Instruction-Based Approach-Avoidance Effects:

Changing Stimulus Evaluation via the Mere Instruction to Approach or Avoid Stimuli

Pieter Van Dessel[1]

Jan De Houwer[1]

Anne Gast[2]

Colin Tucker Smith[3]

[1]Ghent University, Belgium

[2]University of Cologne, Germany

[3]University of Florida, US

Word count: 4.759

Author note: Correspondence regarding 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.

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 BOF09/01M00209 of Ghent University and by the Interuniversity Attraction Poles Program initiated by the Belgian Science Policy Office (IUAPVII/33).

Abstract

Prior research suggests that repeatedly approaching or avoiding a certain stimulus changes the liking of this stimulus. We investigated whether these effects of approach and avoidance training occur also when participants do not perform these actions but are merely instructed about the stimulus–action contingencies. Stimulus evaluations were registered using both implicit (Implicit Association Test and evaluative priming) and explicit measures (valence ratings). Instruction-based approach-avoidance effects were observed for relatively neutral fictitious social groups (i.e., Niffites and Luupites), but not for clearly valenced well-known social groups (i.e., Blacks and Whites). We conclude that instructions to approach or avoid stimuli can provide sufficient bases for establishing both implicit and explicit evaluations of novel stimuli and discuss several possible reasons for why similar instruction-based approach-avoidance effects were not found for valenced well-known stimuli.

Keywords: approach, avoidance, training, instructions, evaluations, implicit attitudes, IAT

Instruction-Based Approach-Avoidance Effects:

Changing Stimulus Evaluation via the Mere Instruction to Approach or Avoid Stimuli

In recent years, it has been argued that there is a bi-directional link between attitudes and approach-avoidance motor actions (Neumann, Förster, & Strack, 2003). On the one hand, attitudes are thought to determine the speed with which people perform approach and avoidance motor actions (Solarz, 1960; Chen & Bargh, 1999). On the other hand, the execution of approach and avoidance actions during stimulus processing is said to influence attitude formation and change (Cacioppo, Priester, & Berntson, 1993). In this paper, we extend research on attitude formation via Approach and Avoidance (AA) training by exploring the possibility that instructions about AA training can have effects without the actual execution of these AA actions.

A number of studies have provided evidence that AA training influences not only explicit (non-automatic) evaluations of stimuli but also implicit (i.e., automatic) evaluations of *novel* stimuli such as unknown persons or fictitious social groups (e.g., Woud, Maas, Becker, & Rinck, 2013; Laham et al., in press). Additionally, Kawakami, Phills, Steele, and Dovidio (2007) observed effects of AA training on implicit evaluations of *well-known* social groups. In a series of studies, they found significant reductions in White people's implicit preference for faces of White people over Black people after they had responded with approach actions to photos of Black faces and with avoidance actions to photos of White faces. In line with these results, typical AA training effects have been reported in studies with other well-known stimuli, such as pictures of familiar alcoholic drinks (Wiers, Eberl, Rinck, Becker, & Lindenmeyer, 2011), insects and spiders (Jones, Vilensky, Vasey, & Fazio, 2013), or contamination-related objects (Amir, Kuckertz, & Najmi, 2013). Not all attempts to find effects of AA training, however, have been

successful (e.g., Vandenbosch & De Houwer, 2011), suggesting that there are as yet undiscovered boundary conditions (Laham et al., in press; Vandenbosch & De Houwer, 2011).

At a mental process level, AA training effects are typically interpreted within the framework of embodied cognition. From this perspective, mental representations are assumed to be grounded in modality specific systems of perception and motor action (Niedenthal, Barsalou, Winkielman, Krauth-Gruber, & Ric, 2005). AA processes are given a special status as they are considered essential for successful adaptation to the environment (Elliot, 2006). Embodiment theories assume that, as a result of this evolutionary benefit, evaluative processing is closely tied to representations of AA behavior. More specifically, they postulate that motivational systems of AA mediate the relation between AA behavior and stimulus evaluations (Cacioppo et al., 1993). Motivational systems of AA are activated automatically during the processing of positive or negative stimuli, thereby triggering AA actions (Chen & Bargh, 1999). In turn, because AA actions are wired into these motivational systems, performing AA actions also leads to the activation of these motivational systems, which can bias the automatic evaluative processing of stimuli (Neumann & Strack, 2000). Most important for the purposes of our paper, approaching or avoiding a stimulus is assumed to have long term effects on the evaluation of that stimulus via the formation of associations in memory (Strack & Deutsch, 2004). Each time that the stimulus is approached or avoided, the corresponding stimulus representation and motivational representation are both activated, thereby gradually strengthening the association between those representations. Consequently, AA training effects are assumed to necessitate a large number of trials in which the AA behavior is performed in response to the stimulus (Woud et al., 2013; Phills, Kawakami, Tabi, Nadolny, & Inzlicht, 2011).

There are, however, reasons to believe that the standard embodiment theory of AA training is incomplete at best. First, it has been argued that AA behavior is not simply hard-wired into motivational systems. Instead, the motivational implication of AA responses seems to depend on how these responses are coded cognitively (e.g., pushing a lever can be coded as moving toward the stimulus or as pushing the stimulus away; Eder & Rothermund, 2008). Even the mere planning or anticipation of the AA response might result in the activation of motivational representations (e.g., Eder & Klauer, 2009; Hommel, 2004). Second, contrary to the standard view that associations are formed in a slow, gradual manner, some have argued that associations in memory can emerge very quickly, even as the result of mere instructions or propositional reasoning (e.g., Fazio, 2007, p. 609; Field, 2006, pp. 867-868). Likewise, recent non-associative accounts of learning allow for learning via the rapid formation of propositions via instructions or inferences (De Houwer, 2009; Mitchell, De Houwer, & Lovibond, 2009). Once acquired, these propositions might even be activated automatically and hence underlie not only explicit but also implicit evaluations (De Houwer, 2014).

Based on these theoretical considerations, we put forward the hypothesis that a stimulus does not actually have to be physically approached or avoided in order for AA training effects to arise. Instead, the mere instruction to approach or avoid a stimulus might suffice to produce changes in the (implicit) evaluation of that stimulus. Although we are the first to examine AA training via instructions, it has already been demonstrated that mere instructions about future events can influence both implicit and explicit evaluations. For instance, in studies on evaluative conditioning via instructions, De Houwer (2006; Gast & De Houwer, 2013) told participants that they would see trials on which a first neutral stimulus is paired with positive pictures and trials on which a second neutral stimulus is paired with negative pictures. Despite the fact that the

participants never actually saw the stimulus pairings, the instructions did result in a preference for the first neutral stimulus over the second one, even on measures of implicit evaluation. Of course, these findings do not imply that instructions about stimulus-action contingencies also induce changes in liking, especially because of the special motivational significance of actually performing AA responses. Nevertheless, if mere instructions about stimulus-stimulus relations can produce changes in liking, than it is at least plausible that mere instructions about stimulus-action relations also produce changes in liking.

In our studies, we therefore adapted the procedure of De Houwer (2006) in such a way that participants received instructions about a later phase in which they would be asked to approach or avoid stimuli. Although the main aim of our work was to examine whether AA instruction can influence implicit and explicit evaluations, we already looked at a first potential boundary condition of these effects, being the type of attitude object. More specifically, we investigated effects both on relatively neutral, fictitious groups (i.e., Niffites and Luupites) and on clearly valenced, well-known social groups (i.e., Whites and Blacks). Previous studies suggest that instructions might be more effective in changing the implicit evaluations of novel, affectively neutral attitude objects than in altering the existing evaluations of known, affectively laden attitude objects. For instance, Gregg, Seibt, and Banaji (2006) observed that implicit evaluations of novel social groups could be induced quite easily on the basis of instructions about the behavior and traits of those groups, but could not be undone by giving additional instructions about those groups. Although Gregg and colleagues did not manipulate directly whether the attitude objects were novel or affect-laden, their results are in line with the common sense idea that instructions might not be powerful enough to change existing (implicit) evaluations of well-known attitude objects. This might hold also for AA instructions. [1]

Main study

Experiments 1 and 2 involved a large number of participants who were assigned to either a condition with neutral, novel social groups or with valenced, well-known social groups as attitude objects. In line with existing AA training research we used Black and White people as well-known social groups (see Kawakami et al., 2007). Following Gregg et al. (2006), Niffites and Luupites were used as relatively neutral, novel social groups. A priori power analyses indicated that, to detect a small effect of type of attitude object (i.e., effect size $d = 0.20$; see Cohen, 1992) with sufficient power (power > .75) approximately 270 participants needed to be included in each between-subjects condition. We were able to recruit this large number of participants by implementing our study on the internet. This also allowed us to subdivide the sample based on different kinds of criteria to gain additional information about the moderators of instruction-based AA effects. To this end, we asked participants to indicate whether they had inferred that the purpose of the experiment was to change their attitudes and to what extent they believed that performing this experiment might have changed their attitudes. This allowed us to investigate whether, in line with AA training studies, effects can be observed even in subgroups of participants who do not infer the purpose of the study, or even believe that approach or avoidance would influence their attitudes. In line with most studies investigating AA training effects (e.g., Kawakami et al., 2007; Phills et al., 2011) Experiment 1 used IAT effects as a measure of implicit evaluations. However, because AA training has not always produced clear effects when other implicit measures were used (e.g., Vandenbosch & De Houwer, 2011) it can be argued that AA training effects on implicit evaluations and instruction-based effects in particular, are due to specific properties of the IAT. In Experiment 2, we therefore investigated

instruction-based AA effects by using a different task to measure implicit evaluations, namely the evaluative priming task (Fazio, Sanbonmatsu, Powell, & Kardes, 1986).

Method

Participants. Participants were visitors of the Project Implicit research website (https://implicit.harvard.edu). Participation was restricted to United States citizens. 949 participants took part in Experiment 1 and 773 participants took part in Experiment 2. Data-exclusion involved removing participants who (a) did not complete all tasks (3.1%; 3.4%), (b) were either African American or of mixed Black-White race (10.4%; 10.1%), or (c) did not correctly answer the memory questions (18.0%; 13.5%) [2]. Additionally, data were discarded following standard procedures of data reduction for Project Implicit IAT scores (2.6%) and evaluative priming scores (4.8%) (see Smith, De Houwer, & Nosek, 2013). The analyses were performed on the data of 625 participants (423 women, mean age = 35, *SD* = 14) in Experiment 1 (i.e., 271 in the Niffites/Luupites condition and 352 in the Whites/Blacks condition) and 533 participants (368 women, mean age = 39, *SD* = 14) in Experiment 2 (i.e., 257 in the Niffites/Luupites condition and 286 in the Whites/Blacks condition).

Procedure. All participants were randomly assigned to the condition with neutral fictitious social groups (i.e., Niffites and Luupites) or with valenced well-known social groups (i.e., Whites and Blacks). In the Niffites/Luupites condition participants were instructed that they would be presented with the names of members of two groups, called Luupites and Niffites. They were told that all the names of Luupites have two consecutive vowels in them and end with "lup". They were then shown two examples of Luupites' names (i.e., Loomalup, Ageelup). Subsequently, participants were told that all the names of Niffites would contain two consecutive consonants and end with "nif". Again, this statement was followed by two Niffites names (i.e., Borrinif, Kennunif). Next, half of the participants were told that they would have to approach

each name of a Luupite and avoid each name of a Niffite. The other participants were given the opposite instruction. In the Whites/Blacks condition, half of the participants were instructed to approach typical names of White people and avoid typical names of Black people. The other participants received reversed instructions (i.e., to approach names of Black people and avoid names of White people). In both conditions, these AA instructions were followed by the information that participants would first complete a reaction time task which would last approximately 10 minutes. They were asked to make sure that they would not forget which action they would later on have to perform in reaction to the different types of names.

After these AA instructions, the implicit evaluation task was administered. In Experiment 1, a standard IAT was performed. We followed the procedure of Smith et al. (2013, Experiment 1), with the only exception that participants in the Niffites/Luupites condition categorized positive words, negative words, five Luupites names (i.e., Meesolup, Naanolup, Omeelup, Wenaalup, Tuuraluup) and five Niffites names (i.e., Cellanif, Eskannif, Lebbunif, Zallunif, Otrannif). In the Whites/Blacks condition the names that were used were five prototypical names of Black men (i.e., Darnell, Leroy, Terrence, Tyrone, Jerome) and five prototypical names of White men (i.e., Alfred, Hank, Edmund, Wilbur, Marty). These names were matched on word familiarity in a US-sample and have been used in previous studies on implicit prejudice (Ottaway, Hayden, & Oakes, 2001). During the critical IAT test blocks participants categorized positive words and names of members of one social group with the same key, and negative words and names of members of the second social group with another key. The order of the test blocks was counterbalanced across participants. In the evaluative priming task that was used in Experiment 2, participants categorized target words as either "Good" or "Bad". Procedural details were identical to Smith & De Houwer (under review), except for the prime stimuli that were presented

before the target words. In the Niffites/Luupites condition primes consisted of the word ‘Niffite’, or the word ‘Luupite’ for participants in the Niffites/Luupites condition. In the Blacks/Whites condition the primes were the five prototypical names of Black men and the five prototypical names of White men that were used in Experiment 1.

After the implicit evaluation task, participants completed an explicit evaluation measure that consisted of four ratings. Participants completed liking ratings and thermometer ratings of self-reported warmth or cold feelings towards Niffites and Luupites or Black and Whites on a 9-point Likert scale (1= not warm/liked at all; 9 = completely warm/liked). In the memory test that followed, participants were asked what action they would have to perform when the name of a Niffite/Luupite or White/Black person would be presented in the next task. Participants chose between the words ‘approach’ or ‘avoid’ for each question. Subsequently, participants answered three additional questions. First, participants indicated whether they thought that the purpose of the experiment was to change their attitude towards the social groups. Then participants indicated when they first started thinking that this was a purpose of the experiment (i.e., when reading the AA instructions, when performing the IAT, or when completing the explicit measure). Finally, they indicated to what extent they believed that performing this experiment could have changed their attitude towards the social groups on a 5-point Likert scale.

Finally, even though performance on this task was irrelevant for our hypotheses, participants performed twenty trials of an AA training task in which they were instructed to act as stated in the instructions they had received at the start of the experiment. During this task participants pushed away names by pressing the up arrow on the keyboard (i.e., avoided) and pulled names towards them by pressing the down arrow on the keyboard (i.e., approached). A

zoom effect enhanced the visual experience of approaching or avoiding. This task was included in order not to deceive participants in the earlier instructions.

Results

In Experiment 1, the IAT-scores were calculated using the D2-algorithm (Greenwald, Nosek, & Banaji, 2003) such that positive scores indicate a preference for Niffites or Whites. Split-half reliability of the IAT score was $r(271) = .63$ for participants in the Niffites/Luupites condition and $r(352) = .53$ for participants in the Whites/Blacks condition.

In Experiment 2, to calculate the evaluative priming score, a first difference score was created for each participant in the Niffites/Luupites condition by subtracting mean latencies for Niffites-positive trials from mean latencies for Niffites-negative trials. A second difference score was created in the same way for Luupites-trials such that, in both cases, higher scores indicate more positive evaluations for the group. Finally, the evaluative priming score was constructed by subtracting the difference score for Luupites-trials from the difference score for Niffites trials such that a positive score indicates a preference for Niffites. For participants in the Whites/Blacks condition the same procedure was used to construct an evaluative priming score that indicates a preference for Whites relative to Blacks. A correlation between the priming scores for the first and second block of 60 trials indicated a split-half reliability of $r(257) = .38$ for participants in the Niffites/Luupites condition and $r(286) = .29$ for participants in the Whites/Blacks condition.

In both experiments, the responses on the explicit measures were collapsed into two scores. The rating scores (i.e., warmth score and liking score) were calculated by subtracting the score rating for Luupites/Blacks from the corresponding score rating for Niffites/Whites. Positive scores indicate a preference for Niffites/Whites.

Performance on implicit and explicit measures in Niffites/Luupites condition. In Experiment 1, analysis of the IAT scores indicated that participants preferred Luupites over Niffites (M = -0.14, SD =0.52), $t(270) = -4.29$, $p < .001$. Crucially, a between-groups t-test revealed a significant effect of the instructions, $t(269) = 7.98$, $p < .001$, $d = 0.97$ (Figure 1). When participants had been instructed to approach Niffites and avoid Luupites, the former was preferred (M = 0.08, SD = 0.51), $t(139) = 1.98$, $p = .050$, and when participants had been instructed to avoid Niffites and approach Luupites, the latter was preferred (M = -0.37, SD = 0.43) , $t(130) = -9.90$, $p < .001$. The explicit liking score revealed a significant preference for Luupites (M = -0.31, SD = 2.34, $t[270] = -2.18$, $p = .03$), whereas no significant difference was observed on the warmth score (M = -0.22, SD = 2.47, $t[270] = -1.47$, $p = .14$). Between-groups t-tests revealed a significant instruction effect both on the warmth score (approach Niffites: M = 0.66, SD = 2.21; approach Luupites: M = -1.16, SD = 2.40), $t(269) = 6.49$, $p < .001$, $d = 0.79$, and the liking score (approach Niffites: M = 0.51, SD = 2.06; approach Luupites: M = -1.19, SD = 2.31), $t(269) = 6.42$, $p < .001$, $d = 0.78$).

In Experiment 2, analysis of the evaluative priming scores in the Niffites/Luupites condition indicated no significant preference for either Niffites or Luupites (M = 3.14, SD =72.61), $t(256) = 0.69$, $p = .49$. The crucial between-subjects t-test did, however, reveal an effect of instructions, $t(255) = 4.26$, $p < .001$, $d = 0.59$ (Figure 2). When participants had been instructed to approach Niffites and avoid Luupites, the former was preferred (M = 22.03, SD = 62.77), $t(126) = 3.96$, $p < .001$, but the latter was preferred when participants had been instructed to avoid Niffites and approach Luupites (M = -15.32, SD = 76.94), $t(129) = -2.27$, $p = .025$. In the Niffites/Luupites condition, both the warmth score and the liking score did not reveal a significant preference for any of the two groups, ts < 0.47, ps > .65. However, both the warmth

score and the liking score indicated a significant instruction effect (warmth score: approach Niffites: M = 1.13, SD = 2.89; approach Luupites: M = -1.19, SD = 2.46, t[255] = 6.94, p < .001, d = 0.86; liking score: approach Niffites: M = 1.01, SD = 2.76; approach Luupites: M = -1.15, SD = 2.53, t[255] = 6.51, p < .001, d = 0.81).

Performance on implicit and explicit measures in Whites/Blacks condition. In Experiment 1, analysis of the IAT performance replicated previous research on implicit prejudice (e.g., Dasgupta, McGhee, Greenwald, & Banaji, 2000), demonstrating that participants displayed a strong implicit preference for Whites (M = 0.42, SD = 0.40) , t(353) = 20.14, p < .001. Crucially, a significant main effect of instructions could not be observed, t(352) = 0.01, p = .99, d < 0.01 (Figure 1). Participants who had been instructed to approach Whites and avoid Blacks did not have a significantly different degree of implicit prejudice (M = 0.42, SD = 0.36) than participants who had been instructed to avoid Whites and approach Blacks (M = 0.42, SD = 0.43). The warmth score also revealed a preference for Whites (warmth score: M = 0.25, SD = 1.51), t(353) = 3.18, p = .002, whereas the liking score did not reveal such a preference (M = 0.08, SD = 1.27), t(353) = 1.13, p = .26. Importantly, a t-test did not reveal a significant main effect of instructions for the liking score or the warmth score, ps > .87, ds < 0.02.

In Experiment 2, analysis of the evaluative priming task indicated a preference for Whites (M = 16.21, SD = 58.57), t(275) = 4.60, p < .001. The between-subjects t-test did not reveal a significant main effect of instructions, t(274) = 0.69, p = .49, d = 0.08 (Figure 2). Participants who had been instructed to approach Whites and avoid Blacks did not have a significantly different degree of implicit prejudice (M = 18.51, SD = 58.83) than participants who had been instructed to avoid Whites and approach Blacks (M = 13.67, SD = 58.39). The warmth score and liking score did not reveal a significant preference for Whites, ts < 0.46, ps > .65. Also, a t-test

did not reveal a significant main effect of instructions for the liking score (approach Whites: M = 0.09, SD = 1.08; approach Blacks: M = -0.06, SD = 1.10) or the warmth score (approach Whites: M = 0.06, SD = 1.19; approach Blacks: M = 0.01, SD = 1.45), $ps > .25$, $ds < 0.14$.

Additional analyses. We conducted a number of additional analyses, some of which are described in more detail in the Supplementary Material that is available online. First, analyses involving the data of both conditions confirmed that the instruction effect was significantly larger in the Niffites/Luupites condition than in the Blacks/Whites condition. Second, correlational analyses indicated that the implicit and explicit evaluation scores were significantly correlated for participants in the Niffites/Luupites condition as well as for participants in the Whites/Blacks condition. Third, analyses including participants' answers on the hypothesis awareness questions indicated no impact of hypothesis awareness. A large effect of AA instructions in the Niffites/Luupites condition was observed even if participants did not think that a purpose of the experiment was to change their attitudes. The results of all these analyses were similar for both Experiment 1 and Experiment 2. We also conducted a combined MANOVA on implicit and explicit evaluations that included the data of both experiments, providing us with sufficient power for detecting small effects (i.e., power = .85 to detect an effect size of d = 0.20). This analysis corroborated the significant instruction effect on implicit measures and explicit measures in the Niffites/Luupites condition, $F(3,576) = 36.93$, $p < .001$, whereas an instruction effect was not observed in the Whites/Blacks condition, $F(3,720) = 0.26$, $p = .85$.

Finally, we performed mediational analyses to investigate whether the effect of instructions on implicit evaluations for participants in the Niffites/Luupites condition was mediated by explicit evaluations, or vice versa. Results indicated that, in Experiment 1, changes in implicit evaluations were partly mediated by corresponding changes in explicit evaluations and

changes in explicit evaluations were partly mediated by corresponding changes in implicit evaluations, Z scores > 3.54, *p*s < .001. However, instruction-based AA effects on implicit and explicit evaluations remained significant after controlling for these mediational influences, Z scores > 3.97, *p*s < .001. Notably, in Experiment 2 we did not observe any mediations, Z scores < 0.81, *p*s > .41.

Discussion

We compared evaluations of stimuli that participants were instructed to either approach or avoid. Our data show that typical AA training effects (i.e., a preference for approached stimuli over avoided stimuli) can be observed even if participants do not have to perform the AA actions. Specifically, when participants were instructed to approach the names of members of fictitious social groups, their evaluations of these social groups were more positive than evaluations of social groups they were instructed to avoid. These findings were observed consistently across a number of experiments regardless of whether evaluations were measured with an explicit self-report measure or when implicit measures were used (IAT and evaluative priming), suggesting that these effects were not due to measurement-related factors or demand compliance. In addition, our data suggest the presence of a boundary condition for effects of AA instructions: We found no evidence that AA instructions changed evaluations of clearly valenced, well-known social groups. In the remainder of this section, we explore the implications of these findings.

The presence of instruction-based AA effects for novel stimuli

First, the fact that AA instructions can influence the (implicit and explicit) evaluation of novel attitude objects has implications for theories of AA training. It challenges a strict embodiment theory of AA training (see introduction) but supports the idea that the link between AA behavior and stimulus evaluation depends on cognitive representations of the action rather

than on actual behavior (e.g., Eder & Rothermund, 2008; Eder & Klauer, 2009). Our results extend earlier research by providing evidence for the possibility that (a) motivational or evaluative representations can be activated by the mere anticipation of an AA response rather than the actual execution of the response and (b) associations involving motivational or evaluative representations can be formed instantly as the result of instructions. As such, our findings put important constraints on any current or future theory of AA training.

Second, the demonstration of instruction-based AA training has implications for theories of attitude formation. Our results confirm that, in line with instruction-based EC effects, implicit and explicit evaluations can result not only from extended training but also from mere instructions about relations in the environment (e.g., Gast & De Houwer, 2012). These results cannot be easily explained by single-process association formation models or dual-process models that assume that (a) associations underlie implicit evaluations and (b) that these associations can only form gradually as the result of repeated experiences (e.g., Baeyens, Eelen, Crombez, & Van den Bergh, 1992; Baeyens, Eelen & Crombez, 1995). Our findings are especially striking given that effects of AA training on implicit evaluations are typically interpreted as stemming from gradual changes in associations (e.g., Phills et al., 2011) that necessitate a substantial amount of training. Our data provide evidence that propositional information (at least partially) influences these effects that are considered prototypical examples of effects that result from automatic processing in an associative system (Strack & Deutsch, 2004).

The absence of instruction-based AA effects for clearly valenced, well-known stimuli.

An important limitation of AA instructions seems to be that changes in evaluations of valenced, well-known social groups cannot be readily induced through this procedure. It might

simply be the case that changes in evaluations of these attitude objects are more difficult to obtain (see also Hofmann et al., 2010), and that AA instructions are simply not potent to produce such changes. In contrast, actual AA practice might result in these changes (e.g., Kawakami et al., 2007; Phills et al., 2011; Jones et al., 2013) because, in addition to propositional knowledge about stimulus-action contingencies, it adds something to the effects (e.g., it gives the new association or proposition more power due to the repeated experience). Alternatively, it can be argued that the observed lack of effects for evaluations of valenced, well-known social groups might result from procedural details, such as the specific evaluation objects that were used in the current study (i.e., Whites and Blacks; also see Footnote 1). Moreover, it remains unclear whether the lack of an effect for these attitude objects is due to the fact that they have a strong pre-existing valence for most participants or to the fact that they are already highly familiar to our participants. These factors could be disentangled in future studies by examining the impact of AA instructions on attitudes towards relatively novel but valenced attitude objects (e.g., unknown words that are said to have a good or bad meaning before AA instructions are presented) or towards well-known but relatively neutral groups (e.g., familiar neutral words).

Concluding remarks

In this study we found evidence that instructions to approach or avoid can influence both implicit and explicit evaluations. These findings provide insight into the mechanisms underlying effects of AA training and open up important new questions about when and how evaluations can be formed and changed by means of instructions and actual AA training. However, explanations for this effect need to take into account that changes in liking for valenced, well-known groups were not easily induced with AA instructions in this study. Future research should investigate effects for evaluations of other types of novel and well-known stimuli and provide a direct

comparison between instruction-based and practice-based AA effects to distinguish the mechanisms that underlie effects of AA training.

References

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

Baeyens, F., Eelen, P., Crombez, G., & Van den Bergh, O. (1992). Human evaluative conditioning: Acquisition trials, presentation schedule, evaluative style and contingency awareness. *Behaviour Research and Therapy, 27*, 279-287.

Baeyens, F., Eelen, P., & Crombez, G. (1995). Pavlovian associations are forever: On classical conditioning and extinction. *Journal of Psychophysiology, 9,* 127-141.doi: 10.1016/0005-7967(92)90136-5

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,* 5-17. doi: 10.1037//0022-3514.65.1.5

Chen, M., & Bargh, J. A. (1999). Consequences of automatic evaluation: Immediate behavioral predispositions to approach or avoid the stimulus. *Personality and Social Psychology Bulletin, 25*, 215-224. doi: 10.1177/0146167299025002007

Cohen, J. (1992). "A power primer". *Psychological Bulletin, 112*, 155–159. doi:10.1037/0033-2909.112.1.155

Dasgupta, N., McGhee, D. E., Greenwald, A. G., & Banaji, M. R. (2000). Automatic preference for white Americans: Eliminating the familiarity explanation. *Journal of Experimental Social Psychology, 36*, 316-328, doi: 10.1006/jesp.1999.1418

De Houwer, J. (2006). Using the Implicit Association Test does not rule out an impact of conscious propositional knowledge on evaluative conditioning. *Learning and Motivation, 37*, 176-187. doi: 10.1016/j.lmot.2005.12.002

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

Eder, A. B., & Klauer, K. C. (2009). A common-coding account of the bi-directional evaluation-behavior link. *Journal of Experimental Psychology: General, 138,* 218-235. doi: 10.1037/a0015220

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*, 262-281. doi: 10.1037/0096-3445.137.2.262

Elliot, A. J. (2006). The hierarchical model of approach-avoidance motivation. *Motivation and Emotion, 30*, 111–116. doi: 10.1007/s11031-006-9028-7

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

Fazio, R. H. (2007). Attitudes as object-evaluation associations of varying strength. *Social Cognition*, 25, 603-637. doi: 10.1521/soco.2007.25.5.603

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

Gast, A., & De Houwer, J. (2012). Evaluative conditioning without directly experienced pairings of the conditioned and the unconditioned stimuli. *Quarterly Journal of Experimental Psychology, 65*, 1657-1674. doi: 10.1080/17470218.2012.665061

Gast, A., & De Houwer, J. (2013). The influence of extinction and counterconditioning

instructions on evaluative conditioning effects. *Learning and motiviation, 44*, 312-325. doi: 10.1016/j.lmot.2013.03.003

Gast, A., De Houwer, J., & De Schryver, M. (2012). Evaluative conditioning can be modulated by memory of the CS-US Pairings at the time of testing. *Learning and Motivation, 43*, 116-126. doi: 10.1016/j.lmot.2012.06.001

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

Hofmann, W., De Houwer, J., Perugini, M., Baeyens, F. & Crombez, G. (2010). Evaluative conditioning in humans: a meta-analysis. *Psychological Bulletin, 136,* 390–421. doi: 10.1037/a0018916

Hommel, B. (2004). Event files: Feature binding in and across perception and action. *Trends in Cognitive Sciences, 8*, 494-500. doi: 10.1016/j.tics.2004.08.007

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

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

Mitchell, C. J., De Houwer, J., & Lovibond, P. F. (2009). The propositional nature of human associative learning, *Behavioral and Brain Sciences, 32,* 183-198. doi: 10.1017/S0140525X09000855

Neumann, R., Förster, J. & Strack, F. (2003). Motor compatibility: The bidirectional link between behavior and evaluation. In J. Musch & K. C. Klauer (Eds.), *The psychology of evaluation: Affective processes in cognition and emotion* (pp. 7-49). Mahwah, NJ: Lawrence Erlbaum.

Neumann, R., & Strack, F. (2000). Approach and avoidance: The influence of proprioceptive and exteroceptive cues on encoding of affective information. *Journal of Personality and Social Psychology, 79,* 39–48. doi: 10.1037/0022-3514.79.1.39

Niedenthal, P.M., Barsalou, L.W., Winkielman, P., Krauth-Gruber, S., & Ric, F. (2005). Embodiment in attitudes, social perception, and emotion. *Personality and Social Psychology Review, 9,* 184-211. doi: 10.1207/s15327957pspr0903_1

Ottaway, S. A., Hayden, D. C., & Oakes, M. A. *(*2001*).* Implicit attitudes and racism: Effects of word familiarity and frequency on the implicit association test. *Social Cognition,* 19, 97-144. doi: 10.1521/soco.19.2.97.20706

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

Smith, C. T., & De Houwer, J. (under review). Hooked on a feeling: Comparing the impact of affective and cognitive persuasive messages on implicit evaluations of smoking.

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

Solarz, A.K. (1960). Latency of instrumental responses as a function of compatibility with the meaning of eliciting verbal signs. *Journal of Experimental Psychology, 59*, 239–45. doi: 10.1037/h0047274

Strack, F., & Deutsch, R. (2004). Reflective and impulsive determinants of social behavior. *Personality and Social Psychology Review, 8*, 220-247. doi: 10.1207/s15327957pspr0803_1

Vandenbosch, K., & De Houwer, J. (2011). Failures to induce implicit evaluations by means of approach-avoid training. *Cognition and Emotion, 25,* 1311-1330. doi: 10.1080/02699931.2011.596819

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

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

Footnotes

1. Three initial lab experiments, with 40 participants each, provided preliminary support for instruction-based AA training effects. The first experiment showed AA instruction effects on implicit evaluations of unfamiliar nonwords (i.e., 'BAYRAM' and 'UDIBNON'), whereas the second experiment found instruction-based effects on implicit and explicit evaluations of fictitious social groups (i.e., Niffites and Luupites). The third experiment showed no effects of AA instructions on evaluations of well-known social groups (i.e., Flemish and Turkish people). A full report of these experiments can be obtained by contacting the first author.
2. In all experiments we observed that, when participants did not correctly remember the instructions, they did not show any effects of AA instructions. This is in line with evidence showing that evaluative conditioning effects are stronger or only existent if participants know which US was paired with which CS (see Gast, De Houwer & De Schryver, 2012). Note that the lack of memory could result from processes involved in the encoding, storage, or recall of the contingencies.

   Importantly, including the data from these participants in the analyses did not result in any shift in significance for the effects of AA instructions on novel or well-known stimuli. Also, we excluded participants who were Black or of mixed White-Black race (in line with Kawakami et al., 2007). Including the data of these participants did not change the conclusions. Including race of the participants as a variable in the ANOVA's did reveal an effect of this variable on implicit and explicit evaluations for participants in the Whites/Blacks condition, Fs > 11, ps < .002, showing that participants of Black or mixed White-Black race had less implicit and explicit preference for White names.

Figure captions

Figure 1. Mean IAT D scores indicating an implicit preference for Niffites names over Luupites names or names of Whites over names of Blacks, respectively, as a function of instructions, for participants in Experiment 1. Error bars represent 95% confidence intervals.

Figure 2. Mean evaluative priming scores indicating an implicit preference for Niffites names over Luupites names or names of Whites over names of Blacks, respectively, as a function of instructions, for participants in Experiment 2. Error bars represent 95% confidence intervals.

Figure 1.

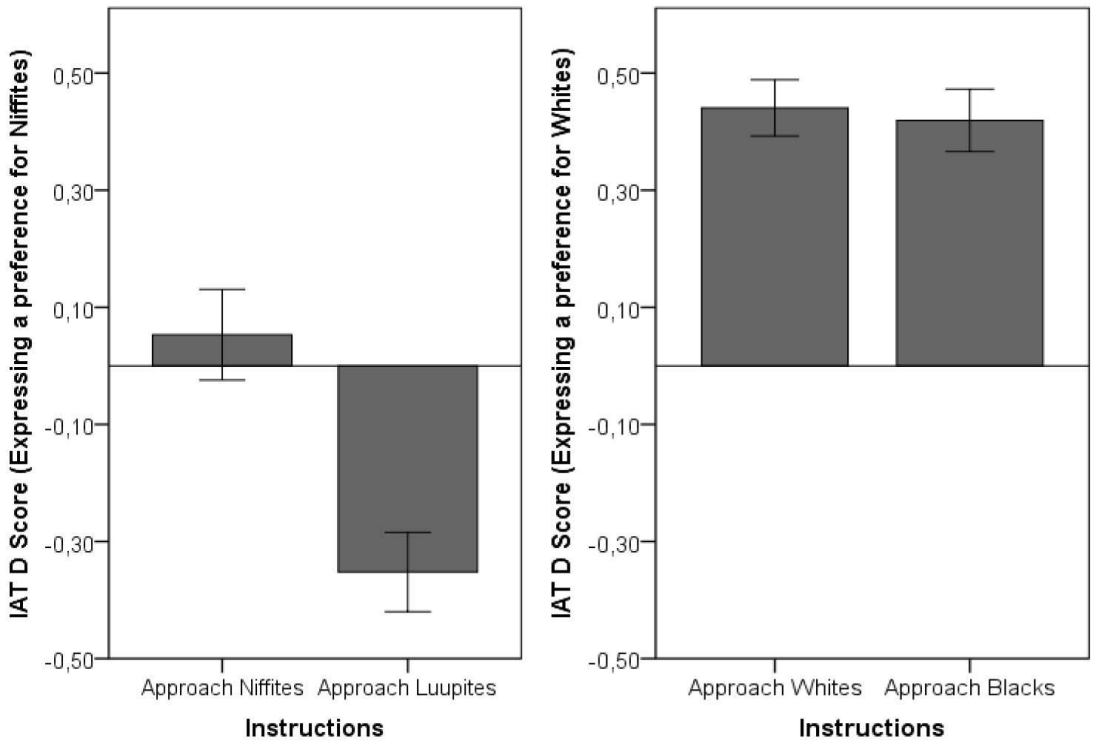
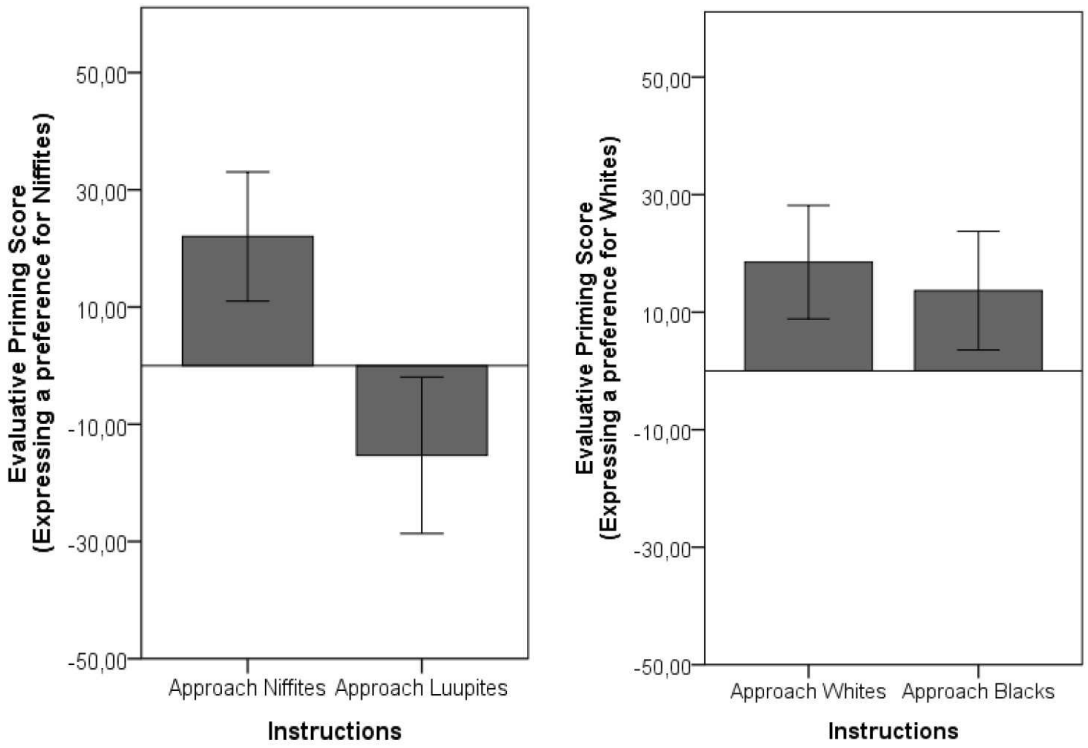

IAT D Score (Expressing a preference for Niffites)
0,50
0,30
0,10
-0,10
-0,30
-0,50
Approach Niffites
Approach Luupites
Instructions
IAT D Score (Expressing a preference for Whites)
0,50
0,30
0,10
-0,10
-0,30
-0,50
Approach Whites
Approach Blacks
Instructions

Figure 2.


Evaluative Priming Score
(Expressing a preference for Niffites)
50,00
30,00
10,00
-10,00
-30,00
-50,00
Approach Niffites
Approach Luupites
Instructions
Evaluative Priming Score
(Expressing a preference for Whites)
50,00
30,00
10,00
-10,00
-30,00
-50,00
Approach Whites
Approach Blacks
Instructions

APPENDIX A: ADDITIONAL ANALYSES

Experiment 1

First, we examined whether the instruction effect that we observed for participants in the Niffites/Luupites condition was significantly larger compared with participants in the Blacks/Whites condition. We performed an Instructions (approach Niffites/Whites vs approach Luupites/Blacks) x Condition (Niffites/Luupites vs Whites/Blacks) Multivariate ANOVA (MANOVA) on IAT and explicit measure scores. In addition to the main effect of Instructions, F (3,619) = 23.39, $p < .001$, and Condition, F (3,619) = 88.87, $p < .001$, we observed a significant interaction effect, F (3,619) = 22.96, $p < .001$. This interaction effect indicated a larger instruction effect for participants in the Niffites/Luupites condition than for participants in the Whites/Blacks condition and was observed on the IAT score and on both explicit measures, ps < .001.

Second, we performed a correlational analysis for participants in both conditions. The scores on the explicit and implicit measures were significantly correlated for participants in the Niffites/Luupites condition (i.e., warmth and liking score: r[271] = .84, $p < .001$; IAT score and warmth score: r[271] = .35, $p < .001$; IAT score and liking score: r[271] = .39, $p < .001$) as well as for participants in the Whites/Blacks condition (i.e., explicit measures: r[354] = .55, $p < .001$; IAT score and warmth score: r[354] = .16, $p = .002$; IAT score and liking score: r[354] = .15, $p = .006$). Additional analyses revealed that correlations between implicit and explicit measures were significantly larger for participants who were instructed to approach Whites than for participants who were instructed to approach Blacks and that correlations were significantly larger for participants who were instructed to approach Luupites than participants who were instructed to approach Niffites. Because evidence suggests that AA training impacts implicit prejudice to a different degree for participants who approach the prejudiced group compared to participants

who approach the group that participants belong to (e.g., Kawakami et al., 2007; Wennekers, 2013), we performed separate correlations for the participants who had been instructed to approach Whites and participants who had been instructed to approach Blacks. This analysis revealed that implicit and explicit prejudice measures were correlated in the approach Whites condition (i.e., IAT score and warmth score: $r[176] = .23$, $p = .002$; IAT score and liking score: $r[176] = .26$, $p = .001$), but not in the approach Blacks condition (i.e., IAT score and warmth score: $r[178] = .11$, $p = .16$; IAT score and liking score: $r[178] = .04$, $p = .56$). We subsequently compared the correlational coefficients for the two groups (Cohen & Cohen, 1983). The difference between the two groups' correlational coefficients of IAT and liking score was statistically significant, $Z = 2.11$, $p = .035$. The difference between the correlational coefficients of IAT and warmth score was not significant, $Z = 1.15$, $p = .25$. Performing separate correlations for participants instructed to approach Niffites and participants instructed to approach Luupites also revealed a different pattern. Implicit and explicit measures were correlated in the approach Luupites condition (i.e., IAT score and warmth score: $r[131] = .36$, $p < .001$; IAT score and liking score: $r[131] = .40$, $p < .001$), but in the approach Niffites condition only IAT score and liking score were significantly correlated ($r[140] = .18$, $p = .037$), whereas IAT score and warmth score were not ($r[140] = .10$, $p = .23$). The difference between the two groups' correlational coefficients was statistically significant (IAT and warmth score: $Z = 2.25$, $p = .024$; IAT and liking score: $Z = 1.97$, $p = .049$).

Third, we compared the instruction effect for participants who indicated that they thought that a purpose of the experiment was to change their attitudes towards the social groups (Niffites/Luupites: 57.6%, Whites/Blacks: 31.9%) and participants who did not indicate this. Additionally, we included the time when participants first believed that a purpose of the

experiment was to change their attitudes (Niffites/Luupites: during the AA instructions: 20.7%, during the IAT: 31.4%, after the IAT or never: 47.9%; Whites/Blacks: during the AA instructions: 16.1%, during the IAT: 12.4%, after the IAT or never: 71.5%) in the analysis as well as participants' ratings about their belief that the experiment could have changed their attitudes (Niffites/Luupites: $M = 2.2$, $SD = 1.1$; Whites/Blacks: $M = 1.6$, $SD = 0.9$). Most importantly, these analyses revealed no impact of the first and second hypothesis awareness factor. For participants in both the Niffites/Luupites and Blacks/Whites conditions, main and interaction effects of the first hypothesis awareness factor (i.e., whether participants thought that a purpose of the experiment was to change their attitudes towards the social groups) and timing factor were not significant, ps $> .44$. Participants in the Niffites/Luupites condition still displayed an instruction effect if they did not think that a purpose of the experiment was to change their attitudes on the IAT score, $t(113) = 5.00$, $p < .001$, $d = 0.97$, and on both explicit measures, ps $< .001$. However, participants' belief ratings (i.e., rating about whether the experiment changed their attitude) were related to the instruction effect, such that the preference for the approached group was larger for participants who had higher belief ratings. This effect was observed only for participants in the Niffites/ Luupites condition, and only on liking ratings, $F(1,268) = 9.42$, $p = .002$, and warmth ratings, $F(1,268) = 6.85$, $p = .009$, but not IAT scores, $F(1,268) = 0.02$, $p = .88$.

Finally, we performed mediational analyses with the lavaan package (version 0.5-16; Rosseel, 2012) to investigate the relationship between implicit and explicit evaluative change. In the Niffites/Luupites condition we observed that changes in implicit evaluations were partly mediated by corresponding changes in explicit evaluations, $Z = 3.55$, $p < .001$. However, the effect of AA instructions on implicit evaluations remained significant after controlling for explicit evaluations, $Z = 6.17$, $p < .001$. Similarly, changes in explicit evaluations were partly mediated

by corresponding changes in implicit evaluations, $Z = 3.81$, $p < .001$, yet the effect of AA instructions on explicit evaluations remained significant after controlling for implicit evaluations, $Z = 3.98$, $p < .001$. In the Whites/Blacks condition no direct or indirect effects of instructions were observed, Zs < 1.06, ps > .27.

Experiment 2

First, we performed an instructions (approach Niffites/Whites vs. approach Luupites/Blacks) x condition (Niffites/Luupites vs. Whites/Blacks) MANOVA on evaluative priming and explicit measure scores. In addition to the main effect of instructions, F (3,527) = 19.26, $p < .001$, we observed a significant interaction effect, F (3,527) = 15.32, $p < .001$. This interaction effect indicated a larger instruction effect for participants in the Niffites/Luupites condition than for participants in the Whites/Blacks condition and was observed on the evaluative priming score and on both explicit measures, ps < .005.

Second, a correlational analysis of the implicit and explicit measures revealed that the scores on the implicit and explicit measures were significantly correlated for participants in the Niffites/Luupites condition (i.e., warmth and liking score: r[257] = .95, $p < .001$; evaluative priming score and warmth score: r[257] = .15, $p = .018$; evaluative priming score and liking score: r[257] = .13, $p = .039$). For participants in the Whites/Blacks condition correlations were significant, except for the correlation between warmth score and the implicit measure score (i.e., explicit measures: r[276] = .57, $p < .001$; evaluative priming score and warmth score: r[276] = .06, $p = .34$; evaluative priming score and liking score: r[276] = .14, $p = .023$). In line with Experiment 1, implicit evaluations were significantly correlated with explicit evaluations in the approach Whites condition (i.e., evaluative priming score and warmth score: r[145] = .22, $p = .007$; evaluative priming score and liking score: r[145] = .19, $p = .021$), but not in the approach

Blacks condition (i.e., evaluative priming score and warmth score: $r[131] = -.09$, $p = .29$; evaluative priming score and liking score: $r[131] = .07$, $p = .40$). The difference between the two groups' correlational coefficients was significant for the IAT and warmth score, $Z = 2.58$, $p = .010$, but not for the IAT and liking score, $Z = 1.00$, $p = .32$. Separate correlations for participants instructed to approach Niffites and participants instructed to approach Luupites did not reveal significant differences between correlations, ps > .55.

Third, we compared the instruction effect for participants who thought that a purpose of the experiment was to change their attitudes towards the social groups (Niffites/Luupites: 46.9%, Whites/Blacks: 20.7%) and participants who did not believe this. Additionally, we included the time when participants first believed that a purpose of the experiment was to change their attitudes (Niffites/Luupites: during the instructions: 20.6%, during the evaluative priming task: 22.2%, after the evaluative priming task or never: 57.2%; Whites/Blacks: during the instructions: 11.6%, during the evaluative priming task: 7.2%, after the evaluative priming task or never: 81.2%) in the analysis as well as participants' ratings about their belief that the experiment could have changed their attitudes (Niffites/Luupites: $M = 2.0$, $SD = 1.1$; Whites/Blacks: $M = 1.5$, $SD = 0.7$). For participants in both the Niffites/Luupites and Blacks/Whites conditions, main and interaction effects including the first two hypothesis awareness factors were not significant, ps > .49. Also, participants in the Niffites/Luupites condition still displayed an instruction effect on the evaluative priming score, $t(133) = 2.62$, $p = .010$, $d = 0.45$, and on both explicit measures, ps < .001, if they had indicated that they did not think that a purpose of the experiment was to change their attitudes. However, participants' belief ratings were related to the instruction-based AA effect for participants in the Niffites/ Luupites condition. We observed an effect of belief on

liking ratings, $F(1,321) = 3.88$, $p = .050$, warmth ratings, $F(1,321) = 7.44$, $p = .007$, and evaluative priming scores, $F(1,254) = 7.99$, $p = .005$.

Finally, mediational analyses indicated that changes in implicit evaluations in the Niffites/Luupites condition were not significantly mediated by corresponding changes in explicit evaluations, $Z = 0.81$, $p = .42$. The effect of AA instructions on implicit evaluations remained significant after controlling for explicit evaluations, $Z = 3.60$, $p < .001$. Similarly, changes in explicit evaluations weren't significantly mediated by corresponding changes in implicit evaluations, $Z = 0.80$, $p = .42$, and the effect of AA instructions on explicit evaluations remained significant after controlling for implicit evaluations, $Z = 6.53$, $p < .001$. In the Whites/Blacks condition no direct or indirect effects of instructions were observed, Zs < 0.67, ps $> .50$.

Cohen, J., & Cohen, P. (1983). *Applied multiple regression/correlation analysis for the behavioral sciences*. Hillsdale, NJ: Erlbaum.

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

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/

Wennekers, A. M. (2013). Embodiment of Prejudice: The Role of the Environment and Bodily States (doctoral dissertation). Retrieved from: http://www.annemariewennekers.com/