Original Papers

Polish Psychological Bulletin
2022, vol. 53(4) 292–301
DOI: 10.24425/ppb.2022.143375

Robert Balas[*]
Adriana Rosocha[*]

Do inconsistent implicit and explicit attitudes have any effect on behavior?

Abstract: Evaluative conditioning (EC) is a change in the evaluation of a neutral stimulus due to its pairing with another affective stimulus. Our Experiment 1 (N = 40) was carried out based on Rydell et al. (2006). During the conditioning stage, participants were presented with pictures of faces (CS) and positive or negative information about their behavior (explicit US). The images were preceded by short verbal primes (implicit US) of opposite valence to behavioral information. In Experiments 2 (N = 122) and 3 (N = 100) we provoked the transfer of implicit and explicit attitudes between USs and CSs by using social objects that potentially carry discrepant implicit and explicit evaluations. The data shows an inconsistency between implicit and explicit attitudes towards The results also confirm that those explicitly assessed attitudes are affected only by explicit information. At the same time, implicit attitudes are influenced not only by automatic processes but also by many other processes and information available to one's conscious mind.

Keywords: *attitude consistency, evaluative conditioning, implicit and explicit evaluations*

Making everyday decisions is strongly influenced by implicit processing, typically characterized as automatic, effortless, and usually unaccompanied by conscious thought. Consider impression formation – a newly met person is almost immediately liked or disliked without much deliberation, even though there is little explicit knowledge about him to form our initial evaluation. However, we sometimes find this first impression confusing, so both positive and negative aspects co-exist temporally. We describe this type of phenomenon as inconsistent attitudes, i.e., those that incorporate positivity and negativity. We put this fascinating research topic in the context of implicit and explicit attitudes. We claim that such an affective inconsistency of attitudes is possible when implicit and explicit evaluation measures show opposite results. In other words, a key question we want to address is whether it is possible that our attitudes towards a specific object/person/situation can be affectively inconsistent and whether this inconsistency results from divergent implicit and explicit attitude measurement.

The literature suggests two functionally different facets of attitudes (Gawronski & Bodenhausen, 2014, 2018). Explicit attitudes are conscious and can be assessed using self-report scales, whereas implicit attitudes are based on automatic, unconscious processes that can only be measured indirectly. Although there is a considerable debate about the basic functional and structural properties of explicit and implicit evaluations (see de Houwer, 2014; Hahn & Gawronski, 2014) or even about their very existence, most researchers agree that they involve different mental processes in acquisition and expression (Hütter & Rothermund, 2020; Hütter & Sweldens, 2018). At times, people may hold differing implicit and explicit attitudes towards the same objects resulting in possible discrepancies between beliefs and behavior (Karpen et al., 2011; R. Rydell et al., 2006).

Although initially, attitudes were treated as relatively stable over time, researchers provided evidence that they can be successfully changed on the spot with relatively simple manipulations (Gawronski et al., 2018; Gawronski & Bodenhausen, 2006; Kurdi & Banaji, 2019; R. J. Rydell et al., 2007). When it comes to explicit attitudes, the modification process is relatively easy. It is determined by, for example, the speaker's attractiveness, logic, the strength of argument, solid emotions, or social context (Krosnick & Petty, 1995). This type of change is based on a modification to the set of beliefs about a person/object/idea due to the presence of explicit evidence that is incongruent with those current beliefs.

[*] **Polish Academy of Sciences, Warsaw, Poland**

Corresponding author: Robert Balas, rbalas@psych.pan.pl

Funding source with grant number
Additional information: This article was prepared within a project (no. 2014/13/N/HS6/03079) awarded by Polish National Science Center to the first Author.

However, changing implicit attitudes is more difficult since people have limited conscious access to its content (but see Hahn et al., 2014). Therefore, it is more difficult, or even impossible, to adapt them to new information voluntarily. Since an implicit attitude is an evaluative response based on associative knowledge acquired through direct repeated experience, the only practical way of changing it is through multiple experiences (like in evaluative conditioning (Halbeisen & Walther, 2015; Langer et al., 2009; Walther et al., 2009).

Gawronski and Bodenhausen's (Gawronski & Bodenhausen, 2006) Associative – Propositional Model assumes the dual nature of attitudes. Explicit attitudes are perceived as reflective processes related to a central information processing strategy. In contrast, implicit attitudes are based on automatic connections and depend on a peripheral information processing strategy. The evaluation of a given object is based on explicit attitudes when motivation and cognitive resources are available. When deprived of these resources, implicit attitudes can affect how we evaluate things (Davies et al., 2012; Gawronski, n.d.; Mierop et al., 2020). Researchers have assumed that although explicit attitudes represent mainly our intentional evaluation, implicit attitudes can temporarily activate an affective category with all its memory associations and thus influence overt evaluations (as in most indirect tests of evaluations, de Houwer et al., 2009).

Explicit and implicit attitudes can be learned or changed through the evaluative conditioning (EC) procedure involving repeated exposure to CS-US pairs (de Houwer et al., 2001; Hughes et al., 2016). As a result of these exposures, an initially neutral CS changes its evaluation in line with or opposite to US affective valence depending on conditions. In any case, a close spatiotemporal presence of affective US changes evaluative responses towards US.

Although many studies have shown that shaped and relatively durable explicit and implicit attitudes can be inconsistent in terms of their valence (Gawronski, Ye et al., 2014; Karpen et al., 2012; Rydell et al., 2008; Shoda et al., 2014), little is known about the mechanisms underlying the creation of such discrepant attitudes. Our experiment was inspired by the data gathered by Rydell et al. (2006), who showed how discrepant attitudes could be formed and changed. The authors attempted to create inconsistent attitudes towards the same person in their experiment. Participants were presented with an image of a person (Bob) along with negative or positive information on his behavior. A priming word preceded each presentation of the pictures. The prime always had an opposite affect than the information given about Bob. Whenever Bob was presented with positive behavioral information, it was preceded by a negative prime, whereas positive prime words preceded the presentation of negative behavioral information. The results showed that it could create divergent attitudes at different levels. The explicit attitude has been shown to be affected by the affective information presented about Bob's behavior, whereas the implicit attitude was formed because of priming. Moreover, in the next stage, they reversed the sign of the conditioned responses.

Rydell et al. (2006) used an evaluative conditioning procedure with the goal of conditioning an ambivalent attitude on two levels – implicit and explicit. The EC seems most relevant here since it has been designed specifically to study attitude acquisition and change in laboratory settings. Most EC research has employed well-controlled experimental designs and stimuli to get to the core of attitudinal processes (de Houwer et al., 2001). However, Rydell et al. conclusions seem limited to artificially created attitudes in laboratory settings. One criticism addressed neither by Rydell nor Heycke (Heycke et al., 2018; Karpen et al., 2011) is that participants formed valence-inconsistent attitudes in the original procedure throughout 200 learning trials. Not only does it extend the number of learning trials typically used in EC procedures, but it also limits the generalization of Rydell's results. To overcome this limitation, we decided to use not only their original stimuli (Experiment 1) but also more real-life affective stimuli to see whether discrepant attitudes may originate from pairing neutral objects with inconsistent explicit and implicit attitudes towards Self and stigmatized outgroup members (Experiments 2 & 3). Therefore, the overarching goal of this research program is to verify Rydell et al.'s claims using his original procedure and more real-life stimuli.

The main goal of Experiment 1 was to replicate Rydell et al. (2006) conceptually. We decided on a partial conceptual replication with specific changes because of technical restrictions. Our replication used only the first part of the original study. Through this replication, we set out to check whether it is possible to create divergent attitudes. To answer this question, the first part of the original procedure was sufficient without involving the subjects in a lengthy procedure. We also changed the exposure time of the priming words from 25 to 33 ms due to hardware requirements. The literature shows no differences in processing semantic stimuli between the times mentioned above (del Cul et al., 2007).

We conducted a two-factor variance analysis to calculate the results, whereas mainly T-student tests were used in the original study. Based on a meta-analytic effect size for EC effects of $d = 0.52$ (Hofmann et al., 2010), we estimated a sample of $N = 40$ per cell that provides a power greater than 80% to detect a significant EC effect for all experiments reported below.

EXPERIMENT 1

METHODS

Participants

Forty participants volunteered for the experiment ($F = 25$) of varying age from 17 to 36 ($M = 27.15$, $SD = 4.58$). Participants were recruited in the hallways of universities and informed of the purpose of the study, the methods used, the estimated completion time, and the ability to withdraw from the study at any time without providing a reason. They were tested individually.

Materials and Procedure

The study used six images showing the neutral faces of six men, a set of prime words, and a set of sentences describing the behaviors of a man called Michal. Prime words and behavioral descriptions were acquired directly from Rydell (Rydell et al., 2006) and translated into Polish using the back-translation method.

In the first conditioning stage, the participants were informed that they would receive information about a person named Michal. Each conditioning trial started with a word prime (positive or negative) presented for 33 ms. The participants then saw a picture of Michal (a randomly selected photo from a set of six images of neutral faces of six men) along with a positive or negative statement about his behavior. Participants were asked to decide whether the information presented about Michal was typical. They recorded their answers by either pressing the "Z" key (if they thought the information was typical) on a keyboard or the "*M*" key (if they thought the information was atypical). The participants were then given feedback on whether their answer was correct. The conditioning phase consisted of 100 trials. Half of the participant pool was presented with the preceding ten negative stimuli (ten times each) along with positive information regarding Michal's behavior. For the other half of the participants, the affective valence of the preceding stimulus and the information presented about Michal's behavior was reversed. In this case, participants were shown ten positive stimuli (ten times each) along with negative information regarding Michal's behavior.

To assess explicit evaluation, Michal (shown in a picture) was rated on a scale from 1 (very unfriendly) to 9 (very friendly). In addition, they also rated Michal using five other 9-level scales on the following dimensions: bad-good, mean-pleasant, agreeable-disagreeable, caring-uncaring, and kind-cruel. The subjects also rated their emotions towards Michal with the help of a feeling thermometer, where they chose their response on a scale of 0 (very cold) to 100 (very warm).

To measure implicit attitudes toward Michal, the *Implicit Association Test* was used (Greenwald et al., 1998). Michal's picture was presented alongside images of five other white men and ten positive and ten negative adjectives. Subjects were asked to categorize the stimuli presented on the screen into groups (Michal vs. Others vs. Positive vs. Negative). To select their choice, the subjects pressed the corresponding key ("q" or "p"). Each block displayed labels in the top left and right corners, reminding participants of the category names. The test consisted of seven blocks. In the first two blocks, subjects were asked to match the adjectives and images presented by men with a positive or negative category (Block 1) and then Michal vs. others (Block 2). In the following two blocks (blocks 3 and 4, presented together), the subjects were shown the adjectives and images in random order and asked to categorize them. In these blocks, participants would select the *q* key to answer the positive adjective and Michal's picture or the *p* key in response to the negative adjectives and images of men other than Michal. In Block 5, the assignment of the reaction to the category of presented images was reversed, that is, the image of Michal required pressing the "p" key, while the image of a different person required pressing the "q" key. In the last Blocks 6 and 7, the participants once again categorized adjectives and pictures, this time with the reversed assignment of the reaction to the photographs. In other words, positive adjectives and other people's images required a response using the "q" key. In contrast, negative adjectives and images of Michal required a reaction using the "p" key. In case of a wrong response, participants received feedback about the error, and the attempt was repeated. The IAT score was calculated as described by Greenwald et al. (2003).

The sequence of consecutive measurements of implicit and explicit attitudes within the above groups was controlled. Finally, the participants were thanked for participating in the experiment.

RESULTS

Analysis of the results in an ANOVA of 2 (affective value of the preceding word: positive vs. negative) x 2 (sequence of attitude measurements) ANOVA did not show significant differences when comparing the order of attitude measurements for any dependent variable ($p = .22$, for measurement on a direct scale, $p = .15$ for the feelings thermometer, $p = .31$ for the semantic differential, and $p = .18$ for the IAT score). No interaction was shown between the affective prime word's valence and the attitude assessment order. Therefore, further analyses are presented without considering the second factor.

Explicit attitude

When analyzing the explicit attitude, we decided to present the analyses for each scale separately since the assessment of Michal on the Likert scale differs significantly from that of the other scales. The evaluation of Michal on the Likert scale was independent of the affective value of the priming word, $F(1,40) = 2.59$, $p = .12$, $h^2 = .06$. Michal's assessments were slightly higher in the group where negative information was preceded by positive adjectives ($M = 5.85$) compared to the group where positive information was preceded by negative adjectives ($M = 4.55$). BF01 was 1.17, showing that both H0 and H1 predicted the data equally well.

Michal's scores differed on the semantic differential, $F(1,40) = 56.2$, $p < .001$, $h^2 = .60$. Michal was rated the highest in the group where a negative priming word preceded the positive information ($M = 7.88$, $SD = 1.44$). Michal was rated lower ($M = 3.41$, $SD = 2.23$) in the group in which a positive priming word preceded the negative information about him.

Analogous results were obtained on the feelings thermometer scale. Michal received warmer feelings in the group in which positive information was preceded by a negative priming word ($M = 66.75$ and $SD = 25.66$). The group members declared colder feelings ($M = 24.90$ and $SD = 26.98$) towards Michal, in which a positive priming word explicitly preceded negative information. $F(1,40) = 25.42$, $p < .001$, $h^2 = .40$.

We then calculated an overall explicit attitude index by averaging Z scores of mean responses on the direct evaluation, semantic differential, and feelings thermometer. The analysis showed that Michal was assessed more positively in the condition where a negative prime was presented before positive verbal information (M = .54, SD = .63) as compared to reverse order (M = -0,54, SD = .62) $F(1,40) = 30.71$; $p < .001$, $h^2 = .45$.

Implicit attitude

Before performing the proper analyzes, the implicit attitude index was calculated according to the algorithm proposed by Greenwald (2003). A one-factor variance analysis was performed for the independent groups to determine whether implicit attitudes differ depending on the condition. As a result, a significant main effect of the prime was obtained, $F(1,40) = 4.56$; $p < .05$, $h^2 = .11$. The implicit assessment of Michal was higher in the group where the negative priming word preceded the positive information (M = .70, SD = .34) than in the group where the positive priming word preceded the negative information about Michal (M = .45, SD = -.39).

Comparison of explicit and implicit attitudes

To compare the difference between the explicit attitude (measured by three scales – direct assessment on the Likert scale, semantic differential, and a thermometer of feelings) and implicit attitude (measured by the IAT) toward the same object, the results obtained on the implicit attitude measurement scales were first standardized.

For Michal's direct assessment on the Likert scale, two-factor analysis of variance with repeated measurement was carried out in a 2 (type of attitude: explicit vs. implicit) x 2 (type of preceding word: negative vs. positive) format. The analysis revealed a significant main effect of attitude type, $F(1,30) = 12.16$, $p < .001$, $h^2 = .25$, and a main effect of the priming word valence, $F(1,38) = 5.15$, $p < .05$, $h^2 = .12$. The average score of the assessment of Michal's photo was higher in the group with a negative priming word (M = .47) than in the group with a positive priming word (M = .10). We also found that the participants rated Michal more negatively on the Likert scale (M = .001) than on the IAT test (M = .58). The interaction of both factors turned out to be insignificant.

For the explicit attitude measured by the semantic differential, an analogous analysis showed a significant main effect of the type of attitude, $F(1,38) = 25.95$, $p < .001$, $h^2 = .40$., and a main effect of the type of preceding word type, $F(1,38) = 53.74$, $p < .001$, $h^2 = .58$. The analysis also revealed an interaction between both factors, $F(1,38) = 31.36$, $p < .001$, $h^2 = .45$ (see Fig. 1).

Simple effects analysis only revealed significant differences between explicit and implicit attitudes only in the group with a positive initial word preceding negative information about Michal, $t(19) = 6.27$, $p < .001$ – explicit attitudes turned out to be more negative (M = -.76) than implicit attitudes (M = .45).

Comparing Michal's scores on the feelings thermometer with those of the IAT test depending on the affective

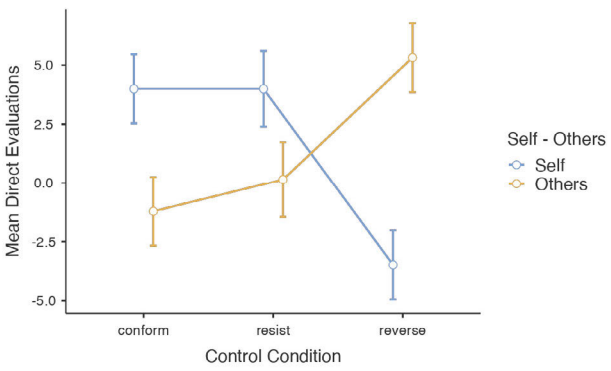

Figure 1. Mean evaluative score as a function of the type of measurement and affective valence of a prime word. IAT – Implicit Association Test, implicit measure; SD – semantic differential, explicit measurement

value of the preceding stimulus showed a significant main effect of the type of attitude measurement, $F(1,38) = 17.83$, $p < .001$, $h^2 = .32$, main effect of the preceding word type, $F(1,38) = 29.91$, $p < .001$, $h^2 = .44$, and effect of interaction of both factors $F(1,38) = 13.25$, $p < .001$, $h^2 = .26$ (see Fig. 2). The simple effects analysis showed significant differences between explicit and implicit attitudes only in the group where positive words preceded negative information about Michal. The explicit attitude was less positive (M = -.62) than the implicit attitude (M = .45).

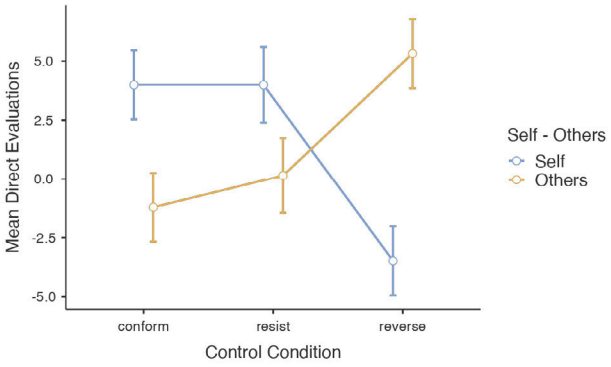

Figure 2. Mean evaluative score as a function of the type of measurement and affective valence of a prime word. IAT – Implicit Association Test, implicit measure; TF – thermometer of feelings, explicit measurement

EXPERIMENT 2

Experiment 1 suggests that the relation between implicit and explicit attitudes and the possibility of changing those is much more complicated than those presented by Rydell and colleagues. In line with recent many-labs replication (Heycke et al., 2018), we could not find convincing evidence of separate attitudinal systems that would respond differently to different EC procedures. One possible caveat of Rydell et al.'s findings is that they used initially neutral CSs that were conditioned and reconditioned. This puts a lot of effort on participants that

may have affected their study (see Heycke et al., 2018). Thus, we decided to use US objects with well-established implicit and explicit attitudes. As previous research has shown (di Pierro et al., 2016; Mattavelli et al., 2019, 2021; Perkins & Forehand, 2012), the Self can serve as such an object because of the possible discrepancy between implicit and explicit self-esteem. Additionally, there are relatively robust methods for measuring self-esteem on both levels (Grumm et al., 2009; Johnson, 2016; Krizan & Suls, 2008).

METHOD

Participants

One hundred twenty-two participants volunteered in this study (70 F). Two participants were not included in the further analyses due to their significantly older age. The mean age of the remaining group was 24.5 (*SD* = 4.02). They were recruited on the University campus and were not compensated for their time and effort. The participants were randomly distributed in three intentional control conditions: conform (*N* = 42), resist (*N* = 35), and reverse (*N* = 42).

Procedure

After informed consent, participants were engaged in the conditioning phase, in which CS-US pairs were shown. Neutral abstract pictures served as CSs, and words associated with Self (me, mine, I) or Others (them, their, they) served as USs. There were 2 CS-US_{self} pairs and 2 CS-US_{others} pairs. Each pair of CS-US was presented in a randomized order ten times simultaneously for 2500 ms (ITI = 1500ms). All groups received the instruction to observe the CS-US pairs with care.

Additionally, as an exploratory manipulation, we have implemented intentional control instructions (Balas & Gawronski, 2012; Gawronski, Balas, et al., 2014) to see whether the hypothesized transfer of explicit and implicit evaluation of the Self to other neutral objects is moderated by intentional effort to control for EC effects. In the "conform" group, participants were explained the essence of the EC effect and asked to form attitudes in line with the expected results. The "resist" group was asked to prevent the desired EC effect intentionally. Finally, the "reverse" group received instructions to form evaluations that oppose the expected EC effect.

After conditioning, two CS measurements were applied. In a direct evaluation, we asked participants to evaluate their impression of each abstract picture on a scale ranging from "-10 – extremely negative" to "+10 – extremely positive". The indirect measurement consisted of the Affective Misattribution Procedure (the AMP, Payne & Lundberg, 2014). In the AMP, participants were required to categorize Chinese ideographs as pleasant or unpleasant. Each AMP trial started with 75 ms presentation of CS followed immediately by a masked presentation of a Chinese ideograph presented for 100 ms. Each CS appeared in AMP twice. In addition, there were two neutral AMP trials in which a grey rectangle preceded a Chinese ideograph. Direct evaluation and AMP were randomized among participants.

Finally, each participant used the Self-IAT (Greenwald & Farnham, 2000) and Rosenberg's scale (Dzwonkowska et al., 2008) to measure implicit and explicit self-esteem levels, respectively.

RESULTS

Direct evaluations

Direct evaluations of abstract CS images were analyzed in 3 (control instruction group) x 2 (CS self vs. CS others) that showed an interaction between the two factors, $F(2,116) = 37.22$, $p < .001$, $h^2 = .32$ (Fig. 3).

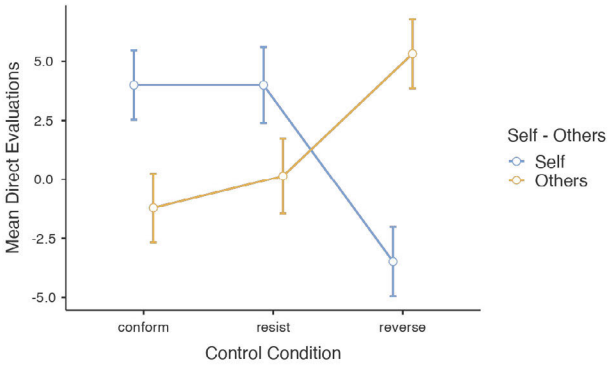

Figure 3. Mean direct evaluations of CS conditioned with Self and Others words as a function of intentional control instructions.

Post hoc tests showed significant differences between Self and Others across all conditions except where participants were asked to resist the influence of US on CS (conform: $p < .001$ and reverse: $p < .001$).

Extensive self-esteem analyses in conditioned image evaluations did not reveal any significant results except a strong negative correlation between the assessment of images associated with Self and those associated with others, $r(119) = -.61$, $p < .001$. The hypothesized correlation between explicit self-esteem and differences in CS assessment with Self and others was close to zero, $p = .65$.

Indirect Evaluations

We calculated the difference between Self and Others conditioned CSs and control CSs responses in the AMP so that positive scores would reflect more positive evaluations, whereas negative scores would reflect negative evaluations of conditioned CS. The results of the indirect AMP measure of evaluations were analyzed in 3 (control conditions) vs. 2 (Self vs. Others) and revealed a main effect of self-other conditioning, $F(1,116) = 27.08$, $p < .001$, $h^2 = .061$. Abstract images conditioned by Self were rated more positive (*M* = .33) than those conditioned by Others (*M* = -.08). Also, there was a significant interaction of both factors, $F(2,116) = 4.66$, $p = .011$, $h^2 = .021$ showing significantly more positive ratings of Self conditioned CSs than Others conditioned CSs in 'conform' and resist conditions but not reverse condition (see Fig. 4).

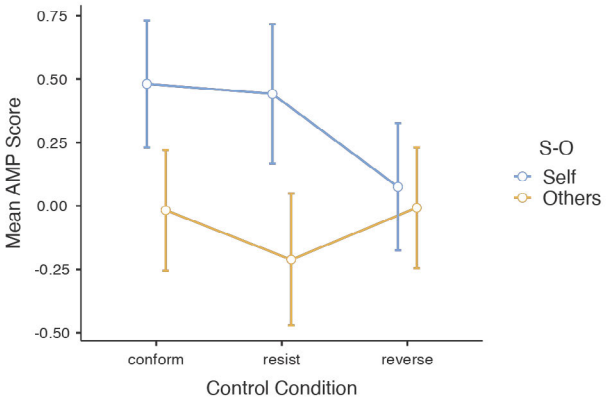

Figure 4. Mean indirect evaluations (the AMP Score) of CS conditioned with Self and Others words as a function of intentional control instructions.

We run regression analyses on the AMP scores with control instructions and the IAT D score as predictors to verify the relationship between implicit self-esteem and implicit evaluations of conditioned abstract images. It showed absolutely no effects on AMP scores for Self conditioned images, but a significant prediction by D score of AMP scores for Others conditioned images, $b = 0.478$, SE = .237, $t = 2.02$, $p = .045$. Since the D score shows Self preference over Others, the latter effect displays more favorable responses for Others conditioned images in higher implicit self-esteem measures.

EXPERIMENT 3

Since the results of Experiment 2 brought inconclusive data, we turned our attention to a particular social group subjected to stereotypization and discrimination, Jews, in the hope of extracting more polarized implicit and explicit attitudes. Previous research has shown that although this group is not explicitly devalued, it might suffer implicit negative attitudes among specific populations (Rudman et al., 1999, 2005). Therefore, we hoped to use this group as the US to show how contrastive implicit and explicit attitudes might be transferred to neutral objects in conditioning.

METHOD

Participants

One hundred participants (53 F) volunteered for the study without compensation. The mean age was 24.83 years (*SD* = 3.96). They were recruited among students on the university campus.

Procedure

After signing informed consent, participants were seated in front of a computer screen and told to follow the instructions carefully. First, they were asked to complete an Affective Priming Task (the APT, as applied and explained in Gawronski et al., 2015) as an indirect measure of implicit attitudes towards Poles and Jews. Each trial started with a red fixation cross centrally displayed on a screen for 500 ms. It was immediately followed by a priming stimulus, either a typical Polish or Jewish first name displayed for 200 ms. Then, positive or negative target words replaced the names. Participants were asked to categorize it as positive or negative as fast and accurately as possible by pressing relevant keys on a computer keyboard (*z* key for negative and *m* key for positive). Positive and negative words were centrally displayed until the response was given. Incorrect responses were followed by a feedback message ("Wrong !!!") displayed for 2000 ms in red. The inter-trial interval was set to 500 ms. Each priming stimulus presented ten positive and ten negative target words, summing up to 80 trials.

The participants were then asked to complete two separate explicit assessments. First, they were asked to express their general feelings towards Poles or Jews on a slider scale presented centrally on a screen (from "very cold" to "very hot"). Then, they were asked to assess their perceived social distance from individuals represented by their first names using a slider scale (from "very far" to "very close"). On both scales, participants were required to move a slider toward the left (described as maximum distance or very cold) or to the right (described as maximum closeness or very hot) and leave it at a point representing their perceived distance or temperature of feelings).

Participants have presented pairs of abstract images (a CS) and Polish or Jewish names (an US) during the conditioning phase. Each pair was shown horizontally (a picture above the name) for 1500 ms. The inter-trial interval was set to 500 ms. Two CSs were randomly assigned to two USs on a participant basis, thus creating 4 CS-US pairs. Each pair was presented eight times, summing up to 32 conditioning trials. Participants were randomly assigned to conform, resist, and reverse groups. The instructions for the groups were identical to those of Experiment 2.

Upon completion of the conditioning phase, participants were randomized to two tests. In the explicit evaluation task, they were asked to assess their impression of each of the four CSs images on a scale from -10 ("very negative") to +10 ("very positive"). In the indirect evaluation task, participants engaged in the Affective Missatribution Procedure (the AMP) precisely as described in Experiment 2. The difference is that the CSs were of Polish and Jewish names. The participants were thanked and briefed upon successful completion of the experiment.

RESULTS

Direct evaluations

Explicit CS evaluations analyzed in 3 (control instructions: conform vs. resist vs. reverse) x 2 (US Name: Polish vs. Jewish) showed no significant effects. Therefore, the CSs images were rated independently of whether they were paired with Polish or Jewish names and whether participants were prompted to conform, resist, or reverse the influence of USs on CSs before conditioning.

Indirect Evaluations

The same lack of effects was present in an indirect measure of evaluation, the AMP. Participants evaluated the Chinese ideographs as a function of the CS images presented as primes in this task, nor as control instructions.

Effects Related to Exhibit and Implicit Preferences towards Names

The paired t-tests on explicit measures of attitudes (feeling's thermometer and perceived social distance) showed more positive feelings toward Poles than Jews ($t(99) = 5.92$, $p < .001$, Cohen's d = .592) and less social distance towards Poles than Jews ($t(99) = 6.52$, $p < .001$, Cohen's d = .653). Since both measures were highly correlated, we calculated a single index of explicit attitude towards Poles and another index for Jews by averaging warmth and social distance ratings separately for both nationalities. Surely, the comparison of these indexes showed more positive explicit ratings of Poles than Jews ($t(99) = 7.30$, $p < .001$, Cohen's d = .730). There was a positive linear correlation between explicit ratings of Poles and direct evaluations of CSs conditioned by Polish names, $r = 0.375$, $p < .001$. Similarly, explicit ratings of Jews were positively correlated with direct evaluations of CSs conditioned by Polish names, r = .243, $p = 0.015$.

The Affective Priming Task was set to measure implicit attitudes towards Poles (as compared to Jews). First, we calculated separate priming indexes for both nationalities by subtracting mean response latencies for positive targets preceded by names from mean response latencies for negative targets preceded by names. Therefore, positive values of this index indicate positive implicit evaluations. Then, we calculated an overall preference index for Poles by subtracting priming indexes for Jewish names from priming indexes for Polish names. Again, positive values of this index indicate a higher preference for Polish names than Jewish ones.

The analysis of implicit measures did not show any preference towards either Poles or Jews. These implicit measures did not correlate significantly with indirect evaluations of CSs conditioned by Polish and Jewish names.

DISCUSSION

Our study aimed to answer whether it is possible to have divergent (on explicit and implicit levels) towards the same object. In Experiment 1, we conducted a partial, conceptual replication of Rydell's and his colleagues' (2006) research, who argued that it is possible to create an attitude with opposite signs at the level of explicit and implicit measurements. In light of our results that only partially replicated original findings as well as their direct replication (Heycke et al., 2018), in Experiments 2 & 3, we tried to test similar hypotheses about the evaluative transfer of effectively discrepant implicit and explicit attitudes, but using well-established attitudes towards real-life social objects.

In many ways, Rydell et al.'s results explain the origin of ambivalent attitudes and the discrepancies between explicit and implicit attitudes described in the literature (e.g., Greenwald et al., 1998). At the same time, these findings are considered the most pervasive evidence for dual-process theories of attitude acquisition. The attitudes would result from the history of acquiring attitudes and separate processes responsible for learning, storing, and expressing explicit and implicit attitudes. Rydell et al. (2006) found these fascinating findings to be difficult to confirm in replication studies (for example, Heycke et al., 2018). Therefore, the objective of our research was to replicate them with some modifications.

Our results confirm the ability to create divergent attitudes towards the same object on explicit and implicit levels. However, these discrepancies do not seem as stable as Rydell et al. (2006) suggested. They postulated that affectively incoherent attitudes occur when a negative prime is paired with positive verbal information, and a positive prime is paired with negative verbal information. The results of our study show that divergent attitudes may exist only when subjects are presented with a positive prime stimulus and explicit negative information. Therefore, we were unable to replicate the results of previous studies fully. Nevertheless, we have shown that evaluative conditioning can have different effects on an explicit and implicit level.

Secondly, we confirmed Heycke et al.'s (Heycke et al. 2018) conclusion that implicit measurement of conditioned attitudes, i.e., the IAT, was not responsive to the affective valence of primes. In our study, an implicit attitude towards Michal always seemed pretty positive. Therefore, the inconsistency between implicit measures occurred only when explicit ratings of Michal's favorability were reduced due to negative behavioral information. All the explicit measures used in our study (direct ratings, thermometer of feelings, and semantic differential) responded strongly to explicit behavioral information but not implicit primes.

In Experiments 2 and 3, we found similar results using more well-established social affective stimuli. Instead of creating divergent attitudes toward initially neutral stimuli, as in Rydell et al., we used real-life stimuli with a documented potential discrepancy between their implicit and explicit measurements. Namely, Experiment 2 utilized the Self as a possible source of conflicting attitudes, and Experiment 3 used the Jews as a social group that has been previously shown to generate discrepant implicit and explicit measurements of attitude. We found significant EC effects in Experiment 2 (which used Self as the US) on both explicit and implicit evaluations of conditioned neutral stimuli suggesting that the hypothesized discrepancy between implicit and explicit attitudes may have been too small to generate a successful transfer of affective valence. Indeed, the explicit and implicit measures of self-esteem did not show considerable differences in their respective values. In Experiment 3 (which used the social group category of Jews as the US), the effective EC effect was present only for explicitly measured attitudes.

We find an explanation within the literature that both types of attitudes are based on different mechanisms. This assumption refers to Gawronski and Bodenhausen's (2006) APE model, which provides a dual understanding of attitudes. Explicit attitudes depend on consciousness and are understood as a reflection process connected with a central information processing strategy, whereas automatic processes are responsible for implicit attitudes. They are associated with a peripheral strategy of information processing. The results confirm the assertion that explicit attitudes are a reflection of the information available to the consciousness. They also provide evidence that the results obtained in the Implicit Association Test might not only be a clear measure of automatic associations but also under the influence of various other active processes, including those that are controllable and introspectively available (see, for example, Hahn et al., 2014). Until now, explicit attitudes in academic explanations have been considered unconscious and unavailable to introspection. However, Hahn et al. (2014) research shows that it can predict certain aspects of implicit attitudes regardless of explicit evaluation.

In future studies, it is worth considering the divergence in both types of attitudes. Rydell, McConnell, and Mackie's studies (Rydell et al., 2008) show that the more significant the gap between attitudes, the greater the risk of cognitive dissonance the individual seeks to reduce, which may result in a lack of difference in their assessments. In contrast, Gawronski and Strack (2004) postulate that the occurrence of cognitive dissonance implies a change only in explicit attitudes. In summary, future studies on attitude inconsistency studies should focus on measuring their divergence and controlling the influence of other variables, e.g., cognitive dissonance, which can be a factor that reduces differences between explicit and implicit assessments.

Apart from the above theoretical considerations, it has to be noted that difficulties in replication of Rydell et al.'s results as well as in finding similar results that would suggest affective transfer of discrepant attitudes may occur due to the specific properties of implicit measures used to capture implicit attitudes. Those measures (like affective misattribution procedure, affective priming task, or the Implicit Association Test) have been criticized recently for their inherent problems with internal stability (Koppehele-Gossel et al., 2020), temporal stability (Dentale et al., 2020), or construct validity and ability to predict behavioral measures (Gawronski et al., 2020). All of the above shortcomings of implicit measures may have contributed to a problematic case of bearing hypothetically divergent attitudes on explicit and implicit levels and, even to more extent, on the question of transfer of those divergent attitudes to new objects with an evaluative conditioning mechanism. As many authors point out (see for example Gawronski et al., 2020) there appears to be a need for a "paradigm change" in light of the growing evidence of problems with implicit / explicit distinction in psychology. All in all, future research should give more consideration to the measurements applied and mechanisms of affective transfer that align with those measurements.

It is worth mentioning that this line of thinking is represented in Experiments 2 and 3, where we used USs that we thought were potentially promising bearers of divergent explicit and implicit evaluations. Although we did not find differences in how implicit and explicit attitudes were transferred from USs to neutral CSs, we do hope that this kind of manipulation can further be used with significant modifications.

Considering the results obtained by other researchers and us, it is easier to understand and refer to the example quoted at the beginning of the article. One plausible way of understanding our possibly mixed feelings about a newly met individual is that the fast, cognitive appraisal based on attended characteristics of that person may not be congruent with the result of implicit processing of even the same characteristics based on memory associations. Therefore, our initial first impression might not occur purely positive or negative but clouded with doubt that we do not fully understand. Because first impressions are crucial in shaping our behaviors and grounding the following more stable attitudes, we think studying inconsistent implicit and explicit attitudes is essential and should be continued.

REFERENCES

Balas, R., & Gawronski, B. (2012). On the intentional control of conditioned evaluative responses. *Learning and Motivation*, *43*(3), 89–98. https://doi.org/10.1016/j.lmot.2012.06.003

Davies, S. R., El-Deredy, W., Zandstra, E. H., & Blanchette, I. (2012). Evidence for the role of cognitive resources in flavor-flavor evaluative conditioning. *QUARTERLY JOURNAL OF EXPERIMENTAL PSYCHOLOGY*, *65*(12), 2297–2308. https://doi.org/10.1080/17470218.2012.701311

de Houwer, J. (2014). A propositional model of implicit evaluation. *Social and Personality Psychology Compass*, *8*(7), 342–353. https://doi.org/10.1111/SPC3.12111

de Houwer, J., Teige-Mocigemba, S., Spruyt, A., & Moors, A. (2009). Implicit measures: A normative analysis and review. *Psychological Bulletin*, *135*(3), 347–368. https://doi.org/10.1037/a0014211

de Houwer, J., Thomas, S., & Baeyens, F. (2001). Associative Learning of Likes and Dislikes: A Review of 25 Years of Research on Human Evaluative Conditioning. In *Psychological Bulletin* (Vol. 127, Issue 6, pp. 853–869). American Psychological Association Inc. https://doi.org/10.1037/0033-2909.127.6.853

del Cul, A., Baillet, S., & Dehaene, S. (2007). Brain Dynamics Underlying the Nonlinear Threshold for Access to Consciousness. *PLOS Biology*, *e260*, 5–10. https://doi.org/10.1371/journal.pbio

Dentale, F., Vecchione, M., Ghezzi, V., Spagnolo, G., Szemenyei, E., & Barbaranelli, C. (2020). Beyond an Associative Conception of Automatic Self-Evaluations: Applying the Relational Responding Task to Measure Self-Esteem. *Psychological Record*, *70*(2), 227–242. https://doi.org/10.1007/S40732-020-00392-4

di Pierro, R., Mattavelli, S., & Gallucci, M. (2016). Narcissistic Traits and Explicit Self-Esteem: The Moderating Role of Implicit Self-View. *Frontiers in Psychology*, *7*. https://doi.org/10.3389/fpsyg.2016.01815

Dzwonkowska, I., Lachowicz-Tabaczek, K., & Łaguna, M. (2008). *SES - Skala Samooceny Rosenberga*. Pracowna Testów Psychologicznych.

Gawronski, B. (n.d.). Attitudinal Effects of Stimulus Co-occurrence and Stimulus Relations: Paradoxical Effects of Cognitive Load. *PER-*

SONALITY AND SOCIAL PSYCHOLOGY BULLETIN. https://doi.org/10.1177/01461672211044322

Gawronski, B., Balas, R., & Creighton, L. A. (2014). Can the Formation of Conditioned Attitudes Be Intentionally Controlled? *Personality and Social Psychology Bulletin*, *40*(4), 419–432. https://doi.org/10.1177/0146167213513907

Gawronski, B., & Bodenhausen, G. v. (2006). Associative and propositional processes in evaluation: An integrative review of implicit and explicit attitude change. *Psychological Bulletin*, *132*(5), 692–731. https://doi.org/10.1037/0033-2909.132.5.692

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. https://doi.org/10.1111/SPC3.12124

Gawronski, B., & Bodenhausen, G. v. (2018). Evaluative Conditioning From the Perspective of the Associative-Propositional Evaluation Model. *Social Psychological Bulletin*, *13*(3). https://doi.org/10.5964/spb.v13i3.28024

Gawronski, B., de Houwer, J., & Sherman, J. W. (2020). Twenty-Five Years of Research Using Implicit Measures. *Social Cognition*, *38*, S1–S25. https://doi.org/10.1521/SOCO.2020.38.SUPP.S1

Gawronski, B., Mitchell, D. G. V., & Balas, R. (2015). Is evaluative conditioning really uncontrollable? A comparative test of three emotion-focused strategies to prevent the acquisition of conditioned preferences. *Emotion*, *15*(5). https://doi.org/10.1037/emo0000078

Gawronski, B., Rydell, R. J., de Houwer, J., Brannon, S. M., Ye, Y., Vervliet, B., & Hu, X. (2018). Contextualized Attitude Change. *Advances in Experimental Social Psychology*, *57*, 1–52. https://doi.org/10.1016/BS.AESP.2017.06.001

Gawronski, B., Ye, Y., Rydell, R. J., & de Houwer, J. (2014). Formation, representation, and activation of contextualized attitudes. *Journal of Experimental Social Psychology*, *54*, 188–203. https://doi.org/10.1016/j.jesp.2014.05.010

Greenwald, A. G., & Farnham, S. D. (2000). Using the implicit association test to measure self-esteem and self-concept. *Journal of Personality and Social Psychology*, *79*(6), 1022–1038. https://doi.org/10.1037/0022-3514.79.6.1022

Greenwald, A. G., McGhee, D. E., & Schwartz, J. L. K. (1998). Measuring Individual Differences in Implicit Cognition: The Implicit Association Test. *Journal of Personality and Social Psychology*, *74*(6), 1464–1480.

Greenwald, A. G., Nosek, B. A., & Banaji, M. R. (2003). Understanding and Using the Implicit Association Test: I. An Improved Scoring Algorithm. *Journal of Personality and Social Psychology*, *85*(2), 197–216. https://doi.org/10.1037/0022-3514.85.2.197

Grumm, M., Nestler, S., & Collani, G. von. (2009). Changing explicit and implicit attitudes: The case of self-esteem. *Journal of Experimental Social Psychology*, *45*(2), 327–335. https://doi.org/10.1016/j.jesp.2008.10.006

Hahn, A., & Gawronski, B. (2014). Do implicit evaluations reflect unconscious attitudes? *Behavioral and Brain Sciences*, *37*(1), 28–29. https://doi.org/10.1017/S0140525X13000721

Hahn, A., Judd, C. M., Hirsh, H. K., & Blair, I. v. (2014). Awareness of implicit attitudes. *Journal of Experimental Psychology: General*, *143*(3), 1369–1392. https://doi.org/10.1037/a0035028

Halbeisen, G., & Walther, E. (2015). Dual-task interference in evaluative conditioning: Similarity matters! *Quarterly Journal of Experimental Psychology*, *68*(10), 2008–2021. https://doi.org/10.1080/17470218.2014.1002506

Heycke, T., Gehrmann, S., Haaf, J. M., & Stahl, C. (2018). Of two minds or one? A registered replication of Rydell et al. (2006). *Cognition and Emotion*, *32*(8), 1708–1727. https://doi.org/10.1080/02699931.2018.1429389

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. https://doi.org/10.1037/a0018916

Hughes, S., de Houwer, J. de, & 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. https://doi.org/10.1037/XGE0000100

Hütter, M., & Rothermund, K. (2020). Automatic processes in evaluative learning. *Cognition and Emotion*, *34*(1), 1–20. https://doi.org/10.1080/02699931.2019.1709315

Hütter, M., & Sweldens, S. (2018). Dissociating Controllable and Uncontrollable Effects of Affective Stimuli on Attitudes and Consumption. *Journal of Consumer Research*, *45*(2), 320–349. https://doi.org/10.1093/jcr/ucx124

Johnson, M. (2016). Relations between explicit and implicit self-esteem measures and self-presentation. *Personality and Individual Differences*, *95*, 159–161. https://doi.org/10.1016/j.paid.2016.02.045

Karpen, S. C., Jia, L., & Rydell, R. J. (2011). Discrepancies between implicit and explicit attitude measures as an indicator of attitude strength. *Eur. J. Soc. Psychol.*, 6.

Karpen, S. C., Jia, L., & Rydell, R. J. (2012). Discrepancies between implicit and explicit attitude measures as an indicator of attitude strength. *European Journal of Social Psychology*, *42*(1), 24–29. https://doi.org/10.1002/ejsp.849

Koppehele-Gossel, J., Hoffmann, L., Banse, R., & Gawronski, B. (2020). Evaluative priming as an implicit measure of evaluation: An examination of outlier-treatments for evaluative priming scores. *Journal of Experimental Social Psychology*, *87*. https://doi.org/10.1016/j.jesp.2019.103905

Krizan, Z., & Suls, J. (2008). Are implicit and explicit measures of self-esteem related? A meta-analysis for the Name-Letter Test. *Personality and Individual Differences*, *44*(2), 521–531. https://doi.org/10.1016/j.paid.2007.09.017

Krosnick, J., & Petty, R. (1995). Attitude strength: An overview. In R. Petty & J. Krosnick (Eds.), *Ohio State University series on attitudes and persuasion, Vol. 4. Attitude strength: Antecedents and consequences* (Vol. 4, pp. 1–24). Lawrence Earlbaum Associates, Inc.

Kurdi, B., & Banaji, M. R. (2019). Attitude change via repeated evaluative pairings versus evaluative statements: Shared and unique features. *Journal of Personality and Social Psychology*, *116*(5). https://doi.org/10.1037/pspa0000151

Langer, T., Walther, E., Gawronski, B., & Blank, H. (2009). When linking is stronger than thinking: Associative transfer of valence disrupts the emergence of cognitive balance after attitude change. *Journal of Experimental Social Psychology*, *45*(6), 1232–1237. https://doi.org/10.1016/j.jesp.2009.07.005

Mattavelli, S., Richetin, J., & Perugini, M. (2019). Not all positive categories are alike: Exploring the superiority of the Self as a positive source for associative attitude change via intersecting regularities. *European Journal of Social Psychology*, *49*(3), 574–588. https://doi.org/10.1002/EJSP.2518

Mattavelli, S., Richetin, J., & Perugini, M. (2021). Using the Self-Referencing Task to Produce Durable Change on Food Evaluations Measured via the IAT. *International Review of Social Psychology*, *34*(1). https://doi.org/10.5334/IRSP.446/GALLEY/233/DOWNLOAD/

Mierop, A., Maurage, P., & Corneille, O. (2020). Cognitive Load Impairs Evaluative Conditioning, Even When Individual CS and US Stimuli are Successfully Encoded. *INTERNATIONAL REVIEW OF SOCIAL PSYCHOLOGY*, *33*(1), 1–7. https://doi.org/10.5334/irsp.339

Perkins, A. W., & Forehand, M. R. (2012). Implicit Self-Referencing: The Effect of Nonvolitional Self-Association on Brand and Product Attitude. *Journal of Consumer Research*, *39*(1), 142–156. https://doi.org/10.1086/662069

Rudman, L. A., Greenwald, A. G., Mellott, D. S., & Schwartz, J. L. K. (1999). Measuring the automatic components of prejudice: Flexibility and generality of the implicit association test. *Social Cognition*, *17*(4), 437–465. https://doi.org/10.1521/SOCO.1999.17.4.437

Rudman, L. A., Joshua Feinberg, J., & Fairchild, K. (2005). Minority Members' Implicit Attitudes: Automatic Ingroup Bias As A Function Of Group Status. *Http://Dx.Doi.Org/10.1521/Soco.20.4.294.19908*, *20*(4), 294–320. https://doi.org/10.1521/SOCO.20.4.294.19908

Rydell, R. J., McConnell, A. R., & Mackie, D. M. (2008). Consequences of discrepant explicit and implicit attitudes: Cognitive dissonance

and increased information processing. *Journal of Experimental Social Psychology*, *44*(6), 1526–1532. https://doi.org/10.1016/j.jesp.2008.07.006

Rydell, R. J., McConnell, A. R., Strain, L. M., Claypool, H. M., & Hugenberg, K. (2007). Implicit and explicit attitudes respond differently to increasing amounts of counterattitudinal information. *European Journal of Social Psychology*, *37*(5), 867–878. https://doi.org/10.1002/ejsp.393

Rydell, R., McConnell, A., Mackie, D., & Strain, L. (2006). Of Two Minds Forming and Changing Valence-Inconsistent Implicit and Explicit Attitudes. *Psychological Science*, *17*, 954–958. https://doi.org/10.1111/j.1467-9280.2006.01811.x

Shoda, T. M., McConnell, A. R., & Rydell, R. J. (2014). Having Explicit-Implicit Evaluation Discrepancies Triggers Race-Based Motivated Reasoning. *Social Cognition*, *32*(2), 190–202. https://doi.org/10.1521/soco.2014.32.2.190

Walther, E., Gawronski, B., Blank, H., & Langer, T. (2009). Changing likes and dislikes through the back door: The US-revaluation effect. *Cognition & Emotion*, *23*(5), 889–917. https://doi.org/10.1080/02699930802212423