The contextual malleability of approach-avoidance training effects: Approaching or avoiding fear conditioned stimuli modulates effects of approach-avoidance training

Gaëtan Mertens[1, 2], Pieter Van Dessel[1] & Jan De Houwer[1]

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

[2]Department of Clinical and Health Psychology, Utrecht University, Utrecht, the Netherlands

Correspondence concerning this article should be addressed to Gaëtan Mertens, Department of Clinical and Health Psychology, Heidelberglaan 1, room H1.29, Utrecht University, 3584CS Utrecht, the Netherlands.

E-mail: g.mertens@uu.nl

Tel: +31 30 253 75 53

Abstract

Previous research showed that the repeated approaching of one stimulus and avoiding of another stimulus typically leads to more positive evaluations of the former stimuli. In the current study, we examined whether approach and avoidance training (AAT) effects on evaluations of neutral stimuli can be modulated by introducing a regularity between the approach-avoidance actions and a positive or negative (feared) stimulus. In an AAT task, participants repeatedly approached one neutral non-word and avoided another neutral non-word. Half of the participants also approached a negative fear-conditioned stimulus (CS+) and avoided a conditioned safe stimulus (CS-). The other half of the participants avoided the CS+ and approached the CS-. Whereas participants in the avoid CS+ condition exhibited a typical AAT effect, participants in the approach CS+ condition exhibited a reversed AAT effect (i.e., they evaluated the approached neutral non-word as more negative than the avoided non-word). These findings provide evidence for the malleability of the AAT effect when strongly valenced stimuli are approached or avoided. We discuss the practical and theoretical implications of our findings.

Keywords: Approach-Avoidance Training; Implicit; Explicit; Evaluations; Context; Fear

The contextual malleability of approach-avoidance training effects: Approaching or avoiding fear conditioned stimuli modulates effects of approach-avoidance training

People's preferences play a pivotal role in many important choices and aspects of their lives such as addictive behaviors (Tibboel et al., 2015), racial attitudes (McConnell & Leibold, 2001) and consumer choices (Maison, Greenwald, & Bruin, 2004). As such, there is a great interest in psychology and related disciplines for identifying effective ways to change preferences. One procedure that has proven to be successful in changing preferences is approach-avoidance training (AAT). Approach and avoidance responses differ from other responses in that they have direction and thus can be described as moving toward or away from a stimulus (Krieglmeyer, De Houwer, & Deutsch, 2013). Research has shown that stimuli that are repeatedly paired with an approach action tend to be more positively evaluated than stimuli that are repeatedly paired with an avoidance action (Van Dessel, De Houwer, & Gast, 2016). AAT may have interesting applied potential given that it can produce changes in racial prejudice (Kawakami, Phills, Steele, & Dovidio, 2007), alcohol-related behavior (Wiers, Eberl, Rinck, Becker, & Lindenmeyer, 2011) and fear responses (Jones, Vilensky, Vasey, & Fazio, 2013).

Though many studies have established the robustness and generality of AAT effects, there are also a number of studies that failed to find AAT effects (Becker, Jostmann, Wiers, & Holland, 2015; Krypotos, Arnaudova, Effting, Kindt, & Beckers, 2015; Van Dessel, De Houwer, Roets, & Gast, 2016; Vandenbosch & De Houwer, 2011). This is at odds with original explanations of AAT effects that postulated that (1) there is an intrinsic connection between approach-avoidance actions and positive and negative feelings, respectively, and (2) the pairing of approach-avoidance actions with stimuli therefore leads to the automatic transfer of valence

from the actions to the stimuli (Cacioppo, Priester, & Berntson, 1993; Neumann, Förster, & Strack, 2003). Rather, the results indicate that the AAT effect might strongly depend on certain boundary conditions and does not automatically arise when AAT actions and stimuli are paired repeatedly (Vandenbosch & De Houwer, 2011).

One important factor that might determine AAT effects is the contextual implications of the approach and avoid actions. This factor is known to be crucial for evaluative compatibility effects that involve approach and avoid actions. In this type of research, participants are instructed to approach or avoid valenced stimuli. Results typically demonstrate that participants find it easier to approach positive stimuli and to avoid negative stimuli than to approach negative and avoid positive stimuli. However, these evaluative compatibility effects do not rely on intrinsic properties of approach and avoid responses such as the muscles that are involved (e.g., the muscles used for the flexion or extension of the arm). Instead, they depend heavily on the fact that the responses are labelled in an evaluative manner (i.e., the fact that the labels "approach" and "avoid" have positive and negative valence, respectively; see Eder & Rothermund, 2008) or on the (ultimate) distance altering effects of the responses (i.e., the fact that they decrease or increase the distance to a stimulus; see Krieglmeyer, De Houwer, & Deutsch, 2011; Krieglmeyer, Deutsch, De Houwer, & De Raedt, 2010). More generally, responses seem to acquire their valence from the context (Eder, Rothermund, & De Houwer, 2013). Based on these findings, one can predict that also AAT effects depend on the contextual implications of approach and avoid actions. In line with this idea, studies have shown that AAT effects depend on how approach-avoidance actions are framed. Laham, Kashima, Dix, Wheeler and Levis (2014) observed AAT effects on evaluations of unfamiliar objects when movements to

pull and push a lever were framed as valenced actions such as collecting and discarding food, but not when the movements were merely framed as pushing or pulling a lever.

These studies might help us understand why AAT is not always successful for changing preferences (Becker et al., 2015; Krypotos et al., 2015; Van Dessel et al., 2016). For instance, AAT effects can be expected to be small or non-existent when the experimental context does not allow for a clearly positive or negative interpretation of the approach and avoidance actions. One area in which this might be especially important is in the context of fear-evoking stimuli. When people encounter fear-evoking stimuli they may evaluate an approach action as less beneficial (and thus more negative) than an avoidance action. In this context, we expect that approached stimuli are not evaluated more positively than stimuli that are repeatedly avoided, even if the approach and avoidance actions are clearly labelled as such and even if they have distance-regulating effects (see van Uijen, van den Hout, & Engelhard, 2015 for related evidence in the context of spider fear).

Therefore, in the current study we investigated whether performing approach and avoidance actions towards fear-evoking stimuli may moderate the AAT effect for other, neutral, stimuli. Therefore, we included two conditioned stimuli (CSs) in an AAT procedure, one of which was paired with an electrical stimulation (CS+) in a preceding conditioning phase and another one which was not paired with the stimulation (CS-). During the AAT that followed, half of the participants performed approach movements in response to one neutral stimulus and the CS + and performed avoidance movements in response to another neutral stimulus and the CS-. The other half of the participants experienced opposite contingencies between the two CSs and the approach-avoidance actions. They performed avoidance movements in response to the CS+

(and a neutral stimulus) and approach movements in response to the CS- (and the other neutral stimulus).

This experimental set-up thus linked approach actions with a fear-evoking stimulus and a neutral stimulus whereas avoidance actions were linked with a safe stimulus and another neutral stimulus (or vice versa). At least two accounts predict that this set-up of stimulus-action contingencies in the AAT phase would change the AAT effect. First, according to an operant evaluative conditioning account (Gast & Rothermund, 2011a, 2011b) it can be predicted that repeatedly approaching a CS+ and avoiding a CS- will make these actions more negative and positive, respectively, due to repeated pairing of these actions with the negative CS+ and positive CS-. This altered valence of the AA actions might modulate the AAT effects for the neutral words such that participants' preference for the approached word over the avoided word is reduced when participants consistently approach the CS+ and avoid the CS-. Second, an intersecting regularities account (Hughes, De Houwer, & Perugini, 2016) also predicts that the AAT effect can be moderated by this contextual manipulation. Previous research demonstrated that an intersection between (1) a regularity involving an action and a valenced stimulus and (2) a regularity involving the same action and a neutral stimulus can allow for a transfer of valence from the valenced stimulus to the neutral stimulus (Hughes et al., 2016). The same logic may apply in the current experimental set-up: Approaching a negative CS+ and a neutral stimulus and avoiding a positive CS- and a second neutral stimulus results in a transfer of the evaluative properties of the CS+ and CS- to the neutral stimuli. However, effects of intersecting regularities have so far only been demonstrated for neutral actions (i.e., a left or right key press), but not for actions that are considered to be valenced such as approach and avoidance actions. Thus, from the perspective of the research on intersecting regularities, it would be interesting to demonstrate

that regularities involving approach and avoidance actions can allow for a transfer of stimulus properties. More generally, our study examines whether a contextual factor (i.e., the inclusion of strongly valenced stimuli in the AAT phase) may impact AAT effects. If we find such a moderation this might shed new light on the inconsistent findings with regard to AAT effects for stimuli with a strong a priori (negative) valence: The contingency of the approach-avoidance actions with these valenced stimuli may change the valence-generating effect of approach and avoidance actions as the result of operant evaluative conditioning or intersecting regularities.

Method

Participants

Seventy-nine native Dutch-speaking undergraduates (58 women) participated in exchange for a monetary reward of 10 euro. This sample size was determined in order to have sufficient statistical power to detect AAT effects in each of the two groups (power > .80 to detect an effect size of $d = 0.40$). All participants had normal or corrected-to-normal vision and were naive with respect to the purpose of the experiment. Participants were randomly assigned to one of the two conditions in the experiment (i.e., approach CS+ group and avoid CS+ group). In line with standard procedures (Spruyt, De Houwer, & Hermans, 2009), we excluded the data from four participants whose error rate in the evaluative priming task was more than 2.5 standard deviations above the population mean (population mean = 5.93%, SD = 10.63%).

Material

Four non-words were used as evaluation stimuli, namely ‘UPUSU’, ‘GIHOJ’, ‘AFUBO, and ‘HUZON’. It was randomized whether these words served as CS+, CS-, neutral word 1 (NW1) or neutral word 2 (NW2). The experiment was programmed and presented using the INQUISIT Millisecond Software package (Inquisit 3.0, 2011) on a PC with a 19-inch monitor

(120 Hz refresh rate), a keyboard, and a joystick (Wingman Attack 2) attached to it. The electric shock was generated by a constant current stimulator (DS7A, Digitimer, Hertfordshire, UK) and applied to the participants through two standard Ag/AgCl electrodes attached to the ankle of the left leg.

Procedure

Electric shock work-up procedure. The intensity of the electric shock was determined individually for each participant to be 'highly unpleasant but not painful' using a shock work-up procedure. For reasons of brevity we refer readers to Mertens and De Houwer (2016) for the details regarding this procedure.

Fear conditioning phase. After participants had given informed consent and went through the electric shock work-up procedure, they were seated in front of a computer screen on which instructions for the conditioning procedure appeared. These instructions specified that in the first part of the experiment participants would see two words appear on the screen, one word would always be followed by the electric shock (CS+) and the other word would never be followed by the electric shock (CS-). Furthermore, participants were told they could sometimes avoid the shock by moving the joystick away from the screen in the event that a white dot appeared underneath the non-word that functioned as CS+.

The fear conditioning phase consisted of 16 trials during which the CS+ and CS- were each presented on eight separate occasions. Stimuli were presented in the center of the computer screen for four seconds. Trial order was semi-randomized, limiting the number of consecutive CS+ or CS- trials to maximally four. The intertrial interval (ITI) was randomly determined to be either three, four or five seconds. On half of the CS+ trials, a white dot appeared in the lower half of the computer screen two seconds after CS+ onset and stayed on the screen until the end of the

trial (in which case participants received an electric shock) or until participants pulled the joystick away from the screen (in which case participants did not receive an electric shock). On the other half of the CS+ trials, a white dot did not appear and participants could not avoid receiving the electric shock. At the end of the conditioning phase, participants were asked whether they had noticed a relationship between the words on the screen and the presence of the electric shock by selecting either ‘yes’, ‘no’ or ‘unsure’. Furthermore, they were asked to indicate which word was paired with the electric shock, by selecting one of the four non-words.

AAT phase. Participants then received instructions for the AAT task which specified that they would now make specific movements each time they would see a particular word. Half of the participants were instructed to make an approach response to the CS+ and to a first neutral word (NW1) by moving the joystick towards the screen and to make an avoidance response to the CS- and to a second neutral word (NW2) by moving the joystick away from the screen (CS+ approach group). The other half of the participants were instructed to avoid the CS+ and NW1 and to approach the CS- and NW2 (CS+ avoid group). Participants were asked to remember this information well as they would need this information to complete the task successfully. In the subsequent AAT phase, participants saw the different words (CS+, CS-, NW1 and NW2) presented in the center of the screen in a random order for a total of 96 trials. Words remained on the screen until participants made an approach or avoidance action with the joystick. When a mistake was made a red ‘X’ was presented in the middle of the screen for 500 ms. The ITI was 200 ms.

Evaluation measurement. After the AAT phase, participants performed an evaluative priming task in which they categorized target words as either "negative" or "positive" using the ‘A’ and ‘P’ keys of an AZERTY computer keyboard, respectively. Participants were instructed

to perform this categorization task as quickly as possible, while making as few mistakes as possible. Participants were further told that they would see non-words presented before the valenced words and that they could look at these words, but that their task was simply to respond on the basis of the valence of the positive or negative word. As was the case in comparable 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 post-prime pause of 50 ms, and the target word in white font for 1500 ms or until participants had given a response. Error feedback was only provided during a first practice session of 10 trials with a neutral prime ('$ùµ=:#'), but was not present in the actual measurement phase. The ITI was set to vary randomly between 500 ms and 1500 ms. Participants completed 160 trials separated into two blocks of 80 trials, each containing 10 trials with each of the four non-words as prime and a positive or negative word as target presented in random order.

Finally, after the priming task, we collected explicit ratings of the four non-words. First, participants were asked to indicate how positive or negative they felt about each of the non-words by moving a slider with the computer mouse over a scale that went from 0 (very negative) over 50 (neutral) to 100 (very positive). Next, participants were asked to rate how fearful they felt when seeing the different non-words by moving the slider over a scale that went from 0 (not fearful) over 50 (neutral) to 100 (very fearful). No response deadlines were imposed for providing these explicit ratings.

Results

Neutral words

Evaluative priming task. Trials with an incorrect response were dropped (4.8 %) as well as any trials in which reaction times (RTs) were at least 2.5 standard deviations removed from an

individual's mean (2.8 %) (Spruyt et al., 2009). In line with previous studies investigating AAT effects (e.g., Van Dessel, De Houwer, & Gast, 2016), analyses were performed with item-based linear mixed effects (lme) models as implemented in R package lme-4 (Bates, Maechler, Bolker, & Walker, 2014). To perform the analysis on evaluative priming task reaction times (RTs) we defined a model with the grouping variables Participant and Target Word as random factors. The random effect of Non-Word was not included in the model because including this factor did not significantly improve model fit, $p > .99$.

We tested a model that contained Prime Word (approached word, avoided word), Target Type (positive, negative) and Condition (approach CS+, avoid CS+) as fixed factors. We observed a main effect of Target Type, $\chi 2(1) = 8.33$, $p = .004$, indicating that participants were faster to respond to positive target words (M = 548 ms, SD = 138 ms) than to negative target words (M = 572 ms, SD = 125 ms). More importantly, this main effect was qualified by an interaction effect of Prime Word and Target Type, $\chi 2(1) = 13.74$, $p < .001$. RTs on trials with a positive target and approached word (M = 543 ms, SD = 125 ms) were faster than RTs on trials with a positive target and avoided word (M = 552 ms, SD = 151 ms), $\chi 2(1) = 4.91$, $p = .027$, 95 % confidence interval (CI) = [-17.40, -1.07], whereas RTs on trials with a negative target were slower when the prime was an approached word (M = 578 ms, SD = 130 ms; avoided word: M = 566 ms, SD = 120 ms), $\chi 2(1) = 8.97$, $p = .003$, 95% CI = [3.83, 18.30]. The 3-way interaction effect was not significant, $\chi 2(1) = 1.41$, $p = .24$. However, given our a priori predictions, we examined the AAT effects in each of the two conditions. The analyses revealed a significant interaction effect of Prime Word and Target Type for participants in the avoid CS+ condition, $\chi 2(1) = 14.31$, $p < .001$, but not for participants in the approach CS+ condition, $\chi 2(1) = 2.66$, $p = .10$. Hence, we found evidence for a typical AAT effect in the avoid CS+ condition, but not in

the approach CS+ condition. We observed no other main or interaction effects, χ2s < 1.03, *p*s > .31.

Evaluative ratings. We tested a model that contained Rated Word (approached word, avoided word) and Condition (approach CS+, avoid CS+) as fixed factors. No significant main effects of Rated Word or Condition were observed, χ2s < 1.44, *p*s > .23. Importantly, we did observe a significant interaction effect of Rated Word and Condition, $\chi2(1) = 17.81$, $p < .001$. Participants in the avoid CS+ condition, preferred the approached word ($M = 59.54$, $SD = 16.29$) over the avoided word ($M = 44.30$, $SD = 17.25$), $\chi2(1) = 15.27$, $p < .001$, 95% CI = [7.47, 23.00]. Conversely, participants in the approach CS+ condition exhibited a reversed AAT effect and evaluated the approached word ($M = 44.24$, $SD = 17.75$) less positively than the avoided word ($M = 52.84$, $SD = 17.83$), $\chi2(1) = 4.44$, $p = .035$, 95% CI = [-16.70, -0.47].

Fear ratings. The lme model on fear rating scores also revealed no significant main effects of Rated Word or Condition, χ2s < 1.87, *p*s > .17, though we did observe a significant interaction effect, $\chi2(1) = 21.60$, $p < .001$. Participants in the avoid CS+ condition indicated less fear for the approached word ($M = 17.03$, $SD = 19.20$) than for the avoided word ($M = 33.51$, $SD = 28.45$), $\chi2(1) = 16.07$, $p < .001$, 95% CI = [-24.80, -8.15]. Participants in the approach CS+ condition, however, exhibited more fear for the approached word ($M = 34.97$, $SD = 24.16$) than for the avoided word ($M = 25.97$, $SD = 21.25$), $\chi2(1) = 6.11$, $p = .013$, 95% CI = [1.62, 16.40].

Conditioned words

Evaluative priming task. We tested a model that contained Prime Word (CS+, CS-), Target Type (positive, negative) and Condition (approach CS+, avoid CS+) as fixed factors. We observed a main effect of Target Type, $\chi2(1) = 25.39$, $p < .001$, indicating that participants were faster to respond to positive target words ($M = 544$ ms, $SD = 122$ ms) than to negative target

words (*M* = 575 ms, *SD* = 133 ms). We also observed a main effect of Prime Word, $\chi 2(1) = 5.02$, $p = .025$. RTs on trials with the CS- prime (*M* = 557 ms, *SD* = 123 ms) were faster than RTs on trials with the CS+ prime (*M* = 562 ms, *SD* = 132 ms). We observed no other main or interaction effects, $\chi 2s < 1.28$, $ps > .25$.

Evaluative ratings. We defined a model with the grouping variables Participant and Non-Word as random factors and Rated Word (CS+, CS-) and Condition (approach CS+, avoid CS+) as fixed factors. We observed a significant main effect of Rated Word, $\chi 2(1) = 36.53$, $p < .001$, and a marginally significant main effect of Condition, $\chi 2(1) = 2.80$, $p = .094$. These main effects were qualified by a significant interaction effect, $\chi 2(1) = 8.08$, $p = .004$. Participants in the avoid CS+ condition exhibited a greater preference for the CS- (*M* = 57.57, *SD* = 17.04) over the CS+ (*M* = 28.49, *SD* = 19.19), $\chi 2(1) = 47.51$, $p < .001$, 95% CI = [20.70, 37.50], than participants in the approach CS+ condition (CS-: *M* = 53.74, *SD* = 19.53; CS+: *M* = 43.26, *SD* = 23.70), $\chi 2(1) = 4.42$, $p = .036$, 95% CI = [0.55, 20.40][1].

Fear ratings. We observed a significant main effect of Rated Word, $\chi 2(1) = 72.91$, $p < .001$. Participants exhibited less fear for the CS- (*M* = 22.27, *SD* = 21.17) than for the CS+ (*M* =

[1] This pattern of results seems to indicate a standard AAT effect for the CSs: the preference for CS- over CS+ was less pronounced in the condition in which participants performed approach actions towards the CS+ and avoid actions towards the CS-. However, an alternative explanation could be that the avoidance action in the avoid CS+ group acted as a reminder cue of avoiding the shock in the presence of the CS+ in the acquisition phase, which therefore resulted in a greater negative evaluation of the CS+ in this group. Currently, we cannot distinguish between these two competing explanations. Importantly however, neither interpretation can account for the reversed AAT effect observed for the neutral words. That is, even if the approach action counteracted the negative evaluation of the CS+, we still observed a diametrically opposed effect of approaching a neutral word depending on whether the CS+ or the CS- was also approached (i.e., a mean evaluative rating for this neutral word of 44.24 (SD = 17.75) and 59.54 (SD = 16.29), respectively). Conversely, even if the avoidance action acted as a reminder cue of avoiding the shock in the presence of the CS+ during the fear conditioning phase, it did not produce a larger negative evaluation of the avoided neutral word in the avoid CS+ condition (M = 44.30, SD = 17.25) compared to the approached word in the approach CS+ condition (M = 44.24, SD = 17.75). The same argument applies for the fear ratings.

56.21, *SD* = 27.22). However, there was no main or interaction effect with Condition, χ2s < 1.24, *p*s > .26.

Discussion

In the current study, we investigated the contextual malleability of the AAT effect. During the AAT phase, we included a non-word that had been paired with an electric shock (i.e., a CS+) and a non-word that had not been paired with an electric shock (i.e., a CS-) as well as two neutral non-words. Half of the participants approached the CS+ and one neutral non-word and avoided the CS- and another neutral non-word. The other half of the participants avoided the CS+ and one neutral non-word and approached the CS- and the other neutral non-word. In line with research on the effects of operant evaluative conditioning and intersecting regularities, but in contrast to the idea that approach and avoidance actions have a context-independent valence, we expected that the AAT effect would depend on the pairing of the approach or avoidance action with CS+/CS- on the one hand and with the neutral stimulus on the other hand.

In line with these predictions, we observed that the AAT effect, as indexed by valence and fear ratings, was modulated by the action that was paired with the CS+ and CS- in the AAT task. Participants who performed avoidance actions in response to the CS+ and approach actions in response to the CS- exhibited a typical AAT effect whereas participants who performed approach actions in response to the CS+ and avoidance actions in response to the CS- exhibited a *reversed* AAT effect. Specifically, the latter group of participants reported less liking and more fear for the approached neutral non-word than for the avoided neutral non-word. This pattern, however, was not observed on implicit evaluations as measured with the evaluative priming task. Such differences between different measures should be interpreted with caution as different measures may be affected by error variance to a different degree and therefore may produce

different results (Shanks & Berry, 2012). Indeed, evaluative priming procedures typically produce relatively small effect sizes and scores that are relatively low in reliability (Wittenbrink, 2007), indicating that it is more affected by error variance than other (self-report) measures. Note also that, although the interaction was not significant, we nevertheless observed that participants in the approach CS+ group did not exhibit a typical AAT effect on implicit evaluations, whereas participants in the avoid CS+ group did exhibit this effect. This pattern of results is suggestive evidence that our manipulations may have indeed reduced the AAT effect for the neutral words in the evaluative priming task for participants in the approach CS+ group. However, our analysis probably lacked sufficient statistical power to detect this subtle difference between the approach CS+ and avoid CS+ groups for the EP task because of the smaller and less reliable effects in this task.

Our study extends previous studies in that the valence-generating effects of approach and avoidance actions were manipulated in an indirect manner (i.e., by introducing regularities between the actions and positive or negative stimuli) rather than through direct labelling or contextual framing of these actions (Laham et al., 2014). Our results confirm that contextual manipulations can change the evaluative effects of approach and avoidance actions that are clearly labeled as such and that have clear distance regulating properties. These results are also in line with earlier research on operant evaluative conditioning and intersecting regularities. According to the operant evaluative conditioning account, repeatedly approaching or avoiding a CS+ changes the evaluative connotation of these actions and consequently alters the evaluative consequences of pairing those actions with stimuli. According to the intersecting regularities account, the valence of the stimuli changes because participants execute the same action towards a valenced CS and a neutral word. As a result, they treat the neutral word as equivalent to the

valenced CS, which includes a change in the evaluation of the initially neutral word (see Hughes et al., 2016, for more details). Note that, contrary to the operant evaluative conditioning account, the intersecting regularities account does not necessarily imply that the valence of the approach and avoid actions changes as a result of their relation to the CSs. Future research could thus disentangle these two accounts by examining whether our contextual manipulation changes the valence of the approach and avoid responses, for instance by using stimulus-response compatibility tasks (Eder et al., 2013).

Regardless of the processes that mediate our effects, the fact that we observed a contextual modulation of AAT effects sheds new light on prior AAT research. As we mentioned earlier, mixed results have been obtained when AAT was used to change the valence of stimuli with a pre-existing valence (Becker et al., 2015; Kawakami et al., 2007; Van Dessel et al., 2016; van Uijen et al., 2015; Wiers et al., 2011). It is not clear from these studies why AAT was sometimes ineffective in changing stimulus evaluations. Some have argued that pre-existing stimulus valence is more resistant to the effects of AAT because changing preferences is more difficult than establishing novel preferences (Woud, Becker, Lange, & Rinck, 2013). Others have argued that AAT can only change evaluations of valenced stimuli when action and stimulus are compatible to the extent that they have the same valence or activate the same motivational system of approach or avoidance (Centerbar & Clore, 2006). That is, according to this motivational congruence account, the effect of approach and avoidance actions on stimulus evaluation depends on the congruency between the motivational orientation evoked by valenced stimuli and the motivational orientation of the approach or avoidance action. Because the experience of motivational congruency is pleasant, both approaching a positive stimulus and avoiding a negative stimulus will lead to more positive evaluations of the stimuli. The results of

our study provide evidence for an alternative explanation why effects are not always found when (valenced) stimuli are repeatedly approached or avoided: Approaching and avoiding valenced stimuli can change the valence-generating effects of approach-avoidance actions. For instance, AAT effects may be hampered when the approach action is linked to negative stimuli (e.g., fear-evoking stimuli: Krypotos et al., 2015; van Uijen et al., 2015). AAT effects might also be hampered when other contextual features link the approach action to negative events. For instance, some AAT studies disambiguate the approach action by including a zoom effect (i.e., an effect where a stimulus becomes larger after performing the action). This might change the impact of the approach action because people often consider such a zoom effect to be unpleasant (Hsee, Tu, Lu, & Ruan, 2014). The presence of a zoom effect might thus explain why in some studies AAT effects were absent (e.g., Becker et al., 2015) or even reversed (e.g., Vandenbosch & De Houwer, 2011). Though future studies are necessary to directly test these ideas, our results suggest that AAT might work better if valence of the actions is disambiguated (i.e., by creating a context where clearly positive stimuli also have to be approached and clearly negative stimuli also have to be avoided). Our data also suggest that it might not be sufficient to disambiguate actions by providing instructions that clearly label the actions as approach and avoidance because we obtained modulated AAT effects despite the fact that our actions were clearly labelled. Hence, our study highlights new possibilities to improve current usage of AAT to change stimulus evaluations.

In summary, we examined the contextual malleability of AAT effects by including both highly valenced and neutral stimuli in an AAT phase. Our results indicate that the intersection of approach-avoidance actions with valenced stimuli modulated the AAT effect for the neutral stimuli: A standard AAT effect for the neutral stimuli was found when participants approached a

positive stimulus and avoided a negative stimulus. However, a reversed AAT effect for the neutral words was found when participants approached a negative stimulus and avoided a positive stimulus. Our results might help us to understand why AAT is not always successful in changing stimulus evaluations of (valenced) stimuli.

Acknowledgements

The research reported in this paper was funded by the Interuniversity Attraction Poles Program initiated by the Belgian Science Policy Office (IUAPVII/33) and by Ghent University Methusalem Grant (BOF09/01M00209) awarded to Jan De Houwer. Pieter Van Dessel is supported by a Postdoctoral fellowship of the Scientific Research Foundation, Flanders (FWO-Vlaanderen).

References

Bates, D., Maechler, M., Bolker, B., & Walker, S. (2014). lme4: linear mixed-effects models using S4 classes. R package version 1.1-7. *R*. https://doi.org/http://CRAN.R-project.org/package=lme4

Becker, D., Jostmann, N. B., Wiers, R. W., & Holland, R. W. (2015). Approach avoidance training in the eating domain: Testing the effectiveness across three single session studies. *Appetite*, *85*(April 2016), 58–65. https://doi.org/10.1016/j.appet.2014.11.017

Cacioppo, J. T., Priester, J. R., & Berntson, G. G. (1993). Rudimentary determinants of attitudes: II. Arm flexion and extension have differential effects on attitudes. *Journal of Personality and Social Psychology*, *65*(1), 5–17. https://doi.org/10.1037/0022-3514.65.1.5

Centerbar, D. B., & Clore, G. L. (2006). Do Approach-Avoidance Actions Create Attitudes? *Psychological Science*, *17*(1), 22–29. https://doi.org/10.1111/j.1467-9280.2005.01660.x

Eder, A. B., & Rothermund, K. (2008). When do motor behaviors (mis)match affective stimuli? An evaluative coding view of approach and avoidance reactions. *Journal of Experimental Psychology: General*, *137*(2), 262–281. https://doi.org/10.1037/0096-3445.137.2.262

Eder, A. B., Rothermund, K., & De Houwer, J. (2013). Affective Compatibility between Stimuli and Response Goals: A Primer for a New Implicit Measure of Attitudes. *PloS One*, *8*(11), e79210. https://doi.org/10.1371/journal.pone.0079210

Gast, A., & Rothermund, K. (2011a). I like it because I said that I like it: Evaluative conditioning effects can be based on stimulus-response learning. *Journal of Experimental Psychology: Animal Behavior Processes*, *37*(4), 466–476. https://doi.org/10.1037/a0023077

Gast, A., & Rothermund, K. (2011b). What you see is what will change: Evaluative conditioning effects depend on a focus on valence. *Cognition & Emotion*, *25*(1), 89–110. https://doi.org/10.1080/02699931003696380

Hsee, C. K., Tu, Y., Lu, Z. Y., & Ruan, B. (2014). Approach aversion: Negative hedonic reactions toward approaching stimuli. *Journal of Personality and Social Psychology*, *106*(5), 699–712. https://doi.org/10.1037/a0036332

Hughes, S., De Houwer, J., & Perugini, M. (2016). Expanding the boundaries of evaluative learning research: How intersecting regularities shape our likes and dislikes. *Journal of Experimental Psychology: General*, *145*(6), 731–754. https://doi.org/10.1037/xge0000100

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*(5), 989–996. https://doi.org/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 attitudes and interracial interactions through approach behaviors. *Journal of Personality and Social Psychology*, *92*(6), 957–971. https://doi.org/10.1037/0022-3514.92.6.957

Krieglmeyer, R., De Houwer, J., & Deutsch, R. (2011). How farsighted are behavioral tendencies of approach and avoidance? The effect of stimulus valence on immediate vs. ultimate distance change. *Journal of Experimental Social Psychology*, *47*(3), 622–627. https://doi.org/10.1016/j.jesp.2010.12.021

Krieglmeyer, R., De Houwer, J., & Deutsch, R. (2013). On the Nature of Automatically Triggered Approach–Avoidance Behavior. *Emotion Review*, *5*(3), 280–284.

https://doi.org/10.1177/1754073913477501

Krieglmeyer, R., Deutsch, R., De Houwer, J., & De Raedt, R. (2010). Being Moved: Valence Activates Approach-Avoidance Behavior Independently of Evaluation and Approach-Avoidance Intentions. *Psychological Science*, *21*(4), 607–613. https://doi.org/10.1177/0956797610365131

Krypotos, A.-M., Arnaudova, I., Effting, M., Kindt, M., & Beckers, T. (2015). Effects of Approach-Avoidance Training on the Extinction and Return of Fear Responses. *PLOS ONE*, *10*(7), e0131581. https://doi.org/10.1371/journal.pone.0131581

Laham, S. M., Kashima, Y., Dix, J., Wheeler, M., & Levis, B. (2014). Elaborated contextual framing is necessary for action-based attitude acquisition. *Cognition and Emotion*, *28*(6), 1119–1126. https://doi.org/10.1080/02699931.2013.867833

Maison, D., Greenwald, A. G., & Bruin, R. H. (2004). Predictive Validity of the Implicit Association Test in Studies of Brands, Consumer Attitudes, and Behavior. *Journal of Consumer Psychology*, *14*(4), 405–415. https://doi.org/10.1207/s15327663jcp1404_9

McConnell, A. R., & Leibold, J. M. (2001). Relations among the Implicit Association Test, Discriminatory Behavior, and Explicit Measures of Racial Attitudes. *Journal of Experimental Social Psychology*, *37*(5), 435–442. https://doi.org/10.1006/jesp.2000.1470

Mertens, G., & De Houwer, J. (2016). Potentiation of the startle reflex is in line with contingency reversal instructions rather than the conditioning history. *Biological Psychology*, *113*, 91–99. https://doi.org/10.1016/j.biopsycho.2015.11.014

Neumann, R., Förster, J., & Strack, F. (2003). Motor compatibility: The bidirectional link between behavior and evaluation. In *The psychology of evaluation: Affective processes in*

cognition and emotion (pp. 371–391).

Shanks, D. R., & Berry, C. J. (2012). Are there multiple memory systems? Tests of models of implicit and explicit memory. *The Quarterly Journal of Experimental Psychology*, *65*(8), 1449–1474. https://doi.org/10.1080/17470218.2012.691887

Spruyt, A., De Houwer, J., & Hermans, D. (2009). Modulation of automatic semantic priming by feature-specific attention allocation. *Journal of Memory and Language*, *61*(1), 37–54. https://doi.org/10.1016/j.jml.2009.03.004

Spruyt, A., De Houwer, J., Hermans, D., & Eelen, P. (2007). Affective Priming of Nonaffective Semantic Categorization Responses. *Experimental Psychology*, *54*(1), 44–53. https://doi.org/10.1027/1618-3169.54.1.44

Tibboel, H., De Houwer, J., Spruyt, A., Brevers, D., Roy, E., & Noël, X. (2015). Heavy social drinkers score higher on implicit wanting and liking for alcohol than alcohol-dependent patients and light social drinkers. *Journal of Behavior Therapy and Experimental Psychiatry*, *48*, 185–191. https://doi.org/10.1016/j.jbtep.2015.04.003

Van Dessel, P., De Houwer, J., & Gast, A. (2016). Approach-Avoidance Training Effects Are Moderated by Awareness of Stimulus-Action Contingencies. *Personality and Social Psychology Bulletin*, *42*(1), 81–93. https://doi.org/10.1177/0146167215615335

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*(1), e1–e15. https://doi.org/10.1037/pspa0000039

van Uijen, S., van den Hout, M., & Engelhard, I. (2015). Active Approach Does not Add to the Effects of in Vivo Exposure. *Journal of Experimental Psychopathology*, *6*(1), 112–125.

https://doi.org/10.5127/jep.042014

Vandenbosch, K., & De Houwer, J. (2011). Failures to induce implicit evaluations by means of approach-avoid training. *Cognition & Emotion*, *25*(7), 1311–30; discussion 1331–8. https://doi.org/10.1080/02699931.2011.596819

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*(4), 490–497. https://doi.org/10.1177/0956797611400615

Wittenbrink, B. (2007). Measuring attitudes through priming. In B. Wittenbrink & N. Schwarz (Eds.), *Implicit measures of attitudes: Progress and controversies* (pp. 17–58). New York: Guilford Press.

Woud, M. L., Becker, E. S., Lange, W.-G., & Rinck, M. (2013). Effects of approach-avoidance training on implicit and explicit evaluations of neutral, angry, and smiling face stimuli. *Psychological Reports*, *113*(1), 199–216. https://doi.org/10.2466/21.07.PR0.113x10z1