The Role of Trait Inferences in Evaluative Conditioning

Tal Moran[1 2], Sean Hughes[1], Pieter Van Dessel[1], & Jan De Houwer[1]

[1]*Ghent University, Belgium*

[2]*The Open University of Israel*

In press: Collabra: Psychology

Evaluative Conditioning (EC) effect is a change in evaluative responding to a neutral stimulus (CS) due to its pairing with a valenced stimulus (US). Traditionally, EC effects are viewed as fundamentally different from persuasion effects. Inspired by a propositional perspective to EC, four studies (N = 1,284) tested if, like persuasion effects, EC effects can also be driven by *trait inferences*. Experiments 1-2 found that promoting trait inferences (by pairing people with trait words rather than nouns) increased EC effects. Experiments 3-4 found that undermining trait inferences (by questioning the validity of those inferences) decreased EC effects. In all experiments, however, EC effects were still significant when trait inferences were invalid. Taken together, our findings (a) suggest that trait inferences can play an important role in EC effects, (b) constrain theoretical models of EC, and (c) have important implications for applied EC interventions.

Keywords: Evaluative Conditioning; Trait Inference; Automatic Evaluation; Propositional Models

A central premise of research and theory in psychological science is that behavior is frequently guided by what we like and dislike (e.g., Allport, 1935). The study of attitudes (i.e., of the environmental moderators and mental mediators of evaluations) shows that our preferences exert a powerful influence over our judgements, decisions, and behaviors (Ajzen & Fishbein, 2005; Glasman & Albarracin, 2006). Given the key role that evaluations play in everyday life, it is not surprising that much attention has been devoted to understanding how and why evaluations are formed and changed (e.g., Wegener et al., 2018).

Correspondence concerning this article should be addressed to tmo286@gmail.com. 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 materials, data and analysis scripts of the whole project are available at the Open Science Framework (osf.io/dphac/).

Contributions: TM contributed to the experimental design, creation of the materials, data collection, analysis, and writing of the paper. SH and PVD 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.

Evaluative Conditioning and Persuasion

A well-studied pathway for establishing and modifying evaluations is known as Evaluative Conditioning (EC). In a typical EC task, a (conditioned) stimulus (CS) is repeatedly paired with a positive or negative (unconditioned) stimulus (US), and as a result, the former typically acquires a similar valence as the latter (i.e., an EC effect; De Houwer, 2006; Hofmann et al., 2010). People encounter stimulus pairings throughout the course of their everyday lives, from exposure to advertisements and media, to social interactions. As such, EC may play a role in many behaviors such as consumer behavior (e.g., Gibson, 2008) and phobias (e.g., Merckelbach et al., 1993).

In the attitude literature, EC is often contrasted with a second evaluative learning pathway: persuasion (e.g., Petty & Cacioppo, 1984; Skowronski & Carlston, 1989). EC is typically viewed as fundamentally different from persuasion and these two pathways have long been studied in isolation from one another (for a discussion, see De Houwer & Hughes, 2016). Whereas

EC research examines changes in evaluative responding that are due to *stimulus pairings*, persuasion research mainly studies changes in evaluative responding that are due to *verbal arguments*.

EC and persuasion are also assumed to strongly differ in the mental processes that underlie their effects. In the past, EC effects were often said to be mediated by the automatic formation and activation of associations in memory, a process that is assumed to take little cognitive effort (e.g., Baeyens et al., 1992, 1995; Levey & Martin, 1975; Martin & Levey, 1994; Olson & Fazio, 2001). In contrast, many persuasion effects were thought to be mediated by inferences that require effortful processing of words and sentences (e.g., Briñol et al., 2009). One well-studied type of inferences in persuasion research is known as *trait inferences*. Trait inferences involve the assignment of traits or attributes (psychological or non-psychological qualities or features that regarded as a characteristic or inherent part of someone or something) to a target person or object based on some (source) information (e.g., Kelly, 1973; Skowronski & Carlston, 1989). Persuasion research on trait inference traditionally focused on a specific type of trait inference: inferring the traits of a person or an object based on information other than merely telling participants about the person possessing this trait. For example, this research finds that people often draw inferences about another person's personality traits or other characteristics based on their physical appearance (e.g., Olivola & Todorov, 2010), their behavior (e.g., Uleman et al., 2008), or their facial expressions (e.g., Knutson, 1996).

Research on trait inferences and their role in persuasion has flourished over the years (e.g., Nisbet & Ross, 1980; Uleman et al., 2008). It is not our goal here to provide a comprehensive literature review of trait inferences research, however, to give a few examples, we now know that people can spontaneously derive trait knowledge (e.g., cruel) about others from behavioral descriptions (e.g., kicking puppies) of those individuals (spontaneous trait inference effect; Carlston & Skowronski, 2005; Uleman et al., 2008). Likewise, behaviors which are considered highly diagnostic (i.e., have high informational value) exert far more impact on trait inferences than those that are less diagnostic (e.g., Menon et al., 1995; Skowronski & Carlston, 1989). Some research examined how people infer specific personality traits (e.g., introvert; Trope & Bassok, 1983) while other research focused on how people infer more general valence traits (e.g., positive; Gregg et al., 2006), based on verbal behavioral descriptions.

Do Trait Inferences Play a Role in EC Effects?

As mentioned above, EC effects were viewed in the past as being mediated by the automatic formation and activation of mental associations between CS and US representations in memory (e.g., Baeyens et al., 1992; Levey & Martin, 1975). Yet, recent theoretical developments have called this idea into question (De Houwer, 2009, 2018). According to a propositional perspective, EC effects are not due to associative processes, but instead reflect the formation and activation of *propositions* in memory. These propositional representations differ from associative representations in that they (a) encode relational information (i.e., they specify how stimuli are related) and (b) have a truth value (i.e., they can be evaluated as being true or false).

Propositional accounts make several assumptions about EC effects that are not shared by associative models. One such assumption is that inferential processing plays a key role in EC effects (e.g., De Houwer et al., 2020). Because propositions have a truth value they allow for inferential reasoning to occur (i.e., the activation or construction of new propositional information based on its compatibility with other activated information). Thus, the propositional accounts make the unique prediction that inferences can underlie EC effects (Van Dessel et al., 2019). For example, when stimuli are paired with one another, a proposition is formed about this pairing (e.g., "*A co-occurs with B*"). This proposition can be combined with pre-existing propositions about the meaning of pairings (e.g., "*similar things tend to co-occur*"), which can lead to the formation of an inference about how those stimuli are related (e.g., "*A and B are similar*"). This then sets the stage for an evaluative inference (e.g., "*A has the same valence as B*"; see De Houwer, 2018) that transfers into evaluative responding.

From this perspective, the spatio-temporal pairing of stimuli, in combination with one's prior knowledge, functions as a sort of argument. For humans, these "regularity-based" arguments can function in the same way as verbal arguments that consist of words and sentences (De Houwer & Hughes, 2016). If this is the case, then EC and persuasion effects may be more alike than previously assumed, insofar as both are driven by inferences. This idea could apply also to

trait inferences given that they are a subclass of inferences that may be crucial when determining valence. Specifically, the proposition that is formed about pairings (e.g., "*an unknown object [A] co-occurs with a positive stimulus [B]*") can be combined with pre-existing propositions about the meaning of pairings ("*co-occurrence indicates similarity*") to form trait inferences (e.g., "*A possesses similar traits as B*") which lead to changes in evaluative inferences (e.g., "*A possesses similar valence as B*"). [1]

Our goal in the present research is to explore the potential role of trait inferences in EC effects. It is important to note that we use the term "trait inference" slightly differently than how it is traditionally used in persuasion research. As mentioned above, persuasion research usually uses the term "trait inference" to refer to inferring the traits of a person or an object based on information that does not involve telling participants about the person possessing the trait itself. In fact, there is typically no mention of the target trait itself. In the present research, in the context of EC, we also use the term "trait inference" to refer to the conclusion that a person or an object holds a trait. Importantly, we argue that this inference can be based on the spatio-temporal pairing of this person or object (CS) with a US. While we also do not tell participants about the person possessing the trait, for our purposes, the US can be a trait word as this can promote inferences of trait possession based on (pairing) information. Indeed, our goal is not to test if people can infer traits based on information that does not mention the trait but rather to test if people infer traits based on spatio-temporal pairing and if these inferences contribute to EC effects.

The idea that trait inferences can influence EC effects seems plausible given that many EC studies provide participants with instructions and cover stories prior to the pairing phase (e.g., Balas & Gawronski, 2012; Gast & Rothermund, 2011; Heycke & Gawronski, 2019; Hu et al., 2017; Hughes et al., 2019; Mitchell et al., 2003; Zanon et al., 2014). These instructions may serve as a (persuasive) context that promotes trait inferences. For example, some researchers (Balas & Gawronski, 2012; Mitchell et al., 2003; Zanon et al., 2014) told participants that the study concerned language acquisition and that they would be presented with words (CSs) together with pictures (USs) illustrating their meaning. Others informed participants that they would see images of pharmaceutical products (CSs) and visual information about their effects (USs) (Heycke & Gawronski 2019; Hu et al., 2017). Such instructions or cover stories may direct participants to infer that the CSs have traits related to the USs (e.g., the pharmaceutical product causes positive health effects).

It is plausible that, even in those EC studies that do not use a cover story or provide pre-pairing instructions, the types of stimuli used (i.e., the specific CSs and USs) create conditions that promote trait inferences (e.g., Gibson, 2008; Kim et al., 1996; Sweldens et al., 2010). For example, it is possible that the presentation of Coca-Cola brand (CS) with the word "awesome" (US; Gibson, 2008) leads participants to infer that Coca-Cola is awesome (and therefore positive). Similarly, it is possible that the presentation of Belgium beers (CSs) with pictures of adults having fun in various ways (USs; Sweldens et al., 2010) leads participants to infer that the Belgium beers are fun and positive. Although there are also EC procedures that do not seem to promote trait inferences via instructions or biased selection of stimuli (e.g., Moran et al., 2021; Olson & Fazio, 2001), the frequent use of EC procedures that do promote trait inferences in this way suggests that trait inferences might have been crucial in many prior EC studies. Interestingly, however, no prior study has attempted to manipulate trait inferences to systematically investigate their potential role in EC effects.

It is worth noting that traits have been used as USs in prior EC studies (e.g., Gibson, 2008; Hughes et al., 2019; Milner et al., 2017). For instance, Wagner et al. (2020) set out to test the effectiveness of an EC procedure that was previously found to change child-related attitudes and expectations in adults (see Milner et al., 2017). To do so, the authors replicated the same EC procedure as Milner et al. but replaced the trait words

[1] It may be that when the proposition "*Person A possesses similar traits as Person or stimulus B*" is generated, people may also use this proposition to make an inference about the valence of Person A (e.g., "*Person A is of similar valence to Person or stimulus B*"). People may spontaneously do so because evaluative inferences are highly useful in everyday life (e.g., they can quickly adapt how they respond to Person A without having to learn about that individual via trial and error; Carlston & Skowronski, 2005; Uleman et al., 2008). It may be that people apply this tendency to the paired stimuli in the EC task as well. Another possibility is that people only make valence inferences whenever valence becomes relevant (e.g., during the evaluative task). Distinguishing between these alternative possibilities is an interesting avenue for future research but it is not the focus of the current research.

that were originally used as USs with emoji. EC effects were found to be weaker (relative to the original research) and the authors concluded that trait inferences mediated the original effects. However, because Wagner et al.'s study did not directly compare the two types of USs, it is difficult to make strong conclusions regarding the role of trait inferences in their EC effects. The present research is the first to systematically test the potential role of trait inferences in EC by directly comparing the effect of different conditions that promote versus undermine trait inferences on EC effects.

Examining the role of trait inferences in EC could shed new light on the moderators of, and the processes underlying, EC as well as the relation between EC and persuasion. Moreover, it can inform us about a potentially important and widespread source of trait inferences that so far received little attention: the spatio-temporal pairing of stimuli.

The Present Research

Across four studies, we tested the idea that trait inferences contribute to EC effects. We did so in two ways. In Experiments 1-2, we manipulated trait inferences by pairing neutral stimuli (male faces) with valenced traits compared to valenced nouns. In Experiment 3, we provided information about the validity of pairings as an information source that can be used for trait inferences. Finally, in Experiment 4, we combined the two manipulations to test their interactive effect.

Traits versus Nouns as USs

Experiments 1-2 used a job-hiring cover story and unfamiliar men (potential job candidates) as CSs. We targeted trait inferences by manipulating the content of the USs by either using as USs (a) positive (e.g., Puppies) and negative nouns (e.g., Cemetery) that are unlikely to give rise to trait inferences about the unfamiliar men (which was also verified in a pretest, see more details below) or (b) positive (e.g., Agreeable) and negative trait adjectives (e.g., Reckless) that are more likely to give rise to trait inferences (also see the pretest described below). The logic behind this manipulation was that that pairing of person with a personality trait word will lead to an inferential leap that involves seeing the personality trait word as being a property of the person (e.g., James is reckless). Such an inferential leap would be less likely to occur when persons are paired with nouns (e.g., James is cemetery).

EC effects were indexed using both self-report and automatic evaluation measures. [2] Discussions regarding the mental processes that might (differentially) impact performance in self-report and automatic evaluation measures are still ongoing (e.g., Bar-Anan & Nosek, 2012; Bar-Anan & Vianello, 2018; Payne et al., 2013; Schimmack, 2019). Therefore, rather than using these different measures to make claims about different mental processes, we used both types of measures primarily to provide convergent evidence for our claims, that is, to generalize the findings beyond one specific measure. However, it is also worth noting that different theoretical perspectives diverge in their predictions regarding the sensitivity of automatic versus self-reported evaluation to inference processes (e.g., De Houwer, 2014; Gawronski & Bodenhausen, 2018). We will return to this in the General Discussion. Experiments 1-2 were identical except for the automatic evaluation measure used. Experiment 1 used a modified version of the Affect Misattribution Procedure (AMP; Mann et al., 2019) and Experiment 2 used the Implicit Association Test (IAT; Greenwald et al., 1998).

The Validity of Making Trait Inferences on the Basis of Pairings

Research on impression formation indicates that the perceived validity of encountered information influences subsequent evaluations (e.g., Golding et al., 1990; Peters & Gawronski, 2011). Specifically, when instructions suggest that a certain piece of information is not true, people usually discount or ignore this information, both when making trait inferences and when forming impressions. Elsewhere, others have found that when the source of information is considered to be low in credibility, the information has less of an impact on trait inferences and impression formation compared to when the source is considered to be high in credibility (e.g., Smith et al., 2013).

If trait inferences can also influence EC effects then similar findings should emerge when neutral male faces (CS) are paired with traits (USs) and the validity of the information source (i.e., pairings) is manipulated. Specifically, inferences that the CS has the US trait should be reduced when the validity of pairings is undermined compared to when it is promoted. With

[2] We view automaticity as an umbrella term that refers to a set of non-overlapping and non-redundant suboptimal conditions for mental processing (e.g., Moors, 2016; Moors & De Houwer, 2006; Van Dessel et al., 2020). Therefore, the automatic evaluation measures used in the current research are argued to capture evaluations under one or more of these suboptimal conditions.

this in mind, Experiment 3 manipulated the validity of pairings as an information source. Participants in the low validity condition read that a hacker had been hired by a competing company to create chaos in their company (e.g., by deleting all the original [genuine] information about the two candidates and inserting fake information in its place). Participants in the high validity condition were additionally informed that the security department had identified this hacking problem and fixed it. As such, they would see the original (genuine) information about the candidates.

To further test the role of trait inference in EC effects, Experiment 4 combined the trait versus nouns as USs manipulation of Experiments 1-2 and the validity manipulation of Experiment 3. Combining these manipulations allowed us to test the boundaries of the trait inference idea. First, we tested whether the effect of using traits rather than nouns as USs is moderated by the perceived validity of pairings as an information source. To illustrate, imagine that people are told prior to an EC procedure that the pairings they will soon encounter are a non-valid piece of information. In such a case, does the content of the USs still matter (i.e., would pairing the men with trait adjectives rather than nouns still moderate EC effects)? If trait inferences contribute to EC effects, then the effect of CS-US pairings on evaluative responding should depend on the assumed relationship between the paired stimuli. If people know in advance that pairings do not indicate possession, then the content of the US (whether it is an adjective or noun) should not matter. Critically, no prior work has tested the mutual effects of validity and US content on EC effects. Finally, we examined if EC effects would still emerge under conditions designed to strongly reduce trait inferences (i.e., when the pairings are said to be an invalid source of information and the USs are unlike to elicit strong trait inferences [nouns]).

Materials, sampling plans, exclusion rules, and analysis plans for all experiments were pre-registered. [3] We report all data exclusions, manipulations, measures, and how we determined our sample sizes. Materials and data of the whole project can be found at the Open Science Framework (osf.io/dphac/).

Experiments 1-2: Traits versus Nouns as USs

In Experiments 1-2, we targeted trait inferences by manipulating the content of the USs. We tested if the manipulation of US content would moderate EC effects. The two experiments were identical with the exception of the automatic evaluation measure used: Experiment 1 employed a modified version of the AMP (Mann et al., 2019). Experiment 2 replicated Experiment 1 while substituting the modified AMP for an IAT (Greenwald et al., 1998).[4]

Method

Participants and design. Participants in Experiment 1-4 were recruited online via the Prolific Academic website (https://prolific.ac). Each experiment took about 14 minutes to complete and all participants were paid £1.40. Pre-registered materials are available at osf.io/p8cvf (Experiment 1), and osf.io/nuhw6 (Experiment 2). [5] In Experiments 1-3, we used a Sequential Bayes Factor (SBF) design with a maximal N to determine sample size (Schönbrodt, & Wagenmakers, 2018). We used a threshold of $BF > 6$ or $BF < 0.16$, a minimum of 60 participants, an addition of 60 participants for each test, and a maximum of 240 participants. In total, 241, and 239 participants completed Experiment 1 and 2, respectively. In line with the preregistrations, in Experiment 1, we excluded participants who had more than 95%, or less than 5%, *pleasant* responses on the AMP (i.e., 14 participants) (Moran et al., 2017). In Experiment 2, we excluded participants who (1) had more than 10% fast trials in the IAT (RT< 300ms; Greenwald et al., 2003) or (2) did not complete all the measures (total 7 participants). The final samples consisted of 227 participants for Experiment 1 (51% Women, M_{age}= 33.63, SD_{age}= 11.60) and 232 participants for Experiment 2 (56% Women, M_{age}= 34.10, SD_{age}= 10.86).

[3] This research project also included four experiments that are not reported in this paper. Two pilot experiments constituted initial (but weak) attempts to manipulate US content and validity and examine their effects. Details about these experiments can be found online (Pilot 1: osf.io/ayf8c/, Pilot 2: osf.io/yu8nb/). Two additional experiments (S1 and S2), due to length considerations, are reported in full in the online supplement (osf.io/qscvm/; see Footnotes 4 and 8 for more details about these experiments).

[4] The online supplement (osf.io/qscvm/) reports another experiment (Experiment S1) that was identical to Experiment 2 but accidently used the same word (Awful) both in the conditioning and IAT procedures.

[5] The preregistered documents can be consulted using the "Files" link on the left of the screen. Note that some of the names of the manipulated factors are different between the preregistration documents and the manuscript. We changed these names because we wanted to be more conceptually accurate in the descriptions of our manipulations.

Experiment 1 involved a three-factor between-subjects design: 2 (*US Content*: nouns vs. adjectives) x 2 (*CS-US assignment*: James paired with positive, Chris paired with positive) x 2 (*Measures order*: self-report measure first, automatic evaluation measure first), with self-reported and automatic evaluations as the dependent variables. The design of Experiment 2 was the same, except for adding a fourth between-subjects factor (*IAT block order*: compatible block first, incompatible block first). The preregistered analytic strategy, in all the experiments, was to use an ANOVA model that includes the main manipulation and all the counterbalancing procedural factors. As specified in the preregistration, the counterbalancing procedural factors were included in the models to control for their influence, but their effects were not analyzed.

Materials. CSs were pictures of two males selected from an open database of facial stimuli (Minear & Park, 2004; pre-tested by Bar-Anan & Amzaleg-David, 2014). We named them Chris and James. The USs were positive and negative words (half of the words were adjectives and half were nouns; see Table 1), selected based on a pretest (N = 92, see the online-supplement osf.io/qscvm/ and osf.io/v6kbj/ for details), wherein participants were asked to rate different adjectives and nouns on valence and the ability to infer evaluative traits from their pairing with a person (question: '*how diagnostic is this word for inferring the true valence of the person? When we ask how 'diagnostic' you consider the word to be, we are asking you to carefully think about how much presenting a person with this word tells you something about him [i.e., how positive or negative he is likely to be*]'). The two positive sets did not differ significantly in valence ratings, $t(91) = -0.45$, $p = .65$, $d = -0.04$, $BF_{10} = 0.12$, but they did differ significantly in trait-inference ratings, $t(91) = 11.82$, $p < .001$, $d = 1.23$, $BF_{10} > 1000$. The positive adjectives were rated as more diagnostic (M = 3.10, SD = 0.75) than the positive nouns (M = 1.67, SD = 0.87). Likewise, the two negative subsets did not differ in valence ratings, $t(91) = 0.26$, $p = .80$, $d = 0.02$, $BF_{10} = 0.11$, but did differ in trait-inference ratings, $t(91) = 11.99$, $p < .001$, $d = 1.25$, $BF_{10} > 1000$. The negative adjectives were rated as more diagnostic (M = 3.20, SD = 0.75) than the negative nouns (M = 1.69, SD = 0.86).

As targets in the AMP (Experiment 1), we used 120 abstract paintings adapted from Mann et al. (2019). As primes in the AMP (Experiment 1) and men target stimuli in the IAT (Experiment 2), we used four different versions of the picture of Chris and of James that varied in their color (colored vs. grayscale) and the presence of a surrounding frame (no frame vs. yellow frame). The IAT also used four negative and four positive words (see Table 1).

Procedure.

Instructions. The full procedure, instructions, and materials of all the experiments are explained in detail in the online-supplement (osf.io/qscvm/). At the beginning of the experiment, participants read a cover story asking them to imagine that they are working for a large company and that it is their job to recruit new employees for the company. We informed them that they would receive information about two potential candidates for the job: Chris and James and that their task was to judge which of the two men they would like to hire.

EC. The EC procedure consisted of two blocks of 24 trials (48 trials in total). Each trial simultaneously presented a male face and his name in compound (CS) on the left side of the screen and a valenced word (US) on the right side of the screen for 2500ms. The inter-trial interval was 1000ms. During each block, one of the men (CS_{pos}) was presented three times with each of four different positive words, whereas the second man (CS_{neg}) was presented three times with each of four negative words. Assignment of the men to positive or negative USs was counterbalanced across participants. We manipulated the type of US words (nouns vs. adjectives) between participants.

Self-reported ratings. Participants rated their liking of each of the two men by answering three questions: '*How good or bad do you consider the above individual?*', '*How much do you like or dislike the above individual?*' and '*How positive or negative do you consider the above individual?*. Response scale was a 9-point Likert scale (-4 = *Very bad/ I dislike him a lot/ Very negative*; +4 = *Very good/ I like him a lot/ Very positive*). Rating scores for each target person were calculated by averaging the three questions (Cronbach's Alpha > .92 in Experiments 1-4). We calculated EC scores on self-report ratings by subtracting the mean score rating for CS_{pos} from the mean score rating for CS_{neg}.

AMP. In Experiment 1, to measure automatic evaluation, we used a modified version of the AMP (Mann et al., 2019) that was designed to reduce intentional responding in the task. Each trial of the AMP displayed stimuli in the following sequence: (1) A photograph of

one of the two targets for 100ms, (2) a blank screen for 100ms, (3) an abstract painting for 100ms, and (4) a pattern mask of black-and-white noise until participants responded. Upon presentation of the mask, participants were prompted to indicate if the painting was more pleasant or more unpleasant than average using two response keys on the keyboard signifying *pleasant* and *unpleasant*. Participants completed three blocks of 40 trials, each with 20 primes for each man. EC scores were computed by subtracting the proportion of *pleasant* responses after primes of the CS_{pos} from the proportion of *pleasant* responses after CS_{neg} primes. The internal consistency, as computed on the basis of two parcels (split-half) of the task, was α = .84.

IAT. In Experiment 2, we used the IAT to measure automatic evaluation. In the IAT, participants categorized stimuli using two computer keys. In the critical blocks, participants responded with the left key to stimuli of two categories (e.g., "James" and "Good"), and with the right key to stimuli of two other categories (e.g., "Chris" and "Bad"). In two of these blocks, *James* and *Good* shared the same response key, and in the other two critical blocks, *Chris* and *Good* shared the same response key. The seven-block IAT followed the procedure described in Nosek et al. (2005). The D4 algorithm was used to create IAT scores (Greenwald et al., 2003). We computed these scores such that they indicate a more positive evaluation for CS_{pos} than for CS_{neg}. The IAT internal consistency, as computed on the basis of two parcels (split-half) of the task, was α = .79.

Exploratory questions. At the end of the experiment, the participants answered exploratory questions related to their experiences in the study (for a full description of these questions and associated findings see the online-supplement osf.io/qscvm/).

Results

Self-report ratings. The mean EC scores on self-report ratings (and automatic evaluations) as a function of US content condition in Experiments 1-2 are detailed in Table 2. The ANOVA model (see Table 3 for details on the ANOVA models in all the experiments) revealed a main effect of US content in Experiment 1: $F(1, 219) = 73.33$, $p < .001$, $\eta_p^2 = .25$, 90%CI [.16, .31], $BF_{10} > 1000$, and in Experiment 2: $F(1, 216) = 53.18$, $p < .001$, $\eta_p^2 = .20$, 90%CI [.11, .25], $BF_{10} > 1000$. In all cases, EC effects were stronger when USs were trait-adjectives (*Ms* = 4.56, 3.76, *SDs* = 2.22, 2.50, in Experiments 1-2, respectively) than nouns (*Ms* = 1.61, 1.51, *SDs* = 2.91, 2.39). In both experiments, EC effects significantly differed from zero both in the *adjectives* and the *nouns* conditions (see Table 2).

Automatic evaluation.

AMP scores. The ANOVA (see Table 3) on AMP scores in Experiment 1 did not show an effect of US content, $F(1, 219) = 1.21$, $p = .27$, $\eta_p^2 = .01$, 90%CI [0, .03], $BF_{10} = 0.26$, indicating no significant difference between the size of EC effects when USs were adjectives ($M = 0.04$, $SD = 0.19$) or nouns ($M = 0.07$, $SD = 0.22$) (see Figure 1). EC effects significantly differed from zero in both conditions (Table 2).

IAT scores. Unlike the results of Experiment 1, the ANOVA (Table 3) on automatic evaluations in Experiment 2 revealed main effects of US content, $F(1, 216) = 7.48$, $p = .007$, $\eta_p^2 = .03$, 90%CI [.004, .07], $BF_{10} = 7.39$, indicating a stronger IAT effect when USs were adjectives ($M = 0.30$, $SD = 0.46$) than nouns ($M = 0.14$, $SD = 0.41$). IAT effects significantly differed from zero in the *adjectives* and *nouns* conditions (Table 2).

Direct comparison between self-reported and automatic evaluation effects. [6] To directly compare the effect of US content on self-reported and automatic evaluations, we first computed a preference for James over Chris for both measures and then standardized these scores. Next, we recoded the standardized scores to reflect a preference for CS_{pos} over CS_{neg}. The mean standardized EC scores as a function of US content condition and measure type, in Experiments 1-2 are illustrated in Figure 1 (panels A-B, respectively). We submitted these standardized scores to mixed ANOVA models (see Table 3). Table 4 summarizes the full results of the ANOVAs. Importantly, the interaction between US content and measure type was significant in Experiment 1, $F(1, 219) = 37.86$, $p < .001$, $\eta_p^2 = .15$, 90%CI [.08, .21], $BF_{10} > 1000$ [7], and Experiment 2, $F(1, 216) = 5.60$, $p = .019$, $\eta_p^2 = .03$, 90%CI

[6] This analysis was not preregistered for Experiment 1. Given the dissociative results of Experiment 1, however, we considered this analysis potentially interesting and decided to preregister it for Experiments 2-4.

[7] For analyses comparing the effect of the trait inference manipulation on self-report ratings and automatic evaluations, in all the experiments, to avoid complexity, the Bayes-factors were computed using a model that exclude the counterbalancing procedural factors

[.002, .06], BF_{10} = 1.05. In both experiments, the interaction indicated that the effect of US content on self-report ratings, (Experiment 1: $F(1, 219) = 73.33$, $p < .001$, $\eta_p^2 = .25$, $BF_{10} > 1000$; Experiment 2: $F(1, 216) = 53.18$, $p < .001$, $\eta_p^2 = .20$, $BF_{10} > 1000$) was stronger than the effect on automatic evaluations, (Experiment 1: $F(1, 219) = 1.21$, $p = .272$, $\eta_p^2 = .01$, $BF_{10} = 0.27$; Experiment 2: $F(1, 216) = 7.48$, $p = .007$, $\eta_p^2 = .03$, $BF_{10} = 7.84$).

Discussion

Results indicate that the nature of the US (nouns vs. trait-adjectives) strongly moderated EC effects as indexed by self-report ratings. Effects were stronger when USs were trait-adjectives compared to when they were nouns. This was also the case for automatic evaluations when measured by an IAT, but not when measured using a modified AMP (Mann et al., 2019). A direct comparison of this moderating effect across self-reported and automatic evaluation measures revealed a stronger effect of US content on the former than on the latter types of measures.

These findings support the idea that trait inferences moderate self-reported and automatic evaluations (at least when measured with the IAT), and the impact of such a manipulation is stronger for self-reported than automatic evaluations. It is also worth noting that, for all measures and in all experiments, EC effects always significantly differed from zero when USs were nouns (Table 2). This suggests that people preferred the CS paired with positive USs over the CS paired with negative USs even under the conditions that did not encourage trait inferences. In Experiment 3, we tested if trait inferences contribute to EC effects by using a different manipulation of trait inferences. Specifically, we manipulated the perceived validity of pairings as an information source. [8]

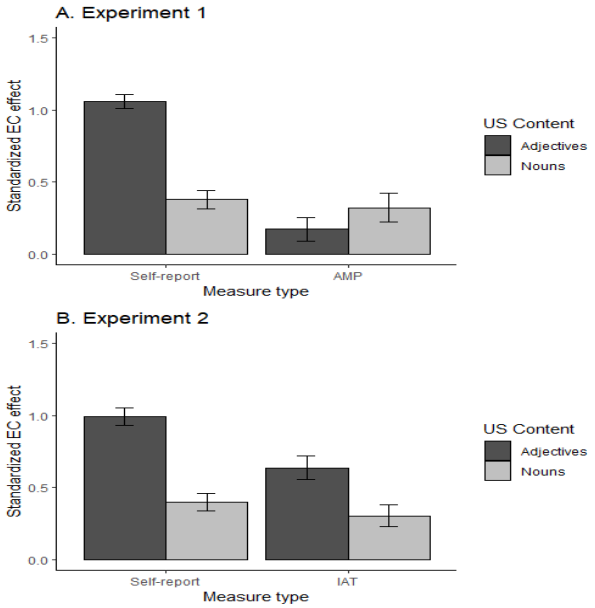

Figure 1. Experiments 1-2: standardized EC effects (preference for the CS that was paired with positive words over the CS that was paired with negative words) as a function of US content (adjectives versus nouns) and measure type (self-report ratings versus AMP/IAT), in Experiment 1 (A) and Experiment 2 (B). Error-bars represent standard errors.

Experiment 3: Validity of Pairings as an Information Source

Method

Participants and design. Preregistered materials for Experiment 3 are available at osf.io/jsg2h. 268 participants completed Experiment 3. We excluded participants who had more than 10% of trials that were too quick in the IAT or those who did not complete all measures (i.e., 28 participants). The final sample included 240 participants (55% Women, M_{age}= 34.20, SD_{age}= 11.40). The experiment involved a four factor between-subjects design: 2 (*Validity*: high vs. low) x 2 (*CS-US assignment*) x 2 (*Measures order*) x 2 (*IAT block order*), with self-reported and automatic evaluations as the main dependent variables.

Materials and procedure.

Instructions. To manipulate the validity of the pairing as an information source, we presented different instructions before the EC phase. In the low validity condition, we modified the cover story to suggest that a hacker, hired by a competing company, deleted the

[8] The online supplement (osf.io/qscvm/) reports another experiment (Experiment S2) that used a weaker manipulation of validity and found weaker results.

original information about Chris and James and inserted new fake information about them. In the high validity condition, the instructions also informed the participants about the hacker but also informed them that the security department identified the problem and fixed it and that the database now included the original and correct information about Chris and James. Prior to the study, the two cover stories were pretested (N = 351, see osf.io/mpyqv/ and the online-supplement osf.io/qscvm/ for more details) for their believability and the extent to which they promote trait inference (i.e., to what extent was the pairing of James and Chris with positive or negative information reflective of something about their true character).

Immediately after reading the validity instructions, participants answered two open-ended questions: "*Please describe - in a few words - what you were just told*", and "*What implications does this information have for what you are soon going to learn about Chris and James?*" The purpose of including these two questions was to increase the chance that participants would elaborate on the implications of the cover story for the pairing they were about to see. We did not analyze the question responses nor exclude participants based on their responding.

Other experiment phases. All other procedural details (EC procedure, self-report and IAT measures) were the same as in Experiment 1-2 except that in the EC procedure, for all participants, USs were always adjectives. Participants also answered the same exploratory questions as in Experiments 1-2 with the addition of a believability question about the cover stories (see online supplement osf.io/qscvm/). The internal consistency of the IAT was α = .80.

Results

Self-report ratings. The ANOVA (see Table 3) revealed a main effect of validity, $F(1, 224) = 53.90$, $p < .001$, $\eta_p^2 = .19$, 90%CI [.11, .25], $BF_{10} > 1000$, indicating a stronger effect when participants read the high validity instructions (M = 4.10, SD = 2.40) than the low validity instructions (M = 1.27, SD = 3.45). Effects significantly differed from zero in the high and low validity conditions (see Table 2).

IAT scores. The ANOVA on the IAT scores (Table 3) revealed a significant effect of validity, $F(1, 224) = 7.41$, $p = .007$, $\eta_p^2 = .03$, 90%CI [.004, .07], $BF_{10} = 5.14$, indicating a stronger effect when participants received the high validity instructions (M = 0.32, SD = 0.44) than low validity instructions (M = 0.18, SD = 0.44). Again, effects significantly differed from zero in both the high and low validity conditions (Table 2).

Direct comparison between self-reported and automatic evaluation effects. Standardized scores were computed and analyzed as in Experiment 1-2 with replacement of the US content factor with the Validity factor. The mean standardized scores as a function of validity condition and measure type, are illustrated in Figure 2, the ANOVA model is specified in Table 3, and the ANOVA results are summarized in Table 4. Importantly, the interaction between validity and measure type was significant, $F(1, 224) = 8.78$, $p = .003$, $\eta_p^2 = .04$, 90%CI [.007, .08], $BF_{10} = 10.38$. The interaction revealed a stronger effect of validity on the self-report ratings, $F(1, 224) = 53.90$, $p < .001$, $\eta_p^2 = .19$, $BF_{10} > 1000$, than on IAT effects, $F(1, 224) = 7.41$, $p = .007$, $\eta_p^2 = .03$, $BF_{10} = 4.22$.

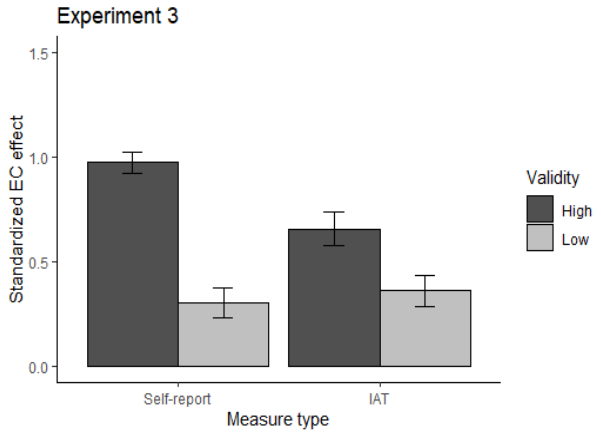

Figure 2. Experiment 3: standardized EC effects as a function of validity condition (high versus low) and measure type (self-report ratings versus IAT). Error-bars represent standard errors.

Discussion

Manipulating how valid pairings are as an information source strongly moderated EC effects as indexed by self-report ratings and IAT scores. A direct comparison between self-reported and automatic evaluative measures revealed a stronger impact of validity on the former relative to the latter. In addition, both automatic and self-reported EC effects were significantly different from zero in the low validity condition. Taken together, the results of Experiment 3 indicate that creating conditions that undermine trait inferences leads to weaker (but still significant) EC effects. In Experiment 4, we sought to test the boundary conditions of trait inferences on EC effects by combining

the trait inference (US content) manipulation of Experiments 1-2 with the trait inference (validity) manipulation of Experiment 3.

Experiment 4: Interactive Effects of US Content and Validity

Method

Participants and design. Preregistered materials are available at osf.io/ys8t4. Based on power analyses, we estimated that 517 participants would provide 90% power to detect a small interaction effect of η_p^2= 0.02. To account for possible exclusions, we decided to collect data from at least 600 participants. As planned, we excluded participants who had more than 10% of trials in the IAT that were too quick, and those who did not complete all measures (i.e., 15 participants). The final sample consisted of 585 participants (56% women, M_{age} = 35.43, SD_{age} = 12.80). The experiment involved a five-factor between-subjects design: 2 (*US Content*: adjectives vs. nouns) x 2 (*Validity*: high vs. low) x 2 (*CS-US assignment*) x 2 (*Measures order*) x 2 (*IAT block order*).

Materials and procedure. To manipulate US content, we used the same stimuli and procedures as in Experiments 1-2. To manipulate validity, we used the same cover stories as in Experiment 3. IAT and self-report tasks were similar to Experiments 2-3. The internal consistency of the IAT was α = .78. Exploratory questions were the same as in Experiment 3.

Results

Analytic strategy. To test the interactive effect of US content and validity on EC effects, we submitted self-reported ratings and IAT scores to a between-participants ANOVA (see details on Table 3). We also preregistered specific hypotheses for a number of comparisons. Details on these comparisons and their results are summarized in Table 5 and in the online-supplement (osf.io/qscvm/).

Self-report ratings. The mean EC scores on self-report ratings (and automatic evaluations) as a function of validity and US content conditions in Experiment 4, are detailed in Table 2. Effects were positive and significantly different from zero in all four conditions (Table 2). The ANOVA (Table 3) revealed a significant interaction between US content and validity, F(1, 552) = 20.77, p < .001, η_p^2 = .04, 90% CI [.01, .06], BF_{10} >1000. The interaction revealed an effect of US content in the high validity condition, F(1, 262) = 56.90, p < .001, η_p^2 = .18, BF_{10} >1000, but not in the low validity condition, F(1, 290) = 1.03, p =.31, η_p^2 < .01, BF_{10} = 0.21. The ANOVA also found main effect of US content, F(1, 552) = 36.00, p < .001, η_p^2 = .06, 90%CI [.03, .09], BF_{10} >1000, indicating stronger effects when USs were adjectives (M = 2.62, SD = 3.26) than nouns (M = 1.29, SD = 2.51). A main effect also emerged for validity, F(1, 552) = 57.93, p < .001, η_p^2 = .09, 90%CI [.05, .12], BF_{10} >1000, indicating a stronger effect when participants received high validity instructions (M = 2.83, SD = 2.86) compared to low validity instructions (M = 1.14, SD = 2.86).

IAT scores. IAT effects were positive and significantly different from zero in all four conditions (see Table 2). The ANOVA (Table 3) revealed a significant interaction between US content and validity, F(1, 552) = 8.61, p = .003, η_p^2 = .02, 90%CI [.002, .03], BF_{10} = 5.49. The interaction revealed a significant impact of US content on IAT effects under high validity conditions, F(1, 262) = 23.64, p < .001, η_p^2 = .08, BF_{10} >1000, but not under low validity conditions, F(1, 290) = 0.19, p = .663, η_p^2 < .01, BF_{10} = 0.14. The ANOVA also revealed a main effect of US content, F(1, 552) = 12.78, p < .001, η_p^2 = .02, 90%CI [.006, .04], BF_{10} = 14.02, indicating a stronger effect when USs were adjectives (M = 0.23, SD = 0.43) than nouns (M = 0.12, SD = 0.40). A main effect also emerged for validity, F(1, 552) = 9.49, p = .002, η_p^2 = .02, 90%CI [.003, .03], BF_{10} = 8.81, indicating a stronger effect when participants received high validity (M = 0.23, SD = 0.40) compared to low validity instructions (M = 0.13, SD = 0.43).

Direct comparison between self-reported and automatic evaluation effects. Standardized scores were computed as in previous experiments. The mean standardized scores as a function of validity, US content, and measure type, are illustrated in Figure 3. We submitted these scores to a mixed ANOVA (see Table 3 for details). Table 4 summarize the ANOVA results. Importantly, the interaction between validity and measure type was significant, F(1, 552) = 8.91, p = .003, η_p^2 = .02, 90%CI [.003, .03], BF_{10} = 5.07, such that validity more strongly impacted self-reported ratings, F(1, 552) = 57.93, p < .001, η_p^2 = .09, than IAT effects, F(1, 552) = 9.49, p = .002, η_p^2 = .02. The interactions between US content and measure type, F(1, 552) = 1.85, p = .18, η_p^2 < .01, 90%CI [0, .01], BF_{10} = 0.29, and the three-way interaction between US content, validity, and measure type, F(1, 552) = 0.71, p = .40, η_p^2 < .01, 90%CI [0, .01], BF_{10} = 0.21, were all non-significant.

Discussion

Our key findings from Experiments 1-3 replicated. Two types of trait inference (US content and validity) manipulations moderated EC effects as indexed by self-report ratings and IAT scores. Validity exerted a stronger impact on EC effects on self-reported compared to IAT scores, whereas the effect of US content was consistent across measures. US content and validity also interacted, such that the former exerted a stronger influence on evaluative responses whenever pairings were considered to be a valid piece of information. When pairings are considered to be an invalid source of information, US content did not moderate evaluative responding. EC effects were significant in all the conditions, including when the pairings were low in validity and US were nouns.

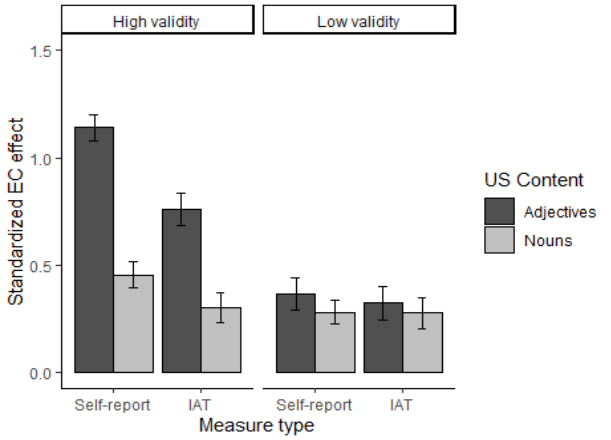

Figure 3. Experiment 4: standardized EC effects as a function of validity (high versus low), US content (adjectives versus nouns) and measure type (self-report ratings versus IAT). Error-bars represent standard errors.

Meta-Analysis

There was variability in how US content and validity manipulations moderated EC effects across our six experiments (the four experiments reported in the main text and the two experiments reported in the online-supplement osf.io/qscvm/). We therefore decided to meta-analyze (fixed effects) those experiments that manipulated US content (Experiments 1, 2, S1, and 4) and validity (Experiments S2, 3 and 4). The meta-analytic effect of *US content* on self-reported ratings, *Hedges' g* = 0.70, *SE* = 0.06, 95%CI [0.58, 0.82], *Z* = 11.51, *p* < .001, and automatic evaluations, *Hedges' g* = 0.23, *SE* = 0.06, 95%CI [0.11, 0.35], *Z* = 3.86, *p* < .001, were both significant. The meta-analytic effect of *validity* on self-reported ratings, *Hedges' g* = 0.65, *SE* = 0.06, 95%CI [0.53, 0.78], *Z* = 10.32, *p* < .001, and automatic evaluation, *Hedges' g* = 0.27, *SE* = 0.06, 95%CI [0.15, 0.39], *Z* = 4.36, *p* < .001, were also significant. In short, US content and validity were genuine and general moderators of EC effects.

General Discussion

Across four studies, we examined the role of inferential processes – and specifically trait inferences – in EC effects. Results consistently supported the idea that manipulations designed to promote (Experiments 1-2) or undermine (Experiment 3) trait inferences influenced the magnitude of self-reported and automatic evaluations, consistently influencing the former more so than the latter. Interestingly, we found evidence for (self-reported and automatic) EC effects under conditions that were designed to undermine trait inferences: when the USs were nouns (Experiments 1-2), when the validity of pairings as an information source was low (Experiment 3), and when these two conditions were combined (Experiment 4).

As is the case for many mental processes, it is difficult if not impossible to demonstrate beyond any doubt that participants made specific trait inferences (e.g., De Houwer, 2011). That said, there are several reasons to substantiate the conclusion that the impact of word type (noun vs. adjective) on EC was mediated by trait inferences. First, the pretest we conducted for choosing the adjective and noun words suggests that the adjectives are more likely to lead to evaluative trait inferences than the nouns. Second, the fact that the validity manipulation influenced the effect of US type (Experiment 4: adjective USs led to stronger EC effects than noun USs but only when the pairings were described as valid but not when the pairings were described as invalid) also fits very well with the idea that inferences were crucial to this effect. Although one could always argue that our results reflect some other difference between adjectives and nouns (e.g., word concreteness), we can find no arguments to support such alternative accounts.

Similarity between EC and Persuasion Effects

As we mentioned in the introduction, our work was inspired by a propositional perspective on EC according to which stimulus pairings function as an argument in much the same way as verbal arguments do and are combined with prior knowledge to inform evaluative responding (see De Houwer, 2018; De Houwer & Hughes, 2016). Building on this idea, we argued that EC and persuasion effects may be more

alike than previously assumed and may therefore also be moderated by manipulations that target trait inferences.

The current findings support this idea on several fronts. First, moderation of EC effects by US content is similar to the moderation of persuasion effects by information diagnosticity. For example, Skowronski and Carlston (1987) argued that the perceived informativeness of behavior is related to its trait *diagnosticity* - the extent to which a behavior promotes the conclusion that an actor possesses one trait and prevents the conclusion that the actor possesses the alternative trait. Being told that "Lisa engages in lots of volunteer work" increases the probability that Lisa is considered generous and lowers the probability that Lisa is considered selfish and therefore is said to be diagnostic for inferences about these traits (Skowronski & Carlston 1987). Other research that focused on stereotypes formation, treated diagnosticity as the extent to which attributes provide the most valuable information for differentiating between groups (Ford & Stangor, 1992). Cone and Ferguson (2015), that tested the role of diagnosticity in the revisions of automatic evaluation, defined diagnosticity as the extent to which information is revealing of a person's true nature, traits, or character. The common ground for all of this research is that it treats diagnosticity as the extent to which information is considered relevant for identifying the nature and cause of something.

Work in this area indicates that information which is considered to be highly diagnostic exert far more impact on judgments than information considered to be less diagnostic (e.g., Cone & Ferguson, 2015; Menon et al., 1995; Skowronski & Carlston, 1992). The manipulation of US content used in the present research can also be viewed as a type of manipulation of diagnosticity of trait inferences. Specifically, both the pretest of Experiments 1-2 and the exploratory questions indicated that participants perceived the adjectives as more diagnostic (relevant and informative) for inferring the men's true character than the nouns. As such, the present research demonstrates that the same factor (i.e., diagnosticity) has similar moderation effects in the contexts of EC as in persuasion.

Second, moderation of EC effects by the validity of pairings as an information source is also similar to the moderation of persuasion effects by validity (and source credibility) information. Persuasion research indicates that validity information exerts a strong impact on evaluations (e.g., Cone et al., 2019; Gregg et al., 2006; Peters & Gawronski, 2011). For example, Cone and colleagues (2019) found that negative information suggesting that a man committed a crime leads to stronger negative evaluations when people were told that this information was derived from police reports than from a coworker who had reason to spread negative rumors. The current work suggests that EC effects, like persuasion effects, are moderated by information about the validity of the evaluative information (regardless if this information constitutes verbal arguments or stimulus pairings). In short, there is marked similarity between the factors that influence EC and those that influence persuasion effects.

Implications for Theoretical Models of EC

In the current research, we used both self-reported and automatic evaluation measures mainly to generalize our findings beyond one specific measure. Moreover, discussions regarding which mental processes underlie self-report versus automatic evaluation measures are still ongoing (e.g., Bar-Anan & Nosek, 2012; Schimmack, 2019). Nevertheless, it is still useful to look at the difference between self-report versus automatic evaluation measures in terms of operating conditions: automatic evaluation measures provide less optimal processing conditions than self-report measures (e.g., Moors et al., 2010; Van Dessel et al., 2020). From this perspective, the results from self-reported and automatic evaluation measures in the current research can impose empirical constraints on evaluation theories (e.g., Gawronski et al., 2020).

The main findings that (1) EC effects on self-reported and automatic evaluation measures are moderated by manipulations that may promote or undermine trait inferences, (2) the moderation effects are stronger on self-report measures, and (3) EC effects on both types of measures are significant also in conditions that are designed to undermine trait inferences, can be used to constrain associative, dual-process, and propositional accounts of EC. First, if EC is mediated purely by the formation (and activation) of mental associations between CS and US (valence), then the specific content of the USs or the credibility of the source who provides the pairings should not influence EC effects. This is because associative mental representations only encode the CS and US information (valence), and not how relevant are those pairings for trait inferences. As such, we find it difficult to reconcile the present findings with such a perspective.

Second, our results constrain dual-process accounts of EC, such as the associative-propositional evaluation (APE) model (Gawronski & Bodenhausen, 2018),

which claim that EC effects can result from two distinct mental processes: associative and propositional. Associative processes refer to the formation and activation of a mental association between CS and US, based on the spatio-temporal relationship between the two. Propositional processes, on the other hand, refer to the formation and activation of mental representations based on inferential processes and the perceived validity of the encoded CS-US relation. A central premise of the APE model is that mere pairings influence mental representations via associative processes even when pairings in the environment are considered invalid during encoding (Gawronski & Bodenhausen, 2018). Moreover, according to the APE model, automatic evaluations mainly reflect the outcome of these associative processes whereas self-reported evaluations mainly reflect the outcome of the propositional processes. From such a perspective, trait inferences (manipulated by US content or validity) should influence self-reported evaluations directly but should not have a direct effect on automatic evaluation. The findings that EC effects were significant also in conditions that are designed to prevent trait inferences, and that the trait inference manipulations had a stronger effect on self-reported evaluations, can be seen as a support for associative processes that influence EC in addition to propositional processes. However, the finding of the present research, that manipulations that promote or undermine trait inferences influenced both self-reported and automatic EC effects do not align with this perspective. The APE model does allow for an indirect influence of trait inferences on automatic evaluation if, during the encoding of information, these inferences lead to the formation of mental associations that reflect diagnosticity or validity information. Critically, however, the APE model, at the moment, does not provide clear predictions regarding the specific conditions under which trait inferences would impact EC effects on automatic evaluation.

Finally, several of our results fit well with the propositional perspective of EC which provides *a priori* predictions regarding a moderating impact of trait inferences on both self-reported and automatic evaluations. The propositional perspective argues that CS evaluation depends on the extent to which participants can infer CS valence or related traits based on CS-US pairings. It therefore assumes that EC effects are mediated by the formation (and activation) of propositions that can encode information about whether stimulus content or pairings are relevant (or credible) for inferring CS valence. From this perspective, both self-reported and automatic evaluations reflect the formation and activation of such propositions (e.g., De Houwer, 2014). According to this perspective, the content of the US and the validity of CS-US pairings should moderate EC effects on both self-reported and automatic evaluations, as was found in the present research.

The finding that EC effects were significant also in conditions that were designed to undermine trait inferences fits less well with the propositional perspective. One could account for such findings by assuming that even in the conditions that were designed to undermine trait inferences some sort of trait inference occurred. Specifically, some of the nouns used in Experiments 1-2 and 4 might have been used to indirectly infer traits. For example, the pairing of Chris with the word “puppies” could have led to the inference that “Chris is like puppies”, or the inference that “Chris hangs around puppies” which both suggest that he is a nice person. It is also possible that in the condition in which the pairing was described as invalid (Experiments 3-4), some participants did not pay attention to the instructions and still used trait inferences.

Alternatively, it is possible that trait inferences are not the only inferences that drive EC effects. Indeed, the propositional perspective also highlights other regularity based inferences (e.g., inferences that co-occurrence of a valenced stimulus and another stimulus indicates similarity in valence; e.g., Hughes et al., 2016) as central to EC effects. Further, the finding that trait inference manipulations had a stronger influence on self-reported evaluations than on automatic evaluation can explained by assuming that when evaluation conditions are suboptimal (i.e., when evaluation) there is less control of which of the inferences will influence performance. However, this alternative account needs to be validated in future research before it can be given much weight.

Implications for Applied EC and Persuasion Research

EC procedures have been used in marketing contexts (e.g., Gibson, 2008; Sweldens et al., 2010). Some have even argued that EC procedures can be used as an intervention to change unwanted attitudes and behavior in areas such as alcohol and snacks consumption (e.g., Houben et al., 2010; Walsh & Kiviniemi, 2014). Until now, inferential processes have rarely – if ever – been taken into account when designing applied EC-based interventions. In many cases, USs are

not chosen based on their relevance to the CSs. Consider, for instance, a study testing whether EC can increase actual fruit (CS) consumption (Walsh & Kiviniemi, 2014). In this case, the USs were generic IAPS pictures (Lang et al., 2008) of a waterfall, a chipmunk, and an astronaut. Other applied studies have sought to minimize awareness of (and thus inferences about) the CS-US pairings, by embedding those pairings in a rapid stream presentation of unrelated stimuli (e.g., Gibson, 2008).

The current findings suggest that EC-based interventions can be 'supercharged' by actively shaping the types of inferences that people will make in the context of an EC procedure, and directing those inferences in a more precise manner. For instance, the effectiveness of an EC task designed to increase fruit (CS) consumption could be improved by choosing USs that are directly relevant for fruit consumption (e.g., words and images that refer to healthiness rather than general positive but irrelevant words or images). Similarly, the effectiveness of EC procedures may increase when those tasks create optimal conditions for inferring that the CS-US pairing is a relevant and credible piece of information instead of directing attention away from that's information source.

From the perspective of persuasion research, our studies highlight a new way of inducing trait inferences and thereby changing evaluations and behavior. Whereas persuasion typically involves verbal message that blatantly communicate information about behaviors or traits, merely pairing stimuli can provide a more subtle way to induce trait inferences. Especially in cases where persuasive messages are likely to evoke reactance, stimulus pairings may provide an alternative pathway for inducing trait inferences that is less likely to evoke reactance (De Houwer & Hughes, 2016).

Future Directions and Limitations

The present research was inspired by a propositional perspective on EC which assumes that EC effects are driven by inferences. Importantly, as we mentioned before, trait inferences are probably not the *only* type of inferences that drive EC effects. For example, conditions that promote inferences about positive or negative appearance (e.g., using "beautiful" and "ugly" as USs) or inferences about positive or negative states (e.g., using "energetic" and "tired" as USs) may increase EC effects relatively to conditions that do not promote inferences. Future research can test if other types of inferences contribute to EC effects.

The current research is subject to a number of limitations. First, our investigation used a restricted set of stimuli (i.e., CSs were men, USs were words), one specific context (i.e., hiring decision) and a specific sample (Prolific participants). Future research should extend this line of research by examining the role of trait inferences in EC with different stimuli, contexts, and samples. Indeed, the observation that trait inferences can play a role in EC does not constitute that they play a role in all types of EC (e.g., in EC procedures which does not mention traits or cover stories: e.g., Moran et al., 2021; Olson & Fazio, 2001). Second, US content and validity effects were observed on only one self-report measure (liking ratings) and one automatic evaluation measure (the IAT). Given that US content did not impact evaluations assessed using another measure (a modified version of the AMP), it is possible that the observed effects on automatic evaluation are due to specific properties of the IAT and will not transfer to other automatic evaluation measures. Future research might examine this question by using different automatic evaluation measures such as the Evaluative Priming task (Fazio et al., 1995) or the original version of the AMP (Payne et al., 2005). Third, the current research design does not allow to test if EC effects and their moderation by trait inference were due to changes in the evaluation of the CS that was paired with positive stimuli, the CS that was paired with negative stimuli, or both. Future research can shed more light on the bi-directionality of the explored effects by including baseline ratings of the CSs before the conditioning procedure.

Summary

Inspired by a propositional perspective of EC, the present research examined the role of inferential processes, and specifically trait inferences, in EC. We found that conditions which promote trait inferences led to stronger EC effects than conditions which undermine such inferences. We also found that conditions that undermine trait inferences do not eliminate EC effects. The results of the present research suggest that inferential processes play an important role in EC effects but also that trait inferences may not be the only pathway to EC effects.

References

Ajzen, I., & Fishbein, M. (2005). The influence of attitudes on behavior. In D. Albarracı´n, B. T. Johnson, & M. P. Zanna (Eds.), *The handbook of attitudes* (pp. 173–221). Mahwah, NJ: Erlbaum.

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

Baeyens, F., Eelen, P., & Crombez, G. (1995). Pavlovian associations are forever: On classical conditioning and extinction. *Journal of Psychophysiology, 9*, 127-141.

Baeyens, F., Eelen, P., Crombez, G., & Van den Bergh, O. (1992). Human evaluative conditioning: Acquisition trials, presentation schedule, evaluative style and contingency awareness. *Behaviour Research and Therapy*, *30*(2), 133-142.

Balas, R., & Gawronski, B. (2012). On the intentional control of conditioned evaluative responses. *Learning and Motivation*, *43*(3), 89-98.

Bar-Anan, Y., & Amzaleg-David, E. (2014). The effect of evaluation on co-occurrence memory judgement. *Cognition and Emotion*, *28*(6), 1030-1046.

Bar-Anan, Y., & Nosek, B. A. (2012). Reporting intentional rating of the primes predicts priming effects in the affective misattribution procedure. *Personality and Social Psychology Bulletin*, *38*(9), 1194-1208.

Bar-Anan, Y., & Vianello, M. (2018). A multi-method multi-trait test of the dual-attitude perspective. *Journal of Experimental Psychology: General*, *147*(8), 1264-1272.

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

Carlston, D. E., & Skowronski, J. J. (2005). Linking versus thinking: evidence for the different associative and attributional bases of spontaneous trait transference and spontaneous trait inference. *Journal of Personality and Social Psychology*, *89*(6), 884-898.

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.

Cone, J., Flaharty, K., & Ferguson, M. J. (2019). Believability of evidence matters for correcting social impressions. *Proceedings of the National Academy of Sciences*, *116*(20), 9802-9807.

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*, 176–187.

De Houwer, J. (2009). The propositional approach to associative learning as an alternative for association formation models. *Learning and Behavior*, *37*, 1-20.

De Houwer, J. (2011). Why the cognitive approach in psychology would profit from a functional approach and vice versa. *Perspectives on Psychological Science*, *6*(2), 202-209.

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., & 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., 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.

Ford, T. E., & Stangor, C. (1992). The role of diagnosticity in stereotype formation: Perceiving group means and variances. *Journal of Personality and Social Psychology*, *63*(3), 356-367.

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

Gawronski, B., & Bodenhausen, G. V. (2018). Evaluative conditioning from the perspective of the associative-propositional evaluation model. *Social Psychological Bulletin*, *13*, e28024.

Gawronski, B., De Houwer, J., & Sherman, J. W. (2020). Twenty-five years of research using implicit measures. *Social Cognition*, *38*(Supplement), s1-s25.

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.

Glasman, L. R., & Albarracin, D. (2006). Forming attitudes that predict future behavior: A meta-analysis of the attitude-behavior relation. *Psychological Bulletin*, *132*(5), 778-822.

Golding, J. M., Fowler, S. B., Long, D. L., & Latta, H. (1990). Instructions to disregard potentially useful information: The effects of pragmatics on evaluative judgments and recall. *Journal of Memory and Language*, *29*(2), 212-227.

Gregg, A. P., Seibt, B., & Banaji, M. R. (2006). Easier done than undone: Asymmetry in the malleability of implicit preferences. *Journal of Personality and Social Psychology*, *90*, 1–20.

Greenwald, A. G., McGhee, D. E., & Schwartz, J. L. K. (1998). Measuring individual differences in implicit cognition: The implicit association test. *Journal of Personality and Social Psychology*, *74*, 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.

Heycke, T., & Gawronski, B. (2020). Co-occurrence and relational information in evaluative learning: A multinomial modeling approach. *Journal of Experimental Psychology: General*, *149*(1), 104-124.

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.

Hu, X., Gawronski, B., & Balas, R. (2017). Propositional versus dual-process accounts of evaluative conditioning: I. The effects of co-occurrence and relational information on implicit and explicit evaluations. *Personality and Social Psychology Bulletin*, *43*(1), 17-32.

Hughes, S., De Houwer, J., & Barnes-Holmes, D. (2016). The moderating impact of distal regularities on the effect of stimulus pairings. *Experimental Psychology, 63*, 20–44.

Hughes, S., Ye, Y., Van Dessel, P., & De Houwer, J. (2019). When people co-occur with good or bad events: Graded effects of relational qualifiers on evaluative conditioning. *Personality and Social Psychology Bulletin*, *45*(2), 196-208.

Kelley, H. H. (1973). The process of causal attribution. *American Psychologist, 28*, 107 – 128.

Kim, J., Allen, C. T., & Kardes, F. R. (1996). An investigation of the mediational mechanisms underlying attitudinal conditioning. *Journal of Marketing Research*, *33*(3), 318-328.

Knutson, B. (1996). Facial expressions of emotion influence interpersonal trait inferences. *journal of Nonverbal Behavior*, *20*(3), 165-182.

Lang, P. J., Bradley, M. M., & Cuthbert, B. N. (2008). *International affective picture system (IAPS): Technical manual and affective ratings*. Gainesville: The Center for Research in Psychophysiology, University of Florida.

Levey, A. B., & Martin, I. (1975). Classical conditioning of human 'evaluative' responses. *Behaviour Research and Therapy*, *13*(4), 221-226.

Martin, I., & Levey, A. (1994). The evaluative response: Primitive but necessary. *Behaviour Research and Therapy*, *32*(3), 301-305.

Merckelbach, H., de Jong, P. J., Arntz, A., & Schouten, E. (1993). The role of evaluative learning and disgust sensitivity in the etiology and treatment of spider phobia. *Advances in Behaviour Research and Therapy*, *15*(4), 243-255.

Mann, T. C., Cone, J., Heggeseth, B., & Ferguson, M. J. (2019). Updating implicit impressions: New evidence on intentionality and the affect misattribution procedure. *Journal of Personality and Social Psychology*, *116*(3), 349-374.

Menon, G., Raghubir, P., & Schwarz, N. (1995). Behavioral frequency judgments: An accessibility-diagnosticity framework. *Journal of Consumer Research*, *22*(2), 212-228.

Milner, J. S., Wagner, M. F., & Crouch, J. L. (2017). Reducing child-related negative attitudes, attributions of hostile intent, anger, harsh parenting behaviors, and punishment through evaluative conditioning. *Cognitive Therapy and Research*, *41*(1), 43-61.

Minear, M., & Park, D. C. (2004). A lifespan database of adult facial stimuli. *Behavior Research Methods, Instruments, and Computers*, *36*(4), 630-633.

Mitchell, C. J., Anderson, N. E., & Lovibond, P. F. (2003). Measuring evaluative conditioning using the Implicit Association Test. *Learning and Motivation*, *34*(2), 203-217.

Moors, A. (2016). Automaticity: Componential, causal, and mechanistic explanations. *Annual Review of Psychology*, *67*, 263-287.

Moors, A., & De Houwer, J. (2006). Automaticity: a theoretical and conceptual analysis. *Psychological Bulletin*, *132*(2), 297-326.

Moors, A., Spruyt, A., & De Houwer, J. (2010). In search of a measure that qualifies as implicit: Recommendations based on a decompositional view of automaticity. In B. Gawronski & B. K. Payne (Eds.), *Handbook of implicit social cognition: Measurement, theory, and applications* (pp. 19-35). New York, NY: Guilford

Moran, T., Bar-Anan, Y., & Nosek, B. A. (2017). The effect of the validity of co-occurrence on automatic and deliberate evaluations. *European Journal of Social Psychology*, *47*(6), 708-723.

Moran, T., Hughes, S., Hussey, I., Vadillo, M. A., Olson, M. A., Aust, F., ... & De Houwer, J. (2021). Incidental attitude formation via the surveillance task: A preregistered replication of the Olson and Fazio (2001) study. *Psychological Science*, *32*(1), 120-131.

Nisbett, R. E., & Ross, L. (1980). *Human inference: Strategies and shortcomings of social judgment*. Englewood Cliffs, NJ: Prentice Hall.

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.

Olivola, C. Y., & Todorov, A. (2010). Elected in 100 milliseconds: Appearance-based trait inferences and voting. *Journal of Nonverbal Behavior*, *34*(2), 83-110.

Olson, M. A., & Fazio, R. H. (2001). Implicit attitude formation through classical conditioning. *Psychological Science*, *12*(5), 413-417.

Payne, B. K., Brown-Iannuzzi, J., Burkley, M., Arbuckle, N. L., Cooley, E., Cameron, C. D., & Lundberg, K. B. (2013). Intention invention and the affect misattribution procedure: Reply to Bar-Anan and Nosek (2012). *Personality and Social Psychology Bulletin*, *39*(3), 375-386.

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.

Peters, K. R., & Gawronski, B. (2011). Are we puppets on a string? Comparing the impact of contingency and validity on implicit and explicit evaluations. *Personality and Social Psychology Bulletin*, *37*(4), 557-569.

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.

Schimmack, U. (2021). The Implicit Association Test: A method in search of a construct. *Perspectives on Psychological Science*, *16*(2), 396-414.

Schönbrodt, F. D., & Wagenmakers, E. J. (2018). Bayes factor design analysis: Planning for compelling evidence. *Psychonomic Bulletin and Review*, *25*(1), 128-142.

Skowronski, J. J., & Carlston, D. E. (1987). Social judgment and social memory: The role of cue diagnosticity in negativity, positivity, and extremity biases. *Journal of Personality and Social Psychology*, *52*(4), 689-699.

Skowronski, J. J., & Carlston, D. E. (1989). Negativity and extremity biases in impression formation: A review of explanations. *Psychological Bulletin*, *105*(1), 131-142.

Smith, C. T., De Houwer, J., & Nosek, B. A. (2013). Consider the source: Persuasion of implicit evaluations is moderated by source credibility. *Personality and Social Psychology Bulletin*, *39*(2), 193-205.

Sweldens, S., Van Osselaer, S. M., & Janiszewski, C. (2010). Evaluative conditioning procedures and the resilience of conditioned brand attitudes. *Journal of Consumer Research*, *37*(3), 473-489.

Trope, Y., & Bassok, M. (1983). Information-gathering strategies in hypothesis-testing. *Journal of Experimental Social Psychology*, *19*(6), 560-576.

Uleman, J. S., Saribay, S. A., & Gonzalez, C. (2008). Spontaneous inferences, implicit impressions, and implicit theories. *Annual Review of Psychology*, *59*, 329-360.

Van Dessel, P., Cummins, J., Hughes, S., Kasran, S., Cathelyn, F., & Moran, T. (2020). Reflecting on 25 Years of Research Using Implicit Measures: Recommendations for Their Future Use. *Social Cognition*, *38*(Supplement), s223-s242.

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.

Walsh, E. M., & Kiviniemi, M. T. (2014). Changing how I feel about the food: experimentally manipulated affective associations with fruits change fruit choice behaviors. *Journal of Behavioral Medicine*, *37*(2), 322-331.

Wegener, D. T., Clark, J. K., & Petty, R. E. (2018). Cognitive and metacognitive processes in attitude formation and change. In *The Handbook of Attitudes, Volume 1: Basic Principles* (pp. 291-331). Routledge.

Wagner, M. F., Skowronski, J. J., Milner, J. S., Crouch, J. L., & Ammar, J. (2020). Exploring Positive Classical Conditioning Procedure Effects on Evaluations of Children, Thoughts About Children, and Behaviors Toward Children: Two Experiments. *Psychological Reports*, *123*(5), 1753-1784.

Zanon, R., De Houwer, J., Gast, A., & Smith, C. T. (2014). When does relational information influence evaluative conditioning?. *Quarterly Journal of Experimental Psychology*, *67*(11), 2105-2122.

Table 1

Words used in the EC and the IAT procedures in Experiments 1-4.

EC				IAT	
Experiments 1-4		Experiments 1-2, and 4		Experiments 2-4	
Positive adjectives	Negative adjectives	Positive nouns	Negative nouns	Positive words	Negative words
Educated	*Foolish*	*Sunset*	*Cemetery*	*Positive*	*Negative*
Amusing	*Reckless*	*Chocolate*	*Funeral*	*Good*	*Bad*
Modest	*Anxious*	*Holiday*	*Injury*	*Pleasant*	*Unpleasant*
Agreeable	*Awful*	*Puppies*	*Feces*	*Great*	*Terrible*

Table 2

Mean (SD) CS evaluation and EC scores in Experiments 1-4 as a function of trait inference condition.

Experiment – Condition	**n**	**Self-reported evaluation**			**Automatic evaluation[a]**
		CS_{pos}[b]	CS_{neg}[c]	EC score[d] / difference from zero	EC score[d] / difference from zero
Experiment 1 – US content					
Adjectives	115	2.56 (1.12)	-2.00 (1.33)	4.56 (2.22) / $t(114) = 22.03$, $p < .001$, $d = 2.05$, $BF_{10} > 1000$	0.04 (0.19) / $t(114) = 2.10$, $p = .038$, $d = 0.20$, $BF_{10} = 0.86$
Nouns	112	1.32 (1.52)	-0.29 (1.72)	1.61 (2.91) / $t(111) = 5.85$, $p < .001$, $d = 0.55$, $BF_{10} > 1000$	0.07 (0.22) / $t(111) = 3.17$, $p = .002$, $d = 0.30$, $BF_{10} = 11.36$
Experiment 2 – US content					
Adjectives	119	2.22 (1.33)	-1.54 (1.39)	3.76 (2.50) / $t(118) = 16.39$, $p < .001$, $d = 1.50$, $BF_{10} > 1000$	0.30 (0.46) / $t(118) = 7.17$, $p < .001$, $d = 0.66$, $BF_{10} > 1000$
Nouns	113	1.05 (1.39)	-0.46 (1.41)	1.51 (2.39) / $t(112) = 6.71$, $p < .001$, $d = 0.63$, $BF_{10} > 1000$	0.14 (0.41) / $t(112) = 3.78$, $p < .001$, $d = 0.36$, $BF_{10} = 73.60$
Experiment 3 – Validity					
High	119	2.25 (1.31)	-1.85 (1.30)	4.10 (2.40) / $t(118) = 18.63$, $p < .001$, $d = 1.71$, $BF_{10} > 1000$	0.32 (0.44) / $t(118) = 8.02$, $p < .001$, $d = 0.74$, $BF_{10} > 1000$
Low	121	0.83 (1.88)	-0.44 (1.79)	1.27 (3.45) / $t(120) = 4.05$, $p < .001$, $d = 0.37$, $BF_{10} = 186.63$	0.18 (0.44) / $t(120) = 4.36$, $p < .001$, $d = 0.40$, $BF_{10} = 578.64$
Experiment 4 – Validity / US content					
High / Adjectives	139	2.31 (1.35)	-1.75 (1.51)	4.06 (2.58) / $t(138) = 18.56$, $p < .001$, $d = 1.57$, $BF_{10} > 1000$	0.33 (0.38) / $t(138) = 10.47$, $p < .001$, $d = 0.89$, $BF_{10} > 1000$
High / Nouns	140	0.99 (1.37)	-0.63 (1.57)	1.62 (2.60) / $t(139) = 7.38$, $p < .001$, $d = 0.62$, $BF_{10} > 1000$	0.13 (0.39) / $t(139) = 4.00$, $p < .001$, $d = 0.34$, $BF_{10} = 158.66$
Low / Adjectives	150	0.80 (1.71)	-0.50 (1.69)	1.29 (3.27) / $t(149) = 4.84$, $p < .001$, $d = 0.40$, $BF_{10} > 1000$	0.13 (0.46) / $t(149) = 3.56$, $p < .001$, $d = 0.29$, $BF_{10} = 36.02$
Low / Nouns	156	0.69 (1.37)	-0.30 (1.32)	0.99 (2.40) / $t(155) = 5.14$, $p < .001$, $d = 0.41$, $BF_{10} > 1000$	0.12 (0.41) / $t(155) = 3.54$, $p < .001$, $d = 0.28$, $BF_{10} = 32.86$

Note. [a]Automatic evaluation was measured with the AMP in Experiment 1 and with the IAT in Experiments 2-4. [b]CS_{pos} = rating of the target person who was paired with positive words. [c]CS_{neg} = rating of the target person who was paired with negative words. [d]EC score = preference for the CS that was paired with positive words over the CS that was paired with negative words. The preference was computed by subtracting the mean score rating for CS_{pos} from the mean score rating for CS_{neg}.

Table 3
ANOVA models in Experiments 1-4.

Experiment	Model for self-report rating and IAT scores	Model for the direct comparison between self-reported and automatic evaluation effects
1	**2 (*US Content*: nouns, adjectives)** x 2 (*CS-US assignment*: Chris with positive, James with positive) x 2 (*Measures order*: self-report first, automatic evaluation first) between-participants ANOVA	**2 (*Measure type:* self-report, IAT) x 2 (*US Content*)** x 2 (*CS-US assignment*: Chris with positive, James with positive) x 2 (*Measures order*: self-report first, automatic evaluation first) mixed ANOVA[1]
2	**2 (*US Content*: nouns, adjectives)** x 2 (*CS-US assignment*: Chris with positive, James with positive) x 2 (*Measures order*: self-report first, automatic evaluation first) x 2 (*IAT block order*: compatible block first, incompatible block first) between-participants ANOVA	**2 (*Measure type:* self-report, IAT) x 2 (*US Content*)** x 2 (*CS-US assignment*: Chris with positive, James with positive) x 2 (*Measures order*: self-report first, automatic evaluation first) x 2 (*IAT block order*: compatible block first, incompatible block first) mixed ANOVA[1]
3	**2 (*Validity*: low, high)** x 2 (*CS-US assignment*: Chris with positive, James with positive) x 2 (*Measures order*: self-report first, automatic evaluation first) x 2 (*IAT block order*: compatible block first, incompatible block first) between-participants ANOVA	**2 (*Measure type:* self-report, IAT) x 2 (*Validity*: low, high)** x 2 (*CS-US assignment*: Chris with positive, James with positive) x 2 (*Measures order*: self-report first, automatic evaluation first) x 2 (*IAT block order*: compatible block first, incompatible block first) mixed ANOVA[1]
4	**2 (*Validity*: low, high) x 2 (*US Content*: nouns, adjectives)** x 2 (*CS-US assignment*: Chris with positive, James with positive) x 2 (*Measures order*: self-report first, automatic evaluation first) x 2 (*IAT block order*: compatible block first, incompatible block first) between-participants ANOVA	**2 (*Measure type:* self-report, IAT) x 2 (*Validity*: low, high) x 2 (*US Content*: nouns, adjectives)** x 2 (*CS-US assignment*: Chris with positive, James with positive) x 2 (*Measures order*: self-report first, automatic evaluation first) x 2 (*IAT block order*: compatible block first, incompatible block first) mixed ANOVA[1]

Note. In all the models the main manipulations are in Bold. As specified in the preregistrations of all the experiments, the counterbalancing procedural factors were included in the models to control for their influence, but their effects were not analyzed. [1]In all the mixed ANOVA models, measure type was the only within-participants factor.

Table 4

ANOVA results for the direct comparison of self-reported and automatic evaluation effects in Experiments 1-4.

Experiment / *Factors*	**Results**	**Interpretation**
Experiment 1		
Measure type	$F(1, 219) = 49.79$, $p < .001$, $\eta_p^2 = .19$, 90% CI [.11, .25], $BF_{10} > 1000$	Overall stronger EC effects for self-report than automatic evaluation measures
US Content	$F(1, 219) = 10.35$, $p = .001$, $\eta_p^2 = .05$, 90%CI [.01, .09], $BF_{10} = 12.71$	A stronger EC effect when USs were adjectives than nouns
Measure type × *US Content*	$F(1, 219) = 37.86$, $p < .001$, $\eta_p^2 = .15$, 90% CI [.08, .21], $BF_{10} > 1000$	The effect of US content on self-report ratings was stronger than the effect on automatic evaluations
Experiment 2		
Measure type	$F(1, 216) = 13.13$, $p < .001$, $\eta_p^2 = .06$, 90% CI [.01, .11], $BF_{10} = 49.96$	Overall stronger EC effects for self-report than automatic evaluation measures
US Content	$F(1, 216) = 34.82$, $p < .001$, $\eta_p^2 = .14$, 90% CI [.07, .20], $BF_{10} > 1000$	A stronger EC effect when USs were adjectives than nouns
Measure type × *US Content*	$F(1, 216) = 5.60$, $p = .019$, $\eta_p^2 = .03$, 90% CI [.002, .06], $BF_{10} = 1.05$	The effect of US content on self-report ratings was stronger than the effect on automatic evaluations
Experiment 3		
Measure type	$F(1, 224) = 4.31$, $p = .039$, $\eta_p^2 = .02$, 90% CI [.0004, .05], $BF_{10} = 0.73$	Overall stronger EC effects for self-report than automatic evaluation measures
Validity	$F(1, 224) = 36.43$, $p < .001$, $\eta_p^2 = .14$, 90% CI [.07, .20], $BF_{10} > 1000$	A stronger EC effect when participants received high validity than low validity instructions
Measure type × *Validity*	$F(1, 224) = 8.78$, $p = .003$, $\eta_p^2 = .04$, 90% CI [.007, .08], $BF_{10} = 10.38$	The effect of validity on self-report ratings was stronger than the effect on automatic evaluations
Experiment 4		
Measure type	$F(1, 552) = 9.87$, $p = .002$, $\eta_p^2 = .02$, 90% CI [.003, .03], $BF_{10} = 10.64$	Overall stronger EC effects for self-report than automatic evaluation measures
Validity	$F(1, 552) = 43.09$, $p < .001$, $\eta_p^2 = .07$, 90% CI [.04, .10], $BF_{10} > 1000$	A stronger EC effect when participants received high validity than low validity instructions
US Content	$F(1, 552) = 35.50$, $p < .001$, $\eta_p^2 = .06$, 90% CI [.03, .09], $BF_{10} > 1000$	A stronger EC effect when USs were adjectives than nouns
Validity × *US Content*	$F(1, 552) = 21.86$, $p < .001$, $\eta_p^2 = .04$, 90% CI [.01, .06], $BF_{10} > 1000$	US content moderated EC effects under high validity, but not under low validity instructions
Measure type × *Validity*	$F(1, 552) = 8.91$, $p = .003$, $\eta_p^2 = .02$, 90% CI [.003, .03], $BF_{10} = 5.07$	The effect of validity on self-report ratings was stronger than the effect on automatic evaluations
Measure type × *US Content*	$F(1, 552) = 1.85$, $p = .18$, $\eta_p^2 < .01$, 90% CI [0, .01], $BF_{10} = 0.29$	
Measure type × *Validity* × *US Content*	$F(1, 552) = 0.71$, $p = .40$, $\eta_p^2 < .01$, 90% CI [0, .01], $BF_{10} = 0.21$	

Note. The counterbalancing procedural factors were also entered into the models to control for their influence, but their effects were not analyzed. For all analyses, in all the experiments, the Bayes-factors were computed using a model that excluded the counterbalancing procedural factors (to simplify the analyses).

Table 5

Specific comparisons in Experiment 4.

Comparison	Results	Interpretation
Self-report ratings		
low validity + nouns vs. high validity + adjectives	$t(283.12) = 10.55$, $p < .001$, $d = 1.24$, $BF_{10} > 1000$	Participants in the low validity/nouns condition showed a weaker effect than participants in the high validity/adjectives condition
high validity + nouns vs. low validity + adjectives	$t(280.99) = 0.95$, $p = .34$, $d = 0.11$, $BF_{10} = 0.20$	
high validity + nouns vs. high validity + adjective	$t(277) = 7.87$, $p < .001$, $d = 0.94$, $BF_{10} > 1000$	Participants in the high validity/nouns condition exhibited a weaker effect than those in the high validity/adjectives condition
low validity + nouns vs. low validity + adjectives	$t(273.07) = 0.92$, $p = .18$, $d = 0.11$, $BF_{10} = 0.19$	
IAT scores		
low validity + nouns vs. high validity + adjectives	$t(292.82) = 4.74$, $p < .001$, $d = 0.55$, $BF_{10} > 1000$	Participants in the low validity/nouns condition showed a weaker effect than participants in the high validity/adjectives condition
high validity + nouns vs. low validity + adjectives	$t(286.21) = 0.004$, $p > .99$, $d = 0.00$, $BF_{10} = 0.13$	
high validity + nouns vs. high validity + adjective	$t(276.63) = 4.37$, $p < .001$, $d = 0.52$, $BF_{10} = 928.96$	Participants in the high validity/nouns condition exhibited a weaker effect than those in the high validity/adjectives condition
low validity + nouns vs. low validity + adjectives	$t(298.12) = 0.32$, $p = .37$, $d = 0.04$, $BF_{10} = 0.13$	