On the Effectiveness of Approach-Avoidance Instructions and Training for Changing Evaluations of Social Groups

Pieter Van Dessel[1]

Jan De Houwer[1]

Anne Gast[2]

Arne Roets[3]

Colin Tucker Smith[4]

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

[2]Social Cognition Center, University of Cologne, Germany

[3]Department of Developmental, Personality and Social Psychology, Ghent University, Belgium

[4]Department of Psychology, University of Florida, US

Author note: Correspondence concerning this article should be addressed to Pieter Van Dessel, Department of Experimental-Clinical and Health Psychology, Ghent University, Dunantlaan 2, 9000 Ghent, Belgium. E-mail: Pieter.VanDessel@UGent.be. The research reported in this paper was funded by the Scientific Research Foundation, Flanders under Grant FWO16/PDO/201 to PVD, by the Deutsche Forschungsgemeinschaft under grant GA 1520/2-1 to AG and by the Ghent University Methusalem Grant BOF09/01M00209 to JDH. The design of all experiments as well as the raw data, experimental and analytic scripts are available on Open Science Framework: https://osf.io/y37te/

Abstract

Prior evidence suggests that White participants who repeatedly approach images of Black people and avoid images of White people can exhibit a reduction in implicit racial bias (Kawakami, Phills, Steele, & Dovidio, 2007). In contrast, a recent study by Van Dessel, De Houwer, Gast, and Smith (2015) showed that mere instructions to perform approach-avoidance training in an upcoming phase produces a similar change in implicit evaluations of unfamiliar but not familiar social groups. We report four experiments that examined the replicability and generalizability of these findings for well-known social groups. Experiment 1 was a replication of the study by Kawakami et al. (2007) in a different domain (i.e., Flemish students' bias towards Turkish people) showing relatively weak evidence for small approach-avoidance training effects on implicit evaluations and explicit liking ratings. Experiment 2 replicated the finding of Van Dessel et al. (2015) that approach-avoidance instructions do not influence implicit evaluations of social out-groups and found no instruction effects even when participants first completed training with non-social stimuli. Experiment 3 established the presence of a small approach-avoidance training effect on implicit (but not explicit evaluations) in a large on-line sample. Experiment 4 directly compared approach-avoidance training and instruction effects, corroborating (1) the effect of training on implicit evaluations which was both small and subject to boundary conditions and (2) the absence of such an effect of instructions. There were again no effects on explicit evaluations. Whereas the current findings provide supportive evidence for training-based approach-avoidance effects (on IAT scores: meta-analytic effect size current experiments: $d = 0.18$, Bayes Factor = 65.22; current and prior experiments: $d = 0.23$, Bayes Factor = 4404.42) and evidence for the absence of instruction-based effects (Bayes Factors < 0.19), they also illustrate that there is still much uncertainty regarding the boundary conditions of these effects and the underlying mental processes.

Keywords: approach-avoidance training, instruction effects, prejudice, implicit bias

On the Effectiveness of Approach-Avoidance Instructions and Training for Changing Evaluations of Social Groups

Although the topic of prejudice has occupied center stage in social psychology for many years, relatively few studies have found strong evidence for effective interventions to reduce prejudice (see Kende, Tropp, & Lantos, 2017; West, Hotchin, & Wood, 2017, for notable exceptions related to intergroup contact: Pettigrew & Tropp, 2006). One of the main areas where prejudice-reducing interventions have been applied often (especially in recent years: see Nelson, Prasad, & Hackman, 2015) and have had some success, is in changing implicit bias or implicit evaluations of well-known social out-groups. We define implicit bias as a behavioral phenomenon (i.e., the more negative automatic evaluation of out-groups compared to in-groups), where "automatic" refers to behavior that occurs under conditions of automaticity (e.g., quickly, with low intention, awareness, or control). This definition deviates from the more traditional perspective that sees implicit bias as a hidden force inside people's minds that causes actions over which the actor may have little to no control (see De Houwer, 2019, for a discussion of the merits of adopting this behavioral perspective). Several interventions such as asking participants to read vivid counter-stereotypical scenarios or repeatedly pairing out-group stimuli with positive words or images have been shown to produce reductions in implicit bias as assessed with measures that probe evaluative responses under automaticity conditions such as the Implicit Association Test (IAT: Greenwald, McGhee, & Schwartz, 1998; see Lai et al., 2014, for an overview). However, the observed changes are often small, unreliable, and fleeting (Jackson, 2018; Lai et al., 2016).

Effects of Approach-avoidance Training and Instructions on Social Group Evaluations

One intervention that has shown promise for the reduction of implicit bias is Approach-Avoidance (AA) Training. In a seminal study, Kawakami, Phills, Steele, and Dovidio (2007) demonstrated that White participants who repeatedly approached images of Black people and avoided images of White people exhibited a reduction in the relatively more negative evaluations of Black people on the IAT, compared to a control group. The authors argued that these effects result from gradual changes in learned associations in memory (e.g., between representations of Black people and negative valence) due to the repeated pairing of valenced approach-avoidance responses and racial stimuli. Phills and colleagues (2011) extended these findings by showing that White participants were faster to categorize Black people and self-related stimuli together in an IAT that used 'Self' and 'Other' (rather than 'Positive' and 'Negative') as attribute categories after training to approach Black people. They concluded that AA training might lead to the automatic formation of associations between representations of social groups (or other targets: Kawakami, Steele, Cifa, Phills, & Dovidio, 2008) and positively valenced representations of the self which influence spontaneous evaluations of these groups.

Recent evidence has, however, challenged the idea that AA training effects are (exclusively) the result of automatic, slow-paced formation of associations on the basis of repeated pairings. First, in contrast with the idea that the association formation that underlies AA training effects is automatic in the sense of unaware (Kawakami et al., 2007), AA training effects strongly depend on participants' awareness of stimulus-action contingencies. For instance, participants who repeatedly approached and avoided novel faces showed a preference for approached over avoided faces only when they were able to report which action they had performed most often in response to the specific faces (Van Dessel, De Houwer, & Gast, 2016). Second, changes in evaluations have been observed when participants do not actually perform AA actions but are merely instructed about stimulus-action contingencies. For instance, participants who learned that they would approach one fictitious social group (e.g., Niffites) and avoid another fictitious social group (e.g., Luupites) in a later phase, exhibited a preference for the former group (Van Dessel, De Houwer, Gast, & Smith, 2015; Van Dessel, De Houwer, Gast, Smith, & De Schryver, 2016). A recent study even found that AA instructions produced larger changes in implicit evaluations of unfamiliar food brands than AA training (Smith, Calanchini, Hughes, Van Dessel, & De Houwer, 2019).

Dissociation between Effects of Approach-avoidance Training and Instructions

Importantly, a recent study found evidence that AA *instructions* produce changes in implicit and explicit evaluations of unfamiliar social groups but not of Black and White people (Van Dessel et al., 2015). In AA *training* research, on the other hand, changes in implicit evaluations of these racial groups were found and have provided the foundations for the widespread adoption of this procedure (Kawakami et al., 2007). Interestingly, the dissociation between AA training and AA instruction effects for well-known social groups contrasts with recent observations that instructions are usually as effective (and sometimes even more effective) as actual experience (in evaluative learning) (Smith et al., 2019). It also contrasts with evaluative conditioning (EC) research that examines changes in evaluations that result from the repeated pairing of stimuli with valenced stimuli (rather than valenced actions) (Hofmann, De Houwer, Perugini, Baeyens, & Crombez, 2010). In this domain, evidence for changes in evaluations based on instructions about (stimulus-stimulus) pairings has also challenged dominant associative explanations (De Houwer, 2006; Gast & De Houwer, 2013) and inspired alternative (inferential) theories (De Houwer, 2018). In contrast to AA research, a recent study established that EC instructions were also effective in reducing implicit bias for existing groups (Kurdi & Banaji, 2017). Following instructions that images of out-group members (i.e., elderly people) would be paired with positive stimuli and images of in-group members would be paired with negative stimuli, participants exhibited a reduction in their relatively more negative implicit evaluations of the former group. This effect of actual instructions was stronger than the effect of pairings. The observation that EC instructions but not AA instructions produce changes in implicit bias for existing groups might indicate that effects of stimulus-stimulus pairings and stimulus-action pairings draw on (qualitatively) different processes. Most importantly, it supports the idea that AA training effects might depend less on propositional processes and more on associative learning processes (Phills et al., 2011).

The observed dissociation – in the context of out-group attitudes – between (the absence of) AA instruction effects in Van Dessel et al. (2015) and (the presence of) AA training effects in Kawakami et al. (2007) could therefore be of great theoretical importance. Specifically, it could indicate that the performance of AA actions sets in motion distinct processes that produce the observed changes in implicit bias based on AA training whereas AA instructions do not. On the one hand, changes in implicit bias might require a gradual re-wiring of robust underlying mental associations on the basis of actual stimulus-action pairings (Kawakami et al., 2007). Whereas instructions could allow for changes in weak associations of unfamiliar groups, only actual training might be strong enough to shift strong associations of well-known groups. Evidence for this idea would be important given the recent problems to find supportive evidence for automatic association formation (Corneille & Stahl, 2018). On the other hand, it is also possible that AA training and AA instructions draw on similar types of processes such as inferential reasoning (see Van Dessel, Hughes, & De Houwer, 2019). The observed dissociation might imply that AA training and instructions lead to distinct inferences and only inferences made on the basis of training might facilitate evaluative inferences that are sufficiently strong to produce a reduction in implicit bias. Evidence for this idea could prompt research that can help elucidate the specific inferences that underlie (changes in) out-group evaluations and that could be targeted in interventions.

Yet, it is important to realize that the observed dissociation could also be the due to things that might have less theoretical value. First, it could indicate that evidence for (1) the presence of the AA training effect or (2) the absence of the AA instruction effect on implicit bias is not robust.

Whereas Van Dessel et al. reported 3 experiments with large sample sizes (total N = 673) drawn from different participant pools (e.g., students and on-line workers), evidence for AA training effects on implicit bias is based on three experiments that used small sample sizes (total N = 156) drawn from a specific pool of undergraduate university students (Kawakami et al; Phills et al., 2011). This is an important limitation, illustrated by the fact that the one experiment of Kawakami et al. in which effects of subliminal AA training on implicit bias were found, did not replicate in a recent set of studies (Van Dessel, De Houwer, Roets, & Gast, 2016). Notably, Bayesian analysis of the critical *t*-tests reported by Kawakami et al. and Phills et al. showed evidence in favor of the hypothesis that AA training causes significant changes in implicit evaluations. However, the relevant Bayes Factors ranged from $BF_1 = 1.53$ to $BF_1 = 5.99$ (Table 1), indicating that evidence was rather weak (Jeffreys, 1961), attesting that replication is warranted to establish the robustness of the AA training effect.

Second, it is possible that the observations that changes in implicit bias toward outgroups can be caused by AA experiences (Kawakami et al., 2007; Phills et al., 2011), but not by AA instructions (Van Dessel et al., 2015) is due to differences in the procedures used by these authors that are unrelated to the critical distinction between AA training and AA instructions (i.e., the actual performance of AA actions). For instance, AA instructions referred to names of Black and White people in Van Dessel et al.'s AA instruction studies and to images of Black and White people in Kawakami et al.'s AA training studies. Also, the evaluation procedures differed between studies. For instance, the IAT procedure that was used to measure implicit bias by Van Dessel et al. used names as target stimuli rather than images.

The Current Studies

Given these limitations, it is important that the robustness of AA instruction and AA training effects on evaluations of well-known social groups is assessed. We therefore conducted four experiments examining AA training (Experiments 1, 3, and 4) and AA instruction effects (Experiments 2 and 4). Experiment 1 is a conceptual replication of Kawakami et al. (2007; Experiment 1). Flemish students first completed a training in which they avoided images of Flemish people and approached images of a specific group that Flemish people are often prejudiced towards (i.e., Turkish people; see Dhont & Van Hiel, 2009) and were then asked for their evaluations of Flemish and Turkish people. Experiment 2 is a replication of the study of Van Dessel et al. (2015), probing effects of AA instructions on out-group evaluations. To create more optimal circumstances for AA instruction effects to occur, we matched the study design to the design of AA training studies (Experiment 1; Kawakami et al.) and included a phase in which participants completed AA training with non-social stimuli before they received AA instructions for the social stimuli. Experiment 3 examined AA training effects on evaluations of Black and White people in a large on-line sample of White United States citizens and Experiment 4 directly compared effects of AA instructions and AA training in a similar sample. The design of all experiments as well as the raw data, experimental and analytic scripts are available on Open Science Framework (Van Dessel, 2020). All studies received approval from the Ghent University ethics committee.

Experiment 1

Experiment 1 tested the hypothesis that training to approach out-group members and avoid in-group members reduces bias in racial evaluations. Procedural details were identical to Kawakami et al. (2007; Experiment 1) with the exception that Flemish and Turkish (rather than Black and White) people were used as the target social groups and that participants also completed measures of explicit prejudice and a question that probed hypothesis awareness. These measures were included to examine the generalizability of AA training effects, which might be important given that changes in implicit bias measured with the IAT are often found to strongly decrease over time (Lai et al., 2016) and to only weakly relate to changes in real-life behavior (under certain conditions) (Forscher et al., 2019).

Method

Participants. Sixty-four native Dutch-speaking undergraduates (53 women) participated in exchange for the payment of 6 euros. Participants were randomly assigned to a condition in which they approached Turkish and avoided Flemish faces or a control condition in which they responded to these faces by performing a joystick movement to the left or right. The sample size was determined by calculating the number of participants needed in each group that would allow .80 power for finding an effect in a one-tailed *t*-test comparing evaluation scores for participants in the experimental and control condition at alpha = .05. The effect-size used for this calculation was *d* = 0.63 as this was the smallest observed effect size in Kawakami et al. (2007) and Phills et al. (2011). The planned sample size was pre-registered on the Open Science Framework as well as the hypothesis that participants would show reduced implicit bias in the approach Black people training compared to the control condition. We did not have a priori hypotheses regarding possible effects of AA training on explicit prejudice measures.

Materials. Thirty images of Turkish faces and thirty images of Flemish faces were obtained from a website that distributes photographs of typical faces of people from different countries, with the photographer's permission (http://www.facesoftomorrow.com). All images depicted male faces with neutral expressions. We restricted the visual stimuli to men to avoid gender influences. We selected 18 images for each nationality category (Turkish and Flemish), based on a pretest in which 56 participants rated all 60 pictures. First, they categorized the depicted person's nationality as "Turkish", "Flemish" or "unclear". Then they rated the emotional neutrality of the depicted person's facial expression and the probability that the person was Turkish or Flemish. Responses were provided on a seven-point Likert scale ranging from “1” (not neutral at all/ highly unlikely) to “7” (totally neutral/ very likely). Three criteria were used for the selection of the faces: First, over 90% of participants categorized the nationality of the face correctly. Second, the probability of the nationality was unequivocally rated as highly probable (Turkish: *M* = 6.36, *SD* = 0.17; Flemish: *M* = 6.04, *SD* = 0.33). Finally, the pictures were rated as emotionally neutral and pictures of both nationalities were matched on this dimension as closely as possible (Turkish: *M* = 4.45, *SD* = 0.49; Flemish: *M* = 4.65, *SD* = 0.82), *t*(46) = 0.91, *p* =. 37.

Procedure. The procedure was matched to the procedure used by Kawakami et al. (2007, Experiment 1), with the main exception that instructions and images involved Flemish and Turkish (rather than White and Black race) people. Other deviations from this procedure are noted below. The experiment was programmed and presented using the Direct RT Empirisoft Software package (DirectRTv2012) on a PC with a 19-inch monitor that had a keyboard and a joystick (Wingman attack 2) attached to it. Upon arrival, participants were seated at a desk in front of the computer and were told that they would respond to images of Turkish and Flemish people by using the joystick. Participants in the approach Turkish condition then received the following instructions (translated from Dutch):

> *These are your instructions –*
>
> *When you see a photo of a Turkish person:*
>
> *Respond by pulling them toward yourself with the joystick*
>
> *When you see a photo of a Flemish person:*
>
> *Respond by pushing them away from yourself with the joystick*

Participants in the sideways control condition received identical instructions except that they were instructed to move the joystick to the right in response to a photo of a Turkish person and to the left in response to a photo of a Flemish person. The assignment of the social groups to either the left or rightward movement was counterbalanced across participants.

In the AA training task, participants completed 480 trials divided into ten blocks. Twelve Turkish and twelve Flemish pictures were each presented two times in each block. Each trial began with the presentation of a Flemish or Turkish face on the computer screen. The face remained on the screen until participants made a response with the joystick. After a correct response, participants were shown a blank screen for 1000ms before the start of the next trial. After incorrect responses, a blank screen was shown for 100ms, followed by the presentation of a red X in the middle of the screen for 800ms and another black screen for 100ms.

After completing the AA training, participants performed an IAT in which they categorized the six remaining pre-rated images of Flemish and Turkish faces along with six negative words (i.e., the Dutch words for evil, pain, hate, sickness, hurt and filth) and six positive words (i.e., the Dutch words for love, cheer, peace, happy, hug and rainbow). Participants first completed two practice blocks in which they categorized the Turkish and Flemish faces or the positive and negative words by pushing a key on the left (Q) or right (M) of an AZERTY keyboard. Category names were presented on the left and right top of the screen. The practice blocks were followed by two critical blocks in which participants categorized both words and pictures with the same set of responses. In one block, response mappings were prejudice-congruent such that Flemish faces and positive words shared one response key and Turkish faces and negative words shared a second response key. In the second block, response key assignment was reversed. IAT block order was counterbalanced across participants. Critical blocks consisted of 60 trials in which a word or image was presented on the screen until the participant provided a response. If the response was correct, the stimulus disappeared, and the next stimulus was presented 1000ms later. If the response was incorrect, the word was replaced by a red “X” for 800ms. The next word appeared 100ms after the red “X” was removed from the screen.

Two additional procedural phases were added to the procedure of Kawakami et al. (2007). First, participants were asked to complete four explicit prejudice measures after performing the IAT. Participants first rated how much they liked Flemish and Turkish people and completed a thermometer rating of self-reported warm or cold feelings towards Turkish and Flemish people (Iyengar, Messing & Hahn, 2011) on 9-point Likert scales (1 = completely not warm/liked; 9 = completely warm/liked). Next, they completed the Subtle Racism Scale (SRS: Pettigrew & Meertens, 1995) adapted to a Flemish-Turkish prejudice context (Van Hiel & Mervielde, 2005). This six-item measure consisted of two items assessing the defense of traditional values (e.g., ‘Turkish immigrants living in Belgium teach their children values and skills different from those required to be successful in our society’), two items assessing the denial of positive emotions (e.g., ‘I admire the Turkish immigrant community members living here under difficult circumstances, reverse scored’), and two items assessing exaggeration of cultural differences (e.g., ‘There are huge differences between Turkish immigrants and Belgians regarding their religion and religious habits’). Second, participants completed four items from the Blatant Racism Scale (BRS: Pettigrew & Meertens, 1995; adapted to a Flemish-Turkish prejudice context: Van Hiel & Mervielde, 2005). This scale includes items like “Turkish immigrants get jobs that actually belong to Flemish people”. Items of both scales were measured on 5-point Likert scales (1 = strongly disagree; 5 = strongly agree).

In a second additional procedural phase, participants were asked to report in an open-response format what they had thought during the experiment was the purpose of the task that involved the joystick responses. Finally, participants were thanked for participation and debriefed.

Results

IAT. IAT D_4-scores were calculated following the procedure by Greenwald, Nosek and Banaji (2003) such that higher scores indicate a stronger preference for Flemish over Turkish people. The Spearman-Brown corrected split-half reliability of the evaluative IAT score, calculated on the basis of an odd-even split, was $r(62) = .78$. Overall, IAT scores indicated a strong preference for Flemish people ($M = 0.29$, $SD = 0.40$), $t(62) = 5.74$, $p < .001$, $d = 0.72$. An ANOVA with AA Condition (approach Turkish and avoid Flemish people, move Turkish and Flemish people left/right) and IAT Block Order (prejudice-congruent IAT block first, prejudice-incongruent IAT block first) as between-subjects factors revealed a main effect of First Block, $F(1,59) = 7.18$, $p = .010$, $\eta^2 = 0.10$. IAT scores were higher for participants who started with the prejudice-congruent block ($M = 0.41$, $SD = 0.38$) than for participants who started with the prejudice-incongruent block ($M = 0.12$, $SD = 0.36$). Notably, the main effect of Condition was non-significant, $F(1,59) = 3.09$, $p = .084$, $\eta^2 = 0.04$, and we also did not observe an interaction of Condition x First Block, $F(1,59) = 0.17$, $p = .68$, $\eta^2 < 0.01$. The planned one-tailed *t*-test that served as the basis for our sampling plan, however, did reveal lower IAT scores for participants who approached Turkish people ($M = 0.20$, $SD = 0.39$) than for participants in the sideways control condition ($M = 0.37$, $SD = 0.39$), $t(61) = 1.72$, $p = .045$, $d = 0.43$. The Bayes Factor, calculated in accordance with Rouder, Speckman, Sun, Morey, and Iverson (2009), indicates that the data provide slightly more evidence for the presence (alternative hypothesis) than for the absence of this effect (null hypothesis) (i.e., $BF_1 > 1$), $BF_1 = 1.77$.

Explicit measures. Warmth and liking scores were calculated by subtracting ratings for Turkish people from ratings for Flemish people. Both scores revealed a preference for Flemish people (warmth score: $M = 1.60$, $SD = 2.07$; liking score: $M = 1.67$, $SD = 2.01$), *t*s > 6.15, $p < .001$, *d*s > 0.77. Planned one-tailed *t*-tests revealed a reduced difference in liking scores for participants who approached Turkish people (M = 1.22, SD = 1.84) as compared to participants in the control condition (M = 2.13, SD = 2.07), $t(61) = 1.83$, $p = .036$, $d = 0.46$, $BF_1 = 2.09$. The difference in warmth scores was not significant (approach Turkish: M = 1.19, SD = 2.04; control: M = 2.13, SD = 2.04), $t(61) = 1.64$, $p = .053$, $d = 0.41$, $BF_1 = 1.57$. Subtle Racism Scale (SRS) and Blatant Racism Scale (BRS) scores were calculated by summing the ratings for the different items and dividing this score by the number of items (Cronbach's Alpha = .76/.70). We did not observe significant differences between the two experimental conditions, ts < 1.30, ps > .099, ds < 0.34, BF_1s < 1.

Hypothesis awareness. Responses on the hypothesis awareness question were coded by two independent raters (interrater agreement: 100%), indicating that eight participants in the approach Turkish condition (25%) reported that the AA training was designed to change prejudice. These eight participants exhibited reduced implicit bias (M = 0.05, SD = 0.32) compared to control participants, $t(37) = 2.63$, $p = .006$, $d = 1.04$, $BF_1 = 8.18$, whereas hypothesis unaware participants did not exhibit this reduction in bias (M = 0.29, SD = 0.33), $t(53) = 0.85$, $p = .20$, $d = 0.23$, $BF_1 =$ 0.62 (Figure 1). We did not observe significant differences in any of the other outcomes when separately comparing hypothesis aware or unaware approach Turkish participants with control condition participants, ts < 1.63, ps > .057, ds < 0.65, BF_1s < 1.84.

Discussion

Experiment 1 replicated previous findings (Kawakami et al., 2007; Phills et al., 2011) that participants who approached faces of out-group members (i.e., Turkish people) and avoided faces of in-group members (i.e., Flemish people) demonstrated a reduction in the relatively more negative implicit evaluation of the out-group compared to participants who completed a control training. In contrast to the earlier studies, however, the observed AA training effect was barely significant on the planned t-test, was non-significant in the ANOVA, and was of small effect size (with Bayes Factors indicating only weak evidence for the effect). Notably, statistical power was low in light of the small size of the observed effects (achieved power = .52 to observe an effect of *d* = 0.43).

Interestingly, results also provided initial evidence for (1) an effect of AA training on one measure of explicit prejudice (i.e., liking rating scale), but not on other explicit measures (e.g., warmth rating scale, SRS, and BRS), and (2) moderation of AA training effects by participants' reported awareness of the experimenter hypotheses. However, this evidence should be interpreted with caution given that it emerged from exploratory analyses and the awareness analyses were correlational in nature and included a comparison between groups of different sizes, one of which involved the data of only eight participants.

Experiment 2

In Experiment 2, we examined the replicability and generalizability of the findings of Van Dessel et al. (2015) that AA instructions do not produce changes in (implicit) out-group evaluations. Experiment 2 used the same stimuli, instructions, and evaluation tasks as Experiment 1 to ensure an experimental design that was more adequately matched to that of experiments in which AA training effects were observed (Kawakami et al., 2007). Moreover, participants completed AA training with unrelated (non-social) stimuli before receiving AA instructions (see Raes et al., 2014, for a similar set-up in the context of instructed fear conditioning). This phase of familiarization with the training task might be important because it allows participants to learn about certain features of the training (e.g., that it is not as invasive as expected) that are not communicated in typical instructions and might therefore increase the opportunity to observe instruction effects (e.g., see Van Dessel, Liefooghe, & De Houwer, 2019).

Participants

To deal with the fact that (1) Experiment 1 indicated that AA training effects might be of small effect size, and (2) effects of instructions might be even more reduced (for known groups: see Van Dessel et al., 2016), we used sequential Bayesian hypothesis testing in Experiment 2. A total of 40 native Dutch-speaking undergraduates were recruited at Ghent University, Belgium. Initial sample size was set at 40 and sample size increases in batches of 10 participants were planned until decisive evidence was obtained for the presence or absence of an AA instruction effect on IAT scores as indicated by the Bayes Factor. The Bayesian criterion for the stopping rule was set at a Bayes Factor larger than 3 or smaller than 1/3 for the crucial *t*-test analysis because this Bayes Factor indicates 'substantial evidence' (i.e., the data supports the null or alternative hypothesis at least three times more strongly; Jeffreys, 1961). This criterion was met after running 40 participants.

Procedure

The procedure was identical to Experiment 1 with the exception that (1) participants first performed an AA training task in which they consistently approached images of chairs and avoided images of tables, (2) this AA training phase was followed by these instructions for participants in the approach Turkish people condition (translated from Dutch): *"You will now perform the same task with photos of Turkish and Flemish faces. You will see photos of 12 Turkish and 12 Flemish people. Each photo will be presented on 20 occasions and you will need to respond to the photo by performing the correct action: Approach Turkish faces by pulling a joystick towards you and avoid Flemish faces by pushing a joystick away from you. Remember these instructions well as you will first perform a different task."* Participants in the control condition received similar instructions except that they learned that they should move Flemish and Turkish faces to the left or to the right, and (3) participants completed the approach Turkish or control training only after completing the evaluation measures.

Results

IAT. IAT scores (reliability: *r*[37] = .82) indicated a strong implicit preference for Flemish over Turkish people ($M = 0.37$, $SD = 0.31$), $t(38) = 7.45$, $p < .001$, $d = 1.19$. We did not observe any significant effects in a 2 (AA Condition) x 2 (IAT Order) ANOVA, *F*s < 0.06, *p*s > .80, η^2s < 0.01. The planned one-tailed *t*-test that served as the basis for our sampling plan revealed no significant reduction in implicit bias for participants who were instructed to approach Turkish people ($M = 0.36$, $SD = 0.33$) compared to participants in the sideways control condition ($M = 0.38$, $SD = 0.25$), $t(37) = -0.24$, $p = .60$, $d = 0.08$, $BF_1 = 0.31$.

Explicit measures. Explicit liking and warmth scores also revealed a preference for Flemish people (warmth score: $M = 1.67$, $SD = 1.95$; liking score: $M = 1.64$, $SD = 1.75$), *t*s > 5.33, $p < .001$, *d*s > 0.85. We did not observe any differences in warmth, liking, BRS (Cronbach's Alpha = .57), or SRS (Cronbach's Alpha = .77) scores for participants who approached Turkish people and participants in the control group, *t*s < 0.46, *p*s > .32, *d*s < 0.16, BF_1s < 0.47.

Discussion

Experiment 2 investigated whether AA instructions can produce changes in implicit or explicit evaluations of out-groups under conditions that were more tightly matched to the conditions under which AA training effects have been observed in previous studies. We found no effects of AA instructions on social group evaluations even when participants first completed AA training with non-social stimuli, corroborating and extending the results of Van Dessel et al. (2015).[1] Notably, the current study used a more lenient stopping rule criterion than recommended in recent papers (Schönbrodt, Wagenmakers, Zehetleitner, & Perugini, 2017) and evidence for the absence of the crucial effect was only 'substantial' ($BF_1 = 0.31$) but not 'strong' (i.e., $BF_1 < 0.10$).

[1] Another lab-based experiment examined effects of AA instructions in a collaboration task in which participants collaborated with another participant who (supposedly) performed the approach out-group training. Results of this experiment further corroborated that AA instructions do not influence out-group evaluations. Full description of this experiment and its results can be found on the Open Science Framework project page (link noted on p.9).

Another limitation of Experiment 2 is that the AA training that was used to familiarize participants with the task included stimuli for which evaluation might be less relevant (chairs and tables). Though AA training effects have been found for this type of non-relevant stimuli (e.g., non-words: Van Dessel et al., 2015), it is possible that using these stimuli shifted attention away from the evaluative implications of the training, precluding AA instruction effects.

Experiment 3

Experiments 1 and 2 examined AA training and instruction effects in lab-based experiments with relatively small student samples. In these experiments, we observed only AA training effects, but even training effects were small and evidence for these effects was weak. Experiment 3 examined whether AA training effects can be found in a large sample of online participants that were recruited via the Project Implicit research website (https://implicit.harvard.edu). In-line with Kawakami et al. (2007) and Phills et al. (2011), and in contrast to Experiment 1, participants were trained to approach Black and avoid White people (or to move Black and White people left/right: control group), and training effects were examined on evaluations of Black and White people. Because the results of Experiment 1 suggested there can be distinct training effects depending on hypothesis awareness, we decided to further explore this effect by experimentally manipulating hypothesis awareness. Half of the participants were instructed that the purpose of the training task was to make participants' attitudes towards Black people more positive, whereas the other participants learned that the training task's purpose was to measure participants' attitudes towards Black people (in Experiment 1, the majority of participants believed this to be the purpose of the task). The main hypothesis was that participants would show reduced implicit bias in the approach Black people training compared to the control condition; analyses including the task goal instruction manipulation were exploratory.

Method

Participants. To ensure sufficient statistical power to observe a small effect, we recruited until at least 1000 White United States citizens had completed the study on Project Implicit. Because the study was not immediately taken off-line, the final sample size was N = 1194. In line with the standard treatment of Project Implicit data (e.g., Smith, De Houwer, & Nosek, 2013), data-exclusion involved removing participants who (a) did not fully complete all questions and tasks (180 participants; i.e., 15.1%), (b) had error rates above 30% when considering all IAT blocks or above 40% for any one of the critical IAT test blocks (17 participants; i.e., 1.4%), or (c) responded faster than 400ms on more than 10% of the IAT trials (15 participants; i.e., 1.3%). Analyses were performed on the data of 982 participants (621 women, mean age = 34, *SD* = 14).

Materials. A total of 28 images of male faces (14 White and 14 Black) were selected from stimuli used in in the AA training studies of Phills et al. (2011). We used 16 images for the AA training task and 12 images for the IAT.

Procedure. Participants were informed that they would see images of faces of Black and White people and that their task would be to make a specific action each time they would see one of these faces. Half of the participants learned they would approach faces of Black people and avoid faces of White people. The other participants were told to respond to faces of Black people by performing a movement to the left, and to faces of White people by performing a movement to the right (or vice versa). It was emphasized that it was very important for participants to remember which action they would perform for the two types of images as they would need this information to complete the task successfully.

Half of the participants were then told that the purpose of performing the described task was to make their attitude towards Black people more positive. The other participants were told that the task purpose was to measure their attitude towards Black people. Participants were then told that we would later explain exactly how they would be able to perform the movements, but they now just needed to remember the stimulus-action contingencies. It was then repeated what actions they would perform in response to which faces.

Participants then performed the AA training task. They were told that, on each trial, a stick figure would appear either above or below an image of a face and they needed to use the up and down arrow keys to approach the type of faces they had been instructed to approach and avoid the type of faces they had been instructed to avoid (approach Black people condition) or use the left and right keys to perform a movement to the right or left in response to the faces (sideways control condition). Participants performed eight practice trials followed by 56 experimental trials. On each trial, an image of a Black or White face was presented together with a manikin that appeared above or below the image. When participants performed the correct movement, the manikin moved towards the Black face or away from the White face or moved to the left or right depending on the condition and image content. Participants could only proceed to the following trial by performing the correct response.

Next, participants completed an IAT identical to the one used in Experiment 1 with the exception that targets were images of Black and White faces and that participants who made an error needed to correct their mistake to continue. Latencies were recorded until a correct response was made. After the IAT, participants first completed liking ratings and thermometer ratings of self-reported warmth or cold feelings towards Black and White people and then completed a standard (7-item) version of the Modern Racism Scale (MRS: McConahay, 1986).

Participants were asked what action they were instructed to perform when they saw White/Black faces during the stick figure task (response options: ‘approach them’, ‘avoid them’, ‘move them left’, ‘move them right’, ‘I can’t remember’) and what we told them was the purpose of this task (response options: ‘To make my attitude towards Black people more positive’, ‘To measure my attitude towards Black people’, ‘I don’t remember’). Next, participants completed hypothesis awareness and reactance scales. They indicated to what extent they (1) believed that the experimenters tried to change their attitudes towards Black people, on a 5- point Likert scale (1 = I did not believe it at all; 5 = I believed it entirely) and (2) felt annoyed that we might have tried to change their attitudes towards Black people (1 = I do not feel annoyed by that at all; 5 = I feel very annoyed by that). Finally, participants were thanked and debriefed.

Results

IAT. IAT scores were calculated using the D_2-algorithm, which is the recommended scoring procedure for IATs in which participants need to correct their mistakes (Greenwald et al., 2003). Split-half reliability of the IAT score was $r(980) = .84$. Across groups, participants displayed an implicit preference for White over Black people ($M = 0.38$, $SD = 0.42$), $t(981) = 27.83$, $p < .001$, $d = 0.89$. A 2 (AA Condition) x 2 (Goal Instructions) x 2 (IAT Order) ANOVA revealed no main effects except for the expected effect of AA Condition, $F(1,974) = 5.35$, $p = .021$, $\eta^2 = 0.01$, indicating that participants in the approach Black people condition ($M = 0.35$, $SD = 0.43$) exhibited less implicit bias than participants in the control condition ($M = 0.41$, $SD = 0.41$), $d = 0.14$, $BF_1 = 2.50$. We also observed an interaction of AA Condition and IAT Order, $F(1,974) = 4.00$, $p = .046$, $\eta^2 = 0.01$, indicating that there was an effect of AA Condition for participants who started with the prejudice-incongruent IAT block, $t(493) = 3.14$, $p < .001$, $d = 0.28$, $BF_1 = 31.50$, but not for participants who started with the prejudice-congruent IAT block, $t(485) = 0.01$, $p = .50$, $d < 0.01$, $BF_1 = 0.16$. Finally, we observed an effect of Goal Instructions x IAT Order, $F(1,974) = 8.54$, $p = .004$, $\eta^2 = 0.01$, indicating that participants who started with the prejudice-congruent IAT block exhibited more implicit bias when they learned that the AA training task purpose was to change their attitudes than when they learned that the purpose was to measure their attitudes, $t(485) = 2.90$,

p = .004, d = 0.26, BF_1 = 5.84. This effect was not observed for participants who started with the congruent IAT block, t(493) = -1.43, p = .15, d = 0.13, BF_1 = 0.27. No other effects were observed in the ANOVA, Fs < 3.45, ps > .063, η^2s < 0.01.

Explicit measures. Warmth scores (M = 0.16, SD = 1.35) revealed a preference for White over Black people, t(952) = 3.70, p < .001, d = 0.12, but liking scores did not (M = 0.05, SD = 1.10), t(981) = 1.35, p = .18, d = 0.04. Separate ANOVA's on warmth, liking, and MRS scores (Cronbach's Alpha = .83), did not reveal any significant effects, Fs < 2.69, ps > .10, η^2s < 0.01, except for an effect of AA Condition on warmth scores, F(1,949) = 5.69, p = .017, η^2 = 0.01, that was qualified by Goal Instructions, F(1,949) = 8.64, p = .003, η^2 = 0.01. Participants who performed approach Black people training exhibited a *stronger* preference for White people compared to the control training condition when they learned that the AA training task purpose was to change their attitudes, t(476) = -3.68, p < .001, d = 0.34, BF_1 = 84.70, but not when they learned that the purpose was to measure their attitudes, t(473) = 0.40, p = .69, d = 0.04, BF_1 = 0.18.

Reactance and hypothesis awareness. For exploratory purposes, we also examined the relation between evaluation scores and reactance and hypothesis awareness. Reactance scores correlated significantly with IAT scores, r(957) = .09, p = .004, warmth ratings, r(937) = .06, p = .043, and MRS scores, r(957) = .26, p < .001, but not with liking rating scores, r(937) = .04 p = .22. Hypothesis awareness scores only correlated significantly with MRS scores, r(964) = 0.13, p < .001. An ANOVA on reactance ratings revealed no effects of AA Condition or Goal Instructions, Fs < 1.53, ps > .21, η^2s < 0.01. The ANOVA on hypothesis awareness scores revealed higher scores for participants in the approach Black people training than for participants in the control condition, F(1,962) = 16.69, p < .001, η^2 = 0.01, and higher scores for participants who learned that the AA training task purpose was to change their attitudes than for participants who learned that the purpose was to measure their attitudes, F(1,962) = 24.81, p < .001, η^2 = 0.02.

Discussion

Experiment 3 replicates the finding that AA training can produce a reduction in implicit bias as measured with the IAT. This effect was small, overall, and it was only observed for participants who completed the prejudice-incongruent IAT block first. Notably, the AA training effect did not generalize to any of the explicit prejudice measures. In contrast, we observed strong evidence for a reversed effect of AA training on warmth ratings when participants were explicitly told about the prejudice reducing goal of the AA training. These results accord with findings that prejudice reducing interventions can lead to a reactance-related increase in prejudice when external control is emphasized (Legault, Gutsell, & Inzlicht, 2011). Notably, a similar increase in more positive evaluation of in-groups compared to out-groups for participants who were informed about the prejudice reducing purpose of AA training was also observed on IAT scores, but this was unrelated to whether participants completed control or approach Black people training.

Experiment 4

Experiment 4 provides a direct comparison of AA instruction and AA training effects on evaluations of racial groups (i.e., Black and White people) in a large sample of White participants recruited via the Project Implicit research website. We manipulated two critical factors between-subjects: AA Condition (approach Black and avoid White people, move Black and White people left/right) and Learning Method (AA training, AA Instructions). Similar to Experiment 3, we also included a manipulation of Goal Instructions. This was implemented in order to match the degree of hypothesis awareness in the AA instruction and training conditions. Because Experiment 3 provides evidence that reactance can influence AA (training) effects when participants learn about the prejudice reducing purpose of AA training, we included this instruction condition but also a condition with instructions that should minimize reactance. That is, we contrasted instructions about the prejudice reducing purpose of the AA task with instructions that did not relate the AA task purpose to prejudice at all (i.e., instructions indicating that the task purpose is to measure general response speed).

Method

Participants. We recruited 969 White United States citizens via Project Implicit. Data-exclusion involved removing participants who (a) did not fully complete all questions and tasks (123 participants; i.e., 12.7%), (b) had error rates above 30% when considering all IAT blocks or above 40% for any one of the critical IAT test blocks (16 participants; i.e., 1.7%), or (c) responded faster than 400ms on more than 10% of the IAT trials (21 participants; i.e., 2.2%). Analyses were performed on the data of 809 participants (528 women, mean age = 34, *SD* = 14).

Procedure. We used the same procedure as Experiment 3, with two exceptions. First, only half of the participants completed AA training before the IAT. The other participants performed the AA training task after completing all measures (i.e., at the end of the experiment). Second, whereas half of the participants were told that the purpose of performing the AA task was to make their attitude towards Black people more positive, the other participants learned that the task purpose was to measure how quickly and accurately they could respond to images.

Results

IAT. Split-half reliability of the IAT D_2 score was $r(807) = .83$. Across groups, participants displayed an implicit preference for White over Black people ($M = 0.37$, $SD = 0.44$), $t(808) = 24.03$, $p < .001$, $d = 0.84$. A 2 (Learning Method) x 2 (AA Condition) x 2 (Goal Instructions) x 2 (IAT Order) ANOVA revealed a main effect of AA Condition, $F(1,793) = 4.68$, $p = .031$, $\eta^2 = 0.01$, which was qualified by Learning Method, $F(1,793) = 4.02$, $p = .045$, $\eta^2 = 0.01$ (Table 2). Participants in the approach Black people condition exhibited less implicit bias than participants in the control condition when they had performed the AA training task (approach Black people: $M = 0.29$, $SD =$

0.45; control: *M* = 0.42, *SD* = 0.43), *t*(362) = 2.76, *p* = .003, *d* = 0.29, BF_1 = 11.77, but not when they had only received AA instructions (approach Black people: *M* = 0.38, *SD* = 0.44; control: *M* = 0.40, *SD* = 0.44), *t*(443) = 0.40, *p* = .34, *d* = 0.04, BF_1 = 0.24. Notably, the Bayes Factor for the interaction effect indicates that the data provide slightly more evidence for the absence than for the presence of a difference in the effect of training and that of instructions (i.e., BF_1 < 1), BF_1 = 0.67.

We also observed an interaction of AA Condition, Goal Instructions, and IAT Order, *F*(1,793) = 9.60, *p* = .002. η^2 = 0.01, For participants who started with the prejudice-incongruent IAT block, there was an effect of AA Condition x Goal Instructions, *F*(1,378) = 7.32, *p* = .007, η^2 = 0.01 Specifically, participants in the approach Black people condition exhibited less implicit bias than participants in the control condition when they had learned that the AA task purpose was to measure response speed (approach Black people: *M* = 0.30, *SD* = 0.40; control: *M* = 0.51, *SD* = 0.40), *t*(194) = 3.49, *p* < .001, *d* = 0.50, BF_1 = 90.31, but not when they had learned the prejudice reduction task purpose, *t*(184) = 0.33, *p* = .63, *d* = 0.05, BF_1 = 0.32. This effect was not observed for participants who started with the congruent IAT block, *F*(1,415) = 2.77, *p* = .097, η^2 < 0.01. No other effects were observed in the ANOVA, *F*s < 3.04, *p*s > .082, η^2s < 0.01.

Explicit measures. Warmth (*M* = 0.00, *SD* = 1.51) and liking scores (*M* = 0.02, *SD* = 1.26) did not reveal a preference for White over Black people, *t*s < 0.40, *ps* > .69, *d*s < 0.02. Separate ANOVA's on warmth, liking, and MRS scores (Cronbach's Alpha = .83), did not reveal any significant effects, *F*s < 2.89, *p*s > .089, η^2s < 0.01, except for an interaction effect of AA Condition x Learning Method on liking scores, *F*(1,771) = 4.41, *p* = .036, η^2 = 0.01. Participants who performed approach Black people training (*M* = -0.11, *SD* = 1.30) exhibited a reduced preference for White people compared to the control training condition (*M* = 0.08, *SD* = 1.28), but this effect was non-significant, *t*(345) = 1.34, *p* = .091, *d* = 0.14, BF_1 = 0.76. Participants who only received

AA instructions exhibited a reversed effect (approach Black people: $M = 0.15$, $SD = 1.36$; control: $M = -0.08$, $SD = 1.09$), $t(430) = -1.93$, $p = .027$, $d = 0.19$, $BF_1 = 1.84$.

Reactance and hypothesis awareness. Reactance scores correlated significantly with all explicit prejudice scores, rs > 0.12, ps < .001, but not with IAT scores, $r(772) = 0.01$, $p = .73$. Hypothesis awareness scores only correlated significantly with MRS scores, $r(791) = 0.13$, $p < .001$. An ANOVA on hypothesis awareness scores revealed only an interaction effect of AA Condition x Learning Method, $F(1,785) = 3.84$, $p = .050$, $\eta^2 = 0.01$. Participants in the approach Black people training condition believed more that the purpose of the experiment was to reduce prejudice than participants in the control training group (approach Black people: $M = 2.63$, $SD = 1.33$; control: $M = 2.26$, $SD = 1.17$), $t(352) = 2.72$, $p = .003$, $d = 0.29$, $BF_1 = 10.69$, but this effect was not observed for participants who had only received AA instructions (approach Black people: $M = 2.47$, $SD = 1.27$; control: $M = 2.47$, $SD = 1.20$), $t(437) = -0.01$, $p = .50$, $d = 0.00$, $BF_1 = 0.17$. An ANOVA on reactance scores revealed no significant effects, Fs < 2.68, ps > .10, η^2s < 0.01.

Discussion

Experiment 4 compared AA training and AA instruction effects in a single experiment, simultaneously replicating previous findings that AA training produces a reduction in implicit bias as measured with the IAT whereas AA instructions do not. This reduction did not generalize to explicit prejudice measures. In fact, we even observed (anecdotal) evidence for a reversed effect of AA instructions on explicit liking scores. Similar to Experiment 3, we also obtained evidence that AA effects on IAT scores are moderated by (1) IAT block order and (2) instructions about the purpose of the AA training. Specifically, we only observed AA training effects when participants first completed the prejudice-incongruent IAT block and did not receive instructions that the purpose of the AA training was to reduce prejudice.

General Discussion

Four experiments investigated whether training and instructions to approach social out-groups can change implicit and explicit evaluations of these groups. First, we replicated previous findings (Kawakami et al., 2007) that AA training can reduce implicit racial bias (Experiments 3-4) and extended it by showing this effect on implicit evaluations of Turkish people in a sample of Flemish participants (Experiment 1). We also observed a reduction in prejudice on explicit liking ratings in Experiment 1, although this effect was not replicated in Experiments 3 and 4. Second, we found consistent evidence (Experiments 2 and 4) that mere instructions to approach social out-groups do not lead to changes in implicit or explicit social group evaluations, even when participants have prior experience with the AA training task (Experiment 2) and when procedural and other features (e.g., hypothesis awareness) of AA instruction conditions are matched with AA training conditions (Experiments 2 and 4). A Bayes Factor meta-analysis indicates that the available evidence favors the hypothesis that, with the current procedures, (1) AA training influences racial evaluations assessed with the IAT (current experiments: $BF_1 = 65.22$; current and prior experiments: $BF_1 = 4404.42$), but not with the explicit measures we used, BF_1s < 0.05 (Table 1) and (2) AA instructions do not influence racial evaluations, BF_1s < 0.19. However, caution is warranted when interpreting the difference in the effect of two interventions, most importantly because the studies that used different interventions also differed with regard to many other aspects of the method. Also, when AA instructions and AA training effects were compared directly in Experiment 4, evidence for a dissociation was inconclusive.

Disentangling the mental processes underlying training- and instruction-based effects could be crucial for our understanding of the processes that underlie (implicit) evaluative learning and learning in general (see Brass, Liefooghe, Braem, & De Houwer, 2017, for a related discussion). The current results provide strong evidence that AA training can influence social group evaluations as measured with the IAT whereas we did not observe AA instructions effects. This can be interpreted in several ways. First, it is possible that instruction- and experience-based AA procedures draw on qualitatively different types of mental processes. Specifically, whereas AA instructions might produce effects on the basis of propositional processes, AA training could lead to the gradual re-training of mental associations (e.g., between representations of the self and the targeted out-group: Phills et al., 2011) on the basis of repeated stimulus-response pairings that may be required to produce changes in implicit evaluations of out-groups. This would be a highly important finding, in light of recent research showing that evidence for this distinct evaluative learning mechanism is weak both in the context of associative (e.g., EC, AA training) and other evaluative learning (see Corneille & Stahl, 2018; Van Dessel et al., 2019).

Yet, given this lack of evidence for associative processes (in AA effects), a purely associative explanation of AA training effects might not be the most convincing. Moreover, it does not readily fit with observations that AA training effects were moderated by hypothesis awareness (in Experiment 1) and goal instructions (in Experiments 3 and 4). If one assumes that (automatic) association formation is based on repeated pairings (Olson & Fazio, 2001; Rydell & McConnell, 2006), these findings speak against the assumption that AA training effects are entirely based on associations. A second interpretation of our results could be provided by dual-process accounts such as the Associative Propositional Evaluation Model (APE model: Gawronski & Bodenhausen, 2006). From this perspective, propositional processes could influence association formation on the basis of AA training or influence implicit evaluations more directly. However, even though these theoretical accounts could explain the current results, they do not clearly specify the moderators of AA training effects and as such they did not predict findings regarding these moderators (see also Van Dessel et al., 2019).

Moreover dissociations between effects of instructions and experience do not necessarily imply different types of learning mechanisms (e.g., Van Dessel, Liefooghe, & De Houwer, 2019). More specifically, a third possible account of our results is that the same type of (e.g., inferential) processes underlie AA training and AA instruction effects but that the involved process components (e.g., the specific inferences) differ. In a recent paper (Van Dessel et al., 2019), the argument was made that AA effects require several inference steps, starting with a Step 1 inference about stimulus-action contingencies (e.g., 'I will approach Black people'). Our results might suggest that not all Step 1 inferences are sufficient to facilitate changes in evaluations of out-groups. Rather, these changes might require the specific Step 1 inference that participants *have* approached (rather than *will* approach) out-group members. The reason for this dissociation could be that only the former inference leads to the required subsequent inference steps in the inferential processing chain. From this perspective, it is possible that, under certain conditions, AA instructions can influence racial evaluations. Indeed, an important limitation of the current study is that the AA instruction and AA training procedures were not matched on theoretically important process components such as the different inferences that participants can make on the basis of these procedures. Experiment 4 directly compared instruction and training-based effects in a single experiment, but even in this experiment, there were still several differences between both procedures that could influence results but that are unrelated to differences in the type of (inferential/associative) mental processes involved.

One possible inference step that is missing in the current implementation of AA instructions is the inference that participants have actually approached out-group members. It is possible that AA effects require this step. For instance, when participants learn that they approached out-group members they might further infer that this AA training can change (or has changed) their out-group evaluations (e.g., because approaching a group is believed to facilitate more positive evaluations of the group). Evidence for this idea is found in Experiment 1 where AA training effects were observed only when participants were able to report that this was the task purpose. Notably, however, AA instruction and AA training effects were not stronger (in fact there was evidence for a reduction) when instructions provided participants with information that the AA task was designed to reduce prejudice (Experiments 2 and 4). On the one hand, this suggests that participants react to the idea that the experimenter tries to change their attitudes by resisting this influence when participants are explicitly informed about this purpose (Brehm, 1966). This reactance-related process could be especially strong when AA training involves out-groups because people are often highly reactant to prejudice reduction interventions (Legault et al., 2011). This might partially explain why AA instructions (and AA Training) produce stronger effects on evaluations of unfamiliar social groups. On the other hand, these results also indicate that knowing the task purpose is not sufficient to observe AA effects on out-group evaluations. Hence, additional inferences (e.g., that the completed intervention actually took place and therefore produced an impact on evaluations: see Carr, Dweck, & Pauker, 2012) are required to explain why AA training but not AA instructions affect out-group evaluations. Future research might test these inferential explanations, for instance, by examining the relation between relevant beliefs and AA effects.

In sum, whereas the current results provide substantial evidence for AA training effects on racial evaluations and do not provide such evidence for AA instruction effects, they also highlight the fact that there is still a lot of uncertainty about (1) the conditions under which AA effects can be observed and (2) the mental processes underlying these effects. This is further illustrated by the fact that strong evidence for AA training effects was found only on the IAT. Although initial (associative) theoretical accounts of AA training effects argued that AA training might specifically influence mental associations which can be best assessed with implicit measures such as the IAT (Kawakami et al., 2007), the idea that IAT scores reflect learned associations is inconsistent with recent evidence (e.g., Bading, Stahl, & Rothermund , 2019; Heycke & Gawronski, 2020). Alternative explanations might relate this result to similarities between the IAT and AA training procedures (e.g., both tasks involve categorization of stimuli by emitting two responses) or to evaluative inferences that influence evaluative responding only when there is limited time to emit responses. Nevertheless, the latter explanations are highly speculative and require further investigation.

It is important to note that, whereas typical AA instructions did not influence (implicit) evaluations of well-known social groups, EC instructions have been found to produce such changes (albeit for young/old and American/foreign target categories rather than Black/White: Kurdi & Banaji, 2017). Moreover, AA instructions have been found to produce changes in evaluations of other attitude objects such as fictitious social groups (e.g., Van Dessel, De Houwer, Gast, et al., 2016). Again, this could indicate limitations in the current implementation of our AA instruction manipulation but it could also indicate important differences in the mental processes underlying these effects (of EC vs. AA instructions and of AA instructions on well-known vs. unfamiliar social groups). Dual-process models could explain the latter dissociation by indicating that the one-time pairing of stimuli with valenced action words (e.g., the word ‘approach’) can create an association between the representation of the stimulus and a valenced representation. However, this newly formed association allows only for the formation but not the change of implicit evaluations (Gregg et al., 2006; but see Van Dessel, Ye, & De Houwer, 2019). Dual-process models could also explain the dissociation between EC and AA instruction effects if they assume that EC instructions (and EC) evoke strong evaluative learning on the basis of propositional processes whereas AA instructions evoke weaker evaluative learning on the basis of propositional processes (and AA training effects primarily depend on processes that involve the formation and change of associations: e.g., with self-representations: Phills et al., 2011).

One possible inferential explanation of the dissociation between EC and AA instruction effects relates to differences in the likelihood of producing the relevant evaluative inferences. Learning via EC instructions that known groups will be paired with positive things in the future might allow for the inference that these groups are positive (e.g., because participants readily infer that a stimulus that will produce positive events in the future is positive). As a result, participants might exhibit better performance in an IAT block that pairs this group with positive stimuli. In contrast, in the context of AA, only participants who believe that they completed (and were influenced by) AA training might infer that the approached group is positive. This inference could be essential to counteract effects on IAT performance of (dominant) evaluative inferences about these groups that are based on more distal learning history. From this perspective, the dissociation of AA instruction effects for well-known compared to novel social groups might indicate that evaluative inferences based on AA instructions influence IAT scores only when other (previously learned) information that evokes inferences about the valence of stimuli is not readily available (see Van Dessel De Houwer, Gast, et al., 2016, for related findings).

The current results also have practical implications. First, they suggest that typical AA instructions are not very effective for changing social group evaluations and that, in contrast, AA training can produce such changes. Most importantly, however, our results indicate that even the effects of AA training are typically small, and they seem to be restricted to an implicit measure of prejudice (the IAT). This is an important limitation in light of recent evidence showing weak stability and generalizability of implicit bias effects on the IAT (e.g., Lai et al., 2016; Forscher et al., 2019). Future research might examine whether AA training can influence implicit evaluations of out-groups also when probed with other measures (e.g., the evaluative priming task, Fazio, Sanbonmatsu, Powell, & Kardes, 1986). This might be especially important because (1) IAT performance can be influenced by non-attitudinal factors such as participants' flexible recoding of

IAT categories (Rothermund, Teige-Mocigemba, Gast, & Wentura, 2009; Bading et al., 2019) and (2) current results provided evidence that effects on the IAT scores are restricted to participants who start with the prejudice-incongruent IAT block. This pattern of responding could indicate that AA training facilitates the immediate categorization of difficult (prejudice-incongruent) mappings in the IAT in-line with the training (which might improve participants' IAT performance in the relevant block) but does not influence their overall (spontaneous) evaluations of social (out-) groups.

Constraint on generality

Results were obtained in a sample of White undergraduate university students (Experiments 1 and 2) and on-line Project Implicit website participants (Experiments 3 and 4). It is possible that effects might not generalize to other participant samples, such as those including non-White people or with higher proportions of people who do not have the opportunity to attend university or access the internet. It is also possible that results might not generalize to evaluations of other well-known social groups than Black and White or Turkish and Flemish people. Although we have no reason to believe that results depend on these or other characteristics, these are limitations that are untestable within the current data. All results were obtained in experiments that were set up to maximize statistical power for the planned analyses. Sample size in Experiment 1 was chosen such that there would be sufficient power (>.80) to find the AA effect observed in Kawakami et al. (2007), whereas Experiment 2 used sequential Bayesian hypothesis testing to determine the sample size and ensure that decisive evidence would be obtained for the presence or absence of the AA instruction effect. Experiments 3 and 4 used a sample size that was sufficient to have more than .90 power to observe the estimated AA effect in the different between-subjects conditions. All materials used in the current study are freely available on Open Science Framework.

References

Bading, K., Stahl, C., & Rothermund, K. (2019). Why a standard IAT effect cannot provide evidence for associative learning: The role of similarity construction. *Cognition & Emotion.*

Bar-on, R. (1997). The emotional intelligence inventory (EQ-i). *Technical Manual.* Toronto Multi-Health Systems.

Brass, M., Liefooghe, B., Braem, S., & De Houwer, J. (2017). Following new task instructions: Evidence for a dissociation between knowing and doing. *Neuroscience & Biobehavioral Reviews, 81*, 16-28

Brehm, J. W. (1966). A theory of psychological reactance. New York: Academic Press

Carr, P. B., Dweck, C. S., & Pauker, K. (2012). "Prejudiced" behavior without prejudice? Beliefs about the malleability of prejudice affect interracial interactions. *Journal of Personality and Social Psychology, 103,* 452–471.

Corneille, O., & Stahl, C. (2018). Associative attitude learning: A closer look at evidence and how it relates to attitude models. *Personality and Social Psychology Review*.

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*(2), 176-187.

De Houwer, J. (2018). Propositional models of evaluative conditioning. *Social Psychological Bulletin, 13 (3).*

De Houwer, J. (2019). Implicit bias is behavior: A functional-cognitive perspective on implicit bias. *Perspectives on Psychological Science.*

Dhont, K., Van Hiel, A. (2009). We must not be enemies: Interracial contact and the reduction of prejudice among authoritarians. *Personality and Individual Differences, 46*, 172–177.

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.

Forscher, P. S., Lai, C. K., Axt, J. R., Ebersole, C. R., Herman, M., Devine, P. G., & Nosek, B. A. (2019). A meta-analysis of change in implicit bias. Manuscript under review.

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.

Gast, A., & De Houwer, J. (2013). The influence of extinction and counterconditioning instructions on evaluative conditioning effects. *Learning and Motivation*, *44*(4), 312-325.

Greenwald, A. G., McGhee, D. E., & Schwartz, J. L. K. (1998). Measuring individual differences in implicit cognition: The implicit association test. *Journal of Personality and Social Psychology, 74*, 1464–1480.

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

Heycke, T., & Gawronski, B. (2019). Co-occurrence and relational information in evaluative learning: A multinomial modeling approach. *Journal of Experimental Psychology: General, 149(1),* 104-124

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

Iyengar, S., Messing, S., Hahn, K. S., Banaji, M., & Dial, C. (2011). Explicit and implicit racial attitudes: A test of their convergent and predictive validity. *APSA 2011 Annual Meeting Paper*. Retrieved from http://ssrn.com/abstract=1901991

Jackson, J. L. (2018). The non-Performativity of implicit bias training. *Radical Teacher, 112*, 46-54.

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

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

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.

Kende, A., Tropp, L., & Lantos, N. A. (2017). Testing a contact intervention based on intergroup friendship between Roma and non-Roma Hungarians: Reducing bias through institutional support in a non-supportive societal context. *Journal of Applied Social Psychology, 47*, 47–55.

Kurdi, B., & Banaji, M. R. (2017). Repeated evaluative pairings and evaluative statements: How effectively do they shift implicit attitudes? *Journal of Experimental Psychology: General, 146*, 194–213.

Lai, C. K., Marini, M., Lehr, S. A., Cerruti, C., Shin, J.-E. L., Joy-Gaba, J. A., et al. (2014). Reducing implicit racial preferences: I. A comparative investigation of 17 interventions. *Journal of Experimental Psychology: General*, 143, 1765-1785.

Lai, C. K., Skinner, A. L., Cooley, E., Murrar, S., Brauer, M., Devos, T., et al. (2016). Reducing implicit racial preferences: II. Intervention effectiveness across time. *Journal of Experimental Psychology: General, 145*, 1001–1016.

Legault, L., Gutsell, J.N., Inzlicht, M. (2011) Ironic effects of antiprejudice messages: How motivational interventions can reduce (but also increase) prejudice. *Psychological Science, 22*, 1472–147

McConahay, J. B. (1986). Modern racism, ambivalence, and the modern racism scale. In J. F. Dovidio & S. L. Gaertner (Eds.), *Prejudice, discrimination and racism* (pp. 91- 126). New York: Academic.

Nelson, S. C., Prasad, S., & Hackman, H. W. (2015). Training providers on issues of race and racism improve health care equity. *Race, Racism, and Medicine, 62*, 915–917.

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

Pettigrew, T. F., & Meertens, R. W. (1995). Subtle and blatant prejudice in Western Europe. *European Journal of Social Psychology, 25*, 57-75.

Pettigrew, T. F., & Tropp, L. R. (2006). A meta-analytic test of intergroup contact theory. *Journal of personality and social psychology, 90* (5), 751.

Phills, C.E., Kawakami, K., Krusemark, D.R. & Nguyen, J. (2019). Does reducing implicit prejudice increase out-Group identification? The downstream consequences of evaluative training on associations between the self and racial categories. *Social Psychological and Personality Science, 10*, 26-34.

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.

Raes, A. K., De Houwer, J., De Schryver, M., Brass, M., & Kalisch, R. (2014). Do CS-US pairings actually matter? A within-subject comparison of instructed fear conditioning with and without actual CS-US pairings. *PLoS ONE, 9*(1): e84888.

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.

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

Schönbrodt, F., Wagenmakers, E.-J., Zehetleitner, M., & Perugini, M. (2017). Sequential hypothesis testing with Bayes factors: Efficiently testing mean differences. *Psychological Methods*.

Smith, C.T., Calanchini, J., Hughes, S., Van Dessel, P., & De Houwer, J. (2019). The Impact of Instruction- and Experience-Based Evaluative Learning on IAT Performance: A Quad Model Perspective. *Cognition & Emotion.*

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.

Van Dessel, P. (2020, February 10). The Effectiveness of Approach-Avoidance Instructions and Training for Reducing Prejudice. https://doi.org/10.17605/OSF.IO/Y37TE

Van Dessel, P., De Houwer, J., & Gast, A. (2016). Approach-avoidance training effects are moderated by awareness of stimulus-action contingencies. *Personality and Social Psychology Bulletin, 42*, 81-93.

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

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

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

Van Dessel, P., Eder, A., & Hughes, S. (2018). Mechanisms underlying effects of approach-avoidance training on stimulus evaluation. *Journal of Experimental Psychology: Learning, Memory and Cognition, 44*, 1224-1241.

Van Dessel, P., Hughes, S., & De Houwer, J. (2018). Consequence-based approach-avoidance training: A new and improved method for changing unwanted behavior. *Psychological Science, 29,* 1899-1910.

Van Dessel, P., Hughes, S., & De Houwer, J. (2019). How Do Actions Influence Attitudes? An Inferential Account of the Impact of Action Performance on Stimulus Evaluation. *Personality and Social Psychology Review, 23*, 267-284.

Van Dessel, P., Liefooghe, B., & De Houwer, J. (2019). The instructed task-switch evaluation effect: Is the instruction to switch tasks sufficient to dislike task switch cues? *Journal of Cognition.*

Van Dessel, P., Ye, Y., & De Houwer, J. (2019). Changing Deep-rooted Implicit Evaluation in the Blink of an Eye: Negative Verbal Information Shifts Automatic Liking of Gandhi. *Social Psychological and Personality Science, 10,* 266-273.

Van Hiel, A. & Mervielde, I. (2005). Authoritarianism and social dominance orientation: Relationships with various forms of racism. Journal of Applied Social Psychology 35, 2323-2344.

West, K., Hotchin, V., & Wood, C. (2017). Imagined contact can be more effective for participants with stronger initial prejudices. *Journal of Applied Social Psychology, 47,* 282-292.

Figures

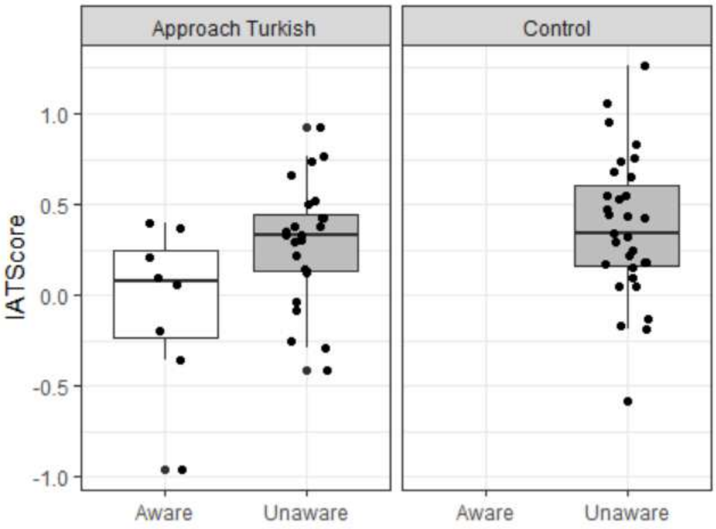

Figure 1. IAT D scores indicating implicit preference for White people over Black people as a function of AA Condition and Hypothesis awareness in Experiment 1.

Tables

Table 1.

Overview of the results of all AA training and AA instruction experiments.

Experiment	Evaluation target	Evaluation task	N	Test statistic	size	Bayes Factor
AA training						
Kawakami et al. (2007, Exp. 1)	Black people	IAT	56	*t*(53) = 2.78, *p* = .007	*d* = 0.76	BF$_1$ = 5.99
Kawakami et al. (2007, Exp. 3)	Black people	IAT	38	*t*(35) = 2.03, *p* = .05	*d* = 0.69	BF$_1$ = 1.53
Phills et al. (2011, Exp. 4)	Black people	IAT	62	*t*(59) = 2.25, *p* = .03	*d* = 0.63	BF$_1$ = 2.10
Current Experiment 1	Turkish people	IAT	64	*t*(61) = 1.72, *p* = .045	*d* = 0.43	BF$_1$ = 1.77
		Warmth rating		*t*(61) = 1.64, *p* = .053	*d* = 0.41	BF$_1$ = 1.57
		Liking rating		*t*(61) = 1.83, *p* = .036	*d* = 0.46	BF$_1$ = 2.09
		SRS		*t*(61) = 1.30, *p* = .10	*d* = 0.33	BF$_1$ = 0.99
		BRS		*t*(61) = 0.80, *p* = .21	*d* = 0.20	BF$_1$ = 0.56
Current Experiment 3 (change goal instructions)	Black people	IAT	492	*t*(489) = 0.28, *p* = .39	*d* = 0.03	BF$_1$ = 0.20
		Warmth rating		*t*(476) = -3.68, *p* < .001	*d* = -0.34	BF$_1$ = 84.70
		Liking rating		*t*(478) = -1.85, *p* = .064	*d* = -0.17	BF$_1$ = 0.80
		MRS		*t*(484) = -0.53, *p* = .60	*d* = -0.05	BF$_1$ = 0.18
Current Experiment 3 (measure goal instructions)	Black people	IAT	492	*t*(489) = 2.96, *p* = .002	*d* = 0.27	BF$_1$ = 18.98
		Warmth rating		*t*(473) = 0.40, *p* = .69	*d* = 0.04	BF$_1$ = 0.18
		Liking rating		*t*(473) = 0.01, *p* = .99	*d* = 0.00	BF$_1$ = 0.16
		MRS		*t*(484) = -0.85, *p* = .40	*d* = -0.08	BF$_1$ = 0.23

Current Experiment 4 (change goal instructions)	Black people	IAT	183	*t*(180) = 2.20, *p* = .015	*d* = 0.33	BF$_1$ = 3.82
		Warmth rating		*t*(173) = -0.99, *p* = .32	*d* = -0.15	BF$_1$ = 0.38
		Liking rating		*t*(171) = 0.75, *p* = .46	*d* = 0.11	BF$_1$ = 0.32
		MRS		*t*(177) = 0.55, *p* = .59	*d* = 0.08	BF$_1$ = 0.29
Current Experiment 4 (measure goal instructions)	Black people	IAT	183	*t*(180) = 1.45, *p* = .075	*d* = 0.11	BF$_1$ = 0.49
		Warmth rating		*t*(173) = 1.27, *p* = .10	*d* = 0.19	BF$_1$ = 0.88
		Liking rating		*t*(172) = 0.98, *p* = .16	*d* = 0.15	BF$_1$ = 0.63
		MRS		*t*(175) = -0.03, *p* = .51	*d* = 0.01	BF$_1$ = 0.25
All current AA training experiments		IAT	1258		*d* = 0.18	BF$_1$ = 65.22
		Warmth rating	1414		*d* = -0.08	BF$_1$ = 0.03
		Liking rating	1414		*d* = -0.04	BF$_1$ = 0.04
		MRS	1350		*d* = -0.04	BF$_1$ = 0.04
All AA training experiments		IAT	1570		*d* = 0.23	BF$_1$ = 4404.42
AA instructions						
Van Dessel et al. (2015, Exp. 3)	Turkish people	IAT	41	*t*(38) = -0.84, *p* = .41	*d* = -0.26	BF$_1$ = 0.19
		Warmth rating		*t*(38) = 1.32, *p* = .19	*d* = 0.42	BF$_1$ = 1.07
		Liking rating		*t*(38) = 1.50, *p* = .14	*d* = 0.47	BF$_1$ = 1.35
Van Dessel et al. (2015, Exp. 5)	Black people	IAT	355	*t*(352) = 0.01, *p* = .99	*d* = 0.00	BF$_1$ = 0.12
		Warmth rating		*t*(352) = 0.16, *p* = .87	*d* = 0.02	BF$_1$ = 0.13
		Liking rating		*t*(352) = 0.05, *p* = .96	*d* = 0.01	BF$_1$ = 0.12
Van Dessel et al. (2015, Exp. 6)	Black people	Priming task	277	*t*(274) = 0.69, *p* = .49	*d* = 0.08	BF$_1$ = 0.25
		Warmth rating		*t*(274) = 0.34, *p* = .73	*d* = 0.04	BF$_1$ = 0.18

		Liking rating		*t*(274) = 1.15, *p* = .25	*d* = 0.14	BF$_1$ = 0.43
Current Experiment 2	Turkish people	IAT	40	*t*(37) = -0.24, *p* = .60	*d* = -0.08	BF$_1$ = 0.31
		Warmth rating		*t*(37) = 0.00, *p* = .50	*d* = 0.00	BF$_1$ = 0.33
		Liking rating		*t*(37) = 0.45, *p* = .33	*d* = 0.16	BF$_1$ = 0.47
		SRS		*t*(37) = -1.37, *p* = .91	*d* = -0.47	BF$_1$ = 0.20
		BRS		*t*(37) = -0.93, *p* = .82	*d* = -0.32	BF$_1$ = 0.24
Current Experiment 4 (change goal instructions)	Black people	IAT	206	*t*(204) = -0.87, *p* = .81	*d* = -0.12	BF$_1$ = 0.14
		Warmth rating		*t*(200) = -0.69, *p* = .49	*d* = -0.10	BF$_1$ = 0.29
		Liking rating		*t*(198) = -0.81, *p* = .42	*d* = -0.11	BF$_1$ = 0.32
		MRS		*t*(203) = -1.27, *p* = .21	*d* = -0.18	BF$_1$ = 0.48
Current Experiment 4 (measure goal instructions)	Black people	IAT	239	*t*(237) = 1.30, *p* = .098	*d* = 0.17	BF$_1$ = 0.83
		Warmth rating		*t*(230) = 1.51, *p* = .13	*d* = 0.20	BF$_1$ = 1.12
		Liking rating		*t*(230) = -1.91, *p* = .057	*d* = -0.25	BF$_1$ = 2.15
		MRS		*t*(236) = 0.91, *p* = .37	*d* = 0.12	BF$_1$ = 0.32
All current AA instruction experiments		IAT	485		*d* = 0.03	BF$_1$ = 0.08
		Warmth rating	476		*d* = 0.06	BF$_1$ = 0.18
		Liking rating	474		*d* = -0.16	BF$_1$ = 0.04
		MRS	444		*d* = -0.02	BF$_1$ = 0.09
All AA instruction experiments		IAT	881		*d* = 0.03	BF$_1$ = 0.07
		Warmth rating	1149		*d* = 0.05	BF$_1$ = 0.16
		Liking rating	1147		*d* = -0.05	BF$_1$ = 0.08

Table 2.

Mean IAT scores in Experiment 4 as a function of AA Condition, Learning Method, IAT Block Order, and Goal Instructions.

	Content of Goal Instructions			
	Reduce prejudice		Measure response speed	
:	Congruent IAT block first	Incongruent IAT block first	Congruent IAT block first	Incongruent IAT block first
AA instructions:				
Control	0.38 (0.47)	0.30 (0.41)	0.37 (0.46)	0.49 (0.40)
Approach Black people	0.38 (0.39)	0.41 (0.50)	0.38 (0.41)	0.34 (0.47)
AA training:				
Control	0.42 (0.43)	0.36 (0.37)	0.37 (0.49)	0.55 (0.39)
Approach Black people	0.20 (0.48)	0.29 (0.50)	0.44 (0.36)	0.25 (0.42)

Note. Standard deviations are in parentheses. Scores reflect a relative preference for White people over Black people.