Changing Deep-rooted Implicit Evaluation in the Blink of an Eye: Negative Verbal Information Shifts Automatic Liking of Gandhi

Pieter Van Dessel*, Yang Ye*, and Jan De Houwer

Ghent University, Belgium

* PVD and YY have contributed equally to the article.

Corresponding Authors: Pieter Van Dessel and Yang Ye, Department of Experimental Clinical and Health Psychology, Henri Dunantlaan 2, B-9000 Ghent (Belgium). E-mail: Pieter.vanDessel@UGent.be; Yang.Ye@UGent.be or Yang.Ye.Psych@gmail.com

Author Contributions: All authors were involved in developing the study concept and contributed to the design. Testing, data collection, and data-analyses were performed by P. Van Dessel and Y. Ye. P. Van Dessel and Y. Ye drafted the manuscript. J. De Houwer provided critical revisions. All authors approved the final version of the manuscript for submission.

Abstract

It is often assumed that, once established, spontaneous or implicit evaluations are resistant to immediate change. Recent research contradicts this theoretical stance, showing that a person's implicit evaluations of an attitude object can be changed rapidly in the face of new counter-attitudinal information. Importantly, it remains unknown whether such changes can also occur for deep-rooted implicit evaluations of well-known attitude objects. We address this question by examining whether the acquisition of negative information changes implicit evaluations of a well-known positive historic figure: Mahatma Gandhi. We report three experiments showing rapid changes in implicit evaluations of Gandhi as measured with an Affect Misattribution Procedure and Evaluative Priming Task but not with an Implicit Association Test (IAT). These findings suggest that implicit evaluations based on deep-rooted representations are subjective to rapid changes in the face of expectancy-violating information, while pointing to limitations of the IAT for assessing such changes.

Keywords: implicit evaluation, attitude change, counter-attitudinal information, IAT

Changing Deep-rooted Implicit Evaluation in the Blink of an Eye:

Negative Verbal Information Shifts Automatic Liking of Gandhi

How novel information changes evaluation is one of the core questions in attitude research. A particularly interesting case is when people learn counter-attitudinal information about attitude objects with which they have had consistently positive or consistently negative experiences over a long period of time. Imagine, for example, that people in romantic relationships find out that their trusted partners have been cheating on them; that loyal customers of a product learn that this product is made by child labor; or that long-time supporters of a presidential candidate wake up to news about a scandal involving the candidate. In these cases, will people update their evaluations on the basis of the new and shocking information, or are their evaluations too deeply rooted to change? Answers to this question are of great significance, as evaluative changes can lead to important behavioral changes (Ajzen & Fishbein, 2005), such as choosing to end a relationship, stop buying a product, or vote for another presidential candidate.

In the present paper, we focus on a particular type of evaluation that has received much attention in contemporary attitude research: implicit evaluation, defined as spontaneous evaluative responses emitted under conditions of automaticity (e.g., unintentional, uncontrolled, unconscious, or fast; De Houwer, 2009). Interestingly, initial research suggested that counter-attitudinal information does not produce observable changes in implicit evaluation, even when such changes are observed in self-reported, explicit evaluation (see Gawronski & Sritharan, 2010, for a review). For instance, Gregg, Seibt, and Banaji (2006) observed that implicit evaluations of novel social groups could be induced easily on the basis of verbal information about the behaviors and traits of those groups, but, in contrast to explicit evaluations, they could not be undone by counter-attitudinal information about those groups. Because implicit evaluations can continue to influence behavior even when one rejects such evaluations at the explicit level (e.g.,

in self-report measures), they are often considered to play a key role in so called irrational or unwanted behavior (e.g., psychopathology: Roefs et al., 2011; addiction: Wiers & Stacy, 2006; racial prejudice: Banaji & Greenwald, 2013).

The findings that implicit evaluations can be resistant to immediate change were often considered as supportive to the notion that implicit evaluation reflects the automatic activation of associations between representations in memory (e.g., representations of positive or negative concepts and attitude objects) (for a review, see Hughes, Barnes-Holmes, & De Houwer, 2011). A key assumption of this view is that, once established, such associations cannot easily be modified or erased. Instead, changes in implicit evaluation might require a gradual "re-wiring" of associations, which occurs mainly on the basis of repeated pairing of the attitude object with counter-attitudinal information (e.g., through evaluative conditioning: Hofmann, De Houwer, Perugini, Baeyens, & Crombez, 2010, or approach-avoidance training: Kawakami, Phills, Steele, & Dovidio, 2007).

The early idea (e.g., Smith & DeCoster, 2000) that implicit evaluation can be changed only via slow and incremental formation of associations was, however, challenged by recent studies demonstrating that, under certain conditions, implicit evaluations of target individuals can be revised on the basis of a single piece of verbal information (Brennan & Gawronski, 2017, Cone & Ferguson, 2015, Mann & Ferguson, 2015, Wyer, 2016). For instance, in Cone and Ferguson's studies, participants were exposed to a large amount of positive behavioral descriptions of a person named Bob so that they formed positive implicit evaluations of Bob. These positive evaluations, however, were immediately revised when the participants encountered a single piece of highly negative information about Bob (e.g., 'Bob was convicted of molesting children'). These results have important implications for understanding the mechanisms underlying implicit evaluation and for developing effective methods for changing

(unwanted) behaviors known to be predicted by automatic evaluations.

Notably, however, studies showing rapid revision of implicit evaluation have, until now, focused exclusively on changing recently formed evaluations of novel attitude objects (e.g., a stranger named Bob). This poses an important limitation to the theoretical implications of the findings, as associations underlying implicit evaluations of novel attitude objects might be weak and therefore susceptible to counter-attitudinal information (Hu, Gawronski, & Balas, 2017). In contrast, implicit evaluations of (certain) well-known attitude objects might be more deep-rooted in that they result from repeated exposure to evaluatively consistent information across multiple contexts and over a long period of time (e.g., over multiple decades). According to associative accounts (e.g., Smith & DeCoster, 2000), deep-rooted implicit evaluations might reflect firmly established associations that have been formed gradually and incrementally during long-term socialization experiences. From this perspective, changes in such evaluations, as opposed to recently learned evaluations of novel objects, should be especially unlikely to arise quickly. At the practical level, the real-life relevance of providing counter-attitudinal verbal information as a means to change implicit evaluation would be severely limited if it would affect only novel attitude objects. Hence, it is important to test whether counter-attitudinal information can also lead to changes in deep-rooted implicit evaluations of well-known attitude objects.

Some studies have already demonstrated rapid changes in (deep-rooted) implicit evaluations of well-known attitude objects that are ambiguous (i.e., with both positive and negative evaluative features). For instance, Smith and De Houwer (2015) observed that smokers' implicit evaluations of smoking became more negative after they read anti-smoking messages. In these cases, however, changes in implicit evaluations might reflect context-dependent activation of existing positive or negative aspects of the overall representation of the ambiguous attitude

object, instead of the integration of new information into this representation (Mitchell, Nosek, & Banaji, 2003; see Blair, 2002, for a review). To make sure that any potential changes in implicit evaluations of well-known attitude objects are not merely instances of context-dependent malleability of implicit evaluations, it is important to examine attitude objects that are unambiguously positive or negative. For such unambiguous targets, any effect of counter-attitudinal information is likely to be due to new learning rather than selective activation of existing old information.

In the present study, we examined whether deep-rooted implicit evaluations of a well-known and unambiguously positive attitude object – Mahatma Gandhi – can be quickly changed on the basis of novel information. We use the term *deep-rooted* to describe an evaluation that is (1) learned a long time ago, (2) held with a strong level of confidence, and (3) highly accessible. In a pre-registered study with a sample of participants (N = 100) that were recruited in the same way as the participants of Experiments 2 and 3, we examined whether participants' evaluations of Gandhi fit this definition of deep-rooted. Results showed that, on average, participants first learned about Gandhi when they were 13 years old (22 years ago on average). Moreover, compared to other famous people (i.e., the ten most famous historical people as proposed by Skiena & Ward, 2013), participants (1) felt more confident about their evaluation of Gandhi and (2) were faster in the speeded evaluation of Gandhi (indicating greater accessibility of the evaluation: Fazio, Powell, & Williams, 1989). In this study, we also asked participants to indicate their liking of Gandhi and to complete an attitude ambivalence measure (Thompson, Zanna, & Griffin, 1995). Results indicated that (1) participants had mostly positive explicit evaluations of Gandhi and (2) participants felt less ambivalent towards Gandhi compared to other famous

people (with low ambivalence scores for Gandhi, overall, according to criteria of Thompson et al.).[1]

For most people, the information that they previously learned about Gandhi is likely to be consistently positive, such as stories about his peaceful movement for civil rights and freedom around the world and numerous positive traits ascribed to him (e.g., faithful, persistent, peaceful). Unbeknownst to most people, however, certain past actions of Gandhi do not accord with this predominantly positive view. We performed three experiments, in which we told participants about such past actions of Gandhi and examined subsequent changes in implicit evaluations.

Method

Participants and design

In Experiment 1, a total of 64 native Dutch-speaking undergraduates (46 women) were recruited from a participant pool at Ghent University, Belgium. In Experiments 2 and 3, 205 and 150 participants (Experiment 2: 84 women; Experiment 3: 56 women) were recruited from the online participant recruitment platform Prolific Academia (https://www.prolific.ac/). In each experiment, we adopted a different implicit evaluation measure that has been widely used in previous research: the Implicit Association Test (IAT, Greenwald, McGhee, & Schwartz, 1998) in Experiment 1, the Affect Misattribution Procedure (AMP, Payne, Cheng, Govorun, & Stewart, 2005) in Experiment 2, and the Evaluative Priming Task (EPT, Fazio, Sanbonmatsu, Powell, & Kardes, 1986) in Experiment 3. The main reason for including different measures is that these measures have been shown to be sensitive to different non-evaluative factors and therefore sometimes produce divergent findings (De Houwer, 2003; Gawronski & De Houwer, 2014).

[1] A more detailed description of the results as well as the pre-registered plan, data, and analysis code for this study are available at https://osf.io/b6t5a/.

Sample sizes were determined on the basis of an a priori power analysis taking into account the results of the previous experiments such that the sample size would provide sufficient power (i.e., power > 0.75) to detect a medium effect in Experiment 1 ($d \approx 0.60$) and a small to medium effect in Experiments 2 ($d \approx 0.35$) and 3 ($d \approx 0.40$). Prior to data-collection, target sample size was pre-registered together with the study design, data-analysis plan and the described hypotheses for all experiments on the Open Science Framework website (https://osf.io). The pre-registered plans, experiment script, stimuli, data, and analysis code are available at https://osf.io/56xmy/.

In line with recommendations by Zhou and Fishbach (2016) to prevent selective attrition, on line participants were informed about the duration of the experiment and warned that dropping out could negatively affect science. Overall dropout rate was 5.4% (Experiment 2) and 4.0% (Experiment 3). The dropout rates were comparable across the two text conditions, χ^2s < 1.01, *ps* > .35. Hence, there was no evidence for condition-dependent attrition.

Following our preregistered data-analysis plans, we excluded the data from participants who (1) had error rates above 30% when considering all critical IAT test blocks or above 40% for any one of these blocks (Experiment 1: 2 participants, 3.1%), (2) responded either “positive” or “negative” to all AMP trials in any of the AMP blocks (Experiment 2: 11 participants, 5.7%), or (3) had error rates in the EPT that exceeded the population mean by more than 2.5 standard deviations (Experiment 3: 4 participants, 2.8%). Analyses were performed on the data of 62, 183, and 140 participants.

Procedure

After providing informed consent, participants were told that they would read a text that describes historically true events. They were instructed to read the text thoroughly so that they

would be able to answer questions about it in a later phase of the experiment. In Experiments 1 and 3, participants were then presented with a text about Mahatma Gandhi. Participants in Experiment 2 completed an AMP assessing implicit evaluations of Gandhi before reading the text. In all experiments, half of the participants read a text that described how, due to religious reasons, Gandhi refused the use of modern medicines on his sick wife who died afterwards, yet allowed the same treatment on himself and recovered when he fell ill shortly thereafter (negative Gandhi text). The other participants read a text that described how Gandhi lived in South-Africa for a short while (neutral Gandhi text). This text was descriptive in nature and did not refer to any of the normatively good or bad actions that Gandhi had done in his life. Texts were adapted versions of historical texts about Gandhi (Chadha, 1997; Wolpert, 2001) and are provided in the Supplementary Material.

The reading task was followed by an implicit evaluation task for half of the participants. The other participants first completed explicit ratings. In Experiment 1, the implicit evaluation task was an IAT in which participants categorized eight attribute words (e.g., the Dutch words for pleasant and evil) as ‘bad’ or ‘good’ and the names of Gandhi (Gandhi or Mahatma Gandhi) and four other Indian people (e.g., Sristavi or Zohar Ganguly) as ‘Gandhi’ or ‘Not Gandhi’. Because there is strong evidence that intentional recoding of IAT categories can produce unwanted IAT effects (Rothermund, Teige-Mocigemba, Gast, & Wentura, 2009), we used an IAT that is designed to disentangle these processes from attitudinal processes with a Process Dissociation Procedure (the ReAL model for IAT performance). The IAT followed the procedure described in more detail by Meissner and Rothermund (2013) with the exception of the target categories and target stimuli that were used.

Participants in Experiment 2 completed the AMP before and after reading the Gandhi text. In line with standard procedures (Payne et al., 2005), the AMP consisted of 80 trials in which

participants were presented with a prime stimulus for 75ms, a blank screen for 125ms, and a Chinese ideograph for 100ms, which was then covered with a black-and-white pattern mask. Participants were asked to indicate if they considered the Chinese ideograph more or less visually pleasant than average by pressing either "E" or "I", respectively. Gandhi-related priming stimuli included the Gandhi names (Gandhi or Mahatma Gandhi) as well as four black and white photographs of Gandhi. Control priming stimuli included two random letter strings of equal length to the Gandhi names (GLjnkO or MfBznXh HgJszr) and four black and white photographs of other men that were rated as evaluatively neutral in a previous validation study.

In Experiment 3, participants completed an EPT that adopted standard procedures (Spruyt, De Houwer, Hermans, & Eelen, 2007). The EPT comprised 8 practice trials and 80 test trials. A single trial consisted of the presentation of a fixation cross for 500 ms, a blank screen for 500 ms, a prime for 200 ms, a post-prime interval for 50 ms, and the presentation of a target word for a maximum of 1500 ms. Targets consisted of 10 positive words (e.g., the words pleasant and positive) and 10 negative words (e.g., the words unkind and annoying). The same prime stimuli were used as in Experiment 2.

Participants indicated their self-reported liking of Gandhi by answering two questions: "How much do you like or dislike Gandhi?" and "To what extent do you have warm feelings towards Gandhi?" on two 9-point Likert scales ranging from 1 (*I dislike him a lot/Not at all*) to 9 (*I like him a lot / a lot*). After completing explicit liking ratings, participants in Experiments 2 and 3 also reported how likely they would be to visit a Gandhi museum if they were in India on a 7-point Likert scale ranging from 1 (*Highly unlikely*) to 7 (*Highly likely*).

Participants were then asked to complete some final questions. First, they indicated to what extent they believed the story they had read to be historically true on a 7-point Likert scale.

The second question asked whether participants had heard about Gandhi before participating in the experiment (response options: yes/no). In Experiments 2 and 3, participants also answered two separate questions asking whether they had previously encountered any positive information about Gandhi or whether they had previously encountered any negative information about Gandhi (response options: yes/no). Finally, participants indicated their previous liking of Gandhi by reporting to what extent they liked Gandhi before participating in the experiment on a 7-point Likert scale. Participants were then thanked for their participation and debriefed. Debriefing included informing participants that the text about Gandhi was adopted from historical sources but that they should still be critical about the content.

Results

Implicit evaluations

IAT (Experiment 1). In accordance with Meissner and Rothermund (2013), trials were excluded when (1) participants emitted an incorrect response (18.0%), (2) response times were faster than 250 ms (0.5%), or (3) response times exceeded the 75th percentile of response times by more than three interquartile ranges (1.7%). IAT latencies were subjected to a 2 × 2 analysis of variance (ANOVA) with Text (negative Gandhi text, neutral Gandhi text) as between-subjects factor and IAT Block (Gandhi-good block, Gandhi-bad block) as within-subjects factor. We observed a main effect of IAT Block, $F(1,60) = 136.80$, $p < .001$, $\eta^2 = 0.11$. Participants were faster in blocks where Gandhi and good were assigned to the same key ($M = 528$, $SD = 45$) than in blocks where Gandhi and bad were assigned to the same key ($M = 560$, $SD = 50$), $d_z = 0.68$, 95% confidence interval of the differences (CI diff) = [-38, -27]. Importantly, however, we did not observe an interaction effect of Text and IAT Block, $F(1,60) = 0.34$, $p = .56$, $\eta^2 < 0.01$. Bayesian analyses indicated substantial evidence for the null hypothesis (reflecting the absence of

a significant effect), $BF_0 = 3.34$. We also calculated IAT scores that integrated errors and RTs using the D4-algortithm recommended by Greenwald, Nosek, and Banaji (2003). On these IAT scores we also observed the compatibility effect (mean D score indicating faster and more accurate responses in the compatible block compared to the incompatible block = 0.65, *SD* = 0.55), $t(61) = 9.26$, $p < .001$, $d_z = 1.18$, but no moderation by the Text condition, $t(60) = 1.03$, $p = .31$, $BF_0 = 2.47$, $\eta^2 < 0.01$.

Fitting the ReAL Model on the IAT data, we obtained a good model fit for participants in both text conditions, median G^2s < 5.32, *ps* > .26. The Re parameter of the model was significant in both conditions, indicating that cognitive recoding of the IAT categories contributed to IAT performance, *t*s > 10.87, *ps* < .001, d_zs > 1.95. Importantly, the A parameter for Gandhi was not significantly higher than 0.5, indicating that IAT performance did not reflect expression of any positive attitude towards Gandhi in the negative Text group (*M* = 0.40, *SD* = 0.16) or in the neutral Text group (*M* = 0.37, *SD* = 0.16), *t*s < –4.74, *p*s > .99. This suggests that recoding processes hindered the IAT from capturing positive (or negative) implicit evaluations of Gandhi in both groups. We did not observe differences in any of the parameters between the two groups, *t*s < 0.79, *p*s > .43.

AMP (Experiment 2). We calculated eight AMP scores as the percentages of 'pleasant' responses for each of the four prime type trials (i.e., trials with Gandhi names, Gandhi photos, neutral letter strings, and control photos as prime) at each time of assessment (i.e., pre-reading and post-reading). The AMP scores were subjected to analyses with item-based linear mixed effects (lme) models. Because these models allow us to control for possible effects of random factors they are often recommended for analyzing performance in (evaluative) priming tasks (e.g., Gast, De Houwer, & De Schryver, 2012). We tested a model that included Text (negative

Gandhi text vs. neutral Gandhi text), Content of Prime (Gandhi vs. control), and Time of Assessment (pre-reading vs. post-reading) as fixed factors and Participant as random factor. The inclusion of the variables Task Order (implicit/explicit evaluation task first) and Type of Prime (picture vs. text) as fixed or random factors did not improve model fit so they were not included in the analyses. The analyses revealed the crucial interaction effect of the three fixed factors, $\chi^2(1) = 8.51$, $p = .004$. To examine this interaction, we calculated an AMP score reflecting the difference between the proportion of pleasant responses after a Gandhi prime and the proportion of pleasant responses after another prime. The AMP score for participants who read the neutral Gandhi text was not significantly different before reading the Gandhi text ($M = 0.20$, $SD = 0.30$) compared to after reading the text ($M = 0.17$, $SD = 0.26$), $t(87) = –0.81$, $p = .42$, $d_z = 0.09$, CI diff = [–0.09, 0.04], $BF_0 = 6.17$. In contrast, the AMP score of participants who read the negative Gandhi text was significantly more positive before reading the text ($M = 0.14$, $SD = 0.24$) compared to after reading the text ($M = –0.03$, $SD = 0.21$), $t(94) = –4.56$, $p < .001$, $d_z = 0.47$, 95% CI diff = [–0.20, –0.08], $BF_1 = 1083.71$.

EPT (Experiment 3). In line with standard procedures for analyzing evaluative priming reaction time data (e.g., Spruyt et al., 2007), trials with an incorrect response were dropped (2.8%) as well as any trials in which reaction times (RTs) were at least 2.5 standard deviations removed from an individual's mean (2.9%). RTs were subjected to an lme analysis with Target (negative Gandhi text vs. neutral Gandhi text), Content of Prime (Gandhi vs. control), and Target Type (positive vs. negative) as fixed factors. We included a random intercept for Participant and Target Word and a random slope for Target Type by participant. The inclusion of the variables Task Order and Type of Prime as fixed or random factor did not improve model fit so they were not included in the analyses. We observed the crucial interaction effect of the three fixed factors,

$\chi^2(1) = 4.78$, $p = .029$. To examine the interaction we compared RTs for trials with positive targets and Gandhi as prime or with negative targets and other primes (compatible trials) to RTs for trials with negative targets and Gandhi as prime or positive targets and other primes (incompatible trials) for the two experimental conditions. For participants who read the neutral Gandhi text, RTs were faster on compatible trials (M = 622 ms, SD = 171 ms) than on incompatible trials (M = 634 ms, SD = 180 ms), $\chi^2(1) = 9.58$, $p = .002$, $d_z = 0.86$, 95% CI diff = [–18, –5], $BF_1 = 17.62$. For participants who read the negative Gandhi text, we did not observe this effect (compatible trials: M = 625 ms, SD = 181 ms; incompatible trials: M = 624 ms, SD = 188 ms), $\chi^2(1) = 0.00$, $p = .97$, $d_z = 0.01$, 95% CI diff = [–9, 9], $BF_0 = 7.51$.

Explicit evaluations

Ratings of pleasantness of and warm feelings for Gandhi were aggregated into a single score by averaging the respective scores (Cronbach's α = [.92, .94]) We submitted these scores to an analysis of covariance (ANCOVA) with Text as between-subjects factor and Previous Liking Score (Experiment 1: M = 5.60, SD = 1.55, Experiment 2: M = 5.22, SD = 1.11, Experiment 3: M = 5.05, SD = 1.27) included as covariate. In all experiments, the main effect of Previous Liking Score was significant, Fs > 13.12, ps < .007, η^2s > 0.18. More importantly, we also observed a main effect of Text, Fs > 17.84, ps < .001, η^2s > 0.23. Participants rated Gandhi more negatively when they had read the negative Gandhi text (Experiment 1: M = 2.97, SD = 1.67; Experiment 2: M = 3.04, SD = 1.43; Experiment 3: M = 3.03, SD = 1.31) than when they had read the neutral Gandhi text (Experiment 1: M = 4.91, SD = 1.77; Experiment 2: M = 5.69, SD = 0.85; Experiment 3: M = 5.42, SD = 1.16), ts > 12.41, ps < .001, ds > 1.98, BF_1s > 10^5.

Behavioral intention

For Experiments 2 and 3 we performed an ANCOVA on participants' intention to visit a Gandhi museum with Text as between-subjects factor and Previous Liking included as covariate. The analysis returned a significant main effect of Previous Liking, Fs > 21.96, ps < .001, η^2s > 0.14, as well as a significant main effect of Text, Fs > 7.53, ps < .007, η^2s > 0.08. Participants who read the neutral text indicated a stronger intention to visit a Gandhi museum (Experiment 2: M = 5.72, SD = 0.96; Experiment 3: M = 5.52, SD = 1.23) than participants who read the negative text (Experiment 2: M = 4.81, SD = 1.57; Experiment 3: M = 4.72, SD = 1.66), ds > 0.69, BF_1s > 296.67.

Additional analyses

All analyses were also performed for the subset of participants (Experiment 1: 75.8%; Experiment 2: 68.3%, Experiment 3: 45.0%) who (1) indicated that they had heard about Gandhi before, (2) gave previous liking ratings above the neutral mid-point of the scale and (3) had heard positive information about Gandhi before the experiment but not any negative information (Experiments 2 and 3). These analyses supported the conclusions of the main analyses, including the effect of Text on implicit evaluations in Experiments 2 and 3 but not in Experiment 1.

Discussion

The current experiments provide the first evidence that deep-rooted positive implicit evaluations of a highly familiar and unambiguously positive attitude object (i.e., Gandhi) can quickly change as the result of the acquisition of counter-attitudinal information. On a theoretical level, this finding contradicts the idea that changes in implicit evaluations require the gradual re-wiring of acquired associations on the basis of repeated pairings of the attitude object with positive or negative events (Rydell, McConnell, Mackie, & Strain, 2006; Smith & DeCoster, 2000). It only took a single piece of negative verbal information to override the effects of

extensive past learning on implicit evaluations of a well-known attitude object. The observed changes in implicit evaluation are unlikely to be attributed to selective activation of existing knowledge (Mitchell et al., 2003), as they were observed even for participants who indicated that they had never heard any negative information about Gandhi prior to their participation.

The current findings suggest that mental representations underlying even deep-rooted implicit evaluation can rapidly integrate novel information, which strongly contrasts with the typical conceptualization of such representations as mental associations (e.g., Smith & DeCoster, 2000). Instead, the findings fit better with the idea that that implicit evaluation is driven by propositional processes. The observed changes in implicit evaluations of Gandhi might reflect the integration of new propositional information (e.g., the proposition 'Gandhi caused the death of his own wife') into a system of propositional knowledge, which is more flexible than slow-learning associative systems. According to the propositional account of implicit evaluation (De Houwer, 2014), such newly acquired propositional knowledge can directly guide implicit evaluation. Alternatively, propositional learning might determine the formation of associations, as suggested by some dual-process accounts of implicit evaluation (e.g., Gawronski & Bodenhausen, 2006). Yet, dual-process accounts cannot explain the current findings unless they incorporate the non-trivial assumption that the impact of newly formed associations via propositional learning can overrule the impact of previously formed associations on implicit evaluations.

The current results accord with (and extend) recent studies that showed rapid changes in (newly formed) implicit evaluations on the basis of counter-attitudinal information (see Cone, Mann, & Ferguson, in press, for an overview). These studies provided evidence that the effectiveness of counter-attitudinal information depends on the diagnosticity of the information

(Cone & Ferguson, 2015) and whether the information facilitates re-interpretation of prior information (Mann & Ferguson, 2015). In the current study, we provided truthful information about Gandhi that we considered diagnostic about the valence of Gandhi. However, we did not probe characteristics of this information. An important direction of future research is to identify and examine properties of counter-attitudinal information that are crucial determinants of effects on deep-rooted implicit evaluations.

On a practical level, the current findings add to research showing that a single piece of verbal information can have powerful effects on (automatic) behavior (e.g., Meiran, Liefooghe, & De Houwer, in press) and, specifically, implicit evaluation (e.g., Kurdi & Banaji, 2017). The findings suggest that not only newly established, but also deep-rooted evaluative behaviors can be changed quickly on the basis of information that contradicts previous knowledge. Importantly, the observed changes are not small and trivial but indicate important revisions of previous liking (e.g., moderate to large effect size differences between conditions: d_z = 0.38-0.85). These findings set the stage for exploring the potential of using verbal information to modify other presumably deep-rooted behaviors that might depend on automatic evaluative processes (e.g., addictive or prejudice behaviors; see Smith & De Houwer, 2015, for an initial attempt in the context of smoking).

The current findings also point to limitations of popular implicit evaluation measures. In contrast to the AMP and EPT, the IAT did not show any effect of the counter-attitudinal information on task performance. It is unlikely that this dissociation is (1) due to a lack of power in Experiment 1, as Bayesian factors indicated substantial evidence for the absence of an effect, or (2) due to sample differences, as the effects on other measures (i.e., explicit rating scales) were of similar size across the three experiments. A possible explanation is that IAT performance was

affected by non-evaluative processes. In line with this assumption, findings from the ReAL modeling analysis did not reveal any contribution of an attitudinal component to IAT performance in either of the text conditions and provided strong evidence for cognitive recoding of the IAT categories (see Rothermund et al., 2009). These findings are consistent with previous research showing that IAT effects can be driven by the ease to categorize concepts with the same key, which can depend on non-evaluative properties of the target stimuli (e.g., salience asymmetries, Rothermund & Wentura, 2004). Another possible explanation is that IAT performance reflects normative or culturally shared knowledge (see Karpinski & Hilton, 2001) instead of personal liking. These confounds might explain why the IAT is sometimes outperformed by other implicit measures in predicting real-life behaviors (e.g., Spruyt et al., 2015), as well as the observed lack of changes in implicit evaluations on the basis of counter-attitudinal information in previous studies with IAT measures of implicit evaluation (e.g. Gregg et al., 2006; Rydell et al., 2006). It is important to note that several studies have shown rapid and robust changes in IAT scores (e.g., Mitchell et al., 2003; Van Dessel, De Houwer, Gast, Smith, & De Schryver, 2016). However, a key difference with the current work is that those studies focused on recently formed evaluations of novel attitude objects or on evaluations of ambivalent attitude objects. It is possible that the mechanism that hindered the IAT from capturing evaluations in Experiment 1 (e.g., recoding) is more active when measuring evaluations of well-known attitude objects with a (previously) predominant valence.

Note that neither IAT, AMP, or EPT scores showed a full reversal of implicit evaluations on the basis of our manipulation. AMP and EPT scores were negative for participants who read the negative Gandhi text in absolute terms (i.e., participants provided more unpleasant responses to Chinese ideographs in the AMP and faster negative than positive target categorizations in the

EPT on Gandhi prime trials than on control prime trials), but this pattern was not statistically significant. This contrasts with these participants' explicit ratings of Gandhi, which were significantly lower than the neutral mid-point of the rating scale. This might suggest that explicit evaluation is more strongly influenced by our manipulation than implicit evaluation, which could reflect differences in the processes underlying learning (or expression) of implicit and explicit evaluation. However, it is also possible that this dissociation is due to procedural or scoring-related differences between implicit and explicit evaluation measurement that do not reflect differences in the underlying construct (e.g., only implicit evaluation scores are computed in comparison with control stimuli, only EPT scores are based on RT differences,…). In fact, interpreting scores from implicit evaluation measures as positive, neutral, or negative in an absolute sense is often considered arbitrary because these scores might not reflect the zero point on an underlying psychological dimension (for a discussion see Blanton & Jaccard, 2006).

There were several limitations to the present research. First, we focused on changes in positive implicit evaluations. Because existing research suggests that it can be more difficult to change recently established negative than positive evaluations (e.g., Cone & Ferguson, 2015), it might be interesting to examine whether counter-attitudinal information can also lead to changes in negative implicit evaluations of highly familiar negative attitude objects (e.g., Hitler). Second, the current findings are silent about whether the observed changes in deep-rooted implicit evaluations persist over time or generalize across contexts. It is possible that negative information becomes highly accessible immediately after acquisition, when its impact on implicit evaluation peeks (but see Mann & Ferguson, 2017, for evidence that changes in newly learned implicit evaluations do persist over time). It is also possible that the observed changes in implicit evaluations are context-dependent (e.g., limited to the context in which they were learned, see

Gawronski, Rydell, De Houwer, Brannon, Ye, Vervliet, et al., 2017). This might occur when the newly formed representations underlying these implicit evaluations are contextualized (rather than integrated with previously existing representations) such that they bias evaluative responses only in the context in which they were acquired. Future research should examine the characteristics of changes in deep-rooted implicit evaluations (e.g., context-dependence, robustness) to provide more insights into boundary conditions as well as underlying cognitive mechanisms.

A third limitation is that, although most participants reported that they could not explicitly recall having learned anything negative about Gandhi before participation, we cannot rule out the possibility that some participants did have (implicit) negative information available about Gandhi. Hence, it is still possible that changes in deep-rooted implicit evaluations resulted from the selective activation of previously learned (rather than novel) negative information. To explore this possibility, we performed a t-test analysis on AMP scores of participants in the negative Gandhi text condition of Experiment 2 (i.e., the only experiment for which we have data about participants' implicit evaluations before and after reading the text). We found that the observed negative shift in AMP scores was not more pronounced for participants who indicated that they had heard negative information about Gandhi before.[2] These results contrast with the idea that participants who have negative information available about Gandhi are more susceptible to changes in implicit evaluations.

[2] The negative shift in AMP scores after reading the negative Gandhi text was not more pronounced for participants who indicated that they had heard negative information about Gandhi before (N = 9, M = -0.03, SD = 0.32) than for participants who indicated that they had never heard negative information about Gandhi (N = 86, M = -0.15, SD = 0.30), $t(93) = -1.22$, $p = .89$. Bayesian factors indicated substantial evidence for this null effect, $BF_0 = 5.81$.

Finally, although Gandhi is a highly familiar target, attitudes towards Gandhi might be of low relevance to most people to the extent that it is unrelated to people's self-interest or not central to their values. Future research could examine whether attitudes that are highly relevant to people can also be changed on the basis of a single piece of counter-attitudinal information.

In sum, the present study demonstrates that the implicit liking of highly-familiar targets can be quickly changed on the basis of a single piece of negative information. Contrary to what some may think, even deep-rooted spontaneous preferences can be readily revised.

Acknowledgments

We thank Franziska Meissner for her help with the ReAL model analyses of IAT data.

References

Ajzen, I., & Fishbein, M. (2005). The influence of attitudes on behavior. In D. Albarracín, B. T. Johnson, & M. P. Zanna (Eds.), *The handbook of attitudes* (pp. 173-221). Mahwah, NJ: Erlbaum.

Banaji, M. R., & Greenwald, A. G. (2013). Blindspot: Hidden biases of good people. New York, NY: Random House.

Blair, I. V. (2002). The malleability of automatic stereotypes and prejudice. *Personality and Social Psychology Review, 6*, 242-261.

Blanton, H., & Jaccard, J. (2006). Arbitrary metrics in psychology. *American Psychologist, 61,* 27–41. doi:10.1037/0003-066X.61.1.27

Brannon, S. M., & Gawronski, B. (2017). A second chance for first impressions? Exploring the context (in)dependent updating of implicit evaluations. *Social Psychological and Personality Science, 3*, 275-283.

Chadha, Y. (1997). Gandhi: A Life. New York: John Wiley and Sons, Inc.

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

Cone, J., Mann, T. C., & Ferguson, M. J. (in press). Can we change our implicit minds? New evidence for how, when, and why implicit impressions can be rapidly revised. *Advances in Social Psychology*.

De Houwer, J. (2003). A structural analysis of indirect measures of attitudes. In J. Musch & K.C. Klauer (Eds.), *The Psychology of Evaluation: Affective Processes in Cognition and Emotion* (pp. 219-244). Mahwah, NJ: Lawrence Erlbaum.

De Houwer, J. (2009). How do people evaluate objects? A brief review. *Social and Personality Psychology Compass, 3*, 36-48.

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

Fazio, R. H., Sanbonmatsu, D. M., Powell, M. C., & Kardes, F. R. (1986). On the automatic activation of attitudes. *Journal of Personality and Social Psychology*, *50*, 229–238. doi: 10.1037/0022-3514.50.2.229.

Gast, A., De Houwer, J., & De Schryver, M. (2012). Evaluative conditioning can be modulated by memory of the CS-US Pairings at the time of testing. *Learning and Motivation, 43*, 116-126. doi: 10.1016/j.lmot.2012.06.001

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

Gawronski, B., & De Houwer, J. (2014). Implicit measures in social and personality psychology. In H. T. Reis, & C. M. Judd (Eds.), *Handbook of research methods in social and personality psychology* (2nd edition, pp. 283–310). New York: Cambridge University Press.

Gawronski, B., Rydell, R. J., De Houwer, J., Brannon, S. M., Ye, Y., Vervliet, B., & Hu, X. (2017). Contextualized attitude change. *Advances in Experimental Social Psychology, 55*.

Gawronski, B., & Sritharan, R. (2010). Formation, change, and contextualization of mental associations: Determinants and principles of variations in implicit measures. In B. Gawronski & B. K. Payne (Eds.), *Handbook of implicit social cognition: Measurement, theory, and applications* (pp. 216–240). New York: Guilford Press.

Greenwald, A. G., McGhee, D. E., & Schwartz, J. L. (1998). Measuring individual differences in implicit cognition: the implicit association test. *Journal of Personality and Social Psychology*, *74*, 1464–80. doi:10.1037/0022-3514.74.6.1464

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

Gregg, A. P., Seibt, B., & Banaji, M. R. (2006). Easier done than undone: asymmetry in the malleability of implicit preferences. *Journal of Personality and Social Psychology, 90*, 1–20. doi: 10.1037/0022-3514.90.1.1

Hofmann, W., De Houwer, J., Perugini, M., Baeyens, F., & Crombez, G. (2010). Evaluative conditioning in humans: a meta-analysis. *Psychological Bulletin, 136*, 390–421. doi:10.1037/a0018916

Hu, X., Gawronski, B., & Balas, R. (2017). Propositional versus dual-process accounts of evaluative conditioning: I. The Effects of Co-Occurrence and Relational Information on Implicit and Explicit Evaluations. *Personality and Social Psychology Bulletin, 43*, 43, 17–32. doi:10.1177/0146167216673351

Hughes, S., Barnes-Holmes, D., & De Houwer, J. (2011). the Dominance of Associative Theorizing in Implicit Attitude Research: Propositional and Behavioral Alternatives. *Psychological Record*, *61*, 465–496.

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

Kawakami, K., Phills, C. E., Steele, J. R., & Dovidio, J. F. (2007). (Close) distance makes the heart grow fonder:Improving implicit racial attitudes and interracial interactions through approach behaviors. *Journal of Personality and Social Psychology, 92*, 957–971. doi: 10.1037/0022-3514.92.6.957

Kurdi, B., & Banaji, M. R. (2017). Repeated evaluative pairings and evaluative statements: How effectively do they shift implicit attitudes? *Journal of Experimental Psychology: General, 146*, 194–213. doi: 10.1037/xge0000239

Mann, T. C., & Ferguson, M. J. (2015). Can we undo our first impressions? The role of reinterpretation in reversing implicit evaluations. *Journal of Personality and Social Psychology, 108*, 823-849.

Mann, T. C., & Ferguson, M. J. (2017). Reversing implicit first impressions through reinterpretation after a two-day delay. *Journal of Experimental Social Psychology, 68*, 122–127.

Meiran, M., Liefooghe, B., & De Houwer, J. (in press). Powerful instructions: Automaticity without practice. *Current Directions in Psychological Science.*

Meissner, F. & Rothermund, K. (2013). Estimating the contributions of associations and recoding in the Implicit Association Test: The ReAL model for the IAT. *Journal of Personality and Social Psychology, 104*,45-69. doi:10.1037/a0030734

Mitchell, J. A., Nosek, B. A., & Banaji, M. R. (2003). Contextual variations in implicit evaluation. *Journal of Experimental Psychology: General, 132*, 455–469.

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*, 277–293. doi: 10.1037/0022-3514.89.3.277

Roefs, A., Huijding, J., Smulders, F. T. Y., MacLeod, C. M., de Jong, P. J., Wiers, R. W., & Jansen, A. T. M. (2011). Implicit measures of association in psychopathology research. *Psychological Bulletin, 137*, 149–193. doi: 10.1037/a0021729

Rothermund, K., Teige-Mocigemba, S., Gast, A., & Wentura, D. (2009). Minimizing the influence of recoding in the Implicit Association Test: The Recoding-Free Implicit Association Test (IAT-RF). *The Quarterly Journal of Experimental Psychology, 62*, 84–98. doi:10.1080/17470210701822975

Rothermund, K., & Wentura, D. (2004). Underlying processes in the Implicit Association Test (IAT): Dissociating salience from associations. *Journal of Experimental Psychology: General, 133*, 139–165. doi:10.1037/00963445.133.2.139

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

Skiena, S. S., & Ward, C. B. (2013). Who's bigger? Where historical figures really rank. Cambridge University Press.

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

Smith, C. T., & De Houwer, J. (2015). Hooked on a feeling: Affective anti-smoking messages are more effective than cognitive messages at changing implicit evaluations of smoking. *Frontiers in Psychology, 6*:1488. doi:10.3389/fpsyg.2015.01488

Spruyt, A., De Houwer, J., Hermans, D., & Eelen, P. (2007). Affective priming of non-affective semantic categorization responses. *Experimental Psychology, 54,* 44–53. doi: 10.1027/1618-3169.54.1.44

Spruyt, A., Lemaigre, V., Salhi, B., Van Gucht, D., Tibboel, H., Van Bockstaele, B., De Houwer, J., Van Meerbeeck, J., & Nackaerts, K. (2015). Implicit attitudes towards smoking predict long-term relapse in abstinent smokers. *Psychopharmacology, 232,* 2551-2561.

Thompson, M. M., Zanna, M.P., & Griffin, D.W. (1995). Let's Not Be Indifferent about (Attitudinal) Ambivalence, in: *Attitude Strength: Antecedents and Consequences*, ed. Petty, R. E. & Krosnick, J.A. Mahwah, NJ: Erlbaum, 361–86.

Van Dessel, P., De Houwer, J., Gast, A., Smith, C. T., & De Schryver, M. (2016). Instructing implicit processes: When instructions to approach or avoid influence implicit but not explicit evaluation. *Journal of Experimental Social Psychology, 63,* 1-9. doi:10.1016/j.jesp.2015.11.002

Wiers, R. W., & Stacy, A. W. (2006). Implicit cognition and addiction. *Current Directions in Psychological Science, 15,* 292–296. doi: 10.1111/ j.1467-8721.2006.00455.x

Wolpert, S. (2001). Gandhi's Passion: The Life and Legacy of Mahatma Gandhi. New York: Oxford University Press.

Wyer, N. A. (2016). Easier done than undone… by some of the people, some of the time: The role of elaboration in explicit and implicit group preferences. Journal of *Experimental Social Psychology, 63,* 77–85.

Zhou, H., & Fishbach, A. (2016). The Pitfall of Experimenting on the Web: How Unattended Selective Attrition Leads to Surprising (Yet False) Research Conclusions. *Journal of Personality and Social Psychology, 11,* 493-504. doi: 10.1037/ pspa0000056