Relational information moderates approach-avoidance instruction effects on implicit evaluation

Pieter Van Dessel⁎, Jan De Houwer, Colin Tucker Smith

A R T I C L E   I N F O

Keywords: Approach-avoidance Instructions Implicit evaluation Attitudes Propositional theory

A B S T R A C T

Previous research demonstrated that instructions to approach one stimulus and avoid another stimulus can result in a spontaneous or implicit preference for the former stimulus. In the current study, we tested whether the effect of approach-avoidance instructions on implicit evaluation depends on the relational information embedded in these instructions. Participants received instructions that they would move towards a certain non-existing word and move away from another non-existing word (self-agent instructions) or that one non-existing word would move towards them and the other non-existing word would move away from them (stimulus-agent instructions). Results showed that self-agent instructions produced stronger effects than stimulus-agent instructions on implicit evaluations of the non-existing words. These findings support the idea that propositional processes play an important role in effects of approach-avoidance instructions on implicit evaluation and in implicit evaluation in general.

1. Introduction

As Zajonc (1980) argued in his seminal paper, people often evaluate stimuli in a spontaneous manner. Research has shown that such spontaneous or implicit evaluations are an important determinant of behavior (e.g., Greenwald, Pochman, Uhlmann, & Banaji, 2009) and play a crucial role in a number of important psychological phenomena including psychopathology (Roefs et al., 2011), addiction (Wiers & Stacy, 2006), and social interaction (Fazio & Olson, 2003). Hence, understanding how implicit evaluations are acquired and activated is an important aim of psychological science. Cognitive theories of evaluation have traditionally assumed that implicit evaluations reflect the automatic activation of associations between representations in memory (for a review, see Hughes, Barnes-Holmes, & De Houwer, 2011). Because associations are assumed to form automatically when two events co-occur, much research on the acquisition and change of implicit evaluations has employed paradigms in which stimuli are repeatedly paired with valenced stimuli (EC: Hofmann, De Houwer, Perugini, Baeyens, & Crombez, 2010) or with valenced actions (approach-avoidance training: Kawakami, Phillips, Steele, & Dovidio, 2007).

Recent studies, however, have established that changes in implicit stimulus evaluations can occur not only as the result of repeated pairings but also on the basis of mere instructions (De Houwer, 2006; Van Dessel, De Houwer, Gast, & Smith, 2015). For example, studies on the effects of approach-avoidance (AA) instructions have shown that participants who are instructed to approach certain stimuli and avoid other stimuli exhibit more positive implicit evaluations of to-be-approached stimuli than of to-be-avoided stimuli even if they never actually perform the AA actions. There is even evidence that these instruction-based effects on implicit evaluation can occur under certain conditions of automaticity. For instance, AA instructions influence implicit evaluations even when participants do not consider the acquired information a valid basis for their evaluation (as indicated by the fact that they do not incorporate this information in their explicit evaluation; Van Dessel, De Houwer, Gast, Smith, & De Schryver, 2016).

Effects of AA instructions on implicit evaluation pose a challenge to a particular type of associative models that assume that (a) implicit evaluations reflect the automatic activation of associations in memory and (b) these associations are formed as the result of a slow-learning process that capitalizes on repeated co-occurrences (Rydell & McConnell, 2006; Smith & DeCoster, 2000). Yet, instruction-based AA effects are consistent with propositional models, which assume that propositions, rather than associations, guide implicit evaluation (e.g., De Houwer, 2009, 2014; Mitchell, De Houwer, & Lovibond, 2009). When participants are instructed to approach or avoid a stimulus, they might generate propositions about
these stimulus-action relations, and these propositions can influence their implicit evaluations of the stimuli (Van Dessel et al., 2016). For example, changes in implicit evaluations may occur as the result of AA instructions when participants infer that to-be-approached stimuli are more positive than to-be-avoided stimuli (e.g., because they know that people typically approach good things and avoid bad things) and the automatic retrieval of this propositional information influences implicit evaluation.

Importantly, a propositional model of implicit evaluation not only predicts that implicit evaluations can form as the result of a single instruction, but also that these effects should depend on the relational information embedded in these instructions (De Houwer, 2014). Propositions store information not only about the strength of the relationship between concepts but also about the nature of the relation (e.g., ‘I approach Stimulus A’; see Shanks, 2007). If propositions mediate implicit evaluation, then changes in implicit evaluation (e.g., due to instructions) could depend on the relational meaning of the acquired information (stored as propositions). Hence, instructions that contain the same concepts (e.g., ‘approach’ and ‘Stimulus A’) might produce dissimilar effects on implicit evaluations if those concepts are related in a different manner.

We tested this prediction by giving participants instructions that differed not in the pairing of concepts (e.g., the stimulus and the AA action word), but in the agency relation specified in the instructions (i.e., who performs the AA action). Half of the participants received typical AA instructions which stated that the participant would perform a specific AA action in relation to a specific stimulus (i.e., move towards or away from a non-existing word). The other half of the participants received instructions which stated that the stimulus (i.e., the non-existing word) would perform the AA action in relation to the participant. We refer to the former instructions as ‘self-agent instructions’ and to the latter as ‘stimulus-agent instructions’. Immediately following these instructions, participants’ implicit evaluations of the two stimuli were registered with an Implicit Association Test (IAT; Greenwald, McGhee, & Schwartz, 1998).1 We examined whether the two types of instructions have a differential impact on implicit stimulus evaluations.

From the perspective of a propositional account, instruction effects on implicit evaluation should depend on the extent to which instructions allow for the acquisition of the propositional information that a specific stimulus is positive or negative. Participants may easily infer that the stimuli they approach are more positive than the stimuli they avoid because this is consistent with their previous learning history (i.e., most often, positive stimuli are approached and negative stimuli are avoided; see also Van Dessel et al., 2016). However, it is less certain that participants infer that an approaching stimulus is more positive than an avoiding stimulus because typically both pleasant and unpleasant stimuli can approach or avoid a person (see also Hsee, Tu, Lu, & Ruan, 2014). Hence, from the perspective of this propositional account, there are good reasons to predict that self-agent instructions should more strongly influence implicit evaluations than stimulus-agent instructions. As we will discuss in more detail later on, such a result would not only reveal an important moderator of AA instruction effects but would also have implications for theoretical accounts of those effects.

---

1 AA instruction effects have been observed on a number of implicit and explicit evaluation measures (see Van Dessel et al., 2015). The current study uses the IAT because this is currently the most widely used method to measure implicit evaluations. The IAT captures implicit evaluations in the sense that it registers evaluative responses under conditions that are typically associated with automatic processes (e.g., under time pressure, in the absence of evaluation goals; ...; see De Houwer, Teige-Mocigemba, Spruyt, & Moors, 2009).

2. Method

2.1. Participants and design

A total of 1306 English-speaking volunteers participated online via the Project Implicit research website ([https://implicit.harvard.edu](https://implicit.harvard.edu)). We stopped the data-collection when at least 1000 participants had completed all measures of the experiment to ensure that we would have sufficient statistical power to detect even small effects after data-exclusion (power > 0.80 to detect an effect size of $d = 0.20$). All data were collected in one shot without intermittent data analysis. Overall dropout rate was 29.5%. The dropout rates were comparable across the two conditions: 30.8% in the self-agent condition and 28.1% in the stimulus-agent condition, $\chi^2(1) = 1.20, p = 0.27$. Hence, there was no evidence for condition-dependent attrition.

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 (190 participants; i.e., 14.6%), (b) had error rates above 30% when considering all IAT test blocks or above 40% for any one of the critical IAT test blocks (31 participants; i.e., 2.4%), or (c) responded faster than 400 ms on > 10% of the IAT trials (29 participants; i.e., 2.2%). Analyses were performed on the data of 1056 participants (653 women, mean age = 33, SD = 14). Table 1 provides the number of included and excluded participants in each of the experimental conditions. The proportion of excluded participants did not differ significantly between conditions, $\chi^2(3) = 3.26, p = 0.45$. Note that including the data from all participants in the analyses did not result in any shift in significance for any of the reported effects. A full description of these results can be found at [https://osf.io/d3tpj/](https://osf.io/d3tpj/). At this online repository we also provide a link to the online study as well as all data of the study and data analysis scripts.

2.2. Procedure

Upon being assigned to this study, participants were informed that they would participate in an experiment that would involve two meaningless words: UDIBNON and BAYRAM. Half of the participants then read the self-agent instructions:

You will perform a task in which you will move towards BAYRAM and you will move away from UDIBNON. It is very important to remember which action belongs to which word. You will need this information to complete the task successfully. Later on we will explain to you exactly how you will be able to perform this task. For now, it is crucial that you remember that you will move towards BAYRAM and move away from UDIBNON. Before we present these words and start the task, you will complete a categorization task. This will last about 5 min. Make sure that during that task you do not forget the instructions of the next task. Please press ‘Continue’ when you have memorized the instructions and are ready to begin the categorization task.

The other half of the participants read the stimulus-agent instructions, which were identical with the exception of the two sentences that specified the agency relation. These sentences now indicated that the participant would perform the AA action in relation to the stimulus (e.g., ‘you and UDIBNON will move away from you and BAYRAM will move towards you’). Participants were prompted to remember this information with the following sentence: “For now, it is crucial that you remember that BAYRAM will move towards you and UDIBNON will move away from you”. Note that the assignment of the words to the approach or avoidance action was counterbalanced across participants and across instruction conditions.

The reaction time task that followed was an IAT in which participants categorized attribute words as ‘positive’ or ‘negative’ and target words UDIBNON and BAYRAM as ‘Udibnon’ or ‘Bayram’. To
avoid that the target stimuli were classified only on the basis of simple perceptual features, the words were presented in eight different combinations of font types (Arial Black and Fixedsys), capitalizations (uppercase and lowercase), and size (16 pt and 18 pt), resulting in 8 different stimuli (also see Zanon, De Houwer, Gast, & Smith, 2014). The IAT consisted of three practice blocks and two experimental blocks. Participants began the IAT with 20 practice trials sorting the target words and 20 practice trials sorting positive and negative stimuli. Next, participants completed 56 trials in which UDIBNON and positive stimuli shared a single response key and BAYRAM and negative stimuli shared a single response key (half of the participants completed the IAT in this way, while the other participants began by sorting BAYRAM and positive with the same key). Participants then practiced sorting target words with the response key assignment reversed for 40 trials and finally participants completed a second set of 56 trials in which UDIBNON shared a response key with negative and BAYRAM shared a response key with positive (or vice versa). If 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.

After the implicit evaluation task, participants rated their liking of each of the words by answering two questions for each word: “To what extent do you like BAYRAM/UDIBNON?” and “To what extent do you have warm feelings for BAYRAM/UDIBNON?” Participants gave their ratings on 9-point Likert scales ranging from 1 (not at all warm; like not at all) to 9 (completely warm; like completely). The ratings were aggregated into a single score for each word by averaging the respective scores (mean Cronbach’s Alpha = 0.80, SD = 0.01).

Next, participants were asked to complete a manipulation check for each word. Participants who had received self-agent instructions were asked what they would have to do according to the instructions when seeing the word UDIBNON or the word BAYRAM. Participants answered by selecting one of three options of a dropdown menu with “move towards the word”, “move away from the word” and “I don’t remember” as possible answers. Participants who had received stimulus-agent instructions were asked what would happen according to the instructions when seeing the word UDIBNON or the word BAYRAM. They answered by selecting one of three options: “the word would move towards me”, “the word would move away from me” or “I don’t remember”.

After completion of the manipulation check, participants were asked to report to what extent they had tried to form a mental image of the task in order to help them remember the instructions of this task. Participants answered by selecting a number between 0 (not at all) and 5 (to a great extent). Participants then indicated whether they were familiar with the words UDIBNON or BAYRAM (because they had previously participated in experiments that used these words as stimuli or because they were familiar with these words as instances of Turkish language) by selecting ‘yes’ or ‘no’ from a dropdown menu. Finally, participants were informed that it was not necessary to complete the previously instructed task and they were thanked for their participation.

3. Results

3.1. IAT scores

IAT scores were calculated using the D2-algorithm (Greenwald, Nosek, & Banaji, 2003), such that higher scores indicate a stronger preference for BAYRAM over UDIBNON. The Spearman-Brown corrected split-half reliability of the evaluative IAT score, calculated on the basis of an odd-even split, was r(1054) = 0.85. Across groups, participants displayed a small implicit preference for BAYRAM over UDIBNON (M = 0.05, SD = 0.48), t(1055) = 3.65, p < 0.001, d = 0.11. We performed a 2 (Type of AA Instructions: self-agent instructions, stimulus-agent instructions) × 2 (Content of AA Instructions: approach BAYRAM and avoid UDIBNON, approach UDIBNON and avoid BAYRAM) × 2 (Instruction Memory: no errors on questions that probed memory for the AA instructions: 77.2%, at least one error: 22.8%) analysis of variance (ANOVA) on the IAT scores. To account for the unbalanced design, we used type III sums of squares in this and all subsequent statistical analyses. The ANOVA revealed a main effect of Instruction Memory, F(1,1048) = 13.19, p < 0.001, η² = 0.010, indicating that participants preferred BAYRAM more when they had correct memory of the AA instructions, and a significant interaction of Instruction Memory and Content of AA Instructions, F(1,1048) = 28.00, p < 0.001, η² = 0.022. Importantly, we also observed a significant three-way interaction effect of Instruction Memory, Content of AA Instructions, and Type of AA Instructions, F(1,1048) = 4.29, p = 0.038, η² = 0.004. No other main or interaction effects were observed, Fs < 3.07, ps > 0.080, η²s < 0.001.

To examine the three-way interaction, we performed separate t-tests for participants with correct and incorrect memory of the instructions in the self-agent and stimulus-agent instruction conditions. In line with previous results (Van Dessel et al., 2015), participants who had received self-agent instructions and who had correct memory of these instructions preferred BAYRAM more when they had received instructions to move towards BAYRAM and move away from UDIBNON (M = 0.32, SD = 0.44) than when they had received instructions to move away from BAYRAM and move towards UDIBNON (M = −0.23, SD = 0.46), t(393) = 12.03, p < 0.001, d = 1.21, 95% confidence interval of the difference (CI diff) [0.46, 0.64]. Participants who had received stimulus-agent instructions and who had correct memory of the instructions preferred BAYRAM more when they had received instructions that BAYRAM would move towards them and UDIBNON would move away from them (M = 0.18, SD = 0.45) than when they had received instructions that BAYRAM would move away from them and UDIBNON would move towards them (M = −0.12, SD = 0.44), t(418) = 6.84, p < 0.001, d = 0.67, 95% CI diff [0.21, 0.38]. As expected on the basis of the propositional account, the effect of stimulus-agent instructions was significantly stronger than the effect of stimulus-agent instructions, F(1811) = 15.92, p < 0.001, dagg = 0.54. According with previous studies (e.g., Van Dessel et al.), there were no significant effects for participants who did not correctly remember the instructions, ts < 0.56, ps > 0.57, ds < 0.10.

We also performed a linear regression analysis to investigate main and interaction effects of the two between-subjects categorical variables and the continuous predictor variable of Mental Simulation (i.e., participants’ response to the mental simulation question: M = 2.48, SD = 1.71). The three-way interaction effect of Content of AA Instructions, Type of AA Instructions, and Instruction Memory remained significant, F(1,1037) = 4.18, p = 0.041, η² = 0.002, and we did not observe any main or interaction effects involving the Mental Simulation variable, Fs < 3.41, ps > 0.064, η²s < 0.001. Moreover, an ANOVA on the data of 164 participants with correct instruction

---

Table 1

<table>
<thead>
<tr>
<th></th>
<th>Self-agent instructions</th>
<th>Stimulus-agent instructions</th>
</tr>
</thead>
<tbody>
<tr>
<td></td>
<td>Approach BAYRAM</td>
<td>Approach UDIBNON</td>
</tr>
<tr>
<td></td>
<td>Approach BAYRAM</td>
<td>Approach UDIBNON</td>
</tr>
<tr>
<td>Included</td>
<td>263 (20.1%)</td>
<td>230 (17.6%)</td>
</tr>
<tr>
<td></td>
<td>297 (22.7%)</td>
<td>266 (20.4%)</td>
</tr>
<tr>
<td>Excluded</td>
<td>69 (5.3%)</td>
<td>62 (4.8%)</td>
</tr>
<tr>
<td></td>
<td>58 (4.4%)</td>
<td>61 (4.7%)</td>
</tr>
<tr>
<td></td>
<td>332 (25.4%)</td>
<td>292 (22.4%)</td>
</tr>
<tr>
<td></td>
<td>355 (27.2%)</td>
<td>327 (25.0%)</td>
</tr>
</tbody>
</table>

---

2. Excluding the data from the 36 participants (i.e., 3.4% of the entire sample) who indicated that they were familiar with the words had no significant influence on any of the analyses reported and thus their data were retained.
memory who had indicated that they had not formed a mental image of the task (response on the mental simulation question = 0) still produced a main effect of Content of AA Instructions, $F(11160) = 23.65, p < 0.001, \eta^2 = 0.129$. In the analysis of this strongly reduced data set, the interaction only showed a weak trend towards significance, $F(1160) = 2.11, p = 0.149, \eta^2 = 0.011$. Both self-agent and stimulus-agent instructions caused significant changes in implicit evaluations (self-agent instructions: $t(78) = 4.10, p < 0.001, d = 0.92$, 95% CI $[0.24, 0.70]$; stimulus-agent instructions: $t(82) = 2.70, p = 0.009, d = 0.59$, 95% CI $[0.07, 0.44]$).3

3.2. Process dissociation

To examine whether (self-agent and/or stimulus-agent) instructions can influence automatic components of IAT performance we also decomposed IAT performance into controlled and automatic components using Process Dissociation Procedures (PDP; Jacoby, 1991). This question is important because IAT scores are known to be sensitive to controlled processes and might therefore not always provide a pure index of implicit evaluations (De Houwer, Beekers, & Moors, 2007). It has been argued that PDP can be used to separate controlled and automatic components of IAT performance (Conrey, Sherman, Gawronski, Hugenberg, & Groom, 2005). While realizing that also PDP have limitations, we opted to employ PDP in the current research in order to provide an additional, more stringent test of the idea that (certain types of) AA instructions can influence implicit evaluations (i.e., in addition to implicit measures).

In accordance with the procedures outlined by Payne (2001), we calculated automatic parameter estimates for IAT performance on the basis of participants’ errors in the IAT (average error rate = 6.7%, $SD = 5.1\%$). The automatic parameter estimate provides an indication of the extent to which participants’ performance indicates that participants automatically evaluate BAYRAM as positive and UDIBNON as negative or vice versa. Estimates above 0.50 indicate that participants automatically respond more in accordance with the former knowledge. An ANOVA on automatic IAT parameters revealed a main effect of Content of AA Instructions, $F(1,1048) = 19.82, p < 0.001, \eta^2 = 0.02$, and an interaction effect of Content of AA Instructions and Type of AA Instructions, $F(1,1048) = 4.22, p = 0.040, \eta^2 = 0.001$. The three-way interaction effect of Content of AA Instructions, Type of AA Instructions, and Instruction Memory was marginally significant, $F(1,1048) = 3.73, p = 0.054, \eta^2 = 0.003$. For participants with correct memory and self-agent instructions, the automatic parameter estimates were significantly larger if they had received instructions to move towards BAYRAM and move away from UDIBNON ($M = 0.51, SD = 0.01$) than if they had received instructions to move away from BAYRAM and move towards UDIBNON ($M = 0.50, SD = 0.01$), $t(393) = 5.13, p < 0.001, d = 0.52$, 95% CI diff $[0.004, 0.010]$. For participants with correct memory and stimulus-agent instructions, we observed only a marginally significant difference in automatic parameter estimates between participants with opposite instructions (BAYRAM moves towards you and UDIBNON moves away from you: $M = 0.50, SD = 0.01$; BAYRAM moves away from you and UDIBNON moves towards you: $M = 0.50, SD = 0.01$), $t(418) = 1.76, p = 0.079, d = 0.17$, 95% CI diff $[-0.003, 0.006]$. There were no significant effects for participants with incorrect memory of the instructions, $tS < 0.69$, $ps > 0.49$, $dS < 0.12$.

4. General discussion

The current experiment was designed to test whether the effects of AA instruction on implicit evaluations depend on the relational information specified in the instructions. Results showed that instructions to move towards or away from certain non-existing words (i.e., self-agent instructions) caused bigger changes in implicit evaluations of these words than instructions in which the words were said to move towards or away from the participant (i.e., stimulus-agent instructions). PDP analyses indicated that only the former type of instructions had a significant effect on the automatic parameter of IAT performance, suggesting that only these typical AA instructions caused genuine changes in implicit evaluations. These results provide the first evidence that instructions that contain the same concepts produce different effects on implicit evaluations if those concepts are related in a different manner.

Our results are consistent with, and predicted by, a propositional account of AA instruction effects. From the perspective of a propositional account, participants may infer on the basis of AA instructions that to-be-approached stimuli are more positive than to-be-avoided stimuli. Because such inferences are less certain when instructions specify that it are the stimuli that approach or avoid the participant, self-agent instructions should produce bigger effects on implicit evaluations than stimulus-agent instructions. The observation that stimulus-agent instructions also had a significant effect on implicit evaluations, albeit to a lesser extent than self-agent instructions, was not predicted a priori by the propositional account. On the one hand, this result has to be interpreted with some caution because we did not observe a significant impact of stimulus-agent instructions on the automatic component of IAT performance as indexed using PDP. This suggests that the effects of stimulus-agent instructions might more strongly depend on controlled, non-automatic processes that involve the intentional use of the acquired information for evaluation (e.g., as the result of demand compliance). On the other hand, the mere presence of such an effect could be explained by the propositional account in a post-hoc fashion. One possible post-hoc explanation for this result is that people believe to some extent that stimuli that move towards them are typically more positive than stimuli that move away from them. Such a belief might be based on certain concrete experiences (e.g., receiving positive things such as gifts or money and removing negative things such as garbage or dirt). This post-hoc account could be tested in future research by explicitly asking participants to report their beliefs on this issue or by manipulating the extent to which participants are likely to believe that approaching stimuli are more positive than stimuli that move away from them.

Whereas the moderating effect of relational information on AA instructions effects was predicted on the basis of propositional models of implicit evaluation, it raises questions about specific types of associative models. First, our results do not fit well with associative accounts of these effects which assume that (automatic) changes in
Propositional and associative processes are involved in AA instructions effects on implicit evaluations. For reasons of parsimony, a propositional account that can explain these results is to be preferred over an account that additionally postulates the existence of an entirely different second mechanism (i.e., association formation).

Regardless of the merits of the described explanations, we can conclude that our results put new and important constraints on any current and future explanation of AA instruction effects. They support the contribution of propositional processes to AA instruction effects on implicit evaluations and rule out specific associative explanations. Of course, our results cannot distinguish between the broad class of single-process propositional and the broad class of associative or dual-process models. Because these models have a high degree of flexibility, distinguishing between these models on the basis of a single set of data is difficult, if not impossible. However, the current results do allow us to further constrain evaluation models and to have greater confidence in predictions derived from these models. The continuing investigation of the effects of AA instructions on implicit evaluation may help to further elucidate the mechanisms involved in (implicit) evaluation.

Acknowledgments

Funding: Pieter Van Dessel is supported by a Postdoctoral fellowship of the Scientific Research Foundation, Flanders (FWO-Vlaanderen). Jan De Houwer is supported by Methusalem Grant BOF16/MET_V/002 of Ghent University and by the Interuniversity Attraction Poles Program initiated by the Belgian Science Policy Office (IUAPVII/33).

Appendix A

Data-analyses on explicit rating scores

We observed a moderate correlation of explicit rating scores with evaluative IAT scores, $r(1054) = 0.35$, $p < 0.001$. Similar to the results of the IAT, explicit rating scores revealed a significant preference for BAYRAM over UDIBNON ($M = 0.67$, $SD = 2.22$), $r(1055) = 9.83$, $p < 0.001$, $d = 0.30$. The ANOVA on explicit rating scores revealed a main effect of Content of AA Instructions, $F(1,1048) = 48.11$, $p < 0.001$, $\eta^2 = 0.004$, an interaction of Content of AA instructions and Instruction Memory, $F(1,1048) = 30.20$, $p < 0.001$, $\eta^2 = 0.026$, and a three-way interaction of Content of AA instructions, Type of AA instructions and Instruction Memory $F(1,1048) = 5.60$, $p = 0.015$, $\eta^2 = 0.005$. Participants with correct instruction memory who had received self-agent instructions preferred BAYRAM more when they received instructions to approach BAYRAM and avoid UDIBNON ($M = 1.45$, $SD = 2.19$) than when they received instructions to avoid BAYRAM and approach UDIBNON ($M = -0.24$, $SD = 2.36$), $t(393) = 7.22$, $p < 0.001$, $d = 0.73$, 95% CI $\text{diff} [0.21, 2.11]$. Participants with correct instruction memory who had received stimulus-agent instructions preferred BAYRAM more when they received instructions that BAYRAM would avoid them and UDIBNON would avoid them ($M = 1.13$, $SD = 2.03$) than when they received instructions that BAYRAM would avoid them and UDIBNON would approach them ($M = 0.14$, $SD = 2.18$), $t(418) = 4.85$, $p < 0.001$, $d = 0.47$, 95% CI $\text{diff} [0.59, 1.40]$. Importantly, similar to IAT scores, the effect of self-agent instructions on explicit ratings was significantly stronger than the effect of stimulus-agent instructions, $F(1,811) = 4.61$, $p = 0.032$, $\eta^2 = 0.005$.

An ANOVA that included Mental Simulation as covariate revealed a main effect of Content of AA Instructions, $F(1,1037) = 46.58$, $p < 0.001$, $\eta^2 = 0.063$, and an interaction of Content of AA instructions and Instruction Memory, $F(1,1037) = 28.12$, $p < 0.001$, $\eta^2 = 0.006$. This interaction was qualified by an interaction of
Content of AA instructions, Instruction Memory, and Mental Simulation, $F(1,1037) = 10.17, p < 0.001, \eta^2 = 0.009$. Participants with correct instruction memory exhibited a larger AA instruction effect when they had higher mental simulation scores, $F(1,804) = 4.87, p = 0.028, \eta^2 = 0.005$, whereas participants with incorrect instruction memory exhibited a smaller AA instruction effect when they had higher mental simulation scores, $F(1,233) = 7.10, p = 0.008, \eta^2 = 0.024$. Importantly, the three-way interaction of Content of AA instructions, Type of AA instructions and Instruction Memory remained significant, $F(1,1037) = 5.78, p = 0.016, \eta^2 = 0.004$, and we observed no other interaction effects with Mental Simulation, $F_S < 0.97, ps > 0.32, \eta^2's < 0.001$. Participants who indicated that they had not formed a mental image of the task still exhibited an AA instruction effect if they received self-agent instructions, $t(78) = 2.11, p = 0.038, d = 0.47, 95\% CI [0.06, 1.88]$, but not if they received stimulus-agent instructions, $t(82) = 0.85, p = 0.40, d = 0.19, 95\% CI [-0.47, 1.17]$. The interaction of Content of AA instructions and Type of AA instructions, however, was not significant, $F(1,160) = 1.00, p = 0.319, \eta^2 = 0.006$.

Mediation analyses

To investigate the relationship between AA instruction effects on implicit and explicit evaluations, we performed mediation analyses with the lavaan package in R (version 0.5-16; Rosseel, 2012). We used the bootstrap method to estimate standard errors for the effects. We tested whether changes in explicit rating scores mediate the effect of self-agent and stimulus-agent AA instructions on IAT scores (see Figs. A1, A2). Towards this end, IAT scores were simultaneously regressed on both AA instructions (approach BAYRAM and avoid UDIBNON, approach UDIBNON and avoid BAYRAM) and explicit rating scores (Baron & Kenny, 1986).

For participants in the self-agent instructions condition, the indirect effect of AA instructions on IAT scores with explicit rating scores as a mediator was also statistically significant, $\beta = 0.12, Z = 3.12, p = 0.009, 95\% CI of \beta = [0.03, 0.18], R^2_{dir} = 0.08$.

For participants in the stimulus-agent instructions condition, the indirect effect of AA instructions on IAT scores with explicit rating scores as a mediator was also statistically significant, $\beta = 0.09, Z = 3.45, p < 0.001, 95\% CI of \beta = [0.07, 0.13], R^2_{dir} = 0.10$. The effect of AA instructions on the IAT score remained statistically significant after controlling for changes in explicit evaluations, $\beta = 0.12, Z = 3.12, p = 0.009, 95\% CI of \beta = [0.03, 0.18], R^2_{dir} = 0.08$.

References


