RUNNING HEAD: Distractors On Implicit-Explicit Dissociation

Diverting the mind from the Self-Referencing effect. Which interference leads to implicit-explicit attitude dissociation?

Simone Mattavelli, Juliette Richetin & Marco Perugini

University of Milan-Bicocca

Corresponding author: Simone Mattavelli, University of Milano-Bicocca,

Department of Psychology,

1, Piazza dell'Ateneo Nuovo, 20126 Milan, ITALY.

Mail to: simone.mattavelli@unimib.it

Abstract

In evaluative learning, changes in implicit evaluations do not always result in explicit. The Self-Referencing (SR) task is an associative learning paradigm that relies on intersecting regularities and self-positivity to transfer valence towards target objects. A recent meta-analysis documented its effectiveness in changing both implicit and explicit attitudes. This contribution tests how interfering elements between the implicit (IAT) and the explicit attitude measures qualify the SR effect on the latter. Study 1 (n = 163) showed that distraction tasks disrupting the procedural flow from implicit to explicit attitude did not lead to implicit-explicit dissociation in the occurrence of SR effect, regardless of structural overlap of the distractor with the IAT. In study 2 (n = 236), the SR effect on explicit attitude was qualified by the content of the distractor. The SR effect occurred on both implicit and explicit attitudes when the distractor described the characteristics of the IAT as a measure, but only on implicit when participants were told that the IAT revealed their cognitive abilities. We discussed the contribution of these findings to extant interpretations of implicit-explicit dissociation.

Diverting the mind from the Self-Referencing effect. Which interference leads to implicit-explicit attitude dissociation?

In the last two decades, the distinction between implicit and explicit attitudes has become a major theme in social and experimental psychological research (Greenwald & Banaji, 1995). Explicit attitudes are usually equated with deliberative, self-reported evaluations. Implicit attitudes are inferred from people's performance on cognitive tasks, such as the Implicit Association Test (IAT, Greenwald, McGhee, & Schwartz, 1998), affective priming (Fazio, Jackson, Dunton, & Williams, 1995), semantic priming (Wittenbrink, Judd, & Park, 1997), the go/no-go association task (Nosek & Banaji, 2001), the Extrinsic Affective Simon Task (De Houwer, 2003), and the Affect Misattribution Paradigm (Payne, Cheng, Govorun, & Stewart, 2005). Such measures of implicit attitudes are based on the concept of automaticity and on the idea that attitudes can be treated as the expressions of automatic processes that occur spontaneously and outside awareness or control (Moors & De Houwer, 2006). As the distinction between implicit and explicit attitudes has gained increasing relevance, understanding how attitude change interventions can affect them is important.

Among the associative pathways that lead to implicit and explicit attitude formation or change, the most known is evaluative conditioning (EC). EC refers to the change (or the formation) of an evaluative preference due to the pairing between a conditioned stimulus (CS) and an unconditioned stimulus (US), and this effect results in the transfer of valence from the US to the paired CS (De Houwer, 2007). Empirical evidence supports the efficacy of EC procedures on attitude change (see Hofmann, De Houwer, Perugini, Baeyens, & Crombez, 2010). A different, though related, type of evaluative learning pathway, namely *intersecting regularities*, has been proposed by Hughes, De Houwer, and Perugini (2016). Hughes and colleagues (2016) demonstrated that when a positive stimulus shares an element with a neutral object (e.g., the appearance of the same outcome symbol on-screen upon the correct categorization of both stimuli), the valence carried by the former is acquired by the latter (see also Ebert, Steffens, von Stülpnagel, & Jelenec, 2009). The Self-Referencing (SR) task (Perkins & Forehand, 2012; Prestwich, Perugini, Hurling, &

Richetin, 2010) is a prime example of an associative learning paradigm based on intersecting regularities. The SR task additionally relies on the positivity of the self (Yamaguchi et al., 2007) to generate attitudinal change. Throughout the task, participants perform a common action for categorizing stimuli related to either the self or a target object (i.e., Target A) and an alternative common action for the categorization of stimuli belonging to either the category 'Others' or another target object (i.e., Target B). Because the same action is performed in response to both Target A and the self (*intersecting regularities*), the former can acquire the positive valence carried by the latter. A recent meta-analysis showed that the SR task is effective in changing both implicit and explicit attitudes (Mattavelli, Richetin, Gallucci, & Perugini, 2017).

Dissociation between implicit and explicit attitude change[1]

In the literature, there is ample evidence of the effectiveness of both EC and intersecting regularities in changing both implicit and explicit attitudes. In their Associative-Propositional Evaluation (APE) model, Gawronski and Bodenhausen (2006, 2011, 2014) provided a comprehensive theoretical framework for implicit and explicit attitude change. The APE postulates that when the nature of the intervention is more likely to affect evaluative associations (e.g., EC and intersecting regularities), these latter serve as an input for deliberative evaluations through a validation process. However, in the last decade, abundant evidence within the EC research has shown that just like explicit, also implicit attitudes can be affected by non-automatic processes (De Houwer, 2006; Gast & De Houwer, 2013; Kurdi & Banaji, 2017). It is therefore important to test the specific conditions under which changes in implicit attitudes driven by an associative learning manipulation either do or do not correspond to changes in explicit attitudes. Research has shown parallel effects for associative evaluations and evaluative judgments (e.g., Hermans, Vansteenwegen, Crombez, Baeyens, & Eelen, 2002; Olson & Fazio, 2001). Olson and Fazio (2001), for example, found corresponding influences of EC on both explicit and implicit attitudes, with the

[1] It should be noted that by "dissociation", we refer to dissociation in attitude change, that is, the presence versus absence of the effect of the same learning procedure on either implicit or explicit outcome measures.

two being highly correlated and the effect on explicit being fully mediated by the implicit change (see also Richetin, Mattavelli, & Perugini, 2016 for similar findings on the SR effect). Critically, however, several studies show that under certain circumstances, attitude change interventions can impact on implicit attitudes without affecting explicit attitudes (e.g., Karpinski & Hilton, 2001, Study 3; Olson & Fazio, 2006, Gawronski & Lebel, 2008). For instance, Gawronski and Lebel (2008, Study 2) showed that repeated pairings of CS with positive or negative US influenced an explicit measure of CS evaluations only when participants focused on their feelings regarding the CS, but not when they focused on their knowledge about the CS.

A similar dissociative pattern in implicit and explicit attitude change has also been found with the SR task. For instance, Perugini, Richetin, and Zogmaister (2014) showed that an effect of the SR manipulation on explicit attitudes was more likely to emerge when the explicit measure was taken after the implicit, whereas the order of administration did not impact the effect of the SR task on the implicit measure. This suggests a greater change for implicit and explicit evaluations when the implicit measure is completed before, as compared to after, the explicit measure. The idea of a sequence in measurement administration that maximizes the chance for the SR task to impact both implicit and explicit attitudes may reflect the flow of some processes undergoing the manipulation, suggesting that the implicit measure might act as a signal for a change in evaluation to be elaborated explicitly.

Even when the outcome measures are administered in an order that maximizes the chance to observe an effect on both implicit and explicit attitudes (i.e., self-report following implicit attitude measure), the extent to which the SR task influences both can be conditional on elements of interference, such as distractors. In the persuasion literature, it has been demonstrated that distraction interferes with the processing of a message. For instance, distraction can reduce the impact of weak versus strong persuasive messages (Petty, Wells, & Brock, 1976; Petty & Brock, 1981; Bless, Bohner, Schwarz, & Strack, 1990). Following a similar logic, the effect of an associative learning manipulation, like the SR task, on explicit attitude might be more difficult to

occur when a distracting element intervenes between the assessment of the implicit and the explicit measures.

Distractions

If distracting elements in the environment can interfere with processing propositionally the newly acquired associations, knowing if all distractions are alike or if some are more disruptive than others is an important issue. We focus on two specific features that might prevent individuals from coming up with propositions that validate the newly formed association, that is, the procedure and the content of the distraction.

At the procedural level, one might argue that the impact of a distracting element on the relation between implicit and explicit attitude change might vary as a function of its relevance to the critical task (i.e., the IAT). If the IAT functions as a signal to one's preference, then performing a distraction task characterized by high or low overlap with the IAT could lead, respectively, to greater or lower dissociation between implicit and explicit attitude change.

At the content level, the type of information characterizing the distraction can have different dissociative effects between implicit and explicit attitudes depending on how much it shifts the focus away from further processing the newly acquired stimuli relations. For example, information concerning what the implicit measure reveals about one's implicit attitude should facilitate further processing; hence, everything else being equal, resulting in consistency between implicit and explicit attitudes. On the contrary, information regarding what the implicit attitude measure reveals about oneself (e.g., one's cognitive skills) should impede further processing as the focus is likely to be shifted away from processing the acquired relationships, hence resulting in dissociation in the SR effect between implicit and explicit attitudes. Finally, the disruptive effect of the content of the information might be qualified further by the direction of the information. For instance, when informed that the implicit attitude measure reflects a personal attitude, one might qualify this information by adding whether it is a more versus less valid measure. In the same vein, if the

implicit attitude measure is presented as revealing one's cognitive skills, this information can be qualified by adding whether it has revealed relatively high versus low cognitive skills.

To summarize, distracting away from newly acquired relationships between stimuli might prevent explicit validation, but not all distractions are likely to be disruptive in the same way. It is reasonable to expect that only some of them might be effective to the extent to which they shift the focus away from further processing the acquired relationships between stimuli.

The present research

The present research aims to test whether the impact of the SR task on implicit and explicit attitude change can be interfered by distractors, therefore creating dissociation between implicit and explicit attitude change. There are different ways in which distractors placed between the implicit and the explicit outcome can lead to dissociation. For instance, distractors might disrupt the validation process through which new relations between concepts revealed by the IAT (i.e., one target object and positive stimuli) are validated propositionally. This should result in a decreased implicit-explicit relation, ultimately leading to dissociation. Alternatively, distractors might create dissociation by simply diverting one's mind from the source of the new relations revealed by the IAT, that is, the fact that the target stimulus shares something with the self.

Prior research on associative learning has focused on the role of distractors at the 'encoding' stage, that is, during the processing of the associative learning procedure (Dawson, Rissling, Schell, & Wilcox, 2007; Field & Moore, 2005). However, no studies tested the impact of distractors at the 'elaboration' stage, that is, between the acquisition of a novel set of stimuli relationships and their explicit expression. Therefore, we present two studies in which different types of distractors are operationalized. In the first study, the distractor was based on a task between the IAT and the measure of explicit attitude. Specifically, we inserted two types of distractors, manipulating their procedural overlap with the critical implicit attitude measure. The first distraction task consisted of an entirely different kind of behavior and cognitive processes than the ones involved in the IAT, whereas the second involved the same behavior and cognitive processes, that is, another IAT albeit

on different targets. In the second study, we manipulated the content of the distractor (i.e., the type of information provided). Participants were assigned randomly to one of four conditions that all provided additional information about the IAT potentially interfering with the transmission of evaluative change from implicit to explicit evaluations. We used two sets of information that focused on the implications of the IAT about oneself (i.e., one's cognitive skills), whereas the other two sets centered on the implications of the IAT concerning one's implicit preference towards the target object. Furthermore, within each set of information, the type of feedback was varied to inspect whether the direction of the information (valid/positive vs. invalid/negative) further qualified the effects. In both Study 1 and 2, the order of administration of the attitude measures was kept constant, with the implicit measure preceding the explicit. We did so because we were interested in one specific implicit-explicit dissociation path, that is, the one that consists of the SR task affecting implicit, but not explicit, attitude.

Based on previous results (Perugini et al., 2014, Study 1), introducing an unrelated task between the measurement of implicit and explicit attitudes should not cause any implicit-explicit dissociation. However, whether the extent to which the distractor mirrors the implicit measure procedurally determines variations in the consistency between implicit and explicit evaluations remains an empirical question. Therefore, in the first study, we did not have a specific hypothesis of dissociation. We had instead a clear hypothesis on the type of content that should create greater dissociation between implicit and explicit attitude change. Namely, in the second study, we expected this to be the case when the interference was due to a distractor characterized by shifting the focus away from the target towards oneself. Under this condition, implicit, but not explicit, attitude change should occur as a consequence of the SR manipulation. We did not predict the direction of the informative feedback to qualify further the interference effect above. Instead, it is an empirical question whether such feedback affects explicit attitude change when the distracting information concerns one's implicit attitude towards the target. Last, we will explore the relation

between implicit and explicit attitudes to test whether any differential impact of the SR on either outcome measure is accompanied by weaker implicit-explicit correlation.

Study 1

The study aimed to test whether a distraction task type of interference between the IAT and the explicit measure would affect the transmission of evaluative change at the explicit level. Similar to Gregg, Seibt, and Banaji (2006), the targets of the SR manipulation were two fictitious groups (Lerriani vs. Dattiani) that were pre-selected as neutral in valence both at the implicit and explicit level (Perugini, Richetin, & Zogmaister, 2012). We tested whether implicit-explicit consistency was still detectable when the flow from one measure to the other was interrupted by a distraction task. Additionally, we wanted to see whether distractors that differed in terms of their procedural overlap with the IAT could affect the transfer of liking from implicit to explicit differently.

Method

Procedure

One hundred and sixty-three participants (94 women and 69 men, *M*age = 21.46, *SD* = 2.87) were recruited for a one-session study with a 2 (SR manipulation) x 2 (distraction tasks) between-subjects design. Participants were university students who either volunteered or received course credits for their participation. The University Ethics committee approved both this study and the next one. We report all experimental conditions, measures and all data exclusion criteria for both studies. The targets of the SR intervention were the groups of Lerriani versus Dattiani. First, participants completed a familiarization task with the names of group members and with the orthographic distinction between them, followed by the Self-Referencing task. Participants were randomly assigned to a condition in which they used the same key to classify stimuli concerning Lerriani and the self or to classify stimuli concerning Dattiani and the self. After the SR task, participants carried out an IAT Lerriani/Dattiani. Next, participants were randomly allocated to one of the two distraction tasks. One distraction task consisted of performing an unrelated proofreading task, whereas the other consisted of completing an unrelated IAT Zimmiani/Craviani (groups

selected as neutral in a pilot study). Both tasks lasted approximately 4 to 5 minutes. Finally, all participants completed an explicit evaluation of Lerriani and Dattiani and a measure of intersecting regularities memory, were paid or given course credit, and debriefed.

Sample size determination

We did not perform a formal power analysis to estimate the required sample sizes for Study 1 and 2 before data collection. However, a sensitivity analysis (Perugini, Gallucci, & Costantini, 2018) showed that, given $\alpha = .05$ and power = .80, for Study 1 ($n = 163$), the minimum effect size detectable for a significant interaction is $f = 0.22$ (equivalent to Cohen's $d = 0.44$) whereas for Study 2 ($n = 236$) is $f = 0.18$ (equivalent to Cohen's $d = 0.36$). The studies, therefore, were sufficiently powered to detect small to medium effect sizes.

Materials

Self-Referencing task. Participants completed two initial blocks of 40 trials. In the first two blocks, Lerriani names and words related to the category Others (e.g., they, them, other) were assigned to one response key on the keyboard (i.e., "E"), whilst Dattiani names and words related to Self (e.g., self, me, my) were categorized with an alternative response key (i.e., "I"). Participants then repeated the two blocks of 40 trials with the keys to which the target categories were assigned switched, whilst the combination target group-personal pronouns remained the same (i.e., Lerriani names and Others-related words assigned to the "I" key, and Dattiani versus Lerriani names and Self-related words to the "E" key). In case of errors, a red X appeared on the screen and remained until correction. The inter-trial interval was 400ms. The order in which participants completed these two blocks was counterbalanced. For each target category, five names were used (see Supplementary Material).

Intersecting Regularities memory. Participants were probed about their source-target recognition. Specifically, the question was the following: "*One of the tasks that you have done consisted in classifying with the same key, words related to the self and words related to the members of one fictitious group. Do you remember which group?*". Participants could indicate one

of the two groups or the option "*I don't know*." Answers were coded as correct IR memory (correct responses) or incorrect IR memory (opposite responses or no recollection).

IAT Lerriani/Dattiani. Participants classified words presented individually and in a random order in the middle of the screen using two keys (i.e., 'E' and 'I'). The target concept was Lerriani, and its contrast was Dattiani, whereas the attribute categories were Positive and Negative. The order of the two critical blocks was counterbalanced between participants, with half of the participants having the combination Lerriani and Positive presented first and the other half having the combination Dattiani and Positive presented first. Practice blocks and critical blocks consisted of 20 and 81 trials (80 + one initial dummy trial), respectively. A red X appeared in the middle of the screen for 200ms if the participant did not answer correctly. There was no built-in penalty and the inter-trial interval was 500ms. For each attribute and target category, five words were used (see Supplementary Material).

Distraction tasks.

a) Proofreading. Participants performed a task that consisted of circling the r's in a passage of English text (an excerpt from an engineering handbook).

b) IAT Zimmiani/Craviani. The IAT was identical to that administered to measure implicit evaluation towards the critical groups. The target concept was Zimmiani, and its contrast was Craviani, whereas the attribute categories were Positive and Negative. For each attribute and target category, five words were used (see Supplementary Material).

Explicit attitude. Participants were instructed to evaluate each group separately on four semantic differential pairs of items (negative/positive, mean/nice, bad/good, unpleasant/pleasant) on a 7-point Likert-type scale. The order of the evaluation of the two groups was counterbalanced following the same order as for the IAT (Lerriani first when such group was first matched with positive in the IAT vs. Dattiani first when such group was first matched with positive).

Data preparation

The data of four participants were excluded because the IAT data revealed a large percentage of errors (over 25%). The main analyses were conducted on the remaining 159 participants. As the SR effect has shown to be moderated by participants' IR memory (Mattavelli et al., 2017), we conducted the same analyses on both participants with correct IR memory (n = 115) and with incorrect IR memory (n = 44).

Results

Descriptive statistics are reported in Table 1. Both the IAT and the explicit attitude measures were reliable (α = .87 and α = .91) and were significantly correlated (r = .16, p = .045). The main analyses involved a 2 one-way (SR condition, Lerriani+Self vs. Dattiani+Self) analysis of variance (ANOVA) for the IAT and a 2 (SR condition, Lerriani+Self vs. Dattiani+Self) x 2 (Distraction task: proofreading vs. unrelated IAT) ANOVA for the explicit attitudinal measure. The one-way ANOVA revealed a main SR effect on implicit attitude, $F(1, 157) = 4.25$, $p = .041$, $\eta^2_p = .03$. Thus, participants assigned to the Lerriani+Self condition showed an implicit preference for Lerriani over Dattiani, and the opposite was true for participants in the Dattiani+Self condition. The analysis on explicit attitude revealed a main effect of the SR manipulation, $F(1, 155) = 7.80$, $p = .006$, $\eta^2_p = .05$. Likewise implicit, also explicit attitudes showed that the group assigned to the self was preferred over that assigned to the category others. Instead, we did not find evidence for an effect of the distraction task, $F(1, 155) = 0.01$, $p = .911$, nor a significant interaction, $F(1, 155) = 0.25$, $p = .616$.[2] Finally, we verified whether there were any effects on the unrelated IAT. In line with expectations, we found no significant SR effects on this measure, $F(1, 157) = 0.90$, $p = .347$. Furthermore, this

[2] We also ran the analyses including the two counterbalanced method factors (Order of blocks in familiarization task and Self Referencing; Order of block presentation in IAT and order of group presentation for explicit attitudes) as covariates in the model. The SR effect remained significant on implicit attitude, $F(1, 155) = 4.77$, $p = .030$, $\eta^2_p = .03$. Concerning the covariates, there was no effect of the Order of blocks in familiarization task and Self Referencing, $F(1, 155) = 0.57$, $p = .451$, while a significant effect of the order of block in which the IAT was administered was found, $F(1, 155) = 22.12$, $p < .001$, $\eta^2_p = .13$. However, the interaction between the latter covariate and the main experimental manipulation was far from significant, $F(1, 155) = 0.03$, $p = .864$. The analysis on explicit attitude revealed a main effect of the SR manipulation, $F(1, 153) = 8.07$, $p = .005$, $\eta^2_p = .05$, no effect of the distraction task, $F(1, 153) = 0.05$, $p = .831$, and no significant interaction, $F(1, 153) = 0.26$, $p = .614$. There was also no effect of the two method factors when included as covariates in the model (ps > .075).

IAT did not significantly correlate neither with the IAT Lerriani/Dattiani (r = .16, p = .174) nor with the corresponding explicit attitudinal measure (r = -.10, p = .409).

The correlation between implicit and explicit attitudes was significant on the whole sample, r = .16, p = .045. When focusing on the two distraction conditions separately, we found a stronger correlation in the proofreading condition, r = .20, p = .069, than in the unrelated IAT, r = .10, p = .387, although the standard level of significance was not reached in either subsample. We also conducted a moderated-mediation analysis using Hayes (2003) Process Macro (Model 14) to test whether (i) the effect of the SR on explicit was mediated by the IAT score and (ii) whether the relation between IAT and explicit varied significantly across distraction condition. No evidence for an effect of the SR on explicit mediated by IAT score emerged either in the unrelated IAT condition, b = .03, SE = .07, 95% CI [-.09, .22], or in the proofreading condition, b = .15, SE = .13, 95% CI [-.04, .44]. Moreover, a non-significant moderated mediation, b = .12, SE = .13, 95% CI [-.11, .41] revealed that the type of distraction task did not impact on the relation between IAT score and explicit attitude measure.

For participants with correct IR memory (N = 115), the results were qualitatively identical to the full sample, but with larger significant effects. Instead, for participants with incorrect IR memory (N = 44), there were no SR effects on both implicit and explicit attitudes (all p's > .69) nor any other main or interaction effects (all p's > .56).

Discussion

Study 1 demonstrated that the inclusion of a distraction task between the implicit and the explicit attitude measures did not show any significant impact on the transmission of the evaluative change generated through the SR from implicit to explicit attitudes. The main effect of the SR manipulation was detected on both outcomes. The impact of the distractor on the manipulation was not detectable even when the task was structurally similar to the IAT and therefore, carrying a potential confounding value. Moreover, for participants who completed the unrelated IAT as a

distraction task, we could control that the SR effect was specific to the targeted groups and did not spill over to unrelated social groups.

Study 2

This second study aimed to test whether the content of interference between the IAT and the explicit measure would affect the transmission of evaluative change at the explicit level. More precisely, we aimed at testing whether feedback about the properties of the measure could create implicit-explicit dissociation. We distinguished between an information condition that merely focused on the properties of the IAT and another one in which the same measure was described as capable of revealing the cognitive abilities of the individuals. For both the task- and the self-focused condition, we provided additional feedback to manipulate the validity of the measure (task-focused condition) and individuals' level of cognitive skills (self-focused condition). The target of the manipulation and the SR manipulation were the same as in Study 1.

Method

Procedure

Two hundred and thirty-six participants (121 women, 114 men and 1 missing, *Mage* = 22.41, *SD* = 3.01) were recruited for a one-session study with a 2 (SR manipulation) x 2 (Information condition) x 2 (Feedback type) between-subjects design. First, participants completed a simple familiarization task with the names of group members and then the SR task. Participants were randomly assigned to a condition in which they used the same key to classify stimuli concerning Lerriani and the self or Dattiani stimuli and the self. After the SR task, participants performed an IAT Lerriani/Dattiani and subsequently were allocated to one of two information conditions. The conditions differed in terms of the focus of attention (task-focused vs. self-focused). Within each condition, there were different types of feedback (positive vs. negative). Then, all participants completed an explicit evaluation of the two groups, and finally, they were probed about their IR memory, paid or given course credit, and debriefed.

Materials

The SR task, the IR memory question, the IAT (Lerriani/Dattiani), the explicit attitude measure as well as the other aspects of the study (e.g., counterbalanced factors) were the same as in Study 1.

Information condition.

a) *Task-focused*. After the IAT, a written text appeared on the screen that presented the task they just completed as either a valid or a less valid measure affected by methodological factors. In the positive feedback condition the text was as follows: *"The task you have just completed measures the strength with which two concepts (e.g., Positive and Dattiani or Lerriani) are associated. If two concepts are associated, it is easier to use the same key for both, and therefore, one does fewer mistakes and answers more quickly. Many studies have demonstrated that speed and errors when doing this test reveal our spontaneous preference for one of the two groups".* Instead, in the negative feedback condition, the text was: *"The task you have just completed should measure the strength with which two concepts (e.g., Positive and Dattiani or Lerriani) are associated. If two concepts are associated, it should be easier to use the same key for both, and therefore, one should make fewer mistakes and answers more quickly. However, many studies have demonstrated that speed and errors depend on details such as the length of the words, their familiarity, the order in which one does the task, and the learning that occurs during the task itself. Therefore, the task you have just done reveals the influence of this type of detail"*. Participants were then asked to press the space bar when they finished reading the text to continue with the experimental session.[3]

b) *Self-focused.* In this case, too, participants were provided information about the IAT. They were informed that the IAT reveals their implicit preference towards the target objects but also that the IAT was a task able to reveal their cognitive skills validly. The first part of the text was as follows: *"The task you have just completed provides two main pieces of information. On the one hand, it reveals your implicit preference towards Dattiani or Lerriani. On the other hand, it*

[3] We refer to a positive versus negative feedback type to be consistent with the two feedback types used in the self-focused condition. Here a positive (vs. negative) feedback is a feedback based on which the IAT is a valid (vs. invalid) measure of preference.

indicates your cognitive functioning efficiency that is obtained by combining the speed and accuracy of your performance. The cognitive functioning efficiency is considered one of the most important aspects of intelligence and, therefore, more in general of success in life. Your average speed in the task was XXX milliseconds. Your percentage of errors was X %. [both values corresponded to the actual time and errors for each participant as computed with Inquisit]". The final part of the text instead was different depending on the positive vs. negative information. In the positive feedback condition, the text was, *"By combining this information, the score of your cognitive functioning efficiency is above the 80th percentile relative to the scores of other people who did this task. This means your performance was better than 80% of people's".* Whereas in the negative feedback condition, the text was *"By combining this information, the score of your cognitive functioning efficiency is below the 30th percentile relative to the scores of other people who did this task. This means your performance was worse than 70% of people's".* For participants assigned to this condition, we made clear in the debriefing phase that none of the information provided about the nature of the IAT as a measure of their cognitive ability was real, and that it only served for the purpose of our experimental manipulation.

A manipulation check measure for the self-focused information condition was inserted immediately after the feedback. The measure consisted of two questions asking participants about their opinions about the task. The questions were, *"Do you think that the task is a valid measure to measure the efficiency of cognitive functioning? [from 1 = Not at all valid, to 7 = very much valid]"* and *"Do you think that the performance at this task will be correlated with other performances in other validated tasks that measure the cognitive functioning efficiency (that is, that those who had a good performance to this task will have a good performance also to other validated tasks of efficiency of cognitive functioning and those who, instead, had a bad performance to this task will have a bad performance also to other validated tasks of efficiency of cognitive functioning)? [from 1 = Not at all correlated, to 7 = Very much correlated]*". The two items were substantially correlated ($r = .47$, $p < .001$), and their scores were therefore, averaged. Based on

previous studies (Pronin, Lin, & Ross, 2002), we would expect a significant difference in the evaluation of the validity of the task depending on whether the feedback was positive or negative such that it should be considered as more valid when the feedback was positive compared to when it was negative.

Data Preparation

The data from three participants were excluded because their IAT data revealed a large percentage of errors (over 25%). Of the remaining 233 participants, 192 were classified as correct, and 41 (20 did not remember, and 21 showed opposite memory) as incorrect IR memory. We inspected whether the manipulation of the type of feedback (positive vs. negative) in the self-focused task was successful. There was a significant effect of the type of feedback, $t(114) = 7.58$, $p < .001$, with a large effect Cohen's $d = 1.42$, with participants in the positive feedback condition who judged the test more valid than in the negative feedback condition ($M = 5.15$, $SD = 0.88$ vs. $M = 3.79$, $SD = 1.04$).

Results

Descriptive statistics are reported in Table 2. Both the IAT and the explicit attitude measures were reliable ($\alpha = .81$ and $\alpha = .93$) and were significantly correlated ($r = .26$, $p < .001$). The main analyses involved a univariate ANOVA (SR condition, Lerriani+Self vs. Dattiani+Self) for the IAT and a 2 (SR condition, Lerriani+Self vs. Dattiani+Self) x 2 (Information condition, task-focused vs. self-focused) x 2 (Feedback type, positive vs. negative) ANOVA for the explicit attitude measure. The SR effect was significant on both implicit, $F(1, 230) = 16.54$, $p < .001$, $\eta^2_p = .08$, and explicit attitudes, $F(1, 224) = 7.99$, $p = .005$, $\eta^2_p = .03$. Again, participants showed both implicit and explicit preference for the group paired with the self. On explicit attitude, neither the information condition nor the feedback type showed significant impact, $F(1, 224) = 0.001$, $p = .979$ and $F(1, 224) = 0.43$, $p = .514$, respectively. Instead, the SR effect was qualified by whether the information focused on either the task or the self, as revealed by the interaction term, $F(1, 224) = 4.80$, $p = .029$, $\eta^2_p = .02$. In line with our hypotheses, the SR effect was significant in the task-focused information condition,

$F(1, 112) = 12.28$, $p = .001$, $\eta^2_p = .10$ ($M = 1.28$, $SD = 2.12$ vs. $M = -0.27$, $SD = 2.51$), but not in the self-focused information condition, $F(1, 112) = 0.21$, $p = .651$ ($M = 0.62$, $SD = 2.02$ vs. $M = 0.43$, $SD = 2.46$). There was no significant interaction between the SR manipulation and the type of feedback, $F(1, 224) < 0.01$, $p = .997$ and between the latter factor and the information condition, $F(1, 224) = 0.28$, $p = .596$. The three-way interaction was not significant either, $F(1, 224) = 0.22$, $p = .638$.[4]

There was an overall significant correlation between IAT score and explicit attitude score, $r = .26$, $p < .001$. Importantly, such correlation was significant (and comparable in magnitude) in both the task- and the self-focused conditions, $r = .26$, $p = .008$ and $r = .29$, $p = .002$, respectively. Different from Study 1, the moderated-mediation analysis revealed a significant mediation effect in both the task-focused, $b = .27$, $SE = .11$, 95% CI [.06, .49], and the self-focused interference conditions, $b = .30$, $SE = .14$, 95% CI [.06, .61]. Moreover, as indexed by the non-significant moderated-mediation, $b = .02$, $SE = .15$, 95% CI [-.24, .37], the relation between the IAT score and the explicit score was not qualified by the type of interference manipulation.

Results were qualitatively identical, with larger effects, for participants with correct memory ($N = 191$). Conversely, among the subsample of participants showing either reversed or no IR memory ($N = 41$), there was a main SR effect on implicit attitude, $F(1, 39) = 6.13$, $p = .018$, $\eta^2_p = .13$. However, no main effect of the SR task emerged on explicit, $F(1, 33) = 0.98$, $p = .330$. Importantly, no other main effects and interactions, including the one between SR manipulation and information condition reached the level of significance (ps > .070).

Discussion

Study 2 confirmed the effectiveness of SR manipulation in changing attitudes. Crucially, the transmission of the evaluative change from the implicit to the explicit evaluation was affected

[4] We included the two counterbalanced factors (Order of blocks in familiarization task and Self Referencing; Order of block presentation in IAT and order of group presentation for explicit attitudes) as covariates in the model. The SR effect stayed significant on implicit attitude, $F(1, 228) = 16.89$, $p < .001$, $\eta^2_p = .07$. Concerning the covariates, we found no significant effect (ps > .052). The analysis on explicit attitude confirmed the main effect of the SR manipulation, $F(1, 222) = 7.90$, $p = .005$, $\eta^2_p = .03$. All the other effects, including the covariates, were not significant (ps > .505).

differently by whether the information was task-focused or self-focused. In the first case, the effect on implicit attitude was accompanied by an explicit change, whereas in the second case, we found no evidence that the change observed on the IAT extended onto the explicit. The specific type of feedback did not affect this interaction effect. In essence, it did not matter whether the feedback about the self was positive or negative or whether the IAT was described as valid or invalid. The only thing that mattered was whether the information was about the task or the self. In this latter case, there was no change at the explicit level. Notably, the observed dissociation between implicit and explicit attitude change in the self-focused condition was detected at the mean level, but not at the correlation level, implying that it was not reflected also in a weaker relation between the two outcome variables, as also suggested by the non-significant moderated-mediation.

General discussion

Across two studies, we successfully replicated the effectiveness of the SR task in changing implicit and explicit attitudes (Mattavelli et al., 2017). After categorizing two fictitious groups with the same action required to categorize either self- or other-related stimuli, participants exhibited both implicit and explicit preference for the target group related to the self. Also, in line with the meta-analytical findings of Mattavelli and colleagues (2017), our results appeared stronger when focusing on participants who correctly learned and remembered the contingency between the self and the target object.

Central for the present investigation, the two studies revealed that distinct types of distractors differentially led to implicit-explicit consistency following the SR manipulation. Specifically, in Study 1 we found that the use of a distractor administered in between the critical IAT and the explicit attitude measure did not affect explicit attitude. The SR manipulation changed both implicit and explicit attitudes regardless of whether the distractor could generate confounding with the critical IAT (i.e., unrelated IAT) or whether it was an entirely different task (i.e., proofreading task). The implicit-explicit attitude correlation was positive and significant. Thus, the effect of the SR manipulation on the explicit attitude was not affected by the type of distraction

task. A different pattern emerged from Study 2. We found a two-way interaction between SR manipulation and the content of the distracting information. Namely, the preference for the fictitious group related to the self in the SR task did not generalize to the explicit level of attitude when the feedback involved self-relevant information. In contrast, it did when the feedback informed participants about the characteristics of the IAT as a measure. Additionally, in both the task- and self-focused condition, we observed that information about either the validity of the measure (IAT-focused) or the personal abilities (self-focused) did not affect the SR effect.

The lack of any implicit-explicit dissociation regarding the SR effect in Study 1 seems to indicate that the generalization from implicit to explicit attitudes change is not sensitive to every type of distractor. Earlier research on attitude change showed that the role of distractors might depend on specific characteristics of the message. Distraction increases attitude change to a simple message (one which is easily understood but not very convincing) but decreases attitude change to a complex message (Regan & Cheng, 1973). In the present investigation, one might argue that the 'message' implied by the SR task (i.e., "only one of the two target groups is paired with the self") is a relatively easy one. Therefore, some types of distractors interposing the implicit and the explicit measure do not prevent the SR to influence explicit evaluations. Besides any speculative interpretation, the results from Study 1 revealed the robustness of the SR effect on explicit attitude change. Whereas the SR effect was detectable on both implicit and explicit attitude measures, the correlation between the two was only marginally significant on the whole sample (and not significant in any of the two subsamples). Moreover, we found no evidence for the IAT to mediate the impact of the SR manipulation on explicit. Both findings are in contrast with prior evidence on the SR effect (Mattavelli et al., 2017; Perugini et al., 2014). Thus, the inclusion of either distracting task (i.e., unrelated IAT or proofreading) might have disrupted the flow that went from the IAT to the explicit measure and therefore weakened their relation, although one cannot exclude the possibility that the results can be due to sampling variations (e.g., the dance of *p*-values, Cumming,

2014). Crucially, however, such a weakened relationship between the two measures did not block the effect of the SR manipulation on explicit attitude change.

In Study 2, providing information about the purpose of an IAT as an attitude measure did not prevent the impact of the SR task on explicit attitude change either, in line with the above results. What emerged from Study 2 was a lack of dissociation in the effect of the SR task on implicit and explicit attitude change when participants were overtly informed about the properties of the IAT after its completion. We also observed that the SR effect on explicit attitude was still there even when the capability of the IAT to reflect one's real implicit preference for one object over the other one was questioned. Instead, we found that implicit-explicit dissociation occurred when information about the measure was presented as referring to personal abilities. Thus, when participants were told that the IAT reflected not only their preferences but also their cognitive skills, the latter information seemed to overshadow the former, preventing participants from focusing on the stimuli relations (e.g., Lerriani related with the self) acquired in the SR task. Study 2 also suggested that when distractors were not operationalized in the form of additional tasks (like in Study 1), but as information that pointed to the nature of the IAT, the correlation between implicit and explicit attitude was evident and comparable across the two conditions (self-focused and task-focused). Moreover, and in contrast to Study 1, we found evidence for an SR effect on explicit mediated by the IAT in both the task- and the self-focused condition, even though in the latter, no total effect was found. This might be because, in both conditions, the descriptive texts included as potential distractors increased the salience of the IAT, and, as a consequence, participants tended to provide explicit responses that were in line with the implicit ones. Overall, our results show that the effectiveness of the SR effect on explicit is not strictly dependent on the extent to which implicit and explicit attitudes show concordance.

A self-perception interpretation seems to fit with the pattern of results that emerged from Study 2. To use Bem's words, self-perception takes place when "*individuals come to know their own attitudes, emotions, and other internal states partially by inferring them from observations of*

their own overt behavior and/or the circumstances in which the behavior occurs" (Bem, 1972, p. 3). Therefore, for a learned attitude to manifest explicitly, individuals might need to focus on two pieces of information. The first concerns what the implicit measure tells them about their internal preferences. The second is how the outcome of such a measure relates to a prior learning experience (i.e., the SR task). Given people's tendency to grant a valuable status to their introspection (Pronin, 2009), the attitudinal signal resulting from an IAT should be considered as a valid expression of preferences, therefore leading to consistent, explicit change, as long as participants can connect that signal to a source of attitude formation. What seems to emerge from Study 2 is that when participants are told that the IAT is also informative concerning their cognitive skills, they stop focusing on what the measure can reveal to them about their attitude. This might prevent them from undertaking introspection to the reason for their 'inner' preferences, that is, a learned relationship between the target group and the self. Future investigations should further explore this hypothesis. For instance, a more stringent test for the idea that participants do not process the outcome of an IAT when the latter is meant to reflect other cognitive abilities of the individuals would be to include a measure of awareness of one's associative evaluations. Participants who do not elaborate further on associations should also be less aware of them.

One limitation of the present studies is the lack of a specific type of control condition, that is, without any distractors, given that all conditions involve different types of distractors. We think that this feature, while arguably a limitation, does not challenge the main inferences from the results. Dozens of other studies have already shown the SR effect on implicit and explicit attitudes, as reported in the meta-analysis by Mattavelli et al. (2017), and many of them on fictitious social groups such as the ones used in these two studies. These studies did not have a distractor condition, and the effect sizes for the SR effect are comparable to the ones found in this research. In other words, SR effects similar in magnitude to those found in these studies can be found in a no distractor condition. Therefore, the results of the two studies can be interpreted as what happens

when different types of distractors are introduced between the measurement of implicit and explicit attitudes.

To summarize, our contribution showed that the effect of the SR task on explicit attitude change is resistant to different types of potential interferences. As far as we are aware, this is the first attempt to look at elements of interference as potential causes of implicit-explicit dissociation in attitude change resulting from associative learning. We demonstrated that the SR effect survives when distraction tasks interrupt the path that goes from the implicit to the explicit measures. The same holds when the interference is represented by information about the validity of the IAT as a measure of implicit preferences. Instead, we showed that providing information that shifts one's focus from one's attitudes to other cognitive functions creates an inconsistency between implicit and explicit attitude change. The idea that shifting the focus from the target (i.e., the implicit measure) towards the self hinders further processing of the newly acquired associative evaluation is novel and can pave the way for future investigations on the relations between implicit and explicit attitude change.

References

Bem, D. J. (1972). Self-perception theory. *Advances in Experimental Social Psychology*, *6*, 1-62.

Bless, H., Bohner, G., Schwarz, N., & Strack, F. (1990). Mood and persuasion: A cognitive response analysis. *Personality and Social Psychology Bulletin*, *16*(2), 331-345.

Cumming, G. (2014). The new statistics: Why and how. *Psychological Science*, *25*(1), 7-29.

Dawson, M. E., Rissling, A. J., Schell, A. M., & Wilcox, R. (2007). Under what conditions can human affective conditioning occur without contingency awareness? Test of the evaluative conditioning paradigm. *Emotion*, 7(4), 755-766.

De Houwer, J. (2003). The extrinsic affective Simon task. *Experimental Psychology*, *50*(2), 77-85.

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. (2007). A conceptual and theoretical analysis of evaluative conditioning. *The Spanish Journal of Psychology*, *10*(2), 230-241.

Ebert, I. D., Steffens, M. C., Von Stülpnagel, R., & Jelenec, P. (2009). How to like yourself better, or chocolate less: Changing implicit attitudes with one IAT task. *Journal of Experimental Social Psychology*, *45*(5), 1098-1104.

Fazio, R. H., Jackson, J. R., Dunton, B. C., & Williams, C. J. (1995). Variability in automatic activation as an unobtrusive measure of racial attitudes: A bona fide pipeline? *Journal of Personality and Social Psychology*, *69*(6), 1013-1027.

Field, A. P., & Moore, A. C. (2005). Dissociating the effects of attention and contingency awareness on evaluative conditioning effects in the visual paradigm. *Cognition & Emotion*, 19(2), 217-243.

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

Gawronski, B., & Bodenhausen, G. V. (2006). Associative and propositional processes in evaluation: an integrative review of implicit and explicit attitude change. *Psychological Bulletin*, *132*(5), 692-731.

Gawronski, B., & Bodenhausen, G. V. (2011). 2 The Associative-Propositional Evaluation Model: Theory, Evidence, and Open Questions. *Advances in Experimental Social Psychology*, *44*, 59-118.

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

Gawronski, B., & LeBel, E. P. (2008). Understanding patterns of attitude change: When implicit measures show change, but explicit measures do not. *Journal of Experimental Social Psychology*, *44*(5), 1355-1361.

Greenwald, A. G., & Banaji, M. R. (1995). Implicit social cognition: attitudes, self-esteem, and stereotypes. *Psychological Review*, *102*(1), 4-27.

Greenwald, A. G., McGhee, D. E., & Schwartz, J. L. (1998). Measuring individual differences in implicit cognition: the implicit association test. *Journal of Personality and Social Psychology*, *74*(6), 1464-1480.

Gregg, A. P., Seibt, B., & Banaji, M. R. (2006). Easier done than undone: asymmetry in the malleability of implicit preferences. *Journal of Personality and Social Psychology*, *90*(1), 1-20.

Hayes, A. F. (2013). Model templates for PROCESS for SPSS and SAS.

Hermans, D., Crombez, G., Vansteenwegen, D., Baeyens, F., & Eelen, P. (2002). Expectancy-learning and evaluative learning in human classical conditioning: Differential effects of extinction. In S. P. Shohov (Ed.), *Advances in psychology research, Vol. 12* (p. 17–40). Nova Science Publishers.

Hofmann, W., De Houwer, J., Perugini, M., Baeyens, F., & Crombez, G. (2010). Evaluative conditioning in humans: a meta-analysis. *Psychological Bulletin*, *136*(3), 390-421.

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

Karpinski, A., & Hilton, J. L. (2001). Attitudes and the Implicit Association Test. *Journal of Personality and Social Psychology, 81,* 774-788.

Kruglanski, A. W., & Thompson, E. P. (1999). Persuasion by a single route: A view from the unimodel. *Psychological Inquiry*, *10*(2), 83-109.

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*(2), 194-213.

Martin, I., & Levey, A. B. (1978). Evaluative conditioning. *Advances in Behaviour Research and Therapy*, *1*(2), 57-101.

Mattavelli, S., Richetin, J., Gallucci, M., & Perugini, M. (2017). The Self-Referencing task: Theoretical overview and empirical evidence. *Journal of Experimental Social Psychology*, *71*, 68-82.

Moors, A., & De Houwer, J. (2006). Automaticity: a theoretical and conceptual analysis. *Psychological Bulletin*, *132*(2), 297-326.

Nosek, B. A., & Banaji, M. R. (2001). The go/no-go association task. *Social Cognition*, *19*(6), 625-666.

Olson, M. A., & Fazio, R. H. (2001). Implicit attitude formation through classical conditioning. *Psychological Science*, *12*(5), 413-417.

Olson, M. A., & Fazio, R. H. (2006). Reducing automatically activated racial prejudice through implicit evaluative conditioning. *Personality and Social Psychology Bulletin*, *32*(4), 421-433.

Payne, B. K., Cheng, C. M., Govorun, O., & Stewart, B. D. (2005). An inkblot for attitudes: affect misattribution as implicit measurement. *Journal of Personality and Social Psychology*, *89*(3), 277-293.

Perkins, A., & Forehand, M. (2012). Implicit Self-Referencing: The effect of nonvolitional self-association on brand and product attitude. *Journal of Consumer Research*, *39*(1)*,* 142–156.

Perugini, M., Gallucci, M., & Costantini, G. (2018). A practical primer to power analysis for simple experimental designs. *International Review of Social Psychology*, *31*(1): 20, 1-23.

Perugini, M., Richetin, J., & Zogmaister, C. (2012). The formation of implicit and explicit attitudes for neutral and valenced stimuli using the self. *Learning and Motivation*, *43*(3), 135-143.

Perugini, M., Richetin, J., & Zogmaister, C. (2014). Indirect measures as a signal for evaluative change. *Cognition & Emotion*, *28*(2), 208-229.

Petty, R. E., & Brock, T. C. (1981). Thought disruption and persuasion: Assessing the validity of attitude change experiments. In R. Petty, T. Ostrom, & T. Brock (Eds.), *Cognitive responses in persuasion* (pp. 55-79). Hillsdale, NJ: Erlbaum.

Petty, R. E., Wells, G. L., & Brock, T. C. (1976). Distraction can enhance or reduce yielding to propaganda: Thought disruption versus effort justification. *Journal of Personality and Social Psychology*, *34*(5), 874-884.

Prestwich, A., Perugini, M., Hurling, R., & Richetin, J. (2010). Using the self to change implicit attitudes. *European Journal of Social Psychology*, *40*(1), 61-71.

Pronin, E. (2009). The introspection illusion. *Advances in Experimental Social Psychology*, *41*, 1-67.

Pronin, E., Lin, D. Y., & Ross, L. (2002). The bias blind spot: Perceptions of bias in self versus others. *Personality and Social Psychology Bulletin*, *28*(3), 369-381.

Regan, D. T., & Cheng, J. B. (1973). Distraction and attitude change: A resolution. *Journal of Experimental Social Psychology*, *9*(2), 138-147.

Richetin, J., Mattavelli, S., & Perugini, M. (2016). Increasing implicit and explicit attitudes toward an organic food brand by referencing to oneself. *Journal of Economic Psychology*, *55*, 96-108.

Wittenbrink, B., Judd, C. M., & Park, B. (1997). Evidence for racial prejudice at the implicit level and its relationship with questionnaire measures. *Journal of Personality and Social Psychology*, *72*(2), 262-274.

Yamaguchi, S., Greenwald, A. G., Banaji, M. R., Murakami, F., Chen, D., Shiomura, K., ... & Krendl, A. (2007). Apparent universality of positive implicit self-esteem. *Psychological Science*, *18*(6), 498-500.