effects occur because instructions about a future pairing of a CS with positive or negative stimuli might facilitate inferences that the CS is positive or negative. This could in turn lead to the formation of new associations in memory between the CS and positive or negative valence, associations that then influence automatic evaluation. One prominent assumption of the APE model is that repeated stimulus pairings directly result in the formation of associations and thus directly influence automatic evaluation whereas inferential processes can influence automatic evaluation only via the impact of propositional beliefs on associations. As such, EC instruction effects on automatic evaluation should be mediated by changes in self-reported evaluation (indirect effect of EC instructions on automatic evaluations) (Whitfield & Jordan, 2009; see Gawronski & Bodenhausen, 2006; Case 4). Accordingly, when more diagnostic evaluative information (e.g., trait instructions) and EC instructions are combined (and self-reported evaluation might reflect only this trait information but not EC instructions), EC instructions should not influence automatic evaluation.

This differential prediction of propositional models and the APE model has already been tested in the context of approach-avoidance (AA) instructions (Van Dessel et al. 2016, 2017). In these studies, participants received AA instructions (i.e., instructions to approach or avoid social group members in the future). Some participants also received, prior to the AA instructions, trait information about the groups. Van Dessel et al. (2016) found that, in the absence of trait instructions, AA instructions influenced automatic evaluation and this change was only partially mediated by corresponding changes in self-reported evaluations (see also Van Dessel et al., 2017). Importantly, in the presence of trait instructions, AA instructions influenced only automatic evaluation but not self-reported evaluation.

These results provide initial support for the prediction of propositional accounts of evaluation. However, the fact that this pattern of results was demonstrated for AA instructions does not necessarily mean that the same pattern will be observed for EC. EC is a fundamentally different learning pathway than AA as it does not require any actions by the learner. Moreover, according to a propositional account, inferences involved in AA effects on evaluation (e.g., “approaching is positive and approached stimuli are therefore also positive”, “pleasant stimuli are typically approached and approached stimuli are therefore also pleasant”) do not apply to EC effects (Van Dessel et al., 2019). Moreover, some empirical evidence suggests that the effects of EC and AA instructions can be different. For example, whereas Van Dessel, De Houwer et al. (2020) found that AA instructions are ineffective in shifting automatic evaluations of existing social groups, Kurdi and Banaji, (2017) found that EC instructions were effective. Finally, the studies of Van Dessel et al. (2016; 2017) focused on the interactive effect of (AA) instructions and trait information, but did not test the interactive effect of actual pairings (without instructions) and trait information. As such, there is merit in a systematic investigation of the effect of EC procedures and EC instructions on evaluation in the absence or presence of trait instructions.

Experiments Overview

In two experiments, one group of participants first received instructions about the traits of two fictitious social groups and then either received instructions about a future pairing of one group with positive images and of the other group with negative images (EC instruction condition) or experienced the actual pairings (EC experience condition). The other participants experienced EC procedures or received EC instructions but no trait instructions. At the end of the experiment, we assessed automatic evaluations of the two social groups with the IAT (Greenwald et al., 1998) and self-reported evaluations with rating scales. Experiment 2 provided a replication of Experiment 1 while modifying the procedure to (1) adopt a different EC manipulation and (2) increase the chances that participants would remember the learned information (because in Experiment 1 a large number of participants reported inaccurate memory).

We pre-registered the materials, sampling plan, exclusion rules, and analysis plans on the Open Science Framework (Experiment 1: osf.io/mwbh8/; Experiment 2: osf.io/x358q/). For Experiment 1 and 2, we pre-registered the hypothesis that, in-line with previous findings (Kurdi & Banaji, 2017), participants in the no-trait instructions group would show an automatic and self-reported preference for the group (instructed to be) paired with positive images and that we would not observe interactions with mode of learning (instructions vs. experience). For Experiment 2, we pre-registered the hypothesis that changes in automatic evaluation due to (instructed) stimulus pairings would not be fully mediated by changes in self-reported evaluations in both trait instruction groups. For both experiments, we report all data exclusions, manipulations, measures, and sample size determinations. All materials, data, and analyses scripts are available at the Open Science Framework (Experiment 1: osf.io/ubjwy/; Experiment 2: osf.io/z5ykp/).

Experiment 1

Method

Participants and design. Participants were online volunteers at the Project Implicit research website (Nosek, 2005). As is typical for studies on this website, participants agreed to participate for educational purposes. They did not receive payment or course credit.

We employed a 2 (Presence of trait instructions: yes, no) × 2 (EC content: Niffites paired with positive images and Luupites paired with negative images or vice versa) × 2 (EC mode: EC experience or EC instructions) between-subjects design (see Figure 1). In total, 1,047 participants completed the experiment. This was slightly more than our target sample size of 1,040 participants which was based on a power-analysis that would allow 90% power to detect a small three-way interaction effect ($\eta_p^2 = .02$) in the crucial ANOVA at alpha = .05. Following Van Dessel et al. (2016), the preregistered data exclusion involved removing participants who (a) did not fully complete all questions and tasks (12 participants; i.e., 1%), (b) had more than 10% fast trials in the IAT (35 participants; i.e., 3%), or (c) made at least one error on the memory questions that probed memory for trait information or pairing information (405 participants; i.e., 40.5%). The final sample included 595 participants (66% women, $M_{age} = 34.09$, $SD = 14.40$).

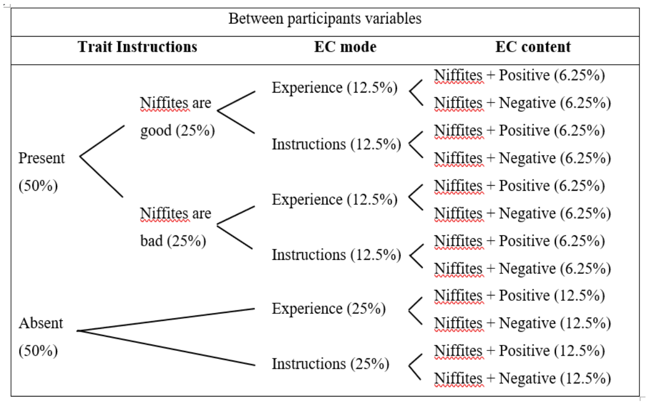
Between participants variables
Trait Instructions
EC mode
EC content
Present (50%)
Niffites are good (25%)
Experience (12.5%)
Niffites + Positive (6.25%)
Niffites + Negative (6.25%)
Instructions (12.5%)
Niffites + Positive (6.25%)
Niffites + Negative (6.25%)
Niffites are bad (25%)
Experience (12.5%)
Niffites + Positive (6.25%)
Niffites + Negative (6.25%)
Instructions (12.5%)
Niffites + Positive (6.25%)
Niffites + Negative (6.25%)
Absent (50%)
Experience (25%)
Niffites + Positive (12.5%)
Niffites + Negative (12.5%)
Instructions (25%)
Niffites + Positive (12.5%)
Niffites + Negative (12.5%)

Figure 1. Experimental design of Experiments 1-2.

Procedure and materials. The full procedure, instructions, and materials are explained in detail in the online supplement. All participants were first informed that they would learn about two social groups (i.e., the Luupites and the Niffites), after which half of the participants read trait instructions. Participants were asked to imagine that the two social groups actually exist and that they have very different characters such that Niffites (or Luupites) are very good people (e.g., peaceful, civilized) whereas Luupites (or Niffites) are very bad people (e.g., violent, savage). Participants in this condition were also instructed to suppose that the two groups consistently behave in line with this information when they interact with each other and with other groups. We counterbalanced across conditions whether participants who received trait instructions learned that Niffites are good and Luupites are bad or vice versa.

Next, half of the participants completed EC procedure (via experience) and the other participants received EC instructions. The procedure in both conditions followed that of Gast and De Houwer (2013). Participants in the experience condition were first instructed that they would see pleasant and unpleasant photos with each photo preceded by a name from one of the groups. Afterwards, they observed the actual pairings. The EC procedure consisted of two blocks of 20 trials. Each trial started with a blank screen for 200ms, and then a CS was presented on the screen for 1500ms. After a stimulus interval of 300ms during which the screen was blank, a US appeared and stayed on the screen for 4000ms. The trial ended with a blank screen for 1800ms. CSs were five Niffites names and five Luupites names (see the online supplement). USs were five positive and five negative images selected from a pool of 16 positive and 16 negative images from the International Affective Picture System (Lang et al., 2008) and photos found in a web search (the USs were adopted from Gawronski et al., 2015). Half of the participants observed pairings of Niffites with positive images and Luupites with negative images, and the other half observed pairings with the opposite assignment.

Participants in the EC instruction condition did not observe actual pairings, but only read general instructions describing the nature of the pairing task. They were informed that (1) they would see pleasant photos (e.g., puppies) and unpleasant photos (e.g., insects) and (2) names from one target group (e.g., Luupites) would always be followed by pleasant photos and names of the other group (e.g., Niffites) would always be followed by unpleasant photos. Half of the participants read that Niffites would be paired with positive images and Luupites with negative images, and the other participants read about the opposite assignment. Importantly, participants in the EC instruction condition did not observe actual CS-US pairings before they completed the evaluation and memory measures.[1]

Next, participants completed self-reported and automatic evaluation measures in a randomized order. To measure self-reported evaluations, participants were

[1] To avoid deception, participants in this condition completed a short EC procedure at the very end of the experiment, in which they observed five pairing trials for each group.

asked to rate their liking of each of the social groups by answering two questions: “How much do you like Niffites/Luupites?” and “How warm or cold are your feelings toward Niffites/Luupites?” Responses were made using a 7-point Likert scale (1 = very cold/strongly dislike; 7 = strongly warm/strongly like). Self-reported evaluation scores were calculated by subtracting the score rating for Luupites from the corresponding score rating for Niffites for each of the questions and then aggregating them to a single score so that positive scores indicate a preference for Niffites over Luupites (Cronbach's Alpha > .84 in Experiments 1-2).

To measure automatic evaluations, participants completed a seven-block IAT (Greenwald et al., 1998; Nosek et al., 2005) in which they categorized stimuli using two computer keys. In the critical blocks, participants responded with the left key to stimuli of two categories (e.g., “Niffites” and “Good”), and with the right key to stimuli of two other categories (e.g., “Luupites” and “Bad”). In two of these blocks, “Niffites” and “Good” shared the same response key, and in the other two critical blocks, “Luupites” and “Good” shared the same response key. Stimuli in the IAT were the same five Niffites’ names and five Luupites’ names used in the experienced EC task, five positive words (“Wonderful”, “Marvelous”, “Excellent”, “Good”, and “Glorious”) and five negative words (“Agony”, “Terrible”, “Evil”, “Poison”, and “Bad”). The D_2 algorithm was used to compute IAT scores (Greenwald et al., 2003) such that positive scores indicate a preference for Niffites over Luupites (split-half internal consistency: $\alpha > .89$ in Experiments 1-2).

Finally, participants completed a set of questions that assessed memory of the information about the two groups. To test if trait information was perceived as more diagnostic than the pairing, we also included a set of questions that estimated how diagnostic participants consider each type of information (trait versus pairing) to be. Specifically, the first two questions were completed only by participants who had received trait instructions. Participants were first asked to remember which trait instructions were presented at the start of the study. Response options were *Niffites are good and Luupites are bad*, *Luupites are good and Niffites are bad*, and *I don't remember*. Then participants were asked to what extent they think that the information about the groups characters was relevant or informative when determining how much they liked them with response options that ranged from “1 = Not at all relevant” to “9 = Very relevant”. The next three questions asked about the pairing contingencies and the diagnosticity of the pairing. For the contingency questions, participants in the EC instruction condition were asked what type of pictures would be followed by names for each of the two groups in the future pairing task. Participants in the EC experience condition were asked what type of pictures had been followed by names for each of the two groups. Response options in both cases were *Positive pictures*, *Negative pictures*, *Positive and Negative pictures*, and *I don't remember*. For the diagnosticity questions, participants in the EC instruction condition were asked to what extent they think that the information about the pairing of the groups with positive or negative pictures was relevant or informative when determining how much they liked them with response options that ranged from “1 = Not at all relevant” to “9 = Very relevant”. Participants in the EC experience condition were asked a similar question with the words “the pairing of the groups*”* instead of *“*the information about the pairing of the groups” in the question.

Results

To test the effect of EC procedures and instructions on evaluation in the absence or presence of diagnostic trait information, we performed separate analyses for participants who did not receive trait instructions and participants who did receive trait instructions.

No-trait instructions condition. Table 1 presents the automatic and self-reported evaluation scores as a function of EC content and EC mode. A 2 (EC content) × 2 (EC mode) ANOVA on the IAT scores revealed a main effect of EC content, $F(1, 306) = 258.24, p < .001$, $\eta_p^2 = .46$, 90%CI [.39, .51], $BF_{10} > 1000$, reflecting a stronger preference for Niffites over Luupites when Niffites were paired with positive valence (and Luupites were paired with negative valence; $M = 0.26$, $SD = 0.44$) than when Niffites were paired with negative valence (and Luupites were paired with positive valence $M = -0.54$, $SD = 0.42$). The main effect of EC mode was not significant ($p = .36$, $\eta_p^2 < .01$, $BF_{10} = 0.15$), but, unexpectedly, the interaction between EC mode and EC content was significant, $F(1, 306) = 9.46$, $p = .002$, $\eta_p^2 = .03$, 90%CI [.006, .06], $BF_{10} = 13.51$. The EC effect was stronger in the EC experience condition, $F(1, 306) = 206.86, p<.001, \eta^2_p = .40$, than in the EC instruction condition, $F(1, 306) = 75.78, p<.001, \eta^2_p = .19$.

The same ANOVA on self-reported evaluation scores revealed similar results. A main effect of EC content, $F(1, 306) = 154.32, p<.001, \eta_p^2 = .34$, 90%CI

[.26, .39], BF_{10} > 1000, reflecting a stronger preference for Niffites over Luupites when Niffites were paired with positive valence (M = 1.43, SD = 2.20) than when Niffites were paired with negative valence (M = -1.68, SD = 2.16). The main effect of EC mode was not significant (p = .184, η_p^2 = .01, BF_{10} = 0.20), and the interaction between EC mode and EC content was significant, F(1, 306) = 27.10, p < .001, η_p^2 = .08, 90%CI [.03, .13], BF_{10} > 1000. Similar to IAT scores, self-reported evaluation scores showed a stronger EC effect in the EC experience condition, F(1, 306) = 175.38, p<.001, η^2_p = .36, than in the EC instruction condition, F(1, 306) = 23.37, p<.001, η^2_p = .07.

Table 1
Mean automatic and self-reported evaluation scores in Experiments 1-2 for participants who did not received trait instructions, as a function of EC content, and EC Mode.

	EC content	
Experiment 1	**Niffites+ Positive**	**Luupites+ Positive**
Automatic evaluation score		
EC instructions	0.16 (0.42)	-0.47 (0.43)
EC experience	0.35 (0.44)	-0.57 (0.41)
Self-reported evaluation score		
EC instructions	0.64 (1.75)	-1.10 (1.84)
EC experience	2.21 (2.33)	-2.04 (2.75)
Experiment 2	**Niffites+ Positive**	**Luupites+ Positive**
Automatic evaluation score		
EC instructions	0.24 (0.53)	-0.43 (0.42)
EC experience	0.37 (0.46)	-0.62 (0.43)
Self-reported evaluation score		
EC instructions	1.17 (1.99)	-1.43 (2.07)
EC experience	1.77 (2.19)	-2.16 (1.98)

Note. Standard deviations are in parentheses. Scores reflect a relative preference for Niffites over Luupites. Range of automatic evaluation score is -2 to +2; self-reported evaluation ranges from -6 to +6.

To investigate whether changes in automatic evaluations are mediated by changes in self-reported evaluations we performed mediation analyses with the LAVAAN package (Rosseel, 2012). We used the bootstrap method to estimate standard errors for effects. Results indicated that changes in automatic evaluations were mediated by corresponding changes in self-reported evaluations, both in the EC instruction condition (Z = 3.37, p = .001, ab_{ps} = .18), and in the EC experience condition (Z = 2.65, p = .008, ab_{ps} = .21). Importantly, however, the EC effect on automatic evaluations remained significant after controlling for changes in self-reported evaluations both in the EC instruction condition (Z = 6.45, p<.001, ab_{ps} = .82) and in the EC experience condition (Z = 7.86, p<.001, ab_{ps} = .79). In other words, changes in automatic evaluations were only partially mediated by changes in self-reported evaluations. Regression coefficients of the performed mediation analyses are provided in the Appendix.

Trait instructions condition. Table 2 presents the automatic and self-reported evaluation scores as a function of EC content, Content of trait instructions, and EC mode. A 2 (EC content) × 2 (EC mode) × 2 (Content of trait instructions: Niffites are good, Luupites are good) ANOVA on the IAT scores revealed a main effect of the content of trait instructions, F(1, 277) = 164.75, p<.001, η_p^2 = .37, 90%CI [.29, .43], BF_{10} > 1000, indicating that participants preferred Niffites more when Niffites were presented as positive (and Luupites as negative; M = 0.23, SD = 0.43) than when Niffites were presented as negative (and Luupites as positive; M = −0.50, SD = 0.47). Importantly, the main effect of EC content was also significant, F(1, 277) = 20.86, p<.001, η_p^2 = .07, 90%CI [.02, .12], BF_{10} > 1000, reflecting a stronger preference for Niffites when Niffites were paired with positive valence (M = 0.07, SD = 0.55) than when Niffites were paired with negative valence (M = -0.31, SD = 0.54). The interaction between EC mode and EC content was also significant, F(1, 277) = 4.47, p = .035, η_p^2 = .02, 90%CI [.0005, .04], but evidence in favor of an effect was only anecdotal, BF_{10} = 2.18. This effect indicated that EC effects were stronger in the EC experience condition, F(1, 277) = 43.79, p<.001, η^2_p = .13, than in the EC instruction condition, F(1, 277) = 13.69, p<.001, η^2_p = .04. All other effects were not significant, ps > .32, η_p^2 < .01, BF_{10} < 0.31.

The same ANOVA on self-reported evaluation scores revealed a main effect of the content of trait instructions, F(1, 277) = 171.54, p<.001, η_p^2 = .38, 90%CI [.30, .44], BF_{10} > 1000, indicating that participants preferred Niffites more when Niffites were presented as positive (M = 1.81, SD = 2.40) than when Niffites were presented as negative (M = −2.11, SD = 2.43). Importantly, unlike with the IAT, the main effect of the EC content was not significant, F(1, 277) = 3.37, p = .067, η_p^2 = .01, 90%CI [0, .04], with anecdotal evidence in favor of the null hypothesis, BF_{10} = 0.59. The interaction between EC mode and EC content was significant, F(1, 277) = 4.47, p = .030, η_p^2 = .02, 90%CI [.0005, .04], but only provided anecdotal evidence in favor of an effect, BF_{10} = 1.97. The effect of EC content on self-reported evaluation was significant in the EC experience condition, F(1, 277) = 21.70, p<.001, η^2_p =

.07, but not in the EC instruction condition, $F(1, 277) = 3.07$, $p = .081$, $\eta^2_p = .01$. All other effects in the ANOVA were not significant, $ps > .35$, $\eta_p^2 < .01$, $BF_{10} < 0.31$.

Table 2

Mean automatic and self-reported evaluation scores in Experiments 1-2 for participants who received trait instructions, as a function of Content of Trait Instructions, EC content, and EC Mode.

	Content of Trait Instructions			
	Niffites good and Luupites bad		**Niffites bad and Luupites good**	
	EC content			
Experiment 1	**Niffites+ Positive**	**Luupites+ Positive**	**Niffites+ Positive**	**Luupites+ Positive**
Automatic evaluation score				
EC instructions	0.27 (0.44)	0.14 (0.40)	-0.36 (0.36)	-0.49 (0.41)
EC experience	0.35 (0.40)	0.04 (0.41)	-0.33 (0.60)	-0.71 (0.40)
Self-reported evaluation score				
EC instructions	1.70 (2.14)	1.98 (2.24)	-1.92 (2.38)	-2.00 (2.08)
EC experience	2.13 (2.36)	1.33 (3.08)	-1.34 (2.79)	-2.80 (2.36)
Experiment 2	**Niffites good and Luupites bad**		**Niffites bad and Luupites good**	
	Niffites+ Positive	**Luupites+ Positive**	**Niffites+ Positive**	**Luupites+ Positive**
Automatic evaluation score				
EC instructions	0.36 (0.46)	-0.04 (0.37)	-0.57 (0.37)	-0.49 (0.40)
EC experience	0.40 (0.43)	0.17 (0.44)	-0.35 (0.50)	-0.72 (0.37)
Self-reported evaluation score				
EC instructions	2.43 (2.23)	1.65 (2.83)	-1.97 (2.78)	-2.62 (2.39)
EC experience	3.33 (2.25)	1.59 (2.56)	-1.72 (3.02)	-2.69 (2.06)

Note. Standard deviations are in parentheses. Scores reflect a relative preference for Niffites over Luupites. Range of automatic evaluation score is -2 to +2; self-reported evaluation ranges from -6 to +6.

Because the results showed a significant effect of EC content on IAT scores but not on self-reported evaluation scores, we performed an additional (not-preregistered) analysis to directly compare the effect of EC content on the two types of evaluations. We standardized the preference scores and submitted them to a 2 (Evaluation type: self-reported, automatic) ×2 (EC content) × 2 (EC mode) × 2 (Content of trait instructions) mixed ANOVA. Most importantly, the ANOVA revealed a significant interaction between evaluation type and content of EC, $F(1, 277) = 4.72$, $p = .031$, $\eta^2_p = .02$, but with only anecdotal evidence in favor of this effect ($BF_{10} = 1.22$).

Mediation analyses showed that changes in automatic evaluations were not significantly mediated by corresponding changes in self-reported evaluations in the EC instruction condition ($Z = -0.31$, $p = .76$, $ab_{ps} = 0$). In contrast, changes in automatic evaluations were significantly mediated by corresponding changes in self-reported evaluations in the EC experience condition ($Z = 2.03$, $p = .042$, $ab_{ps} = .18$). The EC effect on automatic evaluations remained significant after controlling for changes in self-reported evaluations both in the EC instruction condition ($Z = 2.09$, $p = .037$, $ab_{ps} = 1$) and in the EC experience condition ($Z = 3.77$, $p<.001$, $ab_{ps} = .82$). In other words, changes in automatic evaluations in the EC instructions condition were not mediated by changes in self-reported evaluations. Changes in automatic evaluations in the EC experience condition were only partially mediated by changes in self-reported evaluations.

Diagnosticity of traits versus pairing information. Although not preregistered**,** to test our working assumption that trait information is considered more diagnostic than pairing information, we compared the diagnosticity rating of the trait information versus the pairing in the group of participants who received trait instructions. A 2 (information type: trait, pairing, within participants) × 2 (EC mode: experience, instructions, between participants) mixed ANOVA on the diagnosticity ratings found a main effect of information type, $F(1, 283) = 27.26$, $p < .001$, $\eta_p^2 = .09$, 90%CI [.04, .14], $BF_{10} > 1000$. In line with our assumption, trait information was rated as more diagnostic ($M = 5.12$, $SD = 2.88$) than the pairing information ($M = 4.26$, $SD = 2.85$). No other effects were significant ($ps > .12$, $\eta_p^2 < .007$, $BF_{10} < 0.52$).

Discussion

Experiment 1 found that EC procedures and EC instructions influenced both self-reported and automatic evaluation in the absence of trait information. In the presence of trait information, EC procedures influenced both self-reported and automatic evaluation, but EC instructions caused changes only in automatic evaluations and not in self-reported evaluations. The effects of (instructed) EC on automatic evaluation were in no case fully mediated by concurrent changes in self-reported evaluations.

Unexpectedly, we found overall stronger effects of EC procedures than of EC instructions on both self-reported and automatic evaluations. This contrasts with recent studies (Kurdi & Banaji, 2017). One explanation for this discrepancy relates to differences in expectations about the content of the pairings in the two conditions. Whereas participants in the EC instruction condition were told that the pairing task would include positive (e.g., puppy) and negative (e.g., insect) photos, but did not receive details about the content of all the positive and negative photos, participants in the EC experience condition were exposed to detailed positive and negative photos (of puppies and insects but also of feces, landscapes, etc.). This differs from other studies that used similar content information in the two EC conditions (Kurdi & Banaji, 2017) and could have caused participants' expectations of the pairings to be less valenced than actual experienced pairings in the EC instruction condition, allowing for stronger EC effects in the experience condition.

One limitation of Experiment 1 is that a large number of participants (41%) were excluded from analyses because they provided incorrect answers to questions that assessed participants' memory for the trait instructions or the (instructed) EC pairings. This low accuracy could be due to low engagement of the online volunteer participants, but it could also relate to lack of clarity in the instructions or the memory questions. Experiment 2 was designed as a replication of Experiment 1 with the additional aim of increasing the likelihood that participants would remember the information they learned. To this end, we added questions about stimulus pairings and trait instructions to these procedures that participants had to answer correctly in order to complete the study. Moreover, Experiment 2 adopted an EC manipulation that kept the content of the information in the two EC conditions as similar as possible (Kurdi & Banaji, 2017), allowing us to test if the observed differences between effects of EC instructions and EC procedures would hold even when the two types of EC tasks provide more similar information regarding the content of the pairings.

Experiment 2

Method

Participants and design. We sampled 1,300 Project Implicit participants to provide 90% power to detect a small interaction effect ($\eta_p^2 = .02$) when taking into account an estimated 25% exclusions due to inaccurate memory. We excluded the data of participants who (a) did not fully complete all questions and tasks (10 participants; i.e., 0.7%), (b) had more than 10% fast trials in the IAT (32 participants; i.e., 2%), or (c) made at least one error on the memory questions that probed memory for trait or EC information (359 participants; i.e., 28%). The final sample included 915 participants (69% women, M_{age} = 39.44, *SD* = 14.20).

Procedure and materials. The procedure and materials of Experiment 2 were similar to Experiment 1 except for the following changes. First, we adopted the EC procedures from Kurdi and Banaji (2017). Participants in the EC experience condition were instructed that they would watch a pairing task in which they would see pleasant, positive pictures and unpleasant, negative pictures and that each picture would be paired with a name from one of the groups. They were further informed that their task was to learn the association between a certain type of name and a certain type of picture. Participants then saw a presentation of the full set of CSs and USs. Then, participants observed an EC procedure which consisted of one block of 40 trials. On each trial, one CS (a Luupites or Niffites name) and one US were presented simultaneously next to each other in the center of the screen for 2500ms, followed by an intertrial interval of 1000ms, consisting of a blank screen. USs were five positive and five negative images adopted from Kurdi and Banaji, (2017; Experiment 4). Half of the participants observed pairings of Niffites with positive images and Luupites with negative images and the other half observed the opposite assignment. In the EC instruction condition, participants were first informed that they would see parings of names with pictures such that one target group (e.g., Luupites) would always be paired with pictures of pleasant things and the other target group (e.g., Niffites) would always be paired with pictures of unpleasant things. Importantly, they then saw a presentation of the full set of CSs and USs (in different screens, such that participants were not exposed to stimulus pairings).

Second, we added questions after the trait instructions, the EC instructions, and the EC experience task, asking participants about the learned contingencies. Specifically, after the trait instructions, participants were asked to indicate what information the instructions presented about Niffites and Luupites (in two separate questions). Response options were *they are very bad people*, and *they are very good people*. If participants answered one of the questions incorrectly, they read the trait instructions again. In the EC experience condition, after the EC task, participants were asked with what type of pictures Niffites' and Luupites' names were paired. Response options were *Niffites were paired with positive pictures and Luupites were paired with negative pictures*, and *Niffites were paired with negative pictures and Luupites were paired with* *positive pictures*. If participants answered the question incorrectly, they were exposed to another pairing block of 10 trials. In the EC instruction condition, after participants received the instructions about the future pairing task, they were asked what information the instructions presented about Niffites and Luupites names (in two separate questions). Response options were *they will be paired with positive pictures*, and *they will be paired with negative pictures*. If participants answered one of the questions incorrectly, they were redirected to the EC instructions. Evaluation, memory and diagnosticity measures were the same as in Experiment 1.[2]

Results

No-trait instructions condition. Table 1 presents the automatic and self-reported evaluation scores as a function of EC content and EC mode. A 2 (EC content) × 2 (EC mode) ANOVA on the IAT scores revealed a main effect of EC content, $F(1, 465) = 379.27, p < .001, \eta_p^2 = .45$, 90%CI [.39, .49], $BF_{10} > 1000$, reflecting a stronger preference for Niffites over Luupites when Niffites were paired with positive valence ($M = 0.30$, $SD = 0.50$) than when Niffites were paired with negative valence ($M = -0.53$, $SD = 0.43$). The main effect of EC mode was not significant ($p = .522$, $\eta_p^2 < .01$, $BF_{10} = 0.13$), but the interaction between EC mode and EC content was significant, $F(1, 465) = 13.53, p < .001, \eta_p^2 = .03$, 90%CI [.008, .05], $BF_{10} = 85.3$. The EC effect was stronger in the EC experience condition, $F(1, 465) = 206.6, p < .001, \eta^2_p = .35$, than in the EC instruction condition, $F(1, 306) = 27.49, p < .001, \eta^2_p = .21$.

The same ANOVA on self-reported evaluation scores revealed a main effect of EC content, $F(1, 465) = 295.46, p < .001, \eta_p^2 = .39$, 90%CI [.33, .43], $BF_{10} > 1000$, reflecting a stronger preference for Niffites over Luupites when Niffites were paired with positive valence ($M = 1.45$, $SD = 2.10$) than when Niffites were paired with negative valence ($M = -1.80$, $SD = 2.05$). The main effect of EC mode was not significant ($p = .715$, $\eta_p^2 < .01$, $BF_{10} = 0.11$), but the interaction between EC mode and EC content was significant, $F(1, 465) = 12.27, p < .001, \eta_p^2 = .03$, 90%CI [.007, .05], $BF_{10} = 44.02$. Similar to the IAT scores, self-reported evaluation scores revealed stronger EC effects in the EC experience condition, $F(1, 465) = 208.14, p < .001, \eta^2_p = .30$, than in the EC instruction condition, $F(1, 465) = 96.40, p < .001, \eta^2_p = .17$.

Mediation analyses indicated that changes in automatic evaluations were mediated by corresponding changes in self-reported evaluations, both in the EC instruction condition ($Z = 3.45$, $p = .001$, $ab_{ps} = .18$), and in the EC experience condition ($Z = 4.30$, $p < .001$, $ab_{ps} = .26$). Importantly, however, the EC effect on automatic evaluations remained significant after controlling for changes in self-reported evaluations both in the EC instruction condition ($Z = 7.67$, $p < .001$, $ab_{ps} = .82$) and in the EC experience condition ($Z = 8.23$, $p < .001$, $ab_{ps} = .74$). In other words, changes in automatic evaluations were only partially mediated by changes in self-reported evaluations.

Trait instructions condition. Table 2 presents the automatic and self-reported evaluation scores as a function of EC content, content of trait instructions, and EC mode. A 2 (EC content) × 2 (EC mode) × 2 (Content of trait instructions) ANOVA on the IAT scores revealed a main effect of the content of trait instructions, $F(1, 438) = 334.21, p < .001, \eta_p^2 = .43$, 90%CI [.37, .47], $BF_{10} > 1000$, indicating that participants preferred Niffites more when Niffites were presented as positive ($M = 0.24$, $SD = 0.46$) than when Niffites were presented as negative ($M = -0.53$, $SD = 0.46$). The main effect of the EC content was also significant, $F(1, 438) = 30.87, p < .001, \eta_p^2 = .07$, 90%CI [.03, .10], $BF_{10} > 1000$, reflecting a stronger preference for Niffites when Niffites were paired with positive valence ($M = 0.05$, $SD = 0.60$) than when Niffites were paired with negative valence ($M = -0.32$, $SD = 0.52$). Unlike Experiment 1, the interaction between EC mode and EC content was not significant ($p = .08$, $\eta_p^2 = .01$, $BF_{10} = 0.58$), indicating no statistical difference between the EC effect in the EC experience condition, $F(1, 438) = 35.42, p < .001, \eta^2_p = .07$, and the EC instruction condition, $F(1, 438) = 49.95, p < .001, \eta^2_p = .10$. We also observed a two-way interaction between content of trait instructions and EC content, $F(1, 438) = 4.00, p = .046, \eta_p^2 = .01$, 90%CI [.00, .02], $BF_{10} = 0.98$, and a three-way interaction between content of trait instructions, EC content and EC mode, $F(1, 438) = 13.77, p < .001, \eta_p^2 = .03$, 90%CI [.009, .06], $BF_{10} = 133.96$. However, because these effects are not theoretically relevant to the present research, we did not further interpret them. There were no other significant effects, $ps > .11$, $\eta_p^2 = .01$, $BF_{10} < 0.72$.

The same ANOVA on self-reported evaluation scores revealed a main effect of the content of trait instructions, $F(1, 438) = 342.04, p < .001, \eta_p^2 = .44$,

[2] For exploratory reasons, we added a general attention test at the end of the experiment. More details are available at osf.io/gesbv/.

90%CI [.38, .48], $BF_{10} > 1000$, indicating that participants preferred Niffites more when Niffites were presented as positive ($M = 2.28$, $SD = 2.50$) than when Niffites were presented as negative ($M = -2.32$, $SD = 2.55$). Importantly, unlike Experiment 1, the main effect of EC content on self-reported evaluation was significant, $F(1, 438) = 18.09$, $p < .001$, $\eta_p^2 = .01$, 90%CI [.01, .07], with the BF score providing strong evidence in favor of an effect ($BF_{10} = 287.54$). This effect indicates a stronger preference for Niffites when Niffites were paired with positive valence ($M = 0.95$, $SD = 3.37$) than when Niffites were paired with negative valence ($M = -0.85$, $SD = 3.23$). The interaction between EC mode and EC content did not reach significance, $F(1, 438) = 1.73$, $p = .19$, $\eta_p^2 < .01$, 90%CI [0, .01], $BF_{10} = 0.34$, indicating no statistical difference between the EC effect in the EC experience condition, $F(1, 438) = 20.69$, $p < .001$, $\eta^2_p = .04$, and the EC instruction condition, $F(1, 438) = 37.91$, $p < .001$, $\eta^2_p = .07$. We did not observe any other significant effects, ps > .35, $\eta_p^2 < .01$, $BF_{10} < 0.42$.

We also directly compared the effect of content of EC on the two types of evaluations by standardizing the preference scores and submitting them to a 2 (Evaluation type) ×2 (EC content) × 2 (EC mode) × 2 (Content of trait instructions) mixed ANOVA. Unlike Experiment 1, the ANOVA did not find a significant interaction between evaluation type and the EC content, $F(1, 438) = 0.94$, $p = .33$, $\eta^2_p < .01$, with moderate evidence in favor of the absence of an effect ($BF_{10} = 0.20$).

Similar to Experiment 1, mediation analyses showed that changes in automatic evaluations were not significantly mediated by corresponding changes in self-reported evaluations, in the EC instruction condition ($Z = 1.59$, $p = .11$, $ab_{ps} = .11$). Changes in automatic evaluations were significantly mediated by corresponding changes in self-reported evaluations in the EC experience condition ($Z = 2.72$, $p = .006$, $ab_{ps} = .22$). The EC effect on automatic evaluations remained significant after controlling for changes in self-reported evaluations both in the EC instruction condition ($Z = 3.06$, $p = .002$, $ab_{ps} = .89$) and in the EC experience condition ($Z = 4.06$, $p < .001$, $ab_{ps} = .78$). In other words, as in Experiment 1, changes in automatic evaluations in the EC instructions condition were not mediated by changes in self-reported evaluations. Changes in automatic evaluations in the EC experience condition were only partially mediated by changes in self-reported evaluations.

Diagnosticity of traits versus pairing information. A 2 (information type) × 2 (EC mode) mixed ANOVA on the diagnosticity ratings found a main effect of information type, $F(1, 443) = 35.11$, $p < .001$, $\eta_p^2 = .07$, 90%CI [.03, .11], $BF_{10} > 1000$. As in Experiment 1, trait information was rated as more diagnostic ($M = 5.18$, $SD = 2.93$) than the pairing information ($M = 4.32$, $SD = 2.86$). The interaction between information type and EC mode was also significant, $F(1, 443) = 11.47$, $p = .001$, $\eta_p^2 = .03$, 90%CI [.006, .05], $BF_{10} = 30.6$. This interaction reflected a stronger effect on information type in the EC instruction condition, $F(1, 443) = 49.22$, $p < .001$, $\eta_p^2 = .03$ than in the EC experience condition, $F(1, 443) = 2.88$, $p = .091$, $\eta_p^2 = .002$.

Because the lack of information type effect in Experiment 2 for the EC experience condition was unexpected, we tested the combined effect of information type on diagnosticity rating in the two studies using a (fixed effects) meta-analysis. The meta-analysis found an effect of information type in the expected direction (higher rating of diagnosticity for trait information than for pairing information) both in the EC experience condition, *Hedges' g* = 0.16, $SE = 0.06$, 95%CI [0.05, 0.27], $Z = 2.90$, $p = .004$, and in the EC instructions condition, *Hedges' g* = 0.42, $SE = 0.05$, 95%CI [0.31, 0.52], $Z = 7.92$, $p < .001$.

Discussion

Replicating the results of Experiment 1, Experiment 2 found that EC procedures and EC instructions influenced both self-reported and automatic evaluation in the absence of trait information. Unlike Experiment 1, in Experiment 2, in the presence of trait information, both EC procedures and EC instructions influenced both automatic and self-reported evaluations. Again, (instructed) EC effects on automatic evaluation were never fully mediated by changes in self-reported evaluation.

Differences between the effects of EC procedures and EC instructions were also observed (but less consistently than in Experiment 1). Similar to Experiment 1, when no trait information was provided, EC procedures produced overall stronger effects than EC instructions. On the other hand, in contrast to the results of Experiment 1, we did not observe a significant difference in the effects of EC procedures and instructions on (automatic and self-reported) evaluation when trait information was provided.

General Discussion

Two experiments tested the effect of EC procedures and EC instructions on evaluation in the absence or presence of more diagnostic trait information. Both experiments found that when stimulus pairings and EC instructions constitute the only available evaluative information (in the absence of trait information) they influence both self-reported and automatic evaluation. In the presence of trait information, Experiments 1 and 2 showed somewhat different results. Experiment 1 found that whereas EC procedures influence both automatic and self-reported evaluation, EC instructions influence only automatic but not self-reported evaluation. Experiment 2, however, found EC and EC instructions effects on both self-reported and automatic evaluation. Moreover, both Experiments 1 and 2 provided evidence that in the absence of trait information, (instructed) EC effects on automatic evaluation are only partly mediated by changes in self-reported evaluation. In the presence of trait information, both experiments found that changes in the automatic evaluation due to EC procedures are only partly mediated, and changes in the automatic evaluation due to EC instructions are not mediated, by changes in self-reported evaluation.

Implications for the Interaction between EC and Persuasion

EC (stimulus pairings) effects play an important role in social phenomena related to evaluation formation and change including prejudice (Gawronski & Bodenhousen, 2006) and stigmatization (Hebl & Mannix, 2003). However, to understand the role EC plays in attitude formation it is important to understand its influence not only in isolation, but also when other (more diagnostic) evaluative information is presented. In this regard, the results of the present research provide new insights. Importantly, the present findings show that both exposure to repeated pairings (EC) and to instructions about future pairings has a (direct) effect on both automatic and self-reported evaluation and that this effect is present even in the context of more diagnostic trait information. This highlights that exposure to stimulus pairings (or even mere instructions about possible pairing) can have an important role in evaluative learning even when people have more diagnostic information to rely on when making evaluations, highlighting the importance of conditioning processes in evaluation.

Moreover, the fact that both EC procedures and EC instructions influence automatic evaluations in the context of more diagnostic information is important because it provides further evidence that formation and change of automatic evaluations does not require slow-learning on the basis of pairings. In this regard it is also noteworthy that EC procedures had a stronger overall effect on evaluations than EC instructions (in contrast to other studies examining this question: e.g., Kurdi & Banaji, 2017). This suggests that interventions directed at the formation of attitudes via EC (e.g., political campaigns or advertisings that are based on pairing the target object with positive or negative stimuli) can be effective even if they present mere instructions about future pairings. Nevertheless, such interventions might be more effective when presenting actual pairings.

Implications for Theories of Evaluation

The results of the current research provide information that constrains mental process theories of evaluation. First, the finding that EC instructions affect automatic evaluation is difficult to reconcile with a subset of dual-process models of evaluation that make the following assumptions: (a) automatic evaluations reflect the automatic activation of associations in memory and (b) associations are formed only as the result of a slow-learning process that requires repeated pairings (e.g., Rydell & McConnell, 2006; Smith & DeCoster, 2000). Second, the observation that EC instructions influence automatic evaluation without (full) mediation by changes in self-reported evaluation is difficult to reconcile with the highly influential APE model (Gawronski & Bodenhausen, 2006, 2011). According to the APE model, full mediation should occur because EC instruction effects on automatic evaluation are assumed to arise as the result of inferences that first influence self-reported evaluations.

The results of the present research are more in line with the predictions of recent propositional accounts of evaluation (De Houwer, 2014, 2018) which assume that both self-reported and automatic evaluations reflect the activation of propositional representations. From this perspective, inferences about stimulus valence can be readily formed based on EC procedures and EC instructions. Both can have a direct effect on automatic evaluation that is not necessarily mediated by changes in self-reported evaluations. In the context of more diagnostic trait information, EC procedures and EC instructions might have a stronger effect on automatic evaluation than on self-reported evaluation because self-reported evaluation allows for more control over responding which might facilitate effects of diagnosticity. From this perspective, EC procedures and instructions could produce an effect on automatic evaluations but no effect on self-reported evaluation in the context of trait instructions as was observed in Experiment 1. Experiment 2 did not reveal this pattern. Instead, EC procedures and EC instructions both had an effect on automatic evaluations as well as self-reported evaluations. This could imply that the absence of an effect in Experiment 1 was not robust and (in contrast to what is found for AA instructions: Van Dessel et al., 2016 – also see De Houwer et al., 2020), pairing information is considered important enough to influence self-reported evaluation even in the presence of more diagnostic information. Another possible propositional explanation is that the questions added after the pairing instructions in Experiment 2 emphasized the importance of the pairing information to the extent that participants found it sufficiently important to take into account in their self-report ratings.

Implications for EC Research

The current results are also informative for elucidating the cognitive processes underlying EC effects. First, the finding that EC procedures and EC instructions have a direct effect on automatic evaluation that is not fully mediated by changes in self-reported evaluation, might indicate that (instructed) EC effects on evaluation are not solely due to highly controlled processes that involve the intentional use of pairing information for evaluation.

Second, we found that actual pairings can lead to stronger changes in evaluations than EC instructions. In Experiment 1, this finding might be explained by differences in expectations about the content of the pairings as there was much more detailed pairing information in the EC experience than in the EC instruction condition. However, in Experiment 2, we used a procedure that better matches learning about the content of the pairings in the two EC conditions (Kurdi & Banaji, 2017) yet still observed that experienced EC led to a stronger effect on both self-reported and automatic evaluation. This finding contrasts with the findings of Kurdi and Banaji (2017) who found stronger EC effects on automatic evaluation in EC instruction conditions than in EC experience conditions for different types of stimuli, including fictitious social groups. One possible explanation for this discrepancy is that the instructions of Kurdi and Banaji not only provided pairing information (e.g., "Luupites will always be paired with pleasant things and Niffites will always be paired with unpleasant things") but also information about a subsequent inference regarding the evaluative properties of the stimuli (e.g., "Luupites are linked to good things and Niffites are linked to bad things"; "so please remember well: Luupites = pleasant and Niffites = unpleasant"). [3] This inferential step was not communicated in the experience condition and it is possible that this is an important determinant of EC effects (see also Van Dessel et al., 2019). In the current study (Experiment 2), EC instructions included the pairing information but not the additional evaluative information which might explain the dissociation with prior findings.

Limitations and Future Directions

In the present research, we used a statistical approach to test mediation of EC effects on automatic evaluation by changes in self-reported evaluation. The disadvantage of the statistical approach is that the measurement-of-mediation approach is ultimately correlational, and is therefore problematic for establishing causal relations (De Houwer et al., 2013; Spencer et al., 2005). Moreover, when self-reported and automatic evaluations are strongly correlated, like in the present research,[4] this multicollinearity can exaggerate the standard error of the variables in the mediation model and impede the estimation of the indirect effect (Alin, 2010).

Another limitation of the present research is that it used a limited set of stimuli (i.e., fictitious social groups as targets), a specific sample (Project Implicit online volunteers) and one type of automatic evaluation measure (the IAT). It is possible that the observed direct influence of EC instructions on automatic evaluation is due to specific properties of the IAT and does not transfer to other automatic evaluation measures. Future research could extend this line of research by examining the effect of (instructed) EC on evaluation with different stimuli and samples, and by using different automatic evaluation measures such as the Evaluative Priming task (Fazio et al., 1995) or the Affect Misattribution Procedure (Payne et al., 2005).

Conclusion

The present research tested the interactions between important pathways for shaping evaluations: Evaluative Conditioning (EC) and persuasion. We found that

[3] See osf.io/w6qnb/ for the full instructions used in Kurdi and Banaji (2017).

[4] In the current research, the correlations between self-reported and automatic evaluations were r = .58, .57, in the no and yes trait instructions condition, respectively, in Experiment 1, and r = .56, .59, in the no and yes trait instructions condition, respectively, in Experiment 2.

EC procedures and EC instructions have a (direct) effect on automatic and self-reported evaluation both in the absence or presence of more diagnostic trait information. These results highlight the important role that exposure to stimulus pairings (or to instructions about possible pairing) can have in social learning.

Authorship

TM contributed to the experimental design, creation of the materials, data collection, analysis, and writing of the paper. PVD and CTS contributed to the experimental design and writing of the paper. JDH contributed to the writing of the paper. All authors contributed to the conceptualization of the research idea.

References

Alin, A. (2010). Multicollinearity. Wiley Interdisciplinary R views: *Computational Statistics*,*2*, 370–374.

Allport, G. W. (1935). Attitudes. In C. Murchison (Ed.), *A handbook of social psychology* (pp. 798 – 844). Worcester, MA: Clark University Press.

Beggan, J. K. (1992). On the social nature of nonsocial perception: The mere ownership effect. *Journal of personality and social psychology*, *62*(2), 229-237.

Berger, I. E., & Mitchell, A. A. (1989). The effect of advertising on attitude accessibility, attitude confidence, and the attitude-behavior relationship. *Journal of Consumer Research*, *16*(3), 269-279.

Briñol, P., Petty, R. E., & McCaslin, M. J. (2009). Changing attitudes on implicit versus explicit measures: What is the difference? In R. E. Petty, R. H. Fazio, & P. Briñol (Eds.), Attitudes: Insights from the new implicit measures (pp. 285-326). New York: Psychology Press.

Cone, J., & Ferguson, M. J. (2015). He did what? The role of diagnosticity in revising implicit evaluations. *Journal of Personality and Social Psychology*, *108*(1), 37-57.

De Houwer, J. (2006). Using the Implicit Association Test does not rule out an impact of conscious propositional knowledge on evaluative conditioning. *Learning and Motivation*, *37*(2), 176-187.

De Houwer, J. (2014). A Propositional Model of Implicit Evaluation. *Social and Personality Psychology Compass, 8*, 342-353.

De Houwer, J. (2018). Propositional models of evaluative conditioning. *Social Psychological Bulletin*, 13(3), e28046.

De Houwer, J., Gawronski, B., & Barnes-Holmes, D. (2013). A functional-cognitive frameworkfor attitude research. *European Review of Social Psychology, 24*, 252–287.

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.

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), 853-869.

De Houwer, J., Van Dessel, P., & Moran, T. (2020). Attitudes Beyond Associations: On the Role of Propositional Representations in Stimulus Evaluation. *Advances in Experimental Social Psychology, 61*, 127-183.

Fazio, R. H., Jackson, J. R., Dunton, B. C., & Williams, C. J. (1995). Variability in automatic activation as an unobtrusive measure of racial attitudes: A bona fide pipeline?. *Journal of Personality and Social Psychology, 69*(6), 1013 –1027.

Fiske, S. T. (1998). Stereotyping, prejudice, and discrimination. In D. T. Gilbert, S. T. Fiske, & G. Lindzey (Eds.), Handbook of social psychology (4th ed., Vol. 2, pp. 357–411). New York: McGraw-Hill.

Gast, A., & De Houwer, J. (2013). The influence of extinction and counterconditioning instructions on evaluative conditioning effects. *Learning and Motivation, 44*, 312-325.

Gawronski, B., & Bodenhausen, G. V. (2006). Associative and propositional processes in evaluation: an integrative review of implicit and explicit attitude change. *Psychological Bulletin*, *132*(5), 692-731.

Gawronski, B., & Bodenhausen, G. V. (2011). The associative-propositional evaluation model: Theory, evidence, and open questions. *Advances in Experimental Social Psychology, 44*, 59-127.

Gawronski, B., Galdi, S., & Arcuri, L. (2015). What can political psychology learn from implicit measures? Empirical evidence and new directions. *Political Psychology, 36*(1), 1-17.

Gawronski, B., Gast, A., & De Houwer, J. (2015). Is evaluative conditioning really resistant to extinction? Evidence for changes in evaluative judgements without changes in evaluative representations. *Cognition and Emotion*, *29*(5), 816-830.

Gibson, B. (2008). Can evaluative conditioning change attitudes toward mature brands? New evidence from the Implicit Association Test. *Journal of Consumer Research*, *35*(1), 178-188.

Greenwald, A. G., McGhee, D. E., & Schwartz, J. K. L. (1998). Measuring individual differences in implicit cognition: The Implicit Association Test. *Journal of Personality and Social Psychology*, *74*(6), 1464-1480.

Greenwald, A. G., Nosek, B. A., & Banaji, M. R. (2003). Understanding and using the implicit association test: I. An improved scoring algorithm. *Journal of Personality and Social Psychology*, *85*(2), 197-216.

Hebl, M. R., & Mannix, L. M. (2003). The weight of obesity in evaluating others: A mere proximity effect. *Personality and Social psychology Bulletin*, *29*(1), 28-38.

Heycke, T., Gehrmann, S., Haaf, J. M., & Stahl, C. (2018). Of two minds or one? A registered replication of Rydell et al.(2006). *Cognition and Emotion*, *32*(8), 1708-1727.

Hofmann, W., De Houwer, J., Perugini, M., Baeyens, F., & Crombez, G. (2010). Evaluative conditioning in humans: a meta-analysis. *Psychological Bulletin*, *136*(3), 390-421.

Houben, K., Schoenmakers, T. M., & Wiers, R. W. (2010). I didn't feel like drinking but I don't know why: The effects of evaluative conditioning on alcohol-related attitudes, craving and behavior. *Addictive Behaviors, 35*(12), 1161-1163.

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,* 194–213.

Lang, P. J., Bradley, M. M., & Cuthbert, B. N. (2008). International affective picture

system (IAPS): Technical manual and affective ratings. Gainesville, FL: The Center for Research in Psychophysiology, University of Florida.

Mann, T. C., Kurdi, B., & Banaji, M. R. (2019). How effectively can implicit evaluations be updated? Using evaluative statements after aversive repeated evaluative pairings. *Journal of Experimental Psychology: General*.

Nosek, B. A. (2005). Moderators of the relationship between implicit and explicit evaluation. *Journal of Experimental Psychology: General*, *134*(4), 565-584.

Nosek, B. A., Greenwald, A. G., & Banaji, M. R. (2005). Understanding and using the Implicit Association Test: II. Method variables and construct validity. *Personality and Social Psychology Bulletin*, *31*(2), 166-180.

Olson, M. A., & Fazio, R. H. (2006). Reducing automatically activated racial prejudice through implicit evaluative conditioning. *Personality and Social Psychology Bulletin*, *32*(4), 421-433.

Payne, B. K., Cheng, C. M., Govorun, O., & Stewart, B. D. (2005). An inkblot for attitudes: affect misattribution as implicit measurement. *Journal of Personality and Social Psychology*, *89*(3), 277-293.

Petty, R. E., & Cacioppo, J. T. (1984). The effects of involvement on responses to argument quantity and quality: Central and peripheral routes to persuasion. *Journal of Personality and Social Psychology*, *46*(1), 69-81.

Rydell, R. J., & McConnell, A. R. (2006). Understanding implicit and explicit attitude change: A systems of reasoning analysis. *Journal of Personality and Social Psychology, 91*(6), 995–1008.

Rydell, R. J., McConnell, A. R., Mackie, D. M., & Strain, L. M. (2006). Of two minds: Forming and changing valence-inconsistent implicit and explicit attitudes. *Psychological Science*, *17*(11), 954-958.

Rosseel, Y. (2012). lavaan: An R package for structural equation modeling. *Journal of Statistical Software*, *48*(2), 1–36 (URL: http://www.jstatsoft.org/v48/i02/).

Schienle, A., Stark, R., & Vaitl, D. (2001). Evaluative conditioning: A possible explanation for the acquisition of disgust responses?. *Learning and Motivation, 32*(1), 65-83.

Shaw, J. A., Forman, E. M., Espel, H. M., Butryn, M. L., Herbert, J. D., Lowe, M. R., & Nederkoorn, C. (2016). Can evaluative conditioning decrease soft drink consumption? *Appetite, 105*, 60-70.

Smith, C. T., Calanchini, J., Hughes, S., Van Dessel, P., & De Houwer, J. (2020). The impact of instruction- and experience-based evaluative learning on IAT performance: A Quad model perspective. *Cognition and Emotion, 34*, 21-41.

Smith, E. R., & DeCoster, J. (2000). Dual-process models in social and cognitive psychology: Conceptual integration and links to underlying memory systems. *Personality and Social Psychology Review*, *4*, 108–131.

Spencer, S. J., Zanna, M. P., & Fong, G. T. (2005). Establishing a causal chain: Why experiments are often more effective than mediational analyses in examining psychological processes. *Journal of Personality and Social Psychology, 89*, 845–851.

Van Dessel, P., Cone, J., Gast, A., & De Houwer, J. (2020). The impact of valenced verbal information on implicit and explicit evaluation: The role of information diagnosticity, primacy, and memory cueing. *Cognition and Emotion*, *34*(1), 74-85.

Van Dessel, P., De Houwer, J., Gast, A., Roets, A., & Smith, C. T. (2020). On the effectiveness of approach-avoidance instructions and training for changing evaluations of social groups. *Journal of Personality and Social Psychology*.

Van Dessel, P., De Houwer, J., Gast, A., Smith, C. T., & De Schryver, M. (2016). Instructing implicit processes: When instructions to approach or avoid influence implicit but not explicit evaluation. *Journal of Experimental Social Psychology*, *63*, 1-9.

Van Dessel, P., Gawronski, B., Smith, C. T., & De Houwer, J. (2017). Mechanisms underlying approach-avoidance instruction effects on implicit evaluation: Results of a preregistered adversarial collaboration. *Journal of Experimental Social Psychology*, *69*, 23-32.

Van Dessel, P., Hughes, S., & De Houwer, J. (2019). How do actions influence attitudes? An inferential account of the impact of action performance on stimulus evaluation. *Personality and Social Psychology Review*, *23*(3), 267-284.

Whitfield, M., & Jordan, C. H., (2009). Mutual Influence of implicit and explicit attitudes. *Journal of Experimental Social Psychology, 45*(4), 748-759.

Appendix

Mediation Analyses

Experiment 1

No trait instructions condition.

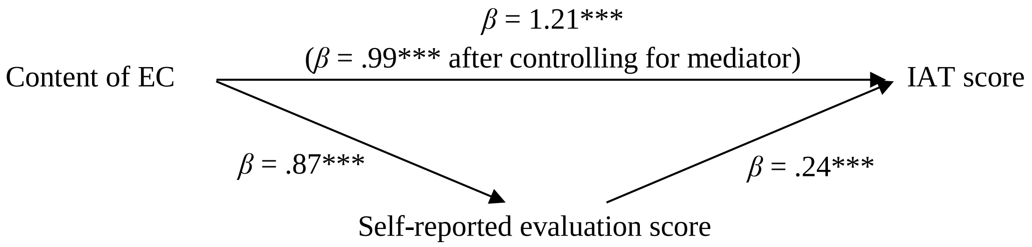
β = 1.21***
(*β* = .99*** after controlling for mediator)
Content of EC
IAT score
β = .87***
β = .24***
Self-reported evaluation score

Figure A1. Standardized Estimates of mediation coefficients for participants in Experiment 1 who received no trait instructions and were in the EC instructions condition. * *p* < .05 ** *p* < .01 *** *p* < .001.

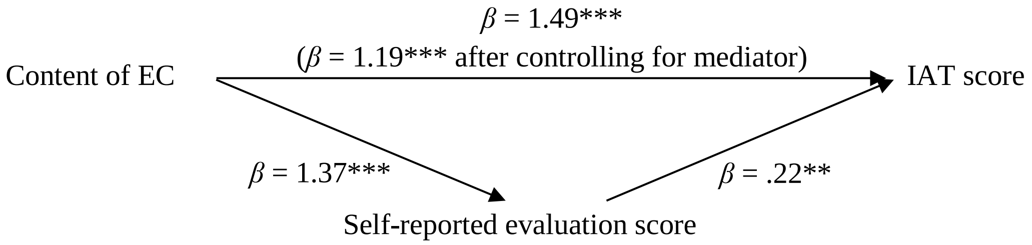
β = 1.49***
(*β* = 1.19*** after controlling for mediator)
Content of EC
IAT score
β = 1.37***
β = .22**
Self-reported evaluation score

Figure A2. Standardized Estimates of mediation coefficients for participants in Experiment 1 who received no trait instructions and were in the EC experience condition. * *p* < .05 ** *p* < .01 *** *p* < .001.

Trait instructions condition.

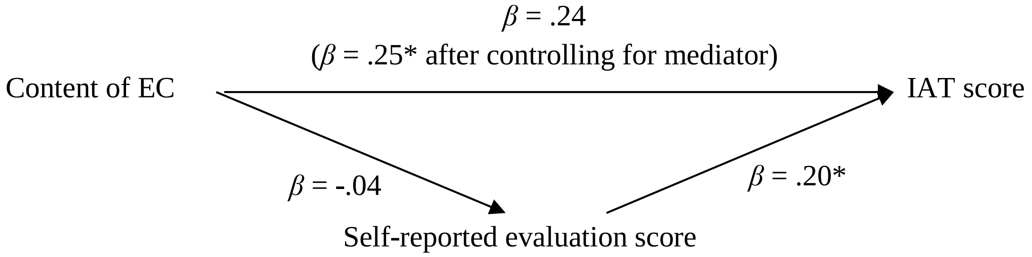
β = .24
(*β* = .25* after controlling for mediator)
Content of EC
IAT score
β = -.04
β = .20*
Self-reported evaluation score

Figure A3. Standardized Estimates of mediation coefficients for participants in Experiment 1 who received trait instructions and were in the EC instructions condition. * *p* < .05 ** *p* < .01 *** *p* < .001.

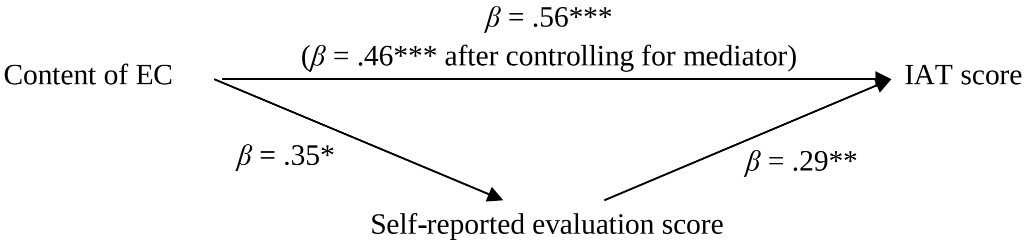
β = .56***
(β = .46*** after controlling for mediator)
Content of EC
IAT score
β = .35*
β = .29**
Self-reported evaluation score

Figure A4. Standardized Estimates of mediation coefficients for participants in Experiment 1 who received trait instructions and were in the EC experience condition. * *p* < .05 ** *p* < .01 *** *p* < .001.

Experiment 2

No trait instructions condition.

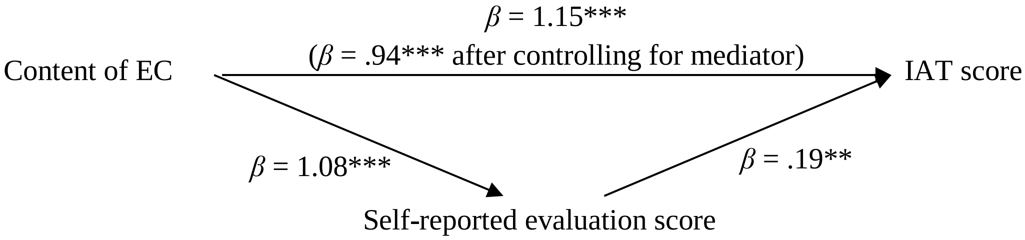
β = 1.15***
(β = .94*** after controlling for mediator)
Content of EC
IAT score
β = 1.08***
β = .19**
Self-reported evaluation score

Figure B1. Standardized Estimates of mediation coefficients for participants in Experiment 2 who received no trait instructions and were in the EC instructions condition. * *p* < .05 ** *p* < .01 *** *p* < .001.

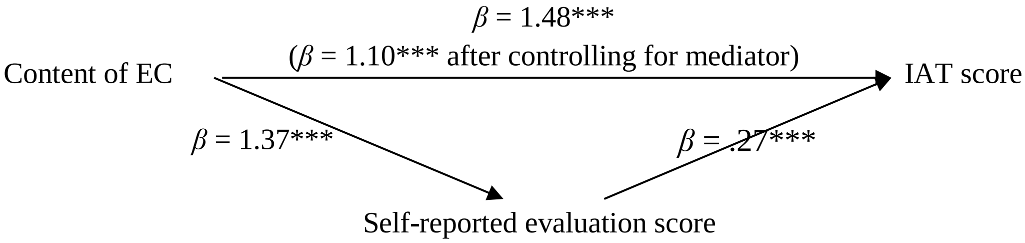
β = 1.48***
(β = 1.10*** after controlling for mediator)
Content of EC
IAT score
β = 1.37***
β = .27***
Self-reported evaluation score

Figure B2. Standardized Estimates of mediation coefficients for participants in Experiment 2 who received no trait instructions and were in the EC experience condition. * *p* < .05 ** *p* < .01 *** *p* < .001.

Trait instructions condition.

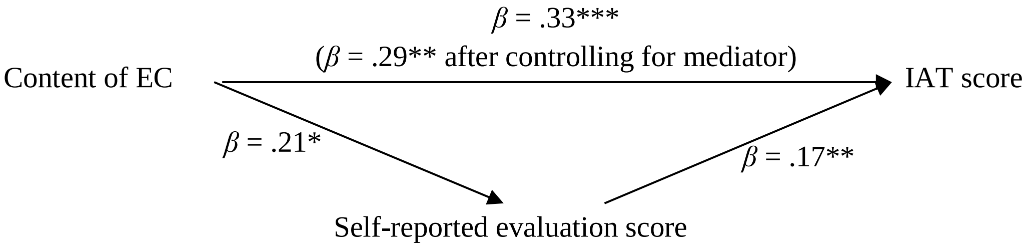
β = .33***
(β = .29** after controlling for mediator)
Content of EC
IAT score
β = .21*
β = .17**
Self-reported evaluation score

Figure B3. Standardized Estimates of mediation coefficients for participants in Experiment 2 who received trait instructions and were in the EC instructions condition. * $p < .05$ ** $p < .01$ *** $p < .001$.

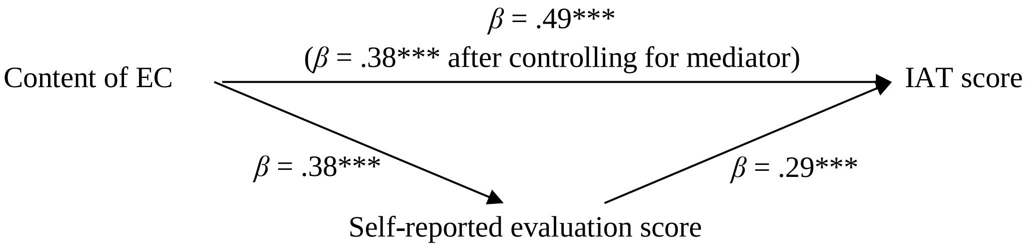
β = .49***
(β = .38*** after controlling for mediator)
Content of EC
IAT score
β = .38***
β = .29***
Self-reported evaluation score

Figure B4. Standardized Estimates of mediation coefficients for participants in Experiment 2 who received trait instructions and were in the EC experience condition. * $p < .05$ ** $p < .01$ *** $p < .001$.