Instructing Implicit Processes: When Instructions to Approach or Avoid Influence Implicit but not Explicit Evaluation

Pieter Van Dessel[a]

Jan De Houwer[a]

Anne Gast[b]

Colin Tucker Smith[c]

Maarten De Schryver[a]

[a]Ghent University, Belgium

[b]University of Cologne, Germany

[c]University of Florida, US

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). The research in this paper has been supported by Grant FWO12/ASP/275 of FWO - Vlaanderen.

Abstract

Previous research has shown that linking approach or avoidance actions to novel stimuli through mere instructions causes changes in the implicit evaluation of these stimuli even when the actions are never performed. In two high-powered experiments (total N = 1147), we examined whether effects of approach-avoidance instructions on implicit evaluations are mediated by changes in explicit evaluations. Participants first received information about the evaluative properties of two fictitious social groups (e.g., Niffites are good; Luupites are bad) and then received instructions to approach one group and avoid the other group. We observed an effect of approach-avoidance instructions on implicit but not explicit evaluations of the groups, even when these instructions were incompatible with the previously obtained evaluative information. These results indicate that approach-avoidance instructions allow for unintentional changes in implicit evaluations. We discuss implications for current theories of implicit evaluation.

Keywords: approach, avoidance, training, instructions, implicit attitudes, evaluation

Instructing Implicit Processes: When Instructions to Approach or Avoid Influence Implicit but not Explicit Evaluation

The way in which humans evaluate stimuli as good or bad has long been a central research topic in various sub-disciplines of psychology (Allport, 1935). In contemporary research on evaluations, researchers often contrast deliberate, explicit evaluations and spontaneous, implicit evaluations (see De Houwer, 2009a; Gawronski & Bodenhausen, 2011). Typically, theorists have postulated distinct underlying processes, with explicit evaluations resulting from belief-based processes that involve the validation of propositional information, and implicit evaluations being the product of processes involving the automatic activation of associations in memory (Gawronski & Bodenhausen, 2011).

Given the unique relation between implicit evaluations and behavior (Greenwald, Poehlman, Uhlmann, & Banaji, 2009), it is vital to understand how implicit stimulus evaluations are acquired and can be changed. Because implicit evaluation is traditionally attributed to the activation of associations between representations in memory and because associations are typically thought to develop gradually over many experiences, it is sometimes assumed that implicit evaluations of stimuli arise exclusively as the result of repeated experiences, such as recurrent pairings of physical stimuli (Rydell & McConnell, 2006). Evaluative conditioning (EC) research provides ample evidence that changes in the implicit evaluation of a stimulus (conditioned stimulus; CS) occur when it is paired with a valenced stimulus (unconditioned stimulus; US; for a review see Hofmann, De Houwer, Perugini, Baeyens, & Crombez, 2010). Moreover, research on approach and avoidance (AA) training has suggested that changes in implicit evaluations can be obtained by pairing a stimulus with a valenced action (i.e., approach or avoidance). Typically, the repeated approaching of one stimulus and avoiding of another stimulus leads to more positive implicit

evaluations for the former stimuli (e.g., Kawakami, Phills, Steele, & Dovidio, 2007; Woud, Maas, Becker, & Rinck, 2013; but see Vandenbosch & De Houwer, 2011).

Recent research has, however, shown that implicit evaluations change even when pairings are not experienced directly, but are implied by the verbal presentation of relational information via instructions. For instance, studies on instructed EC have shown that changes in the implicit evaluation of a CS occur when verbal instructions link a CS with a valenced US even when the CS-US pairings are not experienced directly (De Houwer, 2006; Gast & De Houwer, 2012). Similarly, in a recent study we observed typical AA training effects when participants did not actually perform AA actions, but were merely instructed that they would later have to perform these actions (Van Dessel, De Houwer, Gast, & Smith, 2015). That is, participants who received instructions to approach one fictitious social group (e.g., Niffites) and avoid another fictitious social group (e.g., Luupites) showed a preference for the former group both on implicit measures (i.e., the Implicit Association Test, IAT, Greenwald, McGhee, & Schwarz, 1998; and the evaluative priming task, Fazio, Sanbonmatsu, Powell, & Kardes, 1986) and explicit measures of evaluation.

These findings pose a challenge to associative and dual-process models of evaluation which assume that implicit evaluations result from the gradual formation of associations in memory as the result of actual pairings (Smith & DeCoster, 2000; Rydell & McConnell, 2006). In contrast, contemporary dual-process models in which association formation processes can interact with propositional learning processes allow one to explain effects of instructions on implicit evaluations. For instance, the Associative-Propositional Evaluation (APE) model (Gawronski & Bodenhausen, 2006; 2011; 2014) postulates that associations may sometimes arise as the result of the generation and validation of propositions. More specifically, when people

determine in a propositional manner that a stimulus is either positive or negative this may instigate the proactive construction of new associations between representations of the stimulus and representations of positivity or negativity. As a result, any information that allows participants to consciously entertain the proposition that a stimulus is positive or negative may influence implicit evaluations. In line with this idea, changes in implicit evaluations have been observed when participants are provided with information about the valenced properties of a stimulus (Castelli, Zogmaister, Smith & Arcuri, 2004; Gregg, Seibt, & Banaji, 2006; Whitfield & Jordan, 2009; Cone & Ferguson, 2015).

Importantly, these models predict a specific pattern of mediation such that instruction effects on explicit evaluation should mediate effects on implicit evaluation (see Gawronski & Bodenhausen, 2006; Case 4). That is, instructions should first influence whether participants consider a stimulus positive or negative (which is reflected in explicit evaluations) before this may lead to the formation of novel associations (which is reflected in implicit evaluations). Support for this idea was found by Whitfield and Jordan (2009), who observed that receiving information about the behavior of unknown individuals caused changes in implicit evaluations of these individuals that were fully mediated by changes in explicit evaluations.

Contrasting this result, our previous study on AA instruction effects provided evidence that changes in explicit evaluations do not fully mediate effects of AA instructions on implicit evaluations. Statistical mediation analyses indicated that the impact of AA instructions on implicit evaluations was partly mediated by changes in explicit evaluations, but an effect remained after controlling for changes in explicit evaluation (Van Dessel et al., 2015). This is an intriguing finding because it suggests that mere (AA) instructions may sometimes cause unintentional changes in (implicit) stimulus evaluations. Instructions may have a direct effect on

implicit evaluation (i.e., unmediated by changes in explicit evaluation) and may therefore cause changes in implicit evaluations even when participants do not consider the instructions as a valid basis for their (explicit) evaluation.

However, on the basis of the available evidence it is premature to conclude that AA instructions can influence implicit evaluation without any mediation by changes in explicit evaluation. Most importantly, our earlier AA instruction study (Van Dessel et al., 2015) included only statistical analyses of mediation. This measurement-of-mediation approach, however, is ultimately correlational in nature, and is thus problematic for establishing a causal chain (Spencer, Zanna, & Fong, 2005). This is especially the case when examining patterns of mediation between implicit and explicit evaluations. When a manipulation affects both implicit and explicit measures of evaluation, the particular direction of the obtained mediation pattern is strongly influenced by the internal consistency of the employed measure (Gawronski & Bodenhausen, 2011). Moreover, when implicit and explicit evaluations are strongly correlated (as was the case in our previous study), this creates multicollinearity which inflates the standard error of all variables in the mediation model and compromises the estimation of the indirect effect (Alin, 2010). Hence, when examining mediation of implicit and explicit evaluations, it is strongly recommended to supplement statistical mediation analyses with experimental manipulations (De Houwer, Gawronski, & Barnes-Holmes, 2013). This is particularly true if, as in our case, a theoretical debate requires the precise understanding of the causal relation.

In the current studies, we used both a statistical and an experimental approach to test the extent to which the impact of AA instructions on implicit evaluation is mediated by changes in explicit evaluation. We manipulated the proposed mediating variable (i.e., changes in explicit evaluation) by providing participants with ‘trait instructions’ that should prevent an impact of AA

instructions on explicit evaluation. In line with Gregg et al. (2006), we asked participants to imagine that the members of one fictitious social group had very positive traits and the members of another fictitious social group had very negative traits (e.g., Niffites are peaceful, civilized, benevolent, and law-abiding; Luupites are violent, savage, malicious, and lawless). Subsequently, participants received instructions to approach or avoid these social groups. Whereas trait instructions directly specify the evaluative properties of the social group, AA instructions only provide evaluative information if participants infer that the task to approach or avoid members of a group tells something about the evaluative properties of that group. Participants might rely on this inference when they have no other information about the evaluative properties of the group, but even then they will probably be aware that this inference rests on shaky grounds. Prior research indeed suggests that participants are likely to refrain from using information that has a low diagnostic validity (such as AA instructions) when more valid information (such as instructions about evaluative traits) is available (Lynch, 2005; Cone & Ferguson, 2015). For these reasons, we expected that participants who received trait instructions would not take the AA instructions into account when explicitly evaluating the stimuli. We examined whether, under these circumstances, AA instructions would still cause changes in implicit evaluation. That is, we examined whether an AA instruction effect on implicit evaluation would be observed not only in the absence of mediation by changes in explicit evaluation, but even when there is no impact on explicit evaluation. The latter result would not only confirm that AA instructions can have a direct effect on implicit evaluation (because mediation via changes in explicit evaluation can occur only if there are changes in explicit evaluation) but would also support the novel conclusion that this direct effect can arise even when participants do not have the intention to use the AA instructions for evaluating the stimuli.

If we would find that AA instructions influence implicit evaluation in the absence of (mediation by) changes in explicit evaluation, this is bound to have important theoretical implications. First, it would strongly constrain current and future models of (implicit) evaluation. For instance, it would contradict dual-process models that assume that (1) only directly experienced repeated pairings can influence implicit evaluations (Smith & DeCoster, 2000), and it would contradict dual-process models that assume that (2) instructions can only influence implicit evaluation via the mediation of explicit evaluation (Gawronski & Bodenhausen, 2006). To accommodate these findings, dual-process accounts would need to make additional assumptions (e.g., that strong associations can form as the result of a single pairing of a valenced word and a stimulus even in the absence of changes in explicit evaluation).

Finding an impact of AA instructions on implicit evaluation but not on explicit evaluation would also constrain single-process propositional models of evaluation (De Houwer, 2009b; 2014; Mitchell, De Houwer, & Lovibond, 2009). These models postulate that both implicit and explicit evaluations arise exclusively as the result of propositional processes. Prima facie, these models seem less equipped to explain dissociations between implicit and explicit evaluations (e.g., a change in implicit evaluation in the absence of a similar change in explicit evaluation). However, dissociations do not necessarily mean that different processes underlie these different types of evaluation. Rather, dissociations may arise because implicit and explicit measures of evaluation are differentially sensitive to the truth evaluation of propositional information. For example, when participants are told that a specific stimulus has to be approached, they might consider the possibility that this stimulus is good because it has to be approached. If this newly formed proposition can be activated automatically (e.g., in the sense of unintentional) then it may influence implicit evaluation even when the proposition is not considered valid (De Houwer,

2014). In contrast, explicit evaluation may be more contingent on the outcome of truth validation processes.

Second, finding an AA instruction effect on implicit but not explicit evaluation would provide valuable information about the mechanisms that specifically underlie the acquisition of evaluations by means of AA training, that is, by means of the repeated actual performance of approach and avoidance responses. Currently, there is ample evidence that training-based effects involve changes in implicit evaluation that are not mediated by changes in explicit evaluations (Gawronski & LeBel, 2008; Whitfield & Jordan, 2009). These findings have typically been interpreted as evidence that training directly influences processes of association-formation. However, these effects might also reflect the acquisition of propositional information that specifically influences implicit evaluation (e.g., because it allows for the automatic activation of propositions) but not explicit evaluation (e.g., because the information is not considered a valid basis for evaluation). If we observe an impact of AA instructions on implicit but not explicit evaluations, this would support the idea that propositional information can indeed influence implicit evaluations independently of changes in explicit evaluation.

We conducted two experiments to investigate whether the impact of AA instructions on implicit evaluations is mediated by changes in explicit evaluation. In Experiment 1, half of the participants first received instructions that specified the traits of the fictitious social groups. Subsequently, participants received instructions to approach the names of members of one of the social groups and avoid the names of members of the second social group. For half of the participants, these AA instructions were supplemented with actual AA training. We then assessed implicit and explicit evaluations of the social groups. With this design, two tests are possible of the hypothesis that AA instructions allow for a direct influence on implicit evaluation. First, it

can be tested whether AA instructions influence implicit evaluations even after statistically controlling for changes in explicit evaluations. Second, it can be tested if AA instructions influence implicit evaluations even if trait instructions prevent the effects of AA instructions on explicit evaluations. To investigate this issue, we supplemented standard significance tests with Bayesian analyses. Bayesian analyses were performed according to the procedures outlined by Rouder, Speckman, Sun, Morey, and Iverson (2009). These procedures provide a Bayes Factor (BF) that gives an indication of how strongly the data support either the null hypothesis (BF_0; reflecting the absence of a significant effect) or the alternative hypothesis (BF_1; reflecting the presence of a significant effect). BFs smaller than 1, between 1 and 3, between 3 and 10, respectively designate 'no evidence', 'anecdotal evidence', and 'substantial evidence', for either the null or the alternative hypothesis (Jeffreys, 1961). We examined whether, in the presence of trait instructions, AA instructions do not cause changes in explicit evaluation (i.e., analyses provide substantial evidence for the null hypothesis, $BF_0 > 3$) yet still cause changes in implicit evaluation (i.e., analyses provide substantial evidence for the alternative hypothesis, $BF_1 > 3$).

In Experiment 2, all participants received trait instructions and subsequently received either AA instructions that were compatible with these instructions (e.g., instructions to approach Niffites when participants had been asked to imagine that Niffites have positive traits), AA instructions that were incompatible with these instructions (e.g., instructions to avoid Niffites when participants had been asked to imagine that Niffites have positive traits), or no AA instructions. We examined whether changes in implicit evaluations arise in the absence of changes in explicit evaluations when AA instructions are compatible with the trait instructions (and thus strengthen the previously acquired evaluations) or when they are incompatible with the trait instructions (and thus revise the previously acquired evaluations).

Experiment 1

Method

Participants and Design. In Experiment 1, 1121 English-speaking volunteers participated online via the Project Implicit research website (https://implicit.harvard.edu). We employed a 2 (Presence of Trait instructions: yes, no) x 2 (Content of AA Instructions: approach Niffites, approach Luupites) x 2 (Presence of AA Training: yes, no) between-subjects design (Table 1). Data-exclusion involved removing participants who (a) did not fully complete all questions and tasks (257 participants; i.e., 22.9%), or (b) made at least one error on the memory questions that probed memory for valence or AA instructions (189 participants; i.e., 21.9 %).[1] After removing participants based on the previous two criteria, there were no additional participants who needed to be removed because of IAT error rates above 30% across the entire task, or above 40% for any one of the four critical blocks (Smith, De Houwer, & Nosek, 2013). Analyses were performed on the data of 675 participants (440 women, mean age = 32, *SD* = 13).

Procedure. All participants were first familiarized with the two fictitious social groups (i.e., Luupites and Niffites). They read that all the names of Luupites have two consecutive vowels in them and end with "lup". Then they were shown two examples of Luupites' names (i.e., Loomalup, Ageelup). Subsequently, participants read that all the names of Niffites would contain two consecutive consonants and end with "nif." This statement was followed by two Niffites names (i.e., Borrinif, Kennunif).

Half of the participants were then given trait instructions. Similar to Gregg et al. (2006), participants were asked to imagine that these two social groups actually exist and to suppose that

[1] We excluded participants with incorrect memory because we expected that, in line with previous results (Van Dessel et al., 2015), instructions would impact evaluations only if participants correctly remembered these instructions. Importantly, including the data from all participants in the analyses weakened the main effect of Content of AA Instructions and the main effect of Content of Trait instructions on implicit and explicit evaluations, but did not result in any shift in significance for any of the reported effects.

the two groups have very different characters. They were instructed that one group ‘are very good people; they are peaceful, civilized, benevolent, and law-abiding, whereas the other group ‘are very bad people; they are violent, savage, malicious, and lawless.’ Participants were also instructed to suppose that the two groups consistently behave in ways that justified these descriptions when they interact with each other and with other groups. Participants were asked to try and keep clear in their minds which group is which and which group possesses which characteristics as they would later be asked questions about the groups. Half of the participants who received trait instructions learned that Niffites are good and Luupites are bad, whereas the other half received instructions that conveyed the idea that Luupites are good and Niffites are bad.

Subsequently, all participants received AA instructions. 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. These AA instructions were followed by the information that we would later on explain exactly how they would be able to perform these actions, but that for now it was very important to remember which action they would have to perform with each type of name as they would need this information to complete the task successfully.

Following the AA instructions, only half of the participants actually performed the AA training task. This manipulation was orthogonal to (1) the manipulation of the content of trait instructions (Niffites are good and Luupites are bad / Niffites are bad and Luupites are good) and (2) the content of AA instructions (approach Niffites and avoid Luupites / avoid Luupites and approach Niffites). Participants in the AA training condition performed 80 trials of the AA training task in which 4 Niffites’ names (i.e., Cellanif, Eskannif, Lebbunif, Zallunif) and 4 Luupites’ names (i.e., Meesolup, Naanolup, Omeelup, Wenaalup) were each presented ten times.

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; names that were avoided became smaller and moved off into the perceptual distance, whereas names that were approached became larger and appeared to move toward the participant. Only actions that were in line with the AA instructions were registered as correct and resulted in the zoom effect. Incorrect responses were not registered. Participants always had to perform the correct response to proceed to the following trial. The other half of the participants did not receive AA training and they were instructed that they would complete a reaction time task which would last approximately 10 minutes before they could start the AA task.

The reaction time task that followed was an IAT in which participants categorized positive words, negative words, and the names of members of both social groups into one of four categories: positive, negative, Niffites, or Luupites. The IAT followed the procedure described in more detail in Van Dessel et al. (2015). It consisted of three practice blocks and two experimental blocks. Participants began the IAT with 20 practice trials sorting the names of Niffites and Luupites and 20 practice trials sorting positive and negative stimuli. Next, participants completed 56 trials in which stimuli related to Niffites and positive shared a single response key and stimuli related to Luupites and negative shared a single response key (half of the participants completed the IAT in this way, while the other participants began by sorting Luupites and positive with the same key). Participants then practiced sorting Niffites and Luupites names with the response key assignment reversed for 40 trials and finally participants completed a second set of 56 trials in which Niffites shared a response key with negative and Luupites shared a response key with positive (or vice versa). If the participant made an error in categorizing, a red “X” appeared on the screen and the participant corrected their mistake in order to continue. Latencies were recorded

until a correct response was made. IAT-scores were calculated using the D2-algorithm (Greenwald, Nosek, & Banaji, 2003) so that positive scores indicate a preference for Niffites over Luupites. The Spearman-Brown corrected split-half reliability of the IAT score, calculated on the basis of an odd-even split, was $r(675) = .84$.

After the implicit evaluation task, participants rated their liking of each of the social groups by answering two questions: “To what extent do you like Niffites/Luupites?” and “To what extent do you have warm feelings for Niffites and Luupites?”. Participants gave their ratings by selecting an option on a 9-point Likert scale (1= not warm/liked at all; 9 = completely warm/liked). Rating scores (i.e., warmth scores and liking scores) were calculated by subtracting the score rating for Luupites from the corresponding score rating for Niffites so that positive scores indicate a preference for Niffites over Luupites. Because of high internal consistency (Cronbach’s Alpha = .94), we collapsed these score ratings into one explicit evaluation score by averaging the respective scores. This explicit evaluation score correlated significantly with the IAT score, $r(673) = .43$, $p < .001$.

Finally, participants completed two types of manipulation check questions. The first question was completed only by participants who had received trait instructions. Participants were asked to remember which trait instructions were presented at the start of the study and to answer by selecting an option on a dropdown menu with “That Niffites are good and Luupites are bad”, “That Luupites are good and Niffites are bad”, and “I don’t remember” as possible answers. The next two questions asked what action they would have to perform (or had performed in the case of actual training) according to the instructions when the name of a Niffite/Luupite was presented. Participants answered by selecting an option on a dropdown menu with “Approach”, “Avoid” and “I don’t remember” as possible answers.

Results

We split up the analyses for participants who did not receive trait instructions and participants who did receive trait instructions to separately address (1) whether AA instruction and AA training effects on implicit evaluations are fully mediated by changes in explicit evaluations, and (2) whether AA instructions and AA training cause changes in implicit evaluations even when trait instructions are provided.

No trait instructions condition. We performed a 2 (Content of AA Instructions: approach Niffites, approach Luupites) x 2 (Presence of AA Training: yes, no) analysis of variance (ANOVA) on the IAT scores. Because there was an unequal number of participants per condition (no AA training: $N = 96$ for approach Niffites, $N = 97$ for approach Luupites; AA training: $N = 84$ for approach Niffites, $N = 87$ for approach Luupites), we used type III sums of squares in this and all subsequent statistical analyses. The ANOVA revealed a main effect of Content of AA Instructions, $F(1,360) = 135.93$, $p < .001$. Participants who had been instructed to approach Niffites and avoid Luupites ($M = 0.13$, $SD = 0.43$) preferred Niffites more than participants who had been instructed to approach Luupites and avoid Niffites ($M = -0.27$, $SD = 0.55$), $d = 1.22$, 95% confidence interval (CI) [1.00, 1.45], $BF_1 > 10000$. Neither the main effect of Presence of AA Training nor the interaction with Content of AA Instructions was significant, Fs < 0.93, ps > .33.

An ANOVA on the explicit rating scores revealed a similar pattern. We observed only a main effect of Content of AA Instructions, $F(1,360) = 52.49$, $p < .001$, indicating that participants who had been instructed to approach Niffites and avoid Luupites preferred Niffites ($M = 0.52$, $SD = 1.63$) more than participants who had been instructed to avoid Niffites and approach Luupites

(M = -0.99, SD = 2.29), d = 0.76, 95% CI [0.54, 0.97], BF_1 > 10000. We observed no main or interaction effects involving the Presence of AA Training factor, Fs < 1.33, ps > .24.

To investigate the extent to which changes in implicit evaluation are mediated by changes in explicit evaluations we performed mediation analyses with the LAVAAN package (version 0.5-16; Rosseel, 2012). We used the bootstrap method to estimate standard errors for the effects. Results indicated that changes in implicit evaluations were mediated by corresponding changes in explicit evaluations, both when participants received only AA instructions (Z = 2.31, p = .021), and when they received AA instructions and subsequent AA training (Z = 2.03, p = .042). Importantly, however, the AA effect on implicit evaluations remained significant after controlling for changes in explicit evaluations for participants without (Z = 5.65, p < .001) and with actual training (Z = 8.78, p < .001). Regression coefficients of the performed mediation analyses are provided in Appendix.

Trait instructions condition. To examine AA effects in the context of trait instructions we performed a 2 (Content of AA Instructions: approach Niffites, approach Luupites) x 2 (Presence of AA Training: yes, no) x 2 (Content of Trait Instructions: Niffites are good, Luupites are good) ANOVA on the IAT scores of participants who had received trait instructions. We included the Content of Trait Instructions factor to estimate the effect of trait instructions on evaluations and control for the variance attributable to this factor. We observed a main effect of Content of Trait Instructions, F(1,303) = 183.27, p < .001, indicating that participants preferred Niffites more when Niffites were presented as positive and Luupites as negative (M = 0.23, SD = 0.48) than when Niffites were presented as negative and Luupites as positive (M = -0.45, SD = 0.46), d = 1.45, 95% CI [1.20, 1.70], BF_1 > 10000. This analysis also revealed a main effect of Content of AA Instructions, F(1,303) = 36.78, p < .001, but this effect was qualified by an interaction effect

of Content of AA Instructions x Presence of AA Training, $F(1,303) = 5.02$, $p = .026$ (Table 2). Importantly, a significant effect of Content of AA Instructions was observed for participants who had merely received AA instructions, $F(1,150) = 7.44$, $p = .007$, $d = 0.44$, 95% CI [0.12, 0.77], $BF_1 = 5.36$. This effect was larger for participants who had received additional AA training, $F(1,153) = 33.48$, $p < .001$, $d = 0.74$, 95% CI [0.41, 1.06], $BF_1 = 1961.39$. Finally, an interaction effect of Content of AA Instructions and Content of Trait Instructions, $F(1,303) = 13.22$, $p < .001$, indicated that the effect of Content of AA Instructions was stronger when trait instructions conveyed that Niffites are good and Luupites are bad than when trait instructions conveyed the opposite information.[2]

An ANOVA on the explicit rating scores revealed a main effect of Content of Trait Instructions, $F(1,303) = 222.10$, $p < .001$. This effect was qualified by an interaction effect with Presence of AA Training, $F(1,303) = 5.60$, $p = .019$, which indicated that the effect of trait instructions was smaller for participants who received AA training, $d = 1.47$, 95% CI [1.11, 1.83], than for participants who received no AA training, $d = 1.91$, 95% CI [1.53, 2.30].[3] Most importantly, we observed no main effect of Content of AA instructions, $F(1,303) = 0.01$, $p = .90$, $d = 0.05$, 95% CI [-0.18, 0.28]. The BF score provided substantial evidence in favor of *the null hypothesis* ($BF_0 = 7.19$). We also observed no other main or interaction effects, Fs < 2.66, ps > .10.

[2] This finding relates to the observation that, even in the absence of trait instructions, participants preferred Luupites over Niffites, and may indicate that AA effects are reduced if participants have clearly univalent positive or negative implicit evaluations (e.g., because they find Luupites' names more appealing and they learned that Luupites are positive). Please consult Jones, Vilensky, Vasey, and Fazio (2013), and Woud, Becker, Lange, and Rinck (2013) for reasons why stimuli that have a non-ambivalent valence might be less susceptible to AA effects.

[3] One possible explanation for this is that participants who received actual training may have been distracted from the trait instructions (e.g., because there was a longer delay between receiving these instructions and completing the evaluative rating task) and therefore used these trait instructions to a lesser extent for their evaluative ratings. Receiving the trait instructions, however, still discouraged participants from considering the AA information as a valid source of evaluative information.

Mediation analyses showed that changes in implicit evaluations were not significantly mediated by corresponding changes in explicit evaluations, both for participants who received only AA instructions, $Z = 0.70$, $p = .49$, and participants who received AA instructions in addition to AA training, $Z = -0.07$, $p = .95$. The effect of AA instructions on implicit evaluations remained significant after controlling for changes in explicit evaluations (no training: $Z = 2.66$, $p = .008$; training: $Z = 5.71$, $p < .001$).

Discussion

Experiment 1 provided both correlational and experimental evidence that the impact of AA instructions on implicit evaluation is not fully mediated by changes in explicit evaluation. First, correlational analyses show that changes in explicit evaluation only partly mediated the effects of AA instructions on implicit evaluation. That is, AA instructions (and AA training) caused effects on implicit evaluations that remained significant after controlling for the mediating impact of explicit evaluations. This finding corroborates the correlational results of Van Dessel et al. (2015). Second, and most importantly, we found an experimental dissociation on implicit and explicit evaluations with regard to the impact of AA instructions (and AA training). More specifically, when trait instructions were presented, AA instructions and AA training caused changes in implicit but not explicit evaluations. Participants who received information about the evaluative traits of the social groups did not take the AA instructions or training into account when expressing their explicit evaluation, yet still exhibited an implicit preference for the approached group. This resembles previous findings of changes in implicit, but not explicit evaluations as a result of the repeated pairing of stimuli (e.g., Gawronski & LeBel, 2008) and indicates that both AA instructions and AA training can cause changes in implicit evaluation even when participants do not consider this information as a valid source of evaluative

information. Given the well-known limitations of correlational mediation analyses, our experimental results provide important new evidence for the conclusion that AA instructions can influence implicit evaluations directly, that is, without first changing explicit evaluations. These findings contradict the idea that instructions influence implicit evaluations only if these instructions are considered a valid basis for evaluation and, hence, are incorporated in explicit evaluations (Gawronski & Bodenhausen, 2006; Whitfield & Jordan, 2009).

In addition to showing that instructions can influence implicit evaluations even when they are not considered a valid basis for evaluation, the present findings also provide information about another important research question that has informed research on the nature of implicit evaluation. Specifically, they inform us on whether the formation and change of implicit evaluations can occur rapidly. In line with Van Dessel et al. (2015) and other studies (e.g., De Houwer, 2006; Peters & Gawronski, 2011) our findings challenge the widespread assumption (e.g., Rydell & McConnell, 2006) that implicit evaluations are slow to *build*. Additionally, and more importantly, these findings indicate that existing implicit evaluations can also be *altered* rapidly, as the result of AA instructions. When participants' evaluations were biased in favor of one of the two social groups as the result of trait instructions, subsequent AA instructions still caused changes in the implicit evaluation of these groups. This contrasts with previous findings suggesting that, once established, implicit evaluations cannot be easily changed (Gregg et al., 2006; Rydell, McConnell, Strain, Claypool, & Hugenberg, 2007). Recently, however, research has shown that new valenced information about a stimulus can lead to a rapid revision of implicit evaluations, but only when this information is considered highly diagnostic about the evaluative properties of this stimulus (Mann, Cone & Ferguson, 2015; Cone & Ferguson, 2015). Our findings go beyond this previous research by showing that rapid alterations in implicit

evaluations can occur in the absence of changes in explicit evaluations. AA instructions may thus rapidly alter existing implicit evaluations even if these instructions are not considered diagnostic of the valence of the stimulus.

Experiment 1, however, did not include a control condition to estimate effects of trait instructions and AA instructions separately. Hence, although the results of Experiment 1 confirm our main hypothesis that instructions can cause changes in implicit evaluations in the absence of changes in explicit evaluations, they do not demonstrate conclusively that instructions can also counteract existing implicit evaluations directly. For instance, because of a lack of a control condition, it is theoretically possible that our results were due to the fact that compatible AA instructions *strengthened* the previously acquired implicit evaluations rather than that incompatible AA instructions *revised* them. To examine this question and to ascertain that the finding of a direct influence of AA instructions on implicit evaluations in Experiment 1 was not a chance finding, we performed Experiment 2.

Experiment 2

In Experiment 2 we further explored AA instruction effects in the context of trait instructions. The aim of this experiment was two-fold. First, we aimed to replicate the finding that AA instructions cause a direct influence on implicit evaluation in the absence of changes in explicit evaluation. In contrast to Experiment 1, we counterbalanced the order of the IAT and the explicit rating task to exclude the possibility that performing the implicit evaluation task first, changed the effects on explicit evaluations (see Perugini, Richetin & Zogmaister, 2014). Second, we extended the previous findings by addressing whether AA instructions cause changes in implicit evaluation when AA instructions are compatible or incompatible with the trait instructions. To this end, participants were provided with either compatible AA and trait

instructions, incompatible AA and trait instructions or only trait instructions. Including a condition with only trait instructions allowed us to estimate the effect of trait instructions on evaluations (i.e., the preference for the group that is presented as positive) and examine whether compatible or incompatible AA instructions moderate this effect.

Method

Participants. Participants were 823 English-speaking volunteers who participated online via the Project Implicit research website. Data-exclusion involved removing 195 participants who did not complete all tasks (23.7%), and 156 participants who did not correctly answer the memory questions (24.8%), leaving data from 472 participants (307 women, mean age = 38, *SD* = 13). None of the participants had previously participated in Experiment 1.

Procedure. Experiment 2 was identical to Experiment 1 except for the following points. First, participants were randomly assigned to start with the IAT and then perform the explicit rating task or to perform tasks in the opposite order. Second, participants never received actual AA training. Third, all of the participants received trait instructions. Fourth, not all of the participants received AA instructions. Participants were randomly assigned to receive either (1) no AA instructions, (2) instructions to approach Niffites and avoid Luupites, or (3) instructions to approach Luupites and avoid Niffites. Hence, this experiment employed a 2 (Content of Trait Instructions: Niffites are good, Luupites are good) x 3 (AA Instructions: approach Niffites, approach Luupites, no AA instructions) between-subjects design (Table 3).

Split-half reliability of the IAT score was $r(472) = .92$. Internal consistency of the explicit evaluation score was high (Cronbach's Alpha = .96), and this score correlated significantly with the IAT score, $r(470) = .59$, $p < .001$.

Results

A 3 (AA Instructions: Approach Niffites, Approach Luupites, no AA instructions) x 2 (Content of Trait Instructions: Niffites are good, Luupites are good) ANOVA on the IAT scores revealed a main effect of Content of Trait Instructions, $F(1,466) = 377.50$, $p < .001$, indicating that participants preferred Niffites more when Niffites were presented as positive and Luupites as negative ($M = 0.26$, $SD = 0.48$) than when Niffites were presented as negative and Luupites as positive ($M = -0.54$, $SD = 0.39$), $d = 1.82$, 95% CI [1.60, 2.04], $BF_1 > 10000$. Most importantly, we also observed a main effect of AA Instructions, $F(2,466) = 4.59$, $p = .011$ (Table 4). In line with Experiment 1, participants who had been instructed to approach Niffites and avoid Luupites ($M = -0.02$, $SD = 0.58$) preferred Niffites more than participants who had been instructed to approach Luupites and avoid Niffites ($M = -0.27$, $SD = 0.55$), $F(1,309) = 9.24$, $p = .003$, $d = 0.44$, 95% CI [0.21, 0.66], $BF_1 = 131.22$. Compared to participants who had not received AA instructions ($M = -0.08$, $SD = 0.62$), participants who had received instructions to approach Luupites preferred Luupites more, $F(1,313) = 4.98$, $p = .026$, but we observed no significant difference for participants who had received approach Niffites instructions, $F(1,310) = 0.41$, $p = .52$.

To examine whether compatible or incompatible AA instructions cause changes in evaluations we performed planned tests comparing the main effect of Content of Trait Instructions for participants who received no AA instructions, participants who received compatible AA instructions and participants who received incompatible AA instructions. Importantly, the main effect of Content of Trait Instructions was reduced when AA instructions were incompatible with the trait instructions, $d = 1.39$, 95% CI [1.01, 1.78], compared to when no AA instructions were provided, $d = 1.85$, 95% CI [1.47, 2.23], $F(1,291) = 5.24$, $p = .023$, indicating that incompatible AA instructions influenced implicit evaluations. In contrast, the

main effect of Content of Trait Instructions was not significantly different for participants who received compatible AA instructions, $d = 2.15$, 95% CI [1.78, 2.53] compared to participants who received no AA instructions, $F(1,332) = 0.34$, $p = .56$.

The 3 x 2 ANOVA on explicit ratings revealed only the main effect of Content of Trait Instructions, $F(1,466) = 370.73$, $p < .001$, $d = 1.80$, 95% CI [1.58, 2.01], indicating a larger preference for Niffites when they were presented as positive ($M = 2.76$, $SD = 3.37$) than when they were presented as negative ($M = -3.34$, $SD = 3.42$). We did not observe a significant main effect of AA Instructions, $F(1,466) = 0.36$, $p = .70$, nor an interaction effect with Content of Trait Instructions, $F(1,466) = 0.47$, $p = .63$. Also, the main effect of Content of Trait Instructions did not differ significantly between participants who received compatible, incompatible or no AA Instructions, Fs < 0.37, $ps > .54$, BF_0s > 7.00.

AA instructions condition. In line with Experiment 1, mediation analyses on the data of participants who received both AA and trait instructions showed that AA instruction effects on implicit evaluations were not significantly mediated by corresponding changes in explicit evaluations, $Z = 1.87$, $p = .062$. The effect of AA instructions on implicit evaluations remained significant after controlling for explicit evaluations, $Z = 2.92$, $p = .003$.

Discussion

Results from Experiment 2 provide further support for the idea that the impact of AA instructions on implicit evaluations is not fully mediated by changes in explicit evaluations. Replicating the pattern obtained in Experiment 1, participants who received AA instructions exhibited an implicit, but not an explicit preference for the approached group over the avoided group when prior instructions specified the valence of these groups. Mediation analyses indicated

that AA instruction effects on implicit evaluation were not fully mediated by changes in explicit evaluation in the context of trait instructions.

Additionally, results indicated that AA instructions caused changes in implicit evaluation even when the valence implied by the approach or avoidance action was *incompatible* with the evaluative information provided in the trait instructions. This suggests that AA instructions can (partly) undo recently established implicit evaluations, even in the absence of changes in explicit evaluations. This contrasts evidence that implicit evaluations are more difficult to change than explicit evaluations with verbally presented counter-attitudinal information (Gregg et al., 2006). We found no evidence that AA instructions caused changes in implicit evaluations when these instructions were *compatible* with the trait instructions. This is consistent with previous findings that AA training causes changes in implicit evaluations of social groups only when the training is incompatible with participants' evaluations (Kawakami et al., 2007). It suggests that AA effects may be strongly reduced when participants have clearly univalent positive or negative implicit evaluations and corroborates previous evidence that the effectiveness of instructions to approach or avoid a stimulus may critically depend on specific stimulus properties (e.g., whether a stimulus is novel or well-known; see Van Dessel et al., 2015).

General Discussion

In two experiments, we observed that instructions to approach or avoid members of a fictitious group impact implicit evaluations of these groups. Our results indicate that these changes in implicit evaluation are not fully mediated by changes in explicit evaluations. Experiment 1 provided evidence that participants who merely received AA instructions and participants who received additional AA training exhibited a direct effect on implicit evaluations. Moreover, both procedures caused changes in implicit evaluations even when trait instructions

clearly specified the valence of the groups which canceled any AA effect on explicit evaluative ratings. Experiment 2 corroborated that AA instructions influenced implicit, but not explicit evaluations in the context of trait instructions and extended these findings by showing that AA instructions caused changes in implicit evaluations when AA instructions were incompatible with the trait instructions.

These findings have meaningful theoretical and practical implications. We first discuss implications for theories on the mental processes that underlie implicit evaluation. Afterwards, we discuss implications for mental process theories that account for AA instruction and AA training effects. Finally, we discuss practical implications of the present research.

Implications for theories of implicit evaluation

The current experiments provide important information that constrains current and future models of implicit evaluation. First, the observation that AA instructions have a direct influence on implicit evaluation (i.e., independent of changes in explicit evaluation) is difficult to reconcile with associative and dual-process models of evaluation that only allow for evaluative associations to form (1) gradually as the result of many pairings (e.g., Smith & DeCoster, 2000; Rydell & McConnell, 2006) or (2) rapidly when consciously entertaining the proposition that a stimulus is positive or negative (Gawronski & Bodenhausen, 2006). However, dual-process models can accommodate these findings if they allow for the immediate formation of associations even on the basis of information that is not considered to be valid. Also propositional single-process accounts of evaluation can account for our results if they assume that the automatic activation of propositional information underlies implicit evaluation (De Houwer, 2014). More specifically, receiving AA instructions may allow participants to consider the proposition that the approached social group is positive. A dissociation between implicit and explicit evaluation will arise when

this proposition is judged to be invalid (and thus dismissed when making an explicit evaluation) but still automatically retrieved when the social group is implicitly evaluated.

Second, the observation that incompatible AA instructions reduce effects of trait instructions on implicit, but not on explicit evaluations suggests that implicit evaluations can be updated rapidly. It provides direct evidence against the often entertained idea that implicit evaluations are more difficult to change than explicit evaluations via counter-attitudinal information (Gregg et al., 2006; Rydell & McConnell, 2006). Rather, changes in explicit evaluation seem to critically depend on the perceived validity of the obtained evaluative information (Peters & Gawronski, 2011). When information directly contradicts previous valence information, this causes an immediate reversal of participants' explicit liking of the stimulus (Gregg et al., 2006). Because AA instructions do not invalidate the more diagnostic evaluative trait information, they do not influence explicit evaluation when they contradict trait instructions. In contrast, changes in implicit evaluations may arise as the result of any information that links a stimulus with a specific valence, such as information about its relation with another valenced stimulus (see Zanon, De Houwer, Gast, & Smith, 2014) or with a valenced action (Van Dessel et al., 2015). Immediate changes in implicit evaluation may occur, even when participants do not consider the obtained information as valid.

Note that the present findings do not contradict the idea that the impact of counter-attitudinal information strongly depends on the diagnosticity of this information (Cone & Ferguson, 2015). In fact, our data also suggest that AA instructions have a stronger influence on implicit evaluation if they are more diagnostic. This can be inferred from the fact that we observed a bigger AA instruction effect in the absence of trait instructions, that is, when the AA instructions were the most diagnostic piece of information that was available to the participants.

However, our results extend the previous research by showing that changes in implicit evaluations may occur as the result of instructions even when these instructions provide information that is not considered highly diagnostic of the evaluative properties of the stimulus and therefore do not influence explicit evaluations. This effect is automatic in the sense that, in all likelihood, our participants did not intend to use this information for their evaluation.

In sum, the current findings provide important information for theories that explain how implicit evaluations arise and can be changed. Although our results cannot distinguish between the broad class of single-process propositional and the broad class of dual-process models, they do force these models into adopting specific assumptions without which they cannot account for our effects. In general, we believe that it is difficult, if not impossible, to distinguish between broad classes of models like dual-process or single-process models that have such a high degree of flexibility. Therefore, we believe that, in order to further advance research on evaluation, it is necessary to (1) define specific models (e.g., propositional or association-formation models of AA effects) that make testable predictions and (2) perform research to test these predictions. The data produced by such research will allow us to further constrain these models and to have greater confidence in the assumptions that survive this process.

Implications for accounts of AA instruction and AA training effects

First, the current findings indicate that instructions that link a valenced action and a fictitious social group cause unintentional changes in the implicit evaluation of these groups. This extends knowledge about the effects of AA instructions by showing that these effects are not necessarily the result of controlled, non-automatic processes that involve the intentional use of this information for evaluation (e.g., as the result of demand compliance) (Van Dessel et al., 2015).

Second, our results also constrain ideas about the processes that underlie AA training effects. More specifically, they reveal important similarities between the effects of AA training and those of AA instructions. Not only can both interventions lead to changes in implicit evaluations, they both can have direct effects on implicit evaluations, that is, effects that are not mediated by changes in explicit evaluations. Although these similarities do not prove that both types of effects are due to the same mental processes (e.g., the formation and activation of propositions), they are in line with this idea and hence undermine the position that AA training effects can be due only to low level processes such as the gradual, performance-driven formation of associations in memory (e.g., Woud et al., 2013; Phills, Kawakami, Tabi, Nadolny, & Inzlicht, 2011). Future studies are required to establish whether instructions and pairings are also similar regarding other features, for example regarding uncontrollability (see Gawronski, Balas, & Creighton, 2014).

Finally, our findings suggest that actually performing AA behavior may, under certain conditions, add to the effect of AA instructions on implicit evaluations. Experiment 1 included a direct comparison of AA instruction and AA training effects on implicit and explicit evaluations. For participants who did not receive trait instructions, additional AA training did not have an added effect even though we had sufficient statistical power to detect even a small effect (power = .77 to detect an effect size of d = 0.25). In contrast, participants who received trait instructions exhibited a stronger AA effect on implicit evaluation when AA instructions were supplemented with AA training. Whether this added effect of AA training involves the strengthening of the previously obtained knowledge structures (i.e., associations or propositions) or the acquisition of entirely different knowledge structures requires further research.

Practical Implications

AA training is considered an important procedure for the modification of pathological biases in cognitive functioning (see Woud & Becker, 2014). Repeatedly performing AA movements in response to specific stimuli has proven effective in a number of therapeutic contexts such as the treatment of alcohol addiction (Wiers, Eberl, Rinck, Becker, & Lindenmeyer, 2011), social anxiety (Taylor & Amir, 2012), or contamination-related fear (Amir, Kuckertz, & Najmi, 2013). Given the important relation between implicit evaluation and the dysfunctional behavioral responses under investigation (see Houben, Havermans, & Wiers, 2010), it can be argued that changes in implicit evaluation may (partly) underlie therapeutic effects of AA training. Following this reasoning, our current results may indicate that AA instructions could play an important role in these AA training effects. Preliminary evidence supporting this idea was found in a recent study by Wiers et al. (2014) where therapeutic effects of 'avoid alcohol' training at one month follow-up were more robust if participants had received explicit instructions to push alcohol away in addition to the re-training procedure. Future research might consider whether replacing or complementing AA training with AA instructions may improve the therapeutic effectiveness of AA training.

Concluding remarks

In sum, the present results extend past findings that verbal instructions influence implicit evaluation by showing that AA instruction effects on implicit evaluations occur in the absence of mediation by changes in explicit evaluation. These findings provide insight into the mechanisms underlying implicit evaluation and open up important new avenues for changing implicit evaluations.

References

Alin, A. (2010). Multicollinearity. *Wiley Interdisciplinary Reviews: Computational Statistics*, *2*, 370-374. doi: 10.1002/wics.84.

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

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

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

Cone, J., & Ferguson, M. J. (2015). He Did what?: The Role of Diagnosticity in Revising Implicit evaluations. *Journal of Personality and Social Psychology, 108*, 37-57. doi: 10.1037/pspa0000014

De 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. (2009a). How do people evaluate objects? A brief review. *Social and Personality Psychology Compass, 3*, 36-48. doi: 10.1111/j.1751-9004.2008.00162.x

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

De Houwer, J. (2014). A Propositional Model of Implicit Evaluation. *Social and Personality Psychology Compass, 8*, 342-353. doi: 10.1111/spc3.12111

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

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

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

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

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

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

Gawronski, B., & Bodenhausen, G. V. (2014). Implicit and explicit evaluation: A brief review of the associative-propositional evaluation model. *Social and Personality Psychology Compass, 8*, 448-462. doi: 10.1111/spc3.v8.8

Gawronski, B., & LeBel, E. P. (2008). Understanding patterns of attitude change: When implicit measures show change, but explicit measures do not. *Journal of Experimental Social Psychology, 44*, 1355-1361. doi: 10.1016/j.jesp.2008.04.005

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

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

Greenwald, A. G., Poehlman, T. A., Uhlmann, E., & Banaji, M. R. (2009). Understanding and using the Implicit Association Test: III. Meta-analysis of predictive validity. Journal of *Personality and Social Psychology, 97*, 17–41. doi: 10.1037/ a0015575

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

Houben, K., Havermans, R., & Wiers, R. W. (2010). Learning to dislike alcohol: Conditioning negative implicit attitudes towards alcohol and its effect on drinking behavior. *Psychopharmacology, 211,* 79-86. doi: 10.1007/s00213-010-1872-1

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

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

Kawakami, K., Phills, C. E., Steele, J. R., & Dovidio, J. F. (2007). (Close) distance makes the

heart grow fonder: Improving implicit racial evaluations and interracial interactions through approach behaviors. *Journal of Personality and Social Psychology, 92*, 957-971. doi: 10.1037/0022-3514.92.6.957

Lynch, J. G. Jr. (2005). Accessible but Nondiagnostic Memories about Memory and Consumer Choice. In Griffin, A. & Ottnes, C.C. (Eds.). *16th Paul D. Converse Symposium* (88-115). Chicago: American Marketing Association.

Mann, T., Cone, J., & Ferguson, M. J. (2015). Social-psychological evidence for the effective updating of implicit attitudes. *Behavioral and Brain Sciences.* 38: e15. doi: 10.1017/S0140525X14000223.

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

Peters, K. R., Gawronski, B. (2011). Are We Puppets on a String? Comparing the Impact of Contingency and Validity on Implicit and Explicit Evaluations. *Personality and Social Psychology Bullettin, 37,* 557-569. doi: 10.1177/0146167211400423

Perugini, M., Richetin, J., & Zogmaister, C. (2014). Indirect measures as a signal for evaluative change. *Cognition & Emotion, 28*, 208-229. doi: 10.1080/02699931.2013.810145

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

Rosseel, Y. (2012). lavaan: An R Package for Structural Equation Modeling. *Journal of Statistical Software*, 48(2), 1-36. URL: http://www.jstatsoft.org/v48/i02/

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

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

Rydell, R., McConnell, A. R., Strain, L. M., Claypool, H. M., & Hugenberg, K. (2007). Implicit and explicit attitudes respond differently to increasing amounts of counterattitudinal information. European Journal of Social Psychology, 37, 867–878. doi: 10.1002/ejsp.393

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

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

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

Taylor, C. T., & Amir, N. (2012). Modifying automatic approach action tendencies in individuals with elevated social anxiety symptoms. *Behaviour Research and Therapy, 50*, 529-536. doi: 10.1016/j.brat.2012.05.004

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

Van Dessel, P., De Houwer, J., Gast, A., & Smith, C. T. (2015). Instruction-Based Approach–Avoidance Effects: Changing Stimulus Evaluation via the Mere Instruction to Approach or Avoid Stimuli. *Experimental Psychology, 62*, 161-169. doi: 10.1027/1618-3169/a000282

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

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

Wiers, R.W., Houben, K., Fadardi, J.S., van Beek, P., Rhemtulla, M., & Cox, W.M. (2014) Alcohol Cognitive Bias Modification training for problem drinkers over the web. *Addictive Behaviors.* 40, 21-26. doi: 10.1016/j.addbeh.2014.08.010

Woud, M. L., & Becker, E.S. (2014). Editorial for the Special Issue on Cognitive Bias Modification Techniques: An Introduction to a Time Traveller's Tale. *Cognitive Therapy and Research, 38,* 83-88. doi: 10.1007/s10608-014-9605-0

Woud, M. L., Becker, E. S., Lange, W. G., & Rinck, M. (2013). To like or not to like: The effect of approach-avoidance training on neutral, angry and smiling face stimuli. Psychological Reports, 113, 199-216. doi: 10.2466/21.07.PR0.113x10z1

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

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

Table 1.

Experimental Design of Experiment 1.

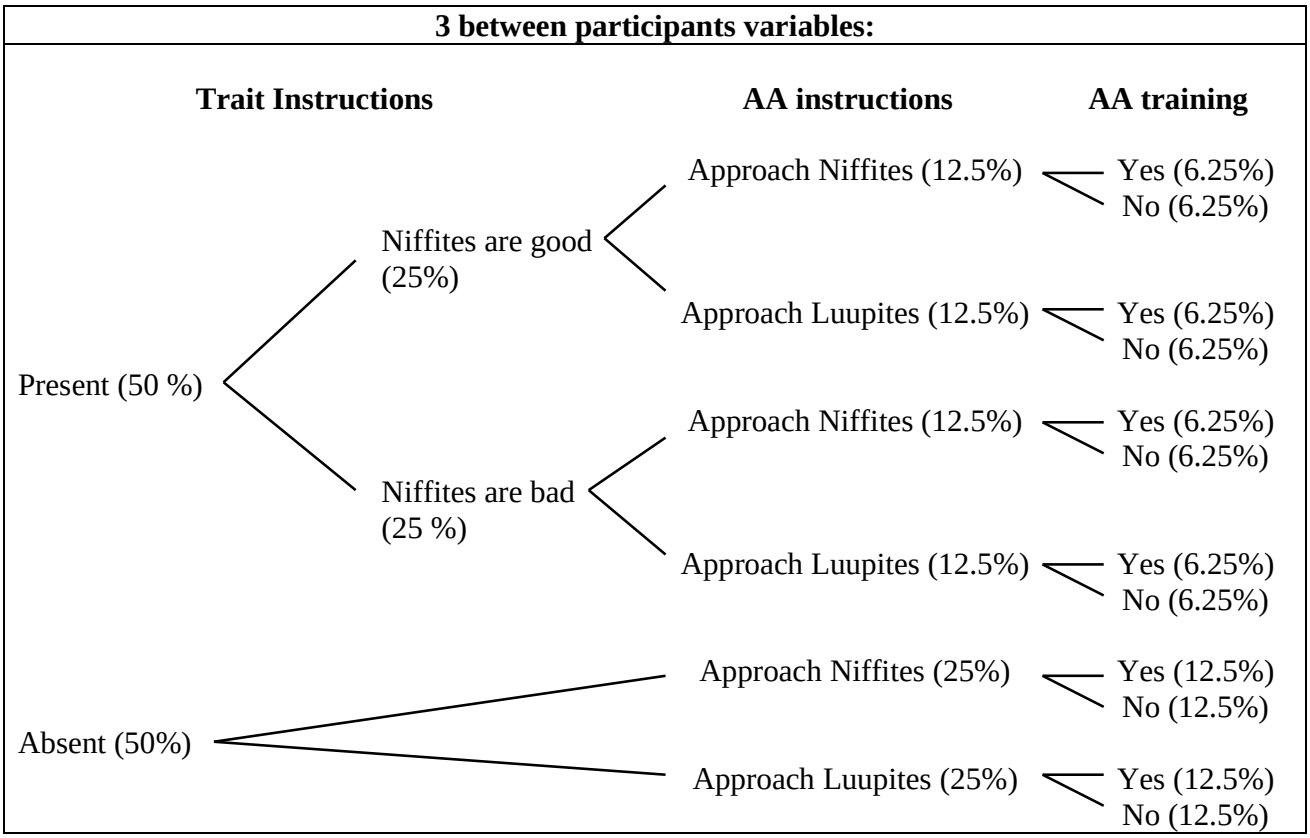

Table 2.

Mean IAT and Explicit Scores in Experiment 1 for participants who received trait instructions as a function of Content of Trait Instructions and Content of AA Instructions.

	Content of Trait Instructions			
	Niffites good and Luupites bad		Niffites bad and Luupites good	
	Approach Niffites	Approach Luupites	Approach Niffites	Approach Luupites
IAT score:				
No AA training	0.42 (0.37)	0.05 (0.47)	-0.40 (0.48)	-0.40 (0.39)
AA training	0.49 (0.37)	-0.09 (0.46)	-0.39 (0.53)	-0.62 (0.39)
Explicit score:				
No AA training	3.20 (3.08)	2.27 (3.49)	-3.99 (3.51)	-3.11 (3.00)
AA training	1.64 (2.89)	1.49 (2.63)	-3.21 (3.51)	-2.79 (3.38)

Note. Standard deviations are in parentheses. Scores reflect a relative preference for Niffites over Luupites.

Table 3.

Experimental Design of Experiment 2.

2 between participants variables:	
Trait Instructions	**AA instructions**
Niffites are good (50 %)	Compatible: Approach Niffites (16.7%) Incompatible: Approach Luupites (16.7%) No AA instructions (16.7%)
Luupites are good (50 %)	Incompatible: Approach Niffites (16.7%) Compatible: Approach Luupites (16.7%) No AA instructions (16.7%)

Table 4.

Mean IAT and Explicit Scores in Experiment 2 as a function of Content of Trait Instructions and AA Instructions.

	Content of Trait Instructions					
	Niffites good and Luupites bad			Niffites bad and Luupites good		
	Approach Niffites	Approach Luupites	No AA instructions	Approach Niffites	Approach Luupites	No AA instructions
IAT score:	0.31 (0.47)	0.15 (0.46)	0.31 (0.50)	-0.46 (0.41)	-0.60 (0.36)	-0.54 (0.40)
Explicit score:	3.07 (3.13)	2.38 (4.02)	2.75 (3.04)	-3.40 (3.80)	-3.34 (3.41)	-3.29 (3.10)

Note. Standard deviations are in parentheses. Scores reflect a relative preference for Niffites over Luupites.