Mechanisms Underlying Approach-Avoidance Instruction Effects on Implicit Evaluation: Results of a Preregistered Adversarial Collaboration

Pieter Van Dessel[1]

Bertram Gawronski[2]

Colin Tucker Smith[3]

Jan De Houwer[1]

[1]Ghent University, Belgium

[2]University of Texas at Austin, USA

[3]University of Florida, USA

Correspondence concerning this article should be addressed to Pieter Van Dessel, Ghent University, Department of Experimental-Clinical and Health Psychology, Henri Dunantlaan 2, B-9000 Ghent (Belgium). E-mail: Pieter.vanDessel@UGent.be.

Abstract

Previous research demonstrated that mere instructions to approach one stimulus and avoid another stimulus result in an implicit preference for the to-be-approached over the to-be-avoided stimulus. To investigate the mechanisms underlying approach-avoidance (AA) instruction effects, we tested predictions of a propositional account and an associative self-anchoring account in a preregistered adversarial collaboration. Consistent with the propositional account, Experiment 1 showed that avoidance instructions had a negative effect on implicit evaluations over and above the positive effect of approach instructions. Consistent with the associative self-anchoring account, Experiment 2 showed that changes in implicit self-stimulus linking mediated AA instruction effects on implicit evaluations. However, mediation was only partial, in that AA instructions showed a significant effect on implicit evaluations after controlling for implicit self-stimulus linking. Together, the results support the contribution of propositional processes to AA instruction effects; the results remain ambiguous regarding an additional contribution of associative self-anchoring.

Keywords: approach, avoidance, instruction, implicit evaluation, self-anchoring, propositional theory

Mechanisms Underlying Approach-Avoidance Instruction Effects on Implicit Evaluation: Results of a Preregistered Adversarial Collaboration

It has been recognized for decades that behavior is shaped by likes and dislikes (Allport, 1935). Hence, understanding how these preferences are acquired is an important endeavor for psychological research. Interestingly, preferences sometimes arise as the result of performing specific behaviors (Olson & Stone, 2005). For example, previous research has shown that the repeated performance of approach and avoidance actions can cause changes in stimulus evaluations. When participants repeatedly approach one stimulus and avoid another stimulus, they typically develop a preference for the approached stimulus over the avoided stimulus (Laham, Kashima, Dix, Wheeler, & Levis, 2014). These approach-avoidance (AA) training effects have been observed for a wide variety of stimuli, such as pictures of unfamiliar faces (Woud, Maas, Becker, & Rinck, 2013), racial groups (Kawakami, Phills, Steele, & Dovidio, 2007), alcoholic beverages (Wiers, Eberl, Rinck, Becker, & Lindenmeyer, 2011), unhealthy foods (Zogmaister, Perugini, & Richetin, in press), insects and spiders (Jones, Vilensky, Vasey, & Fazio, 2013), and contamination-related objects (Amir, Kuckertz, & Najmi, 2013).

In a recent set of studies, Van Dessel, De Houwer, Gast, and Smith (2015) obtained evidence that AA effects can also be observed as a result of mere instructions in the absence of actually performed actions. When participants were instructed to approach certain stimuli and avoid other stimuli, their evaluations of the to-be-approached stimuli were more positive than their evaluations of the to-be-avoided stimuli even though participants never actually performed the AA actions. Effects of AA instructions have been observed for novel non-words, fictitious social groups, and unfamiliar faces (Van Dessel, De Houwer, Roets, & Gast, 2016). Importantly, these AA instruction effects were similar to the effects involving actual AA training in that both

AA instructions and AA training influenced not only explicit (i.e., non-automatic) stimulus evaluations but also implicit (i.e., automatic) stimulus evaluations (Van Dessel, De Houwer, Gast, Smith, & De Schryver, 2016).

Effects of AA instructions on implicit evaluation pose a challenge to a particular type of associative models that assume that (a) implicit evaluations reflect the automatic activation of associations in memory and (b) these associations are formed as the result of a slow-learning process that capitalizes on repeated co-occurrences, such as recurrent pairings of AA actions and stimuli (Rydell & McConnell, 2006; Smith & DeCoster, 2000). Yet, instruction-based AA effects are consistent with propositional models, which assume that implicit evaluations reflect the activation and generation of mental propositions about the relation between objects and events (e.g., De Houwer, 2009, 2014; Mitchell, De Houwer, & Lovibond, 2009). When participants are instructed to approach or avoid a stimulus, they might generate propositions about these stimulus-action relations, and these propositions can influence their implicit evaluations of the stimuli (Van Dessel, De Houwer, Gast, et al., 2016). For example, participants who learn that they will approach a stimulus may infer that this stimulus is positive, and participants who learn that they will avoid a stimulus may infer that this stimulus is negative. These inferences could arise because of the knowledge that positive objects are typically approached and negative objects are avoided (Schneirla, 1959). People may have learned this rule through previous experiences during which they approached liked stimuli and avoided disliked stimuli. Although this knowledge does not logically imply that approached things are good and avoided things are bad, people are known to be prone to affirm the consequent (i.e., conclude that A is true on the basis of the fact that A implies B and B is present). Thus, when participants infer that the to-be-approached stimulus is good and the to-be-avoided stimulus is bad, the (automatic) activation of this mental proposition could impact their implicit evaluations (De Houwer, 2014).

However, AA instruction effects on implicit evaluation are not necessarily incompatible with the view that implicit evaluations reflect the automatic activation of associations in memory (Gawronski & Bodenhausen, 2011). Some dual-process models, such as the associative-propositional evaluation (APE) model (Gawronski & Bodenhausen, 2006), postulate that mental associations can be formed as the result of propositional inferences. According to the APE model, any information that allows participants to entertain the proposition that a stimulus is positive or negative may instigate the proactive construction of new evaluative associations, which in turn may influence implicit evaluations. In line with this idea, changes in implicit evaluations have been observed when participants are provided with verbal information about the evaluative properties of a stimulus (Castelli, Zogmaister, Smith & Arcuri, 2004; Cone & Ferguson, 2015; Gawronski, Walther, & Blank, 2005; Gregg, Seibt, & Banaji, 2006). Importantly, these models predict a specific pattern of mediation such that instruction effects on explicit evaluation should mediate effects on implicit evaluation (e.g., Gawronski & Walther, 2008; Peters & Gawronski, 2011a; Whitfield & Jordan, 2009; see Gawronski & Bodenhausen, 2006; Case 4).

Van Dessel, De Houwer, Gast, et al. (2016) recently performed two experiments that tested the mediating role of explicit evaluations in the effect of AA instructions on implicit evaluations. In these experiments, participants first received information about the evaluative traits of members of two fictitious social groups and were then given instructions to approach or avoid the names of members of these groups. The results showed that trait information eliminated the effects of AA instructions on explicit, but not implicit, evaluations. Statistical mediation analyses further showed that AA instructions had a direct effect on implicit evaluations that was not mediated by changes in explicit evaluations. These findings contradict the idea that AA instructions influence implicit evaluations only if these instructions are considered a valid basis for evaluation and, hence, are incorporated in explicit evaluations (see Gawronski &

Bodenhausen, 2006). Yet, the results are consistent with a propositional explanation of AA instruction effects and support the propositional model of evaluation which postulates that mental propositions, rather than associations, underlie implicit evaluation (De Houwer, 2014). Specifically, AA instructions might allow participants to consider the proposition that a to-be-approached stimulus is positive and a to-be-avoided stimulus is negative. A dissociation between implicit and explicit evaluation will arise when this proposition is judged to be invalid (and thus dismissed when making an explicit evaluation) but still automatically retrieved when the stimuli are implicitly evaluated.

Nevertheless, there is an important alternative explanation of AA instruction effects on implicit evaluation that is compatible with associative theories of implicit evaluation. Effects of AA instructions on implicit evaluation could arise as the result of associative self-anchoring, which involves the transfer of positive valence from the self to a stimulus associated with the self as the result of a newly formed association between the representation of the stimulus and the representation of the self (see Gawronski, Bodenhausen, & Becker, 2007). It is often assumed that approach behaviors are fundamentally related to pulling objects closer to the self (Förster, 2001), which may result in accentuated psychological closeness between approached stimuli and the self (Nussinson, Seibt, Häfner, & Strack, 2010). In line with this idea, it has been argued that the repeated performance of approach behavior in response to a stimulus allows for the formation of a mental association between the representation of the approached stimulus and the positive representation of the self (Kawakami, Steele, Cifa, Phills, & Dovidio, 2008; Phills, Kawakami, Tabi, Nadolny, & Inzlicht, 2011). Once such an association has been established, the positive valence of the self may spread to the approached stimulus, and thereby influence implicit evaluations of the stimulus (Gawronski et al., 2007). This associative transfer of valence is assumed to be driven by processes of spreading activation without requiring any kind of higher-

order propositional processes (Gawronski, Strack, & Bodenhausen, 2009). Although many theories assume that the formation of new associations in memory is a slow, gradual process that requires repeated co-occurrences (e.g., Fazio, Sanbonmatsu, Powell, & Kardes, 1986; Smith & DeCoster, 2000), some researchers have rejected this idea and argued that sufficiently strong associations can be formed as the result of mere instructions (e.g., Field, 2006; Gawronski & Bodenhausen, 2007). From this perspective, mere instructions to approach a given stimulus might allow for the formation of self-stimulus associations, which may lead to more favorable implicit evaluations of the to-be-approached stimulus.

In the current research, we engaged in a preregistered adversarial collaboration to test predictions of a propositional account and an associative self-anchoring account of AA instruction effects in two experiments. Experiment 1 investigated whether both approach instructions and avoidance instructions can cause changes in implicit stimulus evaluations. From the perspective of the associative self-anchoring account, AA instruction effects should occur due to the formation of self-stimulus associations as the result of approach instructions. Processing the semantic meaning of approach instructions should lead to the co-activation of a representation of the self-connected approach action and the to-be-approached stimulus, thereby instigating the automatic formation of an association between the to-be-approached stimulus and the self. Given that most people's implicit self-evaluation is highly positive (Yamaguchi et al., 2001), the subsequent associative transfer of valence should result in a more positive implicit evaluation of the to-be-approached stimulus. In its original formulation, the associative self-anchoring hypothesis does not imply any additional effect of avoidance instructions. Associative self-anchoring is assumed to involve a projection of characteristics of the self to stimuli that are connected to the self but it does not involve a projection of self-characteristics to stimuli that are negatively linked to the self (Gawronski et al., 2007). Thus, even though avoidance can be

construed as distancing the self from a certain stimulus (Nussinson et al., 2010), the associative self-anchoring account does not predict a more negative implicit evaluation for to-be-avoided stimuli. In the context of AA training effects, the operation of positive, but not negative, associative self-anchoring is typically assumed to explain why training to approach a certain stimulus can lead to changes in implicit evaluations whereas avoidance training does not (see Kawakami et al., 2008; Phills et al., 2011).

Note that, even though current theorizing explicitly denies the possibility that avoidance leads to negative stimulus evaluations, it might still be possible to extend the associative self-anchoring account in a manner such that it does predict a negative effect of avoidance actions or avoidance instructions. However, such an account would have to make a number of additional assumptions that seem questionable upon further scrutiny. First, one could argue that avoiding a certain stimulus (or receiving instructions to avoid a certain stimulus) may result in an inhibitory association between the representations of the self and the (to-be-)avoided stimulus. This assumption seems problematic, because a co-activation of these representations when performing the avoidance action or when reading the avoidance instructions should facilitate the formation of an excitatory rather than an inhibitory association. Second, even if avoidance actions or avoidance instructions result in an inhibitory association, it remains unclear why this should lead to a more negative evaluation of the (to-be-)avoided stimulus. According to this extended account, presentation of the stimulus should lead to the inhibition of the representation of the self via the inhibitory association and prevent transfer of positive valence from the self to the stimulus. However, preventing the transfer of positive valence is not the same as triggering the transfer of negative valence. Hence, it is difficult to explain how an inhibitory association would allow for the transfer of negative valence to the (to-be-)avoided stimulus. Third, one could argue that avoidance actions and avoidance instructions create an excitatory association between the

(to-be-)avoided stimulus and a negatively evaluated representation of “not-me.” However, such an account directly contradicts a core assumption of associative theories that negations involve propositional processes and cannot be accomplished via associative processing (Deutsch, Gawronski, & Strack, 2006; Gawronski, Deutsch, Mbirkou, Seibt, & Strack, 2008). Hence, from the perspective of an associative self-anchoring account, AA instructions should lead to more favorable evaluations of the to-be-approached stimulus without affecting evaluations of the to-be-avoided stimulus.

In contrast, a propositional account predicts that both approach and avoidance instructions can influence implicit evaluations. According to this account, participants might infer not only that a to-be-approached stimulus is positive (because they typically approach positive objects), but also that a to-be-avoided stimulus is negative (because they typically avoid negative objects). Once participants have acquired this propositional information about the valence of the stimuli, it may be activated automatically and influence implicit stimulus evaluations (De Houwer, 2014). As a result, AA instructions should not only lead to more favorable implicit evaluations of the to-be-approached stimulus but also to less favorable implicit evaluations of the to-be-avoided stimulus. Because the associative self-anchoring account and the propositional account make different predictions about effects of avoidance instructions on implicit evaluations, we can obtain an estimate of the relative contribution of associative self-anchoring processes and propositional processes in AA instruction effects by comparing the relative magnitude of approach instruction effects and avoidance instruction effects in Experiment 1.

Experiment 2 further investigated whether AA instruction effects are mediated by the formation of a mental association between the representation of the self and the representation of the to-be-approached stimulus. Specifically, we tested whether AA instruction effects on implicit evaluations, as measured with an evaluative Implicit Association Test (IAT; Greenwald,

McGhee, & Schwartz, 1998), are mediated by changes in implicit self-stimulus linking, as measured with a self-stimulus IAT.[1] Such a mediation approach has also been used in previous research to establish the role of associative self-anchoring processes in the context of AA training effects (see Phills et al., 2011). According to the associative self-anchoring account, approach instructions produce changes in self-stimulus associations which, in turn, influence implicit evaluation. Because changes in self-stimulus associations should be reflected in facilitated implicit self-stimulus linking, this account predicts that AA instruction effects on the self-stimulus IAT should mediate AA instruction effects on implicit evaluations. A propositional account, however, does not predict such a mediation. Though participants might more easily relate a to-be-approached stimulus to the self than a to-be-avoided stimulus (e.g., because they infer that a to-be-approached stimulus is more similar to the self than a to-be-avoided stimulus), there is no theoretical basis to assume that AA instruction effects on the self-stimulus IAT would mediate changes in implicit evaluations. By examining the extent to which AA instruction effects are mediated by changes in self-stimulus linking, Experiment 2 can provide a second estimate of the relative contribution of associative self-anchoring processes and propositional processes in these effects.

The described hypotheses as well as the study design and data-analysis plan of Experiment 1 and Experiment 2 were pre-registered on the Open Science Framework prior to data-collection (which was done concurrently for the two experiments). Any deviation from pre-registration is noted in the main text. The pre-registered plan and all code and data are available at https://osf.io/4sajr/. The collaboration between authors qualifies as adversarial in that (a) the

[1] Following recommendations by De Houwer, Gawronski, and Barnes-Holmes (2013), we use the term implicit self-stimulus linking to describe the behavioral phenomenon of automatically connecting the self and a stimulus on an implicit measure (see Ye & Gawronski, 2016).

second author put forward the associative self-anchoring account as an alternative for the propositional account of AA instruction effects developed by the other three authors (De Houwer, 2014; Van Dessel, De Houwer, Gast, et al., 2016) and (b) the four authors jointly devised Experiments 1 and 2 as a way to distinguish between the two competing accounts.

Experiment 1

In Experiment 1, participants received instructions to approach a nonword (to-be-approached word), avoid a second nonword (to-be-avoided word), and to perform no action in response to a third nonword (no-action word). After the instructions, implicit evaluations of the three stimuli were measured with an evaluative priming task (Fazio et al., 1986). We examined whether (a) implicit evaluations of the to-be-approached word were more positive than evaluations of the no-action word, and (b) implicit evaluations of the to-be-avoided word were more negative than evaluations of the no-action word. Following the recommendations of an anonymous reviewer, we also investigated whether implicit evaluations of the to-be-approached word deviated more strongly from implicit evaluations of the no-action word relative to evaluations of the to-be-avoided word.

Method

Participants and design. A total of 1750 English-speaking volunteers participated online via the Project Implicit research website (https://implicit.harvard.edu). In line with the standard treatment of Project Implicit data (e.g., Smith & De Houwer, 2015), we excluded data from participants who (a) did not fully complete all questions and tasks (366 participants; i.e., 20.91%), (b) had error rates in the evaluative priming task that exceeded the population mean by more than 2.5 standard deviations (56 participants; i.e., 4.04%; population mean = 9.06 %, SD = 11.24%), or (c) made at least one error on the questions that probed memory for the AA

instructions (534 participants; i.e., 40.03 %).[2] Analyses were performed on the data of 794 participants (548 women, mean age = 29, *SD* = 13).[3]

Approach-avoidance instructions. All participants were told that the experiment would involve three meaningless words: UDIBNON, BAYRAM, and ENANWAL. Then participants read the following instructions:

> *In this experiment you will see three words with no meaning: UDIBNON, BAYRAM, and ENANWAL. You will perform a task in which you will approach BAYRAM and avoid ENANWAL. It is very important to remember these three words and to remember what you need to do when you see BAYRAM and ENANWAL. You will need all this information to complete the task successfully. Later on we will explain to you exactly how you will be able to perform this task. Before we present the three words and start the task, you will complete a categorization task. This will last about 5 minutes. Make sure that during that task you do not forget the instructions of the next task. Instructions: You will see three words with no meaning: UDIBNON, BAYRAM, and ENANWAL. Approach BAYRAM and avoid ENANWAL. Please press 'Continue' when you have memorized the instructions and are ready to begin the categorization task.*

[2] We excluded participants with incorrect memory because we expected that, in line with previous results (Van Dessel et al., 2015), AA instructions would change evaluations only when participants correctly remembered these instructions. Importantly, including the data from all participants in the analyses reduced the magnitude of the instruction effects, but did not change the statistical significance of any of the reported effects. Yet, when we performed exploratory *t*-tests only on the data of participants who made one error or more on the memory questions, we found no evidence for approach or avoidance instruction effects (all *p*s >.25).

[3] For both Experiments 1 and 2, the sample sizes were determined prior to the data collections and pre-registered together with the respective study designs. In line with the pre-registered sample information, we stopped the data-collections when at least 1000 participants had completed all measures of the experiment to ensure that we would have sufficient statistical power to detect even small effects after excluding data of participants with incorrect instruction memory (power > .80 to detect an effect size of $d = 0.20$). Because the studies could only be taken offline at fixed points in time, the final sample size always exceeded the pre-determined sample size. For both studies, we report all manipulations and measures. All data were collected in one shot without intermittent data analysis.

Assignment of the words UDIBNON, BAYRAM, and ENANWAL to the approach, avoidance, or no action conditions was counterbalanced across participants.

Evaluative priming task. To measure implicit evaluations, we used an evaluative priming task in which participants were asked to categorize target words as either positive or negative using the E and I keys of a computer keyboard. During all trials, the labels “bad” and “good” appeared in the left and right upper corners of the screen, respectively. In line with the procedures of earlier studies (e.g., Spruyt, De Houwer, Hermans, & Eelen, 2007), a single trial consisted of a fixation cross presented for 500 ms, a blank screen for 500 ms, a prime for 200 ms, a blank screen for 50 ms, and the presentation of a target word. All stimuli were presented in white font against a black background. The inter-trial interval was set to vary randomly between 500 ms and 1500 ms. Whenever an incorrect response was made or participants did not respond prior to the response deadline of 1500 ms, a red X was displayed in the center of the screen for 1000 ms before the next trial. Participants were asked to respond as quickly as possible without making too many errors. The three meaningless words UDIBNON, BAYRAM, and ENANWAL were used as prime stimuli. Targets consisted of 14 positive words (e.g., love, pleasure, smile) and 14 negative words (e.g., hate, pain, sadness). With the three primes and the two kinds of targets, there were six types of prime-target combinations. Participants first completed nine practice trials, which were followed by 120 critical test trials. The test trials were separated into two blocks of 60 trials, each containing 10 of the six types of prime-target combinations, presented in random order.

Evaluative rating task. After completion of the evaluative priming task, participants were asked to rate their liking of each of the three nonwords by answering two questions for each nonword: “To what extent do you like BAYRAM/UDIBNON/ENANWAL?” and “To what extent do you have warm feelings for BAYRAM/UDIBNON/ENANWAL?”. Participants gave

their ratings on 9-point Likert scales ranging from 1 (not at all warm; like not at all) to 9 (completely warm; like completely).

Manipulation check. After completion of the evaluative ratings, participants were asked to complete a manipulation check for each nonword. Toward this end, participants were asked what they were instructed to do when seeing the word UDIBNON, BAYRAM, or ENANWAL. Participants answered by selecting one of four options of a dropdown menu with "approach it", "avoid it", "no action was specified", and "I can't remember" as possible answers. After completion of the manipulation check, participants were informed that it was not necessary to complete the previously instructed AA task and they were thanked for their participation.

Results

Latencies from incorrect responses in the evaluative priming task (7.22%) were eliminated and outlier latencies longer than 1000 ms and shorter than 300 ms (6.99% of the correct responses) were truncated.[4] We calculated two evaluative priming scores for each

[4] The current data treatment deviated from our pre-registered data-reduction method, which was originally based on procedures used by Van Dessel et al. (2015). However, after discussion among the authors, we decided to adopt an alternative procedure that was based on previous research by the second author (e.g., Deutsch & Gawronski, 2009; Gawronski, Balas, & Creighton, 2014). This decision was made on the basis of the following arguments. First, the alternative method has produced more reliable evaluative priming scores than the pre-registered method in previous studies as well as in the current study. Second, using the alternative data-reduction method helped to resolve ambiguities in the results that were obtained with the pre-registered method by providing stronger evidence for avoidance instruction effects. Importantly, using the pre-registered data-reduction method reduced the overall magnitude of the instruction effects, but did not result in any shift in significance other than the fact that avoidance instructions had only a marginally significant effect on implicit evaluations $t(794) = -1.93$, $p = .055$, $d = 0.08$, 95% CI diff = [-10.15, 0.28]. Because of this slight inconsistency in the results we decided to also analyze the data with item-based linear mixed effects models as implemented in R package lme-4 (Bates, Maechler, Bolker, & Walker, 2014). This approach allowed us to further investigate the robustness of the approach and avoidance instruction effects by examining raw evaluative priming task reaction times (RTs) rather than compound priming scores. Moreover, it allowed us to control for variance due to unbalanced data and to control for (and test) possible effects of counterbalancing factors. These analyses supported the conclusions of the main analyses, including a significant negative evaluation of the to-be-avoided stimulus relative to the no-action word, $\chi^2(1) = 11.11$, $p < .001$, and revealed no important interactions with counterbalancing factors (see Appendix). We also decided to supplement pre-registered t-test analyses with Bayes factors, calculated according to the procedures outlined by Rouder, Speckman, Sun, Morey, and Iverson (2009) because these Bayes Factors give an indication of how strongly the data support either the null hypothesis (BF_0; reflecting the absence of a significant effect) or the alternative hypothesis (BF_1; reflecting the presence of a significant effect). BFs between 1 and 3, between 3 and 10, and larger than 10,

participant, one for the to-be-approached word and one for the to-be-avoided word. Priming scores were calculated by (a) subtracting the mean latencies on trials with a positive target and a given action-related word prime from the mean latencies on trials with a positive target and the no-action prime, (b) subtracting the mean latencies on trials with a negative target and a given action-related word prime from the mean latencies on trials with a negative target and the no-action prime, and (c) subtracting the second difference score from the first difference score. The Spearman-Brown corrected split-half reliability of this evaluative priming score, calculated on the basis of an odd-even split, was $r(792) = .18$ for the to-be-approached word and $r(792) = .11$ for the to-be-avoided word.

We performed paired-sample *t*-tests on the evaluative priming scores for the to-be-approached word and the to-be-avoided word. First, replicating the results of previous studies (e.g., Van Dessel et al., 2015), implicit evaluations of the to-be-approached word ($M = 5.52$, $SD = 53.66$) were more favorable than implicit evaluations of the to-be-avoided word ($M = -5.90$, $SD = 50.81$), $t(793) = 5.95$, $p < .001$, $d = 0.21$, 95% confidence interval of the difference (CI diff) = [7.65, 15.18]. As predicted by both the associative self-anchoring account and the propositional account, the priming score for the to-be-approached word was significantly larger than zero, indicating that participants preferred the to-be-approached word over the no-action word, $t(793) = 2.90$, $p = .004$, $d = 0.10$, 95% CI diff = [1.78, 9.25], $BF_1 = 5.10$. Second, and most crucially, implicit evaluation scores of the to-be-avoided word were significantly smaller than zero, indicating that participants preferred the no-action word over the to-be-avoided word, $t(793) = -3.27$, $p = .001$, $d = 0.12$, 95% CI diff = [-9.44, -2.36], $BF_1 = 15.95$. Finally, a Bayesian t-test provided strong evidence in favor of the null hypothesis that the avoidance instruction effect on

---

respectively designate ‘anecdotal evidence’, ‘substantial evidence’, and ‘strong evidence’ for either the null (BF_0) or the alternative hypothesis (BF_1) (Jeffreys, 1961).

implicit evaluations is not smaller in magnitude than the approach instruction effect, $t(793) = -0.12$, $p = .90$, $d = -0.004$, 95% CI diff = [-6.61, 5.84], $BF_0 = 27.50$.[5]

Discussion

Experiment 1 provides evidence that (a) instructions to approach a stimulus lead to more positive implicit evaluations of the to-be-approached stimulus, (b) instructions to avoid a stimulus lead to more negative implicit evaluations of the to-be-avoided stimulus, and (c) approach instructions do not produce quantitatively larger effects than avoidance instructions. Bayesian factors indicated that our data provide substantial evidence for the first conclusion and strong evidence for the latter two conclusions. Overall, these conclusions are consistent with the predictions derived from the propositional account of AA instruction effects: both approach and avoidance instructions may allow participants to infer their liking or disliking of the stimulus, which should lead to corresponding changes in implicit evaluations. In contrast, the finding that avoidance instructions influenced implicit evaluations is difficult to reconcile with the associative self-anchoring account. This account implies that instruction effects should be limited to approach instructions, which may lead to a transfer of positive self-evaluations to the to-be-approached stimulus via the formation of self-stimulus associations. However, an exclusive operation of associative self-anchoring does not provide a straightforward explanation for the negative effects of avoidance instructions (see Gawronski et al., 2007, for a discussion). The observation that approach instructions do not produce greater effects than avoidance instructions provides suggestive evidence that associative self-anchoring processes do not play any role in AA instruction effects on stimulus evaluation over and above propositional processes.

Experiment 2

[5] Analyses on participants' explicit rating scores revealed a similar pattern as obtained for implicit evaluations. Because the two competing accounts do not make different predictions for the effects of AA instructions on explicit evaluations, we report the results of these analyses in the Appendix.

In Experiment 2, participants received instructions to approach one nonword and avoid another nonword and then performed an evaluative IAT and a self-stimulus IAT, in counter-balanced order. To test predictions of the associative self-anchoring account and the propositional account, we examined whether AA instruction effects on stimulus evaluations (as measured with an evaluative IAT) are mediated by the effect of AA instructions on implicit self-stimulus linking (as measured with a self-stimulus IAT).

Method

Participants. A total of 1808 visitors of the Project Implicit research website participated in the study. In line with standard treatment of Project Implicit IAT data (e.g., Smith, De Houwer, & Nosek, 2013), we excluded participants who (a) did not fully complete all questions and tasks (440 participants; i.e., 24.34%), (b) had error rates above 30% for any of the IATs (25 participants; i.e., 1.83%), (c) responded faster than 400 ms on more than 10% of the IAT trials for any of the IATs (84 participants; i.e., 6.35%), (d) had error rates above 40% for any of the critical IAT blocks (21 participants; i.e., 1.56%), or (e) did not correctly answer the memory questions (301 participants; i.e., 24.31%).[6] Analyses were performed on the data of 937 participants (636 women, mean age = 38, *SD* = 13).

Procedure. The procedure of Experiment 2 was largely identical to Experiment 1 with a few exceptions. First, the experiment included only two nonwords: UDIBNON and BAYRAM. Participants received AA instructions specifying that they would perform a task in which they would approach UDIBNON and avoid BAYRAM (or vice versa). Second, following the AA instructions, participants completed two IATs instead of an evaluative priming task. In the evaluative IAT, participants categorized eight attribute words (e.g., wonderful, evil) as 'positive'

[6] Including the data from participants who did not correctly answer the memory questions in the analyses did not change the statistical significance of any of the reported effects.

or ‘negative’ and the target words UDIBNON and BAYRAM as ‘Udibnon’ or ‘Bayram’. To avoid that the target stimuli were classified only on the basis of simple perceptual features, these words were presented in different font types (Arial Black and Fixedsys), capitalizations (uppercase and lowercase), and sizes (16pt and 18pt), resulting in 8 different stimuli for each nonword (for a similar procedure, see Zanon, De Houwer, Gast, & Smith, 2014). The attribute words were always presented in Arial Black, font size 16, uppercase. The evaluative IAT consisted of three practice blocks and two experimental blocks. Participants began the IAT with 20 practice trials sorting the target words and 20 practice trials sorting positive and negative stimuli. Next, participants completed 56 trials in which UDIBNON and positive stimuli shared one response key and BAYRAM and negative stimuli shared another response key (or vice versa). Participants then practiced sorting target words on 40 trials with a reversed response key assignment. Finally, participants completed a second set of 56 trials in which UDIBNON shared a response key with negative and BAYRAM shared a response key with positive (or vice versa). If participants made an error in the categorization task, a red “X” appeared on the screen until participants provided the correct response. Latencies were recorded until a correct response was made. In the self-stimulus IAT, participants categorized four self-related words (i.e., I, me, mine, and self) and four other-related words (i.e., they, them, their, and other) as ‘Self’ or ‘Other’ (see Phills et al., 2011) and the target words UDIBNON and BAYRAM as ‘Udibnon’ or ‘Bayram’. All other procedural details of the self-stimulus IAT were identical to the evaluative IAT. The order of the two IATs was counterbalanced across participants.

Results

Evaluative IAT. IAT scores for the evaluative IAT were calculated using the D2-algorithm, which is the recommended scoring procedure for IATs in which participants need to correct their mistakes before moving on to the next trial (Greenwald, Nosek, & Banaji, 2003).

The IAT score was calculated on the basis of the difference in RTs on trials in which UDIBNON shared a response key with positive and UDIBNON shared a response key with negative compared to trials with a reversed response key assignment, such that higher scores indicate a stronger preference for BAYRAM over UDIBNON. The Spearman-Brown corrected split-half reliability of the evaluative IAT score, calculated on the basis of an odd-even split, was $r(935) = .85$. Across groups, participants displayed an implicit preference for BAYRAM over UDIBNON ($M = 0.13$, $SD = 0.50$), $t(936) = 8.05$, $p < .001$. More importantly, a between-groups *t*-test indicated a significant effect of AA instructions, $t(935) = 23.89$, $p < .001$, $d = 1.56$, 95% CI diff = [0.57, 0.67]. Participants who had been instructed to approach BAYRAM and avoid UDIBNON exhibited a stronger implicit preference for BAYRAM over UDIBNON ($M = 0.43$, $SD = 0.38$) than participants who had been instructed to avoid BAYRAM and approach UDIBNON ($M = -0.19$, $SD = 0.41$).

Self-Stimulus IAT. IAT scores for the self-stimulus IAT were calculated using the D2-algorithm, such that higher scores indicate facilitated responses when BAYRAM shared a key with the self than when UDIBNON shared a key with the self. The Spearman-Brown corrected split-half reliability of the self-stimulus IAT score was $r(935) = .81$. Self-stimulus IAT scores showed a significant positive correlation with scores on the evaluative IAT, $r(935) = .17$, $p < .001$. Across groups, self-stimulus IAT scores indicated that participants more easily linked BAYRAM to the self than they linked UDIBNON to the self ($M = 0.16$, $SD = 0.41$), $t(936) = 11.45$, $p < .001$. Crucially, a between-groups *t*-test indicated a significant effect of AA instructions, $t(935) = 15.95$, $p < .001$, $d = 1.04$, 95% CI diff = [0.34, 0.43]. Participants who had been instructed to approach BAYRAM and avoid UDIBNON had higher self-stimulus IAT scores ($M = 0.34$, $SD = 0.37$) than participants who had been instructed to avoid BAYRAM and approach UDIBNON ($M = -0.05$, $SD = 0.36$).

Analysis of Variance (ANOVA). Following the recommendations of an anonymous reviewer, we also performed a mixed ANOVA on IAT scores. This ANOVA included one within-subjects factor: IAT Type (evaluative IAT, self-stimulus IAT), and two between-subjects factors: IAT Order (evaluative IAT first, self-stimulus IAT first) and AA Instructions (approach BAYRAM and avoid UDIBNON, approach UDIBNON and avoid BAYRAM). We observed a main effect of AA instructions, $F(1,1864) = 819.82$, $p < .001$, a two-way interaction of AA Instructions and IAT Type, $F(1,1864) = 44.69$, $p < .001$, and a three-way interaction of AA Instructions, IAT Type, and IAT Order, $F(1,1864) = 37.41$, $p < .001$. Further examination of this three-way interaction revealed that the AA Instruction effect on the evaluative IAT was larger than the AA instruction effect on the self-stimulus IAT for participants who performed the evaluative IAT first (effect on the evaluative IAT: $d = 1.88$; effect on the self-stimulus IAT: $d = 0.93$), $F(1,954) = 85.41$, $p < .001$, but not for participants who performed the self-stimulus IAT first (effect on the evaluative IAT: $d = 1.26$; effect on the self-stimulus IAT: $d = 1.16$), $F(1,908) = 0.04$, $p = .84$.

Mediation analysis. To investigate the relationship between AA instruction effects on the evaluative IAT and the self-stimulus IAT, we performed mediation analyses with the lavaan package (version 0.5-16; Rosseel, 2012). We used the bootstrap method to estimate standard errors for the effects. We first tested whether changes in implicit self-stimulus linking mediate the effect of AA instructions on implicit evaluations (see Figure 1). Toward this end, evaluative IAT scores were simultaneously regressed on both AA instructions (approach BAYRAM and avoid UDIBNON versus approach UDIBNON and avoid BAYRAM) and self-stimulus IAT scores (Baron & Kenny, 1986). Consistent with the predictions of the self-anchoring account, the indirect effect of AA instructions on evaluative IAT scores with self-stimulus IAT scores as a mediator was statistically significant, $\beta = .14$, $Z = 9.17$, $p < .001$, 95% CI of $\beta = [0.11,0.17]$, R^2_{ind}

= 0.15. However, the effect of AA instructions on the evaluative IAT score remained statistically significant after controlling for self-stimulus IAT scores, β = .48, Z = 17.47, p < .001, 95% CI of β = [0.43, 0.54], R^2_{dir} = 0.23, indicating that mediation via implicit self-stimulus linking was only partial rather than full. The proportion mediated (PM) measure was calculated in line with de Heus (2012) and revealed that 21.97% of the effect of AA instructions on implicit evaluations (i.e., 8.42% of the total variance in implicit evaluations) could be accounted for by mediation via changes in self-stimulus linking. The direct pathway accounted for the residual 78.03% of the effect of AA instructions (i.e., 29.84% of the total variance in implicit evaluations). Mediation analyses that were performed separately for participants who first completed the evaluative IAT and participants who first completed the self-stimulus IAT, indicated that, respectively, 15.27% and 37.05% of the AA instruction effect on implicit evaluations could be accounted for by mediation via self-stimulus linking. A mediation model in which the direct path from AA instructions to evaluative IAT scores was constrained to zero did not fit the data for either group of participants, χ^2s > 59, *ps* < .001. The comparative fit index (CFI), which is one of the most common fit indices and least affected by sample size (Fan, Thompson, & Wang, 1999), indicated poor fit of this mediation model (evaluative IAT first: CFI = 0.51; self-stimulus IAT first: CFI = 0.83).

We also tested the reverse mediation model, in which self-stimulus IAT scores were simultaneously regressed on both AA instructions and evaluative IAT scores (Figure 2). In this mediation model, the indirect effect of the AA instructions on self-stimulus IAT scores with evaluative IAT scores as a mediator was also significant, β = .19, Z = 9.37, p <.001, 95% CI of β = [0.15,0.23], R^2_{ind} = 0.11. AA instructions still had a significant effect on self-stimulus IAT scores after controlling for evaluative IAT scores, β =.19, Z = 6.46, p < .001, 95% CI of β = [0.13,0.25], R^2_{dir} = 0.04. Mediation via changes in evaluative IAT scores accounted for 49.48%

the AA instruction effect on self-stimulus IAT scores (evaluative IAT first: 61.86%; self-stimulus IAT: 43.89%). A mediation model in which the direct path from the AA condition variable to the self-stimulus IAT score was constrained to zero did not fit the data, $\chi^2(1) = 44.32$, $p < .001$. However, the CFI indicated good model fit for this restricted model (evaluative IAT first: CFI = 0.98; self-stimulus IAT first: CFI = 0.90).[7]

Discussion

Experiment 2 showed that AA instructions influenced both implicit evaluations, as measured with an evaluative IAT, and implicit self-stimulus linking, as measured with a self-stimulus IAT. Consistent with the associative self-anchoring account, mediation analyses indicated that the effect of AA instructions on implicit evaluation was mediated by corresponding changes in self-stimulus linking. However, the obtained mediation was only partial, in that AA instructions influenced implicit evaluations after controlling for changes in implicit self-stimulus linking. The direct effect on implicit evaluations explained approximately 3.5 times the amount of variance in implicit evaluations due to AA instructions compared to the mediation via changes in self-stimulus IAT scores. Testing the reversed mediation model, we found that the effect of AA instructions on implicit self-stimulus linking was also partially mediated by changes in implicit evaluation. Mediation via implicit evaluations accounted for approximately the same amount of variance in self-stimulus IAT scores due to AA instructions as the direct effect. Thus, although the obtained mediation via implicit self-object linking is consistent with the associative

[7] We also performed t-test analyses on participants' explicit rating scores of the non-words, revealing an AA instruction effect. Similar to previous studies (e.g., Van Dessel et al., 2015), additional mediation analyses showed that AA instruction effects on implicit evaluations were not fully mediated by changes in explicit evaluations. Moreover, the AA instruction effect on evaluative IAT scores remained significant when a multiple mediation model was considered that included explicit evaluations and self-stimulus IAT scores (or evaluative IAT scores) as mediators. It is important to note, however, that the results of these mediation analyses are difficult to interpret because the order of explicit and implicit measures was not counterbalanced in the current study.

self-anchoring account, the current findings suggest that AA instruction effects are also (and more so) driven by processes other than associative self-anchoring.

General Discussion

The current experiments were designed to test predictions of a propositional account and an associative self-anchoring account of AA instruction effects. Toward this end, we probed unique effects of approach and avoidance instructions on implicit evaluation (Experiment 1) and examined the mediating role of implicit self-stimulus linking in AA instruction effects on implicit evaluations (Experiment 2). Overall, the results fit best with a propositional explanation of AA instruction effects.

The results of Experiment 1 indicate that both approach instructions and avoidance instructions can cause changes in implicit evaluations, as predicted by the propositional account. According to the associative self-anchoring account, approach instructions should lead to more favorable implicit evaluations of the to-be-approached stimulus. However, in the absence of additional assumptions, associative self-anchoring fails to explain how avoidance instructions may negatively influence implicit evaluations of the to-be-avoided stimulus. To accommodate the current findings, the associative self-anchoring account could be extended to allow for the possibility that avoidance instructions lead to a transfer of negative valence to the to-be-avoided stimulus either (a) via the formation of an inhibitory association between representations of the self and the to-be-avoided stimulus or (b) via an excitatory association between representations of "not-me" and the to-be-avoided stimulus. The current results would imply that the effects of such negative associative self-anchoring can be of similar magnitude than effects that are obtained via positive self-anchoring. Note, however, that such extensions of the associative self-anchoring account are inconsistent with existing evidence for self-anchoring processes in the context of the ownership effect (see Gawronski et al., 2007, Experiment 3) or AA training effects (Phills et al.,

2011). Moreover, as discussed in the introduction, such extensions of the self-anchoring account must rely on questionable assumptions, such as the assumptions that (a) avoidance results in inhibitory associations and that inhibition of the self-concept results in a negative affective reaction rather than the absence of a positive reaction or (b) associative processes are capable of performing negations. Thus, although the results of Experiment 1 do not rule out a potential contribution of associative self-anchoring to the obtained effect of approach instructions, the obtained effect of avoidance instructions is inconsistent with current ideas and evidence about self-anchoring. Yet, results are consistent with the hypothesized role of propositional processes, which predicts both a positive effect of approach instructions and a negative effect of avoidance instructions.

Suggestive evidence for associative self-anchoring comes from Experiment 2, in which AA instruction effects on implicit evaluations were mediated by changes in implicit self-stimulus linking. This mediation pattern is predicted by the associative self-anchoring account, but it is not predicted by the propositional account. However, the obtained mediation was only partial, in that AA instructions showed a significant effect on implicit evaluations after controlling for implicit self-object linking. A potential explanation for this finding is that propositional inferences and associative self-anchoring jointly contribute to AA instruction effects on implicit evaluations. With the confirmed contribution of propositional processes in Experiment 1, the mediation produced by associative self-anchoring should be only partial (rather than full), in that propositional processes should lead to a direct effect of AA instructions on implicit evaluations that is not mediated by implicit self-object linking. Thus, the indirect effect of AA instructions on implicit evaluations via implicit self-stimulus linking, which accounted for approximately 22% of the variance in AA instruction effects, might reflect the contribution of associative self-anchoring, whereas the direct effect of AA instructions, which accounted for approximately 78%

of the variance in AA instruction effects, might reflect the contribution of propositional processes.

Although a joint contribution of propositional processes and associative self-anchoring is consistent with the obtained pattern of results, it is important to note that the partial mediation makes our data ambiguous about the proposed contribution of associative self-anchoring. From the perspective of the propositional account, one could argue that AA instructions should influence scores on the self-stimulus IAT if participants infer that the to-be-approached stimulus is more similar to the self than a to-be-avoided stimulus. In this case, the evaluative IAT and the self-stimulus IAT should both be affected by AA instructions, as found in Experiment 2. Moreover, because of their shared relation to a common third variable (i.e., AA instructions), the two IATs may show a modest positive correlation, again consistent with the findings of Experiment 2. As a result, mediation analyses should reveal a partial mediation pattern regardless of which variable is treated as the mediator versus the distal outcome. Because we obtained partial mediation in either case, our mediation analyses fail to provide unambiguous evidence for the proposed role of associative self-anchoring. On the one hand, it is possible that the obtained results reflect a joint contribution of propositional processes and associative self-anchoring. On the other hand, it is possible that AA instruction effects are exclusively driven by propositional processes, with the partial mediation patterns being due to the shared relation of implicit evaluations and implicit self-stimulus linking to AA instructions as a common antecedent. Thus, although the current findings provide clear support for the hypothesized role of propositional processes, they remain ambiguous regarding an additional contribution of associative self-anchoring. This ambiguity cannot be addressed with regression-based mediation analyses (e.g.,

Baron & Kenny, 1986), but requires advanced experimental designs to establish the specific structure of the underlying causal chain (e.g., Spencer, Zanna, & Fong, 2005).[8]

Overall, the current findings support the idea that propositional processes play an important role in AA instruction effects on implicit evaluation. This conclusion is consistent with the growing body of evidence showing that (a) verbal instructions can have strong, immediate effects on implicit evaluations (Castelli et al., 2004; Gregg et al., 2006; Whitfield & Jordan, 2009) and (b) instruction-based changes in implicit evaluation depend on the operation of propositional processes (Cone & Ferguson, 2015; Peters & Gawronski, 2011b; Zanon et al., 2014). The current results extend these findings by showing that propositional processes also play a major role in AA instruction effects. Yet, in contrast to instructions that specify evaluative qualities of stimuli (see Whitfield & Jordan, 2009), AA instructions seem to have a direct effect on implicit evaluations that is independent of changes in explicit evaluation (Van Dessel, De Houwer, Gast, et al., 2016).

By uncovering the processes underlying the effects of AA instructions, our research provides important information that constrains mental process models of evaluation. Together with earlier research on AA instruction effects (e.g., Van Dessel, De Houwer, Gast, et al., 2016), the current findings are difficult to reconcile with a particular type of associative or dual-process models which claim that (a) implicit evaluations typically reflect the slow accrual of paired associations in memory (e.g., Rydell & McConnell, 2006) or (b) propositional processes can also

[8] Another problem with the mediation results of Experiment 2 is that indices of model fit are difficult to reconcile with the proposed role of associative-self anchoring. A full mediation model with evaluative IAT scores as a mediator for the impact of AA instructions on self-stimulus IAT scores fit the data better than a full mediation model with self-stimulus IAT scores as a mediator of the impact of AA instructions on evaluative IAT scores. In the current study, the comparative fit index was .94 for the former model and .66 for the latter model. Values close to .95 are generally considered as indicating very good model fit and values below .90 indicate a poor fitting model (Hu & Bentler, 1999).

influence implicit evaluations but only via changes in explicit evaluation (e.g., Gawronski & Bodenhausen, 2006). Instead, the body of research on AA instructions seems to fit better with propositional models which assume that mental propositions can function as the proximal causes of changes in implicit evaluations independent of associative representations (De Houwer, 2014). Of course, distinguishing between broad classes of evaluation theories on the basis of a single set of data is difficult, if not impossible. Proponents of a challenged theory can always make post-hoc assumptions to explain unexpected findings (see Gawronski & Bodenhausen, 2015). For instance, associative accounts of implicit evaluation might explain the current results by postulating that changes in implicit evaluations can occur due to the formation of associations as the result of the pairing of a valenced action word (‘approach’ or ‘avoid’) and a stimulus in the instructions. We believe that scientific progress can be facilitated by pre-specifying the predictions of these theories and testing them in well-controlled studies. By using this method, the current study (a) provides further evidence that (automatic) effects of evaluative learning may depend on propositional processes, and thereby (b) contributes to our understanding of the processes underlying implicit evaluation.

References

Allport, G. (1935). Attitudes. In Murchison, C. (Ed.). *A Handbook of Social Psychology* (pp. 789-843). Worcester, MA: Clark University Press.

Amir, N., Kuckertz, J. M., & Najmi, S. (2013). The Effect of Modifying Automatic Action Tendencies on Overt Avoidance Behaviors. *Emotion, 13,* 478-484. doi:10.1037/a0030443

Baron, R. M., & Kenny, D. A. (1986). The moderator–mediator variable distinction in social psychological research: Conceptual, strategic, and statistical considerations. *Journal of Personality and Social Psychology, 51,* 1173−1182. doi:10.1037/0022-3514.51.6.1173

Bates, D., Maechler, M., Bolker, B., & Walker, S. (2014). lme4: Linear mixed-effects models using Eigen and S4. R package version 1.0-6. http://CRAN.R-project.org/package=lme4

Castelli, L., Zogmaister, C., Smith, E. R., & Arcuri, L. (2004). On the automatic evaluation of social exemplars. *Journal of Personality and Social Psychology, 86,* 373–387. doi:10.1037/0022-3514.86.3.373

Cone, J., & Ferguson, M. J. (2015). He did what? The role of diagnosticity in revising implicit evaluations. *Journal of Personality and Social Psychology, 108,* 37-57. doi:10.1037/pspa0000014

De Heus P. (2012). R squared effect size measures and overlap between direct and indirect effect in mediation analysis. *Behavior Research Methods, 44,* 213-221. doi:10.3758/s13428-011-0141-5

De Houwer, J. (2009). The propositional approach to associative learning as an alternative for association formation models. *Learning & Behavior, 37,* 1–20. doi:10.3758/LB.37.1.1

De Houwer, J. (2014). A propositional model of implicit evaluation. *Social and Personality Psychology Compass, 8,* 342-353. doi:10.1111/spc3.12111

De Houwer, J., Gawronski, B., & Barnes-Holmes, D. (2013). A functional-cognitive framework for attitude research. *European Review of Social Psychology, 24*, 252–287. doi:10.1080/10463283.2014.892320

Deutsch, R., & Gawronski, B. (2009). When the method makes a difference: Antagonistic effects on "automatic evaluations" as a function of task characteristics of the measure. *Journal of Experimental Social Psychology, 45*, 101-114. doi:10.1016/j.jesp.2008.09.001

Deutsch, R., Gawronski, B., & Strack, F. (2006). At the boundaries of automaticity: Negation as reflective operation. *Journal of Personality and Social Psychology, 91*, 385-405.

Fan, X., Thompson, B., and Wang, L. (1999), Effects of sample size, estimation methods, and model specification on structural equation modeling fit indexes, *Structural Equation Modeling, 6*, 56-83. doi:10.1080/10705519909540119

Fazio, R. H., Sanbonmatsu, D. M., Powell, M. C., & Kardes, F. R. (1986). On the automatic activation of attitudes. *Journal of Personality and Social Psychology, 50*, 229–238. doi:10.1037/0022-3514.50.2.229

Förster, J. (2001). Success/failure feedback, expectancies, and approach/avoidance motivation: How regulatory focus moderates classic relations. *Journal of Experimental Social Psychology, 37*, 253–260. doi:10.1006/jesp.2000.1455

Field, A. P. (2006). Is conditioning a useful framework for understanding the development and treatment of phobias? *Clinical Psychology Review, 26*, 857-875. doi:10.1016/j.cpr.2005.05.010

Gawronski, B., Balas, R., & Creighton, L. (2014). Can the formation of conditioned attitudes be intentionally controlled? *Personality & Social Psychology Bulletin, 40*, 419–432. doi:10.1177/0146167213513907

Gawronski, B., & Bodenhausen, G. V. (2006). Associative and propositional processes in evaluation: An integrative review of implicit and explicit attitude change. *Psychological Bulletin, 132,* 692-731. doi:10.1037/0033-2909.132.5.692

Gawronski, B., & Bodenhausen, G. V. (2007). Unraveling the processes underlying evaluation: Attitudes from the perspective of the APE Model. *Social Cognition, 25*, 687-717. doi:10.1521/soco.2007.25.5.687

Gawronski, B., & Bodenhausen, G. V. (2011). The associative-propositional evaluation model: Theory, evidence, and open questions. *Advances in Experimental Social Psychology, 44,* 59-127. doi:10.1016/B978-0-12-385522-0.00002-0

Gawronski, B., & Bodenhausen, G. V. (2015). Theory evaluation. In B. Gawronski, & G. V. Bodenhausen (Eds.), *Theory and explanation in social psychology* (pp. 3-23). New York, NY: Guilford Press.

Gawronski, B., Bodenhausen, G. V., & Becker, A. P. (2007). I like it, because I like myself: Associative self-anchoring and post-decisional change of implicit evaluations. *Journal of Experimental Social Psychology, 43*, 221-232. doi:10.1016/j.jesp.2006.04.001

Gawronski, B., Deutsch, R., Mbirkou, S., Seibt, B., & Strack, F. (2008). When “just say no” is not enough: Affirmation versus negation training and the reduction of automatic stereotype activation. *Journal of Experimental Social Psychology, 44,* 370-377.

Gawronski, B., Strack, F., & Bodenhausen, G. V. (2009). Attitudes and cognitive consistency: The role of associative and propositional processes. In R. E. Petty, R. H. Fazio, & P. Briñol (Eds.), *Attitudes: Insights from the new implicit measures* (pp. 85–117). New York: Psychology Press.

Gawronski, B., & Walther, E. (2008). The TAR effect: When the ones who dislike become the ones who are disliked. *Personality and Social Psychology Bulletin, 9,* 1276-1289. doi:10.1177/0146167208318952

Gawronski, B., Walther, E., & Blank, H. (2005). Cognitive consistency and the formation of interpersonal attitudes: Cognitive balance affects the encoding of social information. *Journal of Experimental Social Psychology, 41,* 618-626. doi:10.1016/j.jesp.2004.10.005

Greenwald, A. G., McGhee, D. E., & Schwartz, J. L. (1998). Measuring individual differences in implicit cognition: the implicit association test. *Journal of Personality and Social Psychology, 74,* 1464–1480. doi:10.1037/0022-3514.74.6.1464

Greenwald, A. G., Nosek, B. A., & Banaji, M. R. (2003). Understanding and using the Implicit Association Test: I. An improved scoring algorithm. *Journal of Personality and Social Psychology, 85,* 197–216. doi:10.1037/0022-3514.85.2.197

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. doi:10.1037/0022-3514.90.1.1

Hu, L., & Bentler, P. M. (1999). Cutoff criteria for fit indexes in covariance structure analysis: Conventional criteria versus new alternatives. *Structural Equation Modeling, 6,* 1-55. doi:10.1080/10705519909540118

Jeffreys, H. (1961). Theory of Probability. Oxford: Oxford University Press.

Jones, C. R., Vilensky, M. R., Vasey, M. W., & Fazio, R. H. (2013). Approach behavior can mitigate predominately univalent negative attitudes: Evidence regarding insects and spiders. *Emotion, 13,* 989-996. doi:10.1037/a0033164

Kawakami, K., Phills, C. E., Steele, J. R., & Dovidio, J. F. (2007). (Close) distance makes the heart grow fonder: Improving implicit racial evaluations and interracial interactions

through approach behaviors. *Journal of Personality and Social Psychology, 92*, 957-971. doi:10.1037/0022-3514.92.6.957

Kawakami, K., Steele, J. R., Cifa, C., Phills, C. E., & Dovidio, J. F. (2008). Approaching math increases math = me and math = pleasant. *Journal of Experimental Social Psychology, 44,* 818–825. doi:10.1016/j.jesp.2007.07.009

Laham, S. M., Kashima, Y., Dix, J., Wheeler, M., & Levis, B. (2014). Elaborated contextual framing is necessary for action-based attitude acquisition. *Cognition & Emotion, 28,* 1119-1126. doi:10.1080/02699931.2013.867833

Mitchell, C. J., De Houwer, J., & Lovibond, P. F. (2009). The propositional nature of human associative learning. *The Behavioral and Brain Sciences, 32,* 183–198. doi:10.1017/s0140525x09000855

Nussinson, R., Seibt, B., Häfner, M., & Strack, F. (2010).Come a bit closer: Approach motor actions lead to feeling similar and behaviorally assimilating to others. *Social Cognition, 28*, 40-58. doi:10.1521/soco.2010.28.1.40

Olson, J. M., & Stone, J. (2005). The influence of behavior on attitudes. In D. Albarracin, B. T. Johnson, & M. P. Zanna (Eds.), *Handbook of attitudes and attitude change* (pp. 223-271). Mahwah, NJ: Erlbaum.

Peters, K. R., & Gawronski, B. (2011a). Mutual influences between the implicit and explicit self-concepts: The role of memory activation and motivated reasoning. *Journal of Experimental Social Psychology, 47*, 436-442. doi:10.1016/j.jesp.2010.11.015

Peters, K. R., & Gawronski, B. (2011b). Are we puppets on a string? Comparing the impact of contingency and validity on implicit and explicit evaluations. *Personality and Social Psychology Bulletin, 37*, 557–569. doi:10.1177/0146167211400423

Phills, C. E., Kawakami, K., Tabi, E., Nadolny, D., & Inzlicht, M. (2011). Mind the gap: Increasing associations between the self and blacks with approach behaviors. *Journal of Personality and Social Psychology, 100,* 197–210. doi:10.1037/a0022159

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

Rouder, J. N., Speckman, P. L., Sun, D., Morey, R. D., & Iverson, G. (2009). Bayesian t tests for accepting and rejecting the null hypothesis. *Psychonomic Bulletin & Review, 16,* 225–237. doi: 10.3758/PBR.16.2.225

Rydell, R. J., & McConnell, A. R. (2006). Understanding implicit and explicit attitude change: A systems of reasoning analysis. *Journal of Personality and Social Psychology, 91,* 995–1008. doi:10.1037/0022-3514.91.6.995

Schneirla, T. (1959). An evolutionary and developmental theory of biphasic processes underlying approach and withdrawal. In Jones, M. (Ed.). *Nebraska Symposium on Motivation* (pp.1-42). Lincoln: University of Nebraska Press.

Smith, C. T., De Houwer, J., & Nosek, B. (2013). Consider the source: Persuasion of implicit evaluations is moderated by source credibility. *Personality and Social Psychology Bulletin*, 39, 193-205. doi:10.1177/0146167212472374

Smith, C. T., & De Houwer, J. (2015). Hooked on a feeling: Affective anti-smoking messages are more effective than cognitive messages at changing implicit evaluations of smoking. *Frontiers in Psychology, 6*:1488. doi:10.3389/fpsyg.2015.01488

Smith, E. R., & DeCoster, J. (2000). Dual-process models in social and cognitive psychology: conceptual integration and links to underlying memory systems. *Personality and Social Psychology Review*, *4*, 108–131. doi:10.1207/S15327957PSPR0402_01

Spencer, S. J., Zanna, M. P., & Fong, G. T. (2005). Establishing a causal chain: Why experiments are often more effective than mediational analyses in examining psychological processes. *Journal of Personality and Social Psychology, 89*, 845–851. doi:10.1037/0022-3514.89.6.845

Spruyt, A., De Houwer, J., & Hermans, D. (2007). Modulation of semantic priming by feature-specific attention allocation. *Journal of Memory and Language, 61,* 37 - 54. doi:10.1016/j.jml.2009.03.004

Van Dessel, P., De Houwer, J., Gast, A., & Smith, C. T. (2015). Instruction-based approach–avoidance Effects: changing stimulus evaluation via the mere instruction to approach or avoid stimuli. *Experimental Psychology, 62,* 161-169. doi:10.1027/1618-3169/a000282

Van Dessel, P., De Houwer, J., Gast, A., Smith, C. T., & De Schryver, M. (2016). Instructing implicit processes: When instructions to approach or avoid influence implicit but not explicit evaluation. *Journal of Experimental Social Psychology, 63,* 1-9. doi:10.1016/j.jesp.2015.11.002

Van Dessel, P., De Houwer, J., Roets, A. & Gast, A. (2016). Failures to change stimulus evaluations by means of subliminal approach and avoidance training. *Journal of Personality and Social Psychology, 110,* e1-e15. doi:10.1037/pspa0000039

Wiers, R. W., Eberl, C., Rinck, M., Becker, E. S., & Lindenmeyer, J. (2011). Retraining automatic action tendencies changes alcoholic patients' approach bias for alcohol and improves treatment outcome. *Psychological Science, 22,* 490–497. doi:10.1177/0956797611400615

Whitfield, M., & Jordan, C. H. (2009). Mutual influence of implicit and explicit attitudes. *Journal of Experimental Social Psychology, 45,* 748–759. doi:10.1016/j.jesp.2009.04.006

Woud, M. L., Maas, J., Becker, E.S., & Rinck, M. (2013). Make the manikin move: Symbolic approach-avoidance responses affect implicit and explicit face evaluations. *Journal of Cognitive Psychology, 25*, 738-744. doi:10.1080/20445911.2013.817413

Yamaguchi, S., Greenwald, A. G., Banaji, M. R., Murakami, F., Chen, D., Shiomura, K.,. Kobayashi, C., & Cai, H (2007). Apparent universality of positive implicit self-esteem. *Psychological Science, 18*, 498–500. doi:10.1111/ j.1467-9280.2007.01928.x

Ye, Y., & Gawronski, B. (2016). When possessions become part of the self: Ownership and implicit self-object linking. *Journal of Experimental Social Psychology, 64*, 72-87. doi:10.1016/j.jesp.2016.01.012

Zanon, R., De Houwer, J., Gast, A., Smith, C. T. (2014). When does relational information influence evaluative conditioning? *Quarterly Journal of Experimental Psychology, 67*, 2105-2122. doi:10.1080/17470218.2014.907324

Zogmaister, C, Perugini, M., Richetin, J (2016). Motivation modulates the effect of approach on implicit preferences. *Cognition & Emotion, 30*, 890-911. doi:10.1080/02699931.2015.1032892

Acknowledgments

Pieter Van Dessel is supported by a Ph.D. fellowship of the Scientific Research Foundation, Flanders (FWO-Vlaanderen). Jan De Houwer is supported by Methusalem Grant BOF16/MET_V/002 of Ghent University and by the Interuniversity Attraction Poles Program initiated by the Belgian Science Policy Office (IUAPVII/33). The research in this paper has been supported by Grant FWO12/ASP/275 of FWO - Vlaanderen.

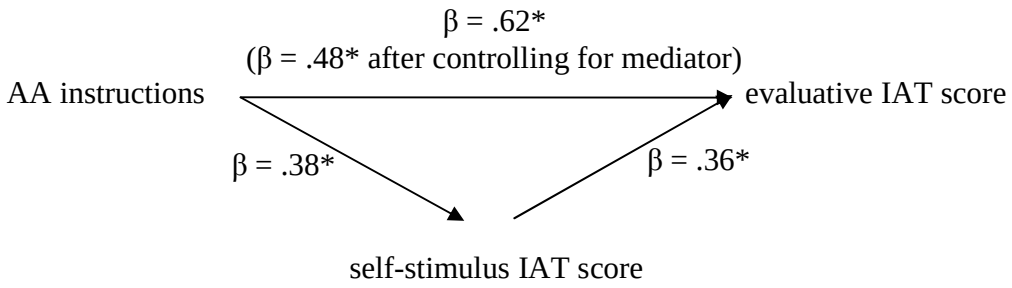

Figure 1. Standardized estimates of mediation coefficients for mediation of AA instruction effects on evaluative IAT scores by changes in self-stimulus IAT scores. * $p < .001$.

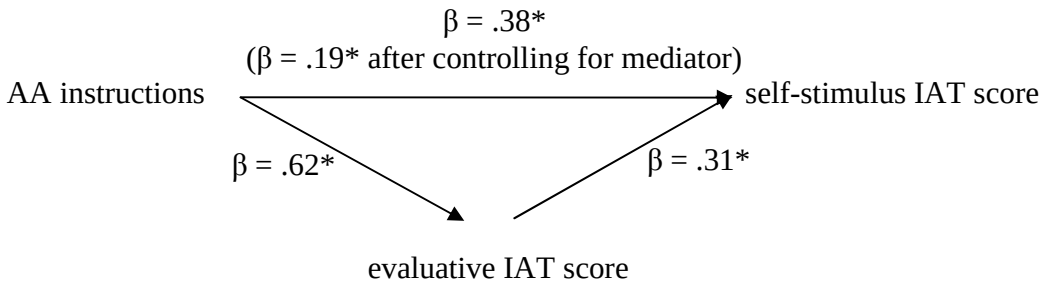

Figure 2. Standardized estimates of mediation coefficients for mediation of AA instruction effects on self-stimulus IAT scores by changes in evaluative IAT scores. * $p < .001$.