In Press, *Journal of Experimental Psychology: General*

How effectively can implicit evaluations be updated?: Using evaluative statements after aversive repeated evaluative pairings

Thomas C. Mann, Benedek Kurdi, and Mahzarin R. Banaji

Harvard University

Cambridge, Massachusetts

Author Note

Thomas C. Mann, Department of Psychology, Harvard University; Benedek Kurdi, Department of Psychology, Harvard University; Mahzarin R. Banaji, Department of Psychology, Harvard University.

Benedek Kurdi is now at the Department of Psychology, Cornell University.

Parts of the data reported in this article have been presented at the 20th Annual Meeting of the Society for Personality and Social Psychology, Portland, OR, in February 2019. The experiments are presented in their current positions for narrative purposes, but were run chronologically in the following order: 2, 1, 4, 3. The datasets, code, stimuli, and text of all

experiment measures in this paper are available at the following URL: https://osf.io/43an6/. This research was supported by a National Science Foundation Postdoctoral Research Fellowship (SPRF-FR) awarded to T. C. Mann [SBE-1714930].

Correspondence concerning this article should be addressed to Thomas C. Mann, Department of Psychology, Harvard University, Cambridge, MA 02138, email: thomasmann@fas.harvard.edu

Version date 10/12/19

Abstract

Implicit evaluations (attitudes) are often described as resistant to change, especially when they were initially formed in a seemingly associative manner, such as via *repeated evaluative pairings* (REP), and new learning is created via propositional material, such as *evaluative statements* (ES). The present research (total $N = 2124$) tested the responsiveness of implicit evaluations instantiated via REP to updating via different types of ES. In Experiment 1, initial learning was created via repeatedly pairing a novel target with strongly negative stimuli (screams) in an aversive REP (A-REP) task. Subsequent ES of opposing valence providing diagnostic information about the target's behavior substantially updated implicit (IAT) evaluations. In Experiment 2, behavioral ES resulted in successful updating after A-REP whether or not they provided an explanation for the initial A-REP learning. A previously unobtained result emerged in Experiment 3 showing that updating was durable even after 1 day. Finally, in Experiment 4, implicit evaluations were updated via diagnostic behavioral ES, but not via an ES instruction to suppose that different pairings had occurred during A-REP. Taken together, these experiments challenge associative theories of implicit evaluation by demonstrating that diagnostic behavioral statements can durably override the effects of initial learning on implicit evaluations, even if such initial learning is aversive and involves direct experience with stimulus pairings. Moreover, by showing that verbal manipulations based on diagnostic behavior but not a mere supposition instruction had impact, the present project advances theory by starting to identify the nature of learning that can adaptively update social impressions.

Keywords: Associative learning; Evaluative statements; Implicit Association Test; Propositional learning; Repeated evaluative pairings

How effectively can implicit evaluations be updated?

Using evaluative statements after aversive repeated evaluative pairings

Successfully navigating the complexities of the social world requires a robust ability to adapt to new information about other people. Impression updating serves to enhance the accuracy of person perception, which in turn contributes to well-calibrated expectations about the degree to which others will be helpful or harmful to us (e.g., Cacioppo, Gardner, & Berntson, 1997; Mende-Siedlecki & Todorov, 2016; Skowronski & Carlston, 1989; Tamir & Thornton, 2018; Wojciszke, 2005). It is clear that humans use even the most minimal, and sometimes even arguably irrelevant, information to form an impression (e.g., Ambady, Bernieri, & Richeson, 2000; Todorov, Olivola, Dotsch, & Mende-Siedlecki, 2015; Todorov & Uleman, 2002; Willis & Todorov, 2006). However, once formed, impressions can be resistant to change, embodied in the aphorism "you never get a second chance to make a first impression." In other words, even when new information emerges that ought to update an existing impression, it can fail to do so. Indeed, research has uncovered numerous factors that contribute to such resistance, including the ways in which the first impression frames how new evidence is interpreted (Higgins, 1996), confirmation bias and motivated reasoning (Darley & Fazio, 1980; Kunda, 1990), and lack of exposure to disconfirming information (Fazio, Eiser, & Shook, 2004).

In addition, research in social cognition has identified yet another potential hurdle to effective impression updating. In particular, experiments have shown that the relative "stickiness" of first impressions depends on how such impressions are measured: Whereas people seem quite able to set aside a first impression in their *explicit* (self-reported) judgments, first impressions can continue to linger in *implicit* (indirectly-measured) evaluations (attitudes), even after they have been overturned at the explicit level (e.g., Gregg, Seibt, & Banaji, 2006;

Petty, Tormala, Briñol, & Jarvis, 2006; Rydell & McConnell, 2006; Rydell, McConnell, Strain, Claypool, & Hugenberg, 2007; Wilson, Lindsey, & Schooler, 2000).[1] Gregg et al. (2006), for instance, found that participants who were informed that previously learned information was baseless (caused by a computer glitch) did report a new, updated evaluation on a self-report measure but failed to show similar updating on an implicit measure of evaluation. The original evaluation, once learned, continued to be expressed.

Why might implicit evaluations be particularly resistant to updating? From the perspective of dual-systems theories, implicit and explicit evaluation have been construed as the output of distinct mental systems that are subject to different learning mechanisms and contain different kinds of representations (Rydell & McConnell, 2006; Smith & DeCoster, 2000; Strack & Deutsch, 2004). In their Systems of Evaluation Model (SEM), Rydell and McConnell (2006) proposed that implicit evaluations generally form and change slowly through associative learning, requiring the repeated experience of stimulus-evaluation pairings over time. Explicit evaluations, on the other hand, are expected to reflect propositional learning of rule-based knowledge, rendering them amenable to being formed and quickly changed in response to reasoning about stimuli. As one illustrative example, Rydell and colleagues (2007) presented participants with an initial set of 100 behavioral statements that painted a uniform (positive or negative) impression of a novel person named Bob (e.g., "Bob helped a lost child find his way home"). From this experience, participants formed implicit impressions of Bob consistent with the valence of the behaviors. When participants subsequently received a varying number (0-100) of statements implying an opposite impression of Bob (e.g., "Bob refused to help a child fix his bike"), their self-reported explicit beliefs readily changed. However, their implicit impressions

updated more gradually, and were dependent on the amount of counter-attitudinal information to which participants had been exposed (see also Rydell & McConnell, 2006).

There are, however, alternative perspectives to the dual-systems approach. Specifically, in the past decade, De Houwer and colleagues have put forth a theoretical proposal under which both implicit and explicit evaluations draw upon propositional knowledge in memory, such that effects of learning on either kind of measure are directly driven by inferences in memory about how stimuli are related (De Houwer, 2014, 2018; Mitchell, De Houwer, & Lovibond, 2009). Under the propositional view, implicit evaluations can form and update in-line with inferences actively drawn during learning. For example, studies have shown that minimal instructions informing participants that novel targets soon will be paired with positive or negative stimuli (even when such repeated pairings never occur) can effectively create corresponding implicit evaluations (De Houwer, 2006), and that such instructions can be even more powerful than actual repeated experience (Kurdi & Banaji, 2017). Other research has shown that drawing new inferences about earlier learning, such as that a person's seemingly negative actions were in fact heroic (Mann & Ferguson, 2015) or were unfounded rumors (Cone, Flaharty, & Ferguson, 2019), updated previously formed implicit evaluations consistent with the new inferences.

Procedures for Implicit Evaluation Formation and Updating

The studies just discussed hint at the variety of procedures that have been used to study the formation and change of implicit evaluations, providing diverse evidence for which any theory must account. In the following sections, we discuss the consistency of dual systems vs. propositional theories with empirical evidence by examining the responsiveness of implicit evaluations to two seemingly very different types of learning procedures: *repeated evaluative pairings* (REP) and *evaluative statements* (ES).[2] In the present project we probe whether

implicit evaluations acquired via a particularly powerful instantiation of repeated evaluative pairings can be overcome by subsequent learning as a result of evaluative statements. As discussed below, updating of REP-based implicit evaluations via ES is seen as highly unlikely from the perspective of dual-systems theories (McConnell & Rydell, 2014; Rydell & McConnell, 2006). These theories posit a particularly tight fit between REP-based learning and implicit evaluations and, as such, do not foresee the possibility that implicit evaluations are amenable to updating via ES, especially following a particularly powerful instantiation of the REP intervention. Furthermore, as shown in Table 1, empirical evidence for this type of updating is almost entirely absent from the relevant literature. As such, the current work has the potential to reveal as yet unidentified conditions under which implicit evaluations can respond with flexibility or remain resistant to relevant, new information. In doing so, the project aims to provide tests of longstanding debates between dual-systems (McConnell & Rydell, 2014; Rydell & McConnell, 2006) and propositional (De Houwer, 2014, 2018) perspectives.

Repeated evaluative pairings (REP). *Repeated evaluative pairings* refers to a learning procedure in which a conditioned stimulus (CS; e.g., an image of the face of a novel individual, like "Bob" in the work of Rydell et al., 2006) is repeatedly paired with a valenced unconditioned stimulus (US; e.g., electric shocks, unpleasant sounds, or words or pictures that denote clear valence). The REP method usually involves multiple repetitions of individual exemplars of the CS (e.g., repeated presentation of Bob's face) and valenced US stimuli that are paired with it in spatiotemporal proximity (De Houwer, Thomas, & Baeyens, 2001; Levey & Martin, 1975). This is the standard procedure used in many studies of evaluative conditioning (Hofmann, De Houwer, Perugini, Baeyens, & Crombez, 2010) to create evaluative learning through CS-US pairings. Importantly, in this particular learning procedure no explicit proposition is offered such

as “Bob is a good person.” In previous studies, the REP method has successfully created initial implicit evaluations of the CS that correspond to the valence of the US (Hu, Gawronski, & Balas, 2017a; 2017b; Kurdi & Banaji, 2017; Olson & Fazio, 2001; Rydell, McConnell, Mackie, & Strain, 2006).

Evaluative statements (ES). In contrast to REP, the term *evaluative statements* is used to refer to the experimental procedure in which a target is described in the form of propositions that denote evaluative information. ES can take many forms, including behavioral descriptions (e.g., “Bob helped a lost child find his way home”), direct character summaries (e.g., “Suppose that Bob is a nice person”), or other descriptions that imply an evaluation (e.g., “Bob will be paired with positive images”). Regardless of the specific type of ES, the essence of any ES procedure is to use propositional statements to relay evaluative information. Like REP, ES has also been shown to successfully form initial implicit evaluations (De Houwer, 2006; Gregg et al., 2006; Kurdi & Banaji, 2017; Rydell & McConnell, 2006; Van Dessel, De Houwer, Gast, & Smith, 2015).

Procedures of updating. As shown in Table 1, the use of REP versus ES can be manipulated both for the formation of evaluations from initial learning, and independently for the subsequent updating of evaluations from new learning. In other words, it is possible to study whether initial implicit evaluations that arose through either REP or ES can be updated by counter-attitudinal learning that also takes the form of either REP or ES. This creates a conceptual 2x2 with four possible procedures for formation and subsequent change: initial REP followed by REP, REP followed by ES, ES followed by REP, and ES followed by ES. For three of these combinations, both dual-systems and propositional theories converge in their prediction of successful updating. As such, those three combinations do not provide clear tests to

differentiate the theories. Importantly, however, procedures in which initial REP are followed by subsequent ES ("pairings-statements") are a combination of strong theoretical interest, in which the predictions of the different theoretical perspectives differ.

Theoretical agreement: pairings-pairings, statements-pairings, and statements-statements. For procedures in which implicit evaluations are updated via REP (pairings-pairings or statements-pairings), or both initially form and are updated via statements (statements-statements), dual-systems and propositional theories converge in their predictions about the resulting implicit evaluations.

Pairings-pairings. Under the dual-systems perspective (Rydell & McConnell, 2006; Smith & DeCoster, 2000; Strack & Deutsch, 2004), REP are capable of successfully forming implicit evaluations over time because repeated coactivation of the CS and US directly builds up a new association in memory. Subsequently reversing the contingency between CSs and USs (counterconditioning) is then expected to update the association in memory in the opposite direction via the same mechanism. Thus, when an evaluation created initially through REP is followed by subsequent REP that suggest the opposite evaluation ("pairings-pairings" procedures), strong updating is predicted to occur. Under the propositional perspective (De Houwer, 2014), such updating is thought to be equally possible, but is posited to be driven by inferences formed about stimulus relations rather than by automatic association formation. Indeed, counterconditioning REP procedures have resulted in the updating of implicit evaluations under such conditions (Table 1, top left).

Statements-pairings. The REP procedure is also assumed to be able to effectively update implicit evaluations that arose initially from ES, though there is little direct evidence for this ("statements-pairings"; Table 1, bottom left). After all, to the extent that implicit evaluations can

gradually form via exposure to many statements, it is because exposure to such statements builds up novel associations through repeated exposure of a target with positive or negative valence. Likewise, propositional theories (De Houwer, 2014) concur that new REP can update prior implicit evaluations, because participants draw evaluative propositional inferences from the pairings (e.g., that the CS and US are similar in meaning, or that the CS causes the US; De Houwer, 2018). Because such inferences underlie *all* learning in these models, the procedures through which the first evaluations initially arose do not moderate the prediction of updating.

Statements-statements. Similar to the cases described above, dual-process and propositional theories can both arrive at similar predictions for updating via ES when the initial evaluations formed from ES (statements-statements; Table 1, bottom right). Propositional theories readily accommodate the idea that the inferences drawn from new ES can immediately be reflected in implicit evaluations, especially if the new statements are inferred to be more valid or important than previous information. From the dual-systems perspective, implicit evaluations can gradually update from the repeated presentation of new statements implying a new evaluation of the target, consistent with the idea that change is mediated via repeated experience over time (Rydell & McConnell, 2006; Rydell et al., 2007). The dual-systems framework can accommodate findings of the sensitivity of implicit evaluations to evaluative statements by proposing that when all experience with a target has taken the form of propositional learning from statements—precisely the "statements-statements" case discussed here—implicit cognition can be sensitive to information presented in a propositional format (McConnell & Rydell, 2014). In other words, if the *only* information that has been learned about a person has come in the form of evaluative statements, then even dual-process theories could allow for implicit evaluations to be shaped by them. Consistent with these proposals from dual-systems and propositional

theories, many studies have shown robust evidence for updating of ES-based implicit impressions through subsequent counter-attitudinal ES. This includes gradual change from the presentation of many new ES (Rydell & McConnell, 2006) as well as more rapid updating from a small amount of ES that convey highly diagnostic information about the character of the target (Cone, Mann, & Ferguson, 2017).

Theoretical divergence: the case of initial learning via pairings and updating via statements. Though the two theories have largely convergent predictions for the three procedural combinations discussed above, their predictions more clearly diverge for procedures in which initial implicit evaluations that arose through REP are followed by countervailing ES (Table 1, upper right). From their dual-systems perspective, McConnell and Rydell (2014) proposed that implicit evaluations would be particularly sensitive to associative learning if associative and propositional learning were pitted against one another. Under this framework, REP involving pairings of a novel stimulus with intrinsically positive or negative stimuli (e.g., screams, pleasant or unpleasant images) may be especially difficult to overturn with ES (McConnell & Rydell, 2014; Rydell et al., 2006; see also Smith & DeCoster, 2000; Strack & Deutsch, 2004). Only in the absence of associative cues in one's experience with the target, such as when both initial and later learning about the target come from ES (like "Bob helped a lost child find his way home" or "Bob refused to help a child fix his bike"), would ES have a controlling influence on implicit evaluations (Rydell & McConnell, 2006; Rydell et al., 2007).

Recent propositional perspectives on implicit evaluation, however, open the door to the possibility of greater impact of propositional reasoning on REP-based implicit evaluations by positing that propositional learning may underlie both ES and REP effects (De Houwer, 2014, 2018; Mitchell, De Houwer, & Lovibond, 2009). With both REP-based and ES-based learning

construed as arising from propositional processes, this perspective suggests that implicit evaluations that formed via REP should not, in principle, be particularly immune to subsequent ES. Instead, as discussed by Van Dessel, Hughes, and De Houwer (2019), different learning experiences may systematically differ in their effectiveness at driving updating of implicit evaluations depending on the relative strength of the inferences drawn during each. To the extent that participants conclude evaluative properties of stimuli (i.e., "Bob is good" or "Bob is bad") from learning experiences of any format (whether pairings, statements, or one of many other forms of learning), their implicit evaluations of those stimuli will shift proportionally. Under this inferential account, REP may actually be a particularly weak form of learning in that the strength of evaluations formed from pairings hinges on ancillary beliefs, like the premise that stimuli that co-occur have compatible evaluative significance. Following a REP experience, then, an ES that conveys clear evaluative properties of the same stimulus could be highly impactful. As such, the propositional approach can accommodate robust evidence for immediate ES-based updating after REP.

With the strict dual-systems and propositional theories offering such different predictions about the likelihood of updating from ES following REP, it is also instructive to consider the predictions made for this case by models that retain the dual-process framework while positing a greater degree of interaction between propositional and associative processes compared to the strict dual-systems approach. Such models have adapted away from the dual-systems view to accommodate the accumulation of evidence for greater impact of minimal ES on implicit evaluations than would be expected under the standard dual-systems formulation (e.g., Gawronski & Bodenhausen, 2011; Petty, Tormala, Briñol, & Jarvis, 2006). The Associative-Propositional Evaluation model (APE; Gawronski & Bodenhausen, 2006), for instance, posits

that implicit evaluations are the product of activated associations (like the dual-systems approach), but holds that the strength of associations can be impacted by propositional learning from statements. When a statement leads to the activation of new associations in memory, implicit evaluations can shift correspondingly. Another dual-process approach, the Metacognitive Model (Petty et al., 2006; Petty, Briñol, & Demarree, 2007), goes even further by specifying that learning about the true or false nature of ideas can be directly represented in associative structure as validity "tags", and can thereby impact association-driven implicit measures.

Despite the greater ability of such models to accommodate the updating of implicit evaluations from ES following REP, the strength of such updating predicted by the models is unclear. For one, repetition during REP is construed as directly building the associations that underlie implicit measures through associative learning. The influence of propositional learning on implicit evaluations, on the other hand, is indirect in such models, either mediated through the strengthening of new associations (APE) or creating an association between the previously-learned information and the concept "false" (MCM). This may result in more minimal impact of ES on implicit evaluations compared to earlier REP (e.g., Hu et al., 2017b), especially if the ES is not elaborated or repeated in a manner that could marshal associative learning (Petty et al., 2006; Wyer, 2010). As a result, although these models leave ample room for updating of REP-based implicit evaluations through ES, the strength of such updating and the conditions under which it can be accomplished (particularly without repetition) are unclear.

Prior evidence for updating from ES following REP. With the divergent and mixed predictions of various models, what is the state of the evidence for updating REP-based implicit evaluations with subsequent statements? The existing evidence for such updating is, in fact, quite

weak. As such, a primary purpose of this research is to place such learning to a series of new stringent tests.

Indeed, many studies have found that ES have failed to fully (or even partially) overcome earlier REP to drive implicit evaluations. A variety of types of statements have been used across studies to test for their impact on implicit evaluations following initial REP-based learning. Some studies have tested the effect of instructing participants that they will next complete a new REP task presenting pairings of the target with stimuli opposite in valence to those presented in the first REP task (e.g., that they will now experience pairings of the target with positive images if the target was previously paired with negative images). Although instructions about upcoming pairings have readily resulted in the formation of novel implicit evaluations (e.g., Kurdi & Banaji, 2017; Gast & De Houwer, 2013), they have had more muted effects when presented *after* REP, even within a single lab session (Gast & De Houwer, 2013; Hu et al., 2017b). As a result, such interventions have resulted in no or minimal change in initial REP-based implicit evaluations. For example, Gast and De Houwer (2013) presented participants with either 10 pairings each of novel targets and positive or negative images, or verbal instructions that the novel targets would soon be paired with such images. They found that both procedures succeeded in creating initial implicit evaluations (see also Kurdi & Banaji, 2017). Following this initial acquisition, however, instructions that the targets would subsequently be presented alone (extinction instructions) had no effect on implicit evaluations. Instructions that the targets would subsequently be paired with opposite-valence images (counter-conditioning instructions) found a reduction in the strength, but not a reversal, of the initial implicit evaluations. As a result, across the conditions, implicit evaluations continued to reflect the initial learning even after an ES instructing participants about a reversal in the pairings.

Similarly, Hu and colleagues (2017b) examined the effect of instructions about upcoming pairings on implicit evaluations. They first presented participants with 10 pairings of novel target stimuli with positive or negative images in an initial REP task. This produced initial implicit and explicit evaluations of the targets consistent with the valence of the pairings. They then found that these initial conditioned implicit evaluations were successfully reversed by 10 directly experienced new counter-attitudinal pairings. However, mere instructions (ES) that such pairings *would* subsequently occur succeeded only in reversing explicit, but not implicit, evaluations. This result replicated across both simultaneous and sequential pairings of the conditioned stimulus with unconditioned stimulus during REP.

Other studies have made use of a different kind of ES by examining the impact of clarifications about the nature of the relationship between the novel and valenced stimuli during REP, such as that the stimuli are similar vs. opposite in meaning. Once again, instructions of this sort have strongly qualified effects of REP when presented *prior* to or *concurrently* with the REP task (Moran, Bar-Anan, & Nosek, 2015; Zanon, De Houwer, Gast, & Smith, 2014; see also Peters & Gawronski, 2011), but have had more muted effects when presented *after* REP, even within a single lab session (Kurdi & Banaji, 2019; Zanon et al., 2014). For example, Zanon and colleagues (2014) showed that instructing participants that novel nonwords paired with positive or negative stimuli only produced implicit evaluations of corresponding valence when such relational qualifiers were presented prior to the REP task, but not after. Likewise, Kurdi and Banaji (2019) demonstrated that a verbal manipulation of the perceived diagnostic value of REP (i.e. that they were meant to inform participants about the underlying nature of two novel groups, or were randomly generated by the computer) influenced implicit evaluations when presented before REP, but had no effect when presented after REP.

Another version of ES presented after REP involves individuating information about the beliefs or behaviors of the target person. In support of the dual-systems prediction of greater sensitivity of implicit evaluations to REP over ES when the two forms of learning are pitted against one another, Rydell and colleagues (2006) found that when a novel target person was presented with ES suggesting a positive or negative impression and concurrent opposite-valence subliminal REP, implicit evaluations both formed and updated in accordance with the valence of REP while explicit evaluations were consistent with the impression implied by ES (but see Heycke, Gehrmann, Haaf, & Stahl, 2018). This finding is consistent, then, with the position that implicit evaluations are preferentially sensitive to REP over ES. In another experiment, participants learned that the target previously presented during REP had views similar or dissimilar to their own on a range of topics (Petty et al., 2006); in another, participants learned moderate behaviors of the target that were compatible or incompatible with the valence associated with that target person during REP (Whitfield & Jordan, 2009). In the latter two studies, implicit evaluations in the "mismatch" conditions were neutral, suggesting that ES may have negated, even if not reversed, the effects of earlier REP; however, the absence of a first implicit measure following the initial REP in both studies makes the strength of REP-based formation and the extent of change difficult to assess.

On the whole, then, the state of the evidence seems to lend credence to the dual-process idea that implicit evaluations may be especially resistant to ES-based information if they originally arose from repeated pairings during REP (McConnell & Rydell, 2014; Rydell et al., 2006; see also McConnell et al., 2008). Importantly, after all, studies that have shown effective, rapid updating of implicit evaluations via ES also used ES for both initial learning and new

learning (Cone et al., 2017). In none of these studies were implicit evaluations created initially by REP, followed by an attempt to update them via ES.

Updating from Evaluative Statements after Pairings: A New Look

Collectively, the evidence suggests that it may be particularly difficult to effectively update REP-based implicit evaluations via ES (pairings-statements), especially compared to implicit evaluations that were initially created via ES (statements-statements). This set of results is seemingly consistent with the dual-systems approach to evaluation that posits special sensitivity of implicit measures to pairings over statements (Rydell & McConnell, 2006) and leaves unrealized the theoretical room for greater updating offered by propositional models (De Houwer, 2014).

However, experiments testing the updating of REP-based implicit evaluations may have been suboptimal for showing robust updating. This is because work on updating REP-based implicit evaluations via statements has largely presented nonbehavioral information, such as relational qualifiers about the meanings of the REP pairings or instructions about upcoming counter-attitudinal pairings (Hu et al., 2017b; Gast & De Houwer, 2013; Zanon et al., 2014) or more moderate behavioral ES (Petty et al., 2006; Whitfield & Jordan, 2009). However, other lines of work suggest that presenting extreme, diagnostic information about the behaviors of the targets may be particularly effective in updating initial ES-based implicit evaluations (Cone, Mann & Ferguson, 2017), whereas more moderate behaviors, negations of earlier learning, or instruction to suppose alternate evaluations are less effective (Gregg et al., 2006; Peters & Gawronski, 2011; Rydell & McConnell, 2006; Rydell et al., 2007). For example, Rydell and McConnell (2006) found that updating ES-based implicit evaluations through subsequent ES occurred only gradually (see also Rydell et al., 2007), whereas robust updating has emerged

when new ES were particularly diagnostic, such as that a person previously described as good had committed a highly immoral act (Cone & Ferguson, 2015) or that a person who performed seemingly negative actions (breaking and entering) had done so for a positive reason (to save children from a fire; Mann & Ferguson, 2015).

This leaves open the possibility that, across formats of initial learning (REP or ES), subsequent ES can effectively update implicit evaluations, but that different ES differ in the effectiveness with which they do so. It may be the case that counter-conditioning ES instructions or qualifications about the nature of earlier pairings following REP are less effective because they seem less diagnostic and persuasive to participants than strong behavioral information would be (Hu et al., 2017b; Gast & De Houwer, 2013; Zanon et al., 2014), rather than due to inherent limitations of ES after REP. If so, such evidence would provide compelling support for the posited importance of propositional inferences for the updating of implicit evaluations, regardless of the procedures through which initial implicit evaluations arose (Van Dessel, Hughes, et al., 2019).

Overview of the Present Work

As a whole, the present experiments attempt to address previously unexamined questions about whether human minds can change in response to newer, more accurate information, so as to correct previously erroneous impressions. In particular, the experiments aimed to provide clear tests of whether implicit evaluations created through an ostensibly associative procedure like REP can be updated via ES, and to isolate the type(s) of such ES that may do so most effectively. A compelling possibility is that the type of ES used in prior work may have had weak effects because it did not carry the kind of content that recent experiments have suggested might be most effective in driving updating: diagnostic behavioral information (Cone, Mann, & Ferguson,

2017). If such information were presented about a target that had previously been presented during REP, which would dominate on implicit evaluations – the initial REP, or the subsequent ES?

In examining this question from the perspective that the *content* of ES may determine how effectively implicit evaluations based in REP are updated, the current work aimed to join with experiments on ES-based implicit evaluations (Cone, Mann, & Ferguson, 2017) to test the idea that the strength of inferences drawn from new information may be a critical determinant of updating regardless of the procedures through which initial evaluations arose, a core element of the propositional approach (De Houwer, 2014, 2018; Van Dessel, Hughes, et al., 2019). In doing so, we begin to answer the question of *when* ES will produce the most change, going beyond the question of whether they can at all.

Across four experiments, the present work tested whether presentation of ES about a novel target can update initial implicit evaluations of that target stemming from prior REP. Due to negativity biases that can sometimes make it harder to overturn initial implicit negative impressions than positive ones (Cone & Ferguson, 2015), we opted to examine negative to positive change for a particularly conservative test of the possibility of updating. To create a strongly affective, viscerally *aversive* REP (A-REP) experience, we used an unpleasant, auditory human scream as the US, a stimulus common in fear conditioning research (e.g., Oyarzún et al., 2012), which prior work has found to be resistant to even concurrently presented relational ES-based information (Moran & Bar-Anan, 2013; Moran, Bar-Anan, & Nosek, 2016; but see Moran et al., 2015). In all experiments, participants were first exposed to a novel target person paired with the human scream in an A-REP procedure, and a control person who was never paired with any sound. If implicit evaluations of the target person relative to the control person draw

especially strongly on associative learning, this choice should allow for a challenging test. Successful updating, on the other hand, would constitute the first evidence that REP-based implicit evaluations can be overturned by evaluative statements, calling the dual-systems account into question. Moreover, to ensure that the dependent measure also provides a conservative test of updating, we relied on the Implicit Association Test (IAT) throughout the present experiments, which some studies suggest may be particularly resistant to updating (Van Dessel, Ye, & De Houwer, 2019; see also Moran, Bar-Anan, & Nosek, 2017).

Given the possibility that the nature of the ES might impact the extent of updating of implicit evaluations, we compared the effects of different types of ES drawn from prior work. Specifically, within and across experiments, we manipulated whether (*a*) the evaluative statements provided diagnostic information about the target (Cone & Ferguson, 2015) or simply instructed participants to suppose that the contingencies in REP had been reversed (e.g., Gregg et al., 2006) and (*b*) the diagnostic information provided to participants prompted a reinterpretation of the original REP experience (Mann & Ferguson, 2015, 2017). As such, the present project goes beyond existing work by systematically exploring the kinds of ES that may be particularly effective in driving updating of implicit evaluations while controlling for extraneous design elements within a single experiment.

In addition, one of the experiments (Experiment 3) tested a different form of strength of updating from ES, examining whether positive implicit evaluations after behavioral ES would persist over a 1-day delay. Prior work has shown that manipulations that produce immediate changes in implicit evaluations (Lai et al., 2014) may have weaker effects after a delay (Lai et al., 2016), which could prompt radically different conclusions about the general effectiveness of procedures for the updating of implicit evaluations. Particularly because recent evidence has

found that the impact of ES relative to REP may be stronger immediately (Kurdi & Banaji, 2017) than after even a 15-minute delay (Kurdi & Banaji, 2019), it seemed especially important to examine the durability of ES-based updating after A-REP.

All studies were run with IRB approval and conformed to APA ethical standards. Experiments 1 and 4 were preregistered. The preregistration form for Experiment 1 is available at https://aspredicted.org/6yq5t.pdf. The preregistration form for Experiment 4 is available at https://aspredicted.org/ds243.pdf. In addition, the datasets, code, stimuli, and text of all experiment measures in this paper are available at https://osf.io/43an6.

Experiment 1

We first assessed the ability of evaluative statements (ES) to update an initial implicit evaluation that formed via aversive repeated evaluative pairings (A-REP), using repeated presentation of a human scream sound to make initial learning viscerally unpleasant (e.g., Oyarzún et al., 2012). To test the possibility of updating following the A-REP task, we selected a behavior that previous work has found to be effective in reversing negative, ES-based implicit evaluations (Mann & Ferguson, 2015, 2017): The man had heroically rushed into a burning building to save a child. We compared effects of this information on implicit evaluations with two control statement conditions: One in which we described a neutral behavior of the target person, and one in which we presented information on a completely unrelated topic. The inclusion of both allowed a test of whether even neutral behavioral information might shift implicit evaluations, which might occur if it is inferred to be more informative about the target than inferences drawn from the A-REP pairings.

Method

Participants. We aimed to recruit 800 US participants from Prolific Academic (https://prolific.ac) in return for $2.00 cash payment. Out of 829 participants who began the experiment, the following exclusions were made: 34 for not finishing all parts of the experiment, 24 for responding faster than 300ms on more than 10% of trials on at least one IAT, 19 for failing a manipulation check, 12 for reporting that the sound was turned off at some point during the A-REP task, and 8 for reporting an inability to hear the sounds during the A-REP task. Eleven of these participants were excluded for multiple reasons, leaving 743 participants in the analysis (50% women, Age M = 32 years, SD = 11 years).

Materials and procedure.

Overview. The procedure consisted of (1) an initial learning phase, (2) an initial test phase, (3) a second learning phase, and finally (4) a second test phase. In the initial learning phase, all participants first completed an A-REP task designed to create an initial, negative implicit evaluation of one novel individual ("target person") relative to another ("control person"). In the initial test phase, this was followed by a measure of their implicit and explicit evaluations of the two individuals, along with a question probing their beliefs about the relationship between the target person and the sounds with which he had been paired. They each then received one of three ES conditions (vignettes) in a second learning phase. Finally, in the second test phase, their implicit and explicit evaluations were measured again to assess the degree of updating following the ES.

Initial learning phase (A-REP). For each participant, photographs of two individuals were randomly selected from a set of four individuals drawn from the Radboud Faces Database (Langner et al., 2010), with one individual assigned to be the target person and the other assigned to be the control person. For each of the two individuals, three images with neutral expressions

and forward eye gazes were used as conditioned stimuli during the A-REP task: one with a frontal camera angle, and two taken from a 45 degree offset from frontal to the left and right. The images were cropped to include only the faces and upper necks, dropping the backgrounds and most clothing.

At the start of the A-REP task, participants were instructed to turn their sound to a loud but comfortable volume, and were required to press a button that played a test bell tone in order to proceed. They were informed that they would see a variety of images of people on the screen, and would sometimes hear sounds that could be unpleasant, but that their task was simply to pay attention to what they saw and heard.

Twenty A-REP trials were presented in random order. On 10 of these, one of the three photographs of the target person was chosen at random and presented in the center of the screen while an audio track of a human scream played for 1.75 seconds (Oyarzún et al., 2012), with the face and scream commencing and terminating simultaneously. On the other 10 trials, one of the three photographs of the control person was chosen at random and presented without accompanying sound. The intertrial interval was 1 second, and trials advanced automatically without participant input.

Initial test phase.

Implicit Association Test. Immediately after the A-REP task, participants completed an Implicit Association Test (IAT; Greenwald, McGhee, & Schwartz, 1998) designed to measure their implicit evaluation of the target person relative to the control person. The IAT consisted of 5 blocks (20, 20, 40, 20, and 40 trials), with Blocks 3 and 5 being the critical test blocks. Stimuli to represent the target and control persons consisted of the frontal images serving as category labels and the images from all three angles serving as the items to be classified. The attribute

categories were *Good* (wonderful, excellent, good, great, outstanding, lovely, fantastic, amazing) and *Bad* (horrible, terrible, awful, bad, mean, dreadful, vicious, offensive). The test blocks combined the sorting of person and attribute categories, such that a common key was used to sort images of one individual and words from one valence set, and the other key was used to sort images of the other individual and words from the other valence set. The order in which each of the two specific mappings was completed was randomized between participants, such that some participants sorted images of the target person and positive words with a common key and images of the control person and negative words with a different key on the first test block (and the opposite mapping on the second), and other participants completed the two mappings in the reverse order. The degree to which responses are faster on the block when the target person and negative words were categorized using a common key then on the block with the opposite configuration serves as an indication of the extent to which the target person is implicitly evaluated negatively relative to the control person, in-line with the A-REP experience. The average split-half reliability of this IAT was $r = .706$ (95% CI: [.705, .708]) across 1000 random permutations (average Spearman-Brown adjusted $r = .828$).

Explicit evaluation measure. A six-item scale drawn from prior research was used to measure explicit evaluations of the target person (Cone & Ferguson, 2015; Mann & Ferguson, 2015; Rydell & McConnell, 2006). Participants were shown the frontal image of the target person and responded to the question "How likable is this person?" on a scale from 1 (*very unlikable*) to 7 (*very likable*). They then placed that person on five 7-point scales from *bad* to *good*, *mean* to *pleasant*, *disagreeable* to *agreeable*, *uncaring* to *caring*, and *cruel* to *kind*, in random order. The six scale items were highly correlated (Cronbach's $\alpha = .97$ at both Time 1 and Time 2) and so were averaged into a single index score.

Because the main hypotheses in the present work focused exclusively on patterns of updating of implicit evaluations, the manner in which explicit evaluations varied by the manipulations within each experiment is discussed in the Supplemental Material.

Sound-related beliefs. Research suggests that how participants interpret pairings in REP tasks moderates the extent of learning on implicit measures, with causal interpretations (i.e., the target person is causing the screams) likely to produce the largest effects of pairings (Hughes, Ye, & De Houwer, 2019; Hughes, Ye, Van Dessel, & De Houwer, 2019). To measure the degree to which participants formed inferences during the REP task that might predict their implicit and explicit evaluations for a secondary analysis, we presented participants with the frontal image of the target person and asked them to select the one statement out of three options that best reflected what they had been thinking about the connection between the target person and the screams during the A-REP task, or to indicate "Other/None of the above". The three other options were presented in random order, and consisted of: "This person was **causing** the screams in some way", "This person was **trying to stop** the screams in some way", and "This person was **not connected** to the screams". For reasons of space, results involving this measure (including some preregistered analyses) in the current and subsequent studies are presented in the Supplemental Material.

Second learning phase (ES). Participants were assigned to one of three ES conditions: *positive behavior*, *neutral behavior*, or *neutral unrelated*. Consistent with our preregistered plan, participants had a 50% chance of assignment to the positive behavior condition and a 25% chance of assignment to each of the neutral conditions, so as to allow for a more precise estimate of mean implicit evaluations in the critical positive behavior condition and with the intention of collapsing the two neutral conditions if they did not significantly differ. In the two behavior

conditions, participants were shown the image of the target person and reminded that he was one of the two individuals presented during the A-REP task. Further, they were asked to use the information to inform their impression of the person, an instruction which has previously been shown to improve learning in similar tasks (Moran et al., 2015). In the positive behavior statement condition, participants read:

> "Recently, this man rushed into a burning house to save a baby, just as the fire took a turn for the worse. Running through flames to reach the wailing infant in the bedroom, he grabbed the baby and rushed outside to safety a split-second before the walls collapsed in the blaze. The baby was unhurt, and this man was hailed as a hero."

In the neutral behavior condition, participants read:

> "This person regularly rides a bus around the city where he lives. He often stops in a store near the bus station to buy a snack a few minutes before his ride. If his bus is late, he waits for a while. He usually rides for several blocks before getting off the bus and walking a few minutes from there to where he lives."

In the neutral unrelated condition, participants were asked simply to read the information, and the image of the target person was not presented. They read:

> "Otters are small mammals, and some species live in cold waters. They have high metabolic rates to help them keep warm. Fish is often central to their diet, but they have been known to eat a variety of other things too. They are playful and clever animals, and active hunters. Most otters live beside water. There are thirteen different species of otter around the world."

Second test phase. After reading the ES, participants completed the IAT a second time. This administration of the task was identical to the first, but the block order—whether the target

person was paired with positive or negative words on the first combined block—was counterbalanced independently of the order of blocks on the first task. The average split-half reliability of this IAT was $r = .609$ (95% CI: [.607, .610]) across 1000 random permutations (average Spearman-Brown adjusted $r = .757$). After the second IAT, the same six-item scale measure used during the first test phase was administered a second time ($\alpha = .97$).

Final questions. Participants provided demographic information (age and gender), entered their best guess for how many times (out of the 20 A-REP trials) the target person had been paired with the negative sounds (for exploratory purposes), indicated whether they could hear the unpleasant sounds during the experiment, and reported whether the sound had been turned off at any point. As a manipulation check, they were asked to identify the statement they had been presented out of 6 options that included the 3 verbatim ES used across the conditions as well as one matched distractor for each. Finally, they were given an option for providing open-ended feedback to the researchers, and were then debriefed and compensated.

Results

Data preparation. Implicit evaluations of the target person relative to the control person were computed both after the initial A-REP task (Time 1) and after the presentation of the final ES (Time 2) for each participant, using the D2 scoring algorithm (Greenwald, Nosek, & Banaji, 2003). Following our pre-registered analysis procedure, no differences in IAT scores were found between the neutral-behavior and neutral-unrelated statement conditions, with no main effect of ES condition, $F(1, 362) = 0.23$, $p = .63$, $\eta^2_G = .004$, or interaction between ES and time, $F(1, 362) = .12$, $p = .73$, $\eta^2_G < .0001$, in an analysis comparing the two neutral ES conditions. Therefore, we collapsed them into a single “neutral” ES condition for subsequent analysis.

Main analysis. IAT preference for the target person over the control person was assessed using a linear mixed-effects model within a 2 (evaluative statement: positive or neutral) × 2 (time: Time 1 and Time 2) design, with fixed factors for statement, time, and their interaction, and a random intercept for participants to account for participant-level dependencies in implicit evaluations.[3]

The results suggested that updating of implicit evaluations unfolded differently depending on the kind of ES that participants received. Although there was a significant main effect of statement, $F(1, 741) = 10.70$, $p = .001$, $\eta^2_G = .010$, as well as a significant main effect of time, $F(1, 741) = 77.60$, $p < .0001$, $\eta^2_G = .03$, these were qualified by a significant interaction between time and statement, indicative of differential updating depending on statement type, $F(1, 741) = 11.06$, $p = .0009$, $\eta^2_G = .005$. Figure 1 displays the mean IAT scores in each statement condition before and after the presentation of the ES.

Time 1. The A-REP task was successful in instilling initial, negative implicit evaluations of the target person. Specifically, one-sample t-tests comparing the mean IAT score at each time to zero revealed that implicit evaluations of the target relative to the control were negative after the initial aversive REP experience in both the neutral statement condition ($M = -0.20$, $SD = 0.46$), $t(363) = -8.53$, $p < .0001$, Cohen's $d = 0.45$, $BF_{10} = 1.13 \times 10^{13}$, and the positive statement condition ($M = -.18$, $SD = .46$), $t(378) = -7.58$, $p < .0001$, $d = 0.39$, $BF_{10} = 1.99 \times 10^{10}$.

Time 2. After the presentation of the statements, there was evidence for the specific effect of the positive ES beyond mere attenuation as a result of repeated testing. Participants who received neutral information continued to have negative implicit evaluations of the target person ($M = -0.10$, $SD = 0.41$), $t(363) = -4.87$, $p < .0001$, $d = 0.26$, $BF_{10} = 5.25 \times 10^3$. These implicit evaluations were less negative than at Time 1, likely indicative of a small effect of retesting,

$t(363) = 4.09$, $p < .001$, $d_z = 0.21$, $BF_{10} = 195$. However, for participants who received a positive statement about the target person, implicit evaluations were no longer negative ($M = .041$, $SD = .40$), showing a more pronounced positive shift from Time 1, $t(378) = 8.21$, $p < .001$, $d_z = 0.42$, $BF_{10} = 1.40 \times 10^{12}$, an effect twice as large as in the neutral statement condition. As a result, participants who received the positive statement had significantly more positive implicit evaluations of the target than participants who received a neutral statement, $t(741) = 4.91$, $p < .001$, $d = 0.36$, $BF_{10} = 9.59 \times 10^{3}$.

Collectively, these results show that initially negative implicit evaluations toward the target person took hold after the A-REP procedure, but were overturned by subsequent positive behavioral ES. Testifying to the strength and non-ephemeral nature of the initial negative evaluations, however, they were not similarly overturned by neutral information and retesting, but remained significantly negative in the absence of the positive behavior. The shift over time in the positive ES condition was large enough that there was some indication that implicit evaluations of the target person even became positive overall at Time 2, as a t-test suggested that mean IAT scores were significantly more positive than zero, $t(378) = 2.00$, $p = .047$, $d = .10$, consistent with a pre-registered prediction. However, this result should be interpreted with caution, as the 95% CI derived from the mixed model included zero ([-0.0029, 0.0842]) and the Bayes Factor suggested that evidence for this difference was equivocal, $BF_{10} = 0.412$.

Discussion

The results of this first experiment demonstrated that implicit evaluations that were formed through a highly aversive repeated evaluative pairings (A-REP) procedure were overturned when participants were presented with a few sentences of behavioral evaluative statements (ES) about the target person. The finding that the initial negative evaluations were not

similarly altered by the mere passage of time and retesting after neutral ES shows that the A-REP based evaluations were not fleeting, and that the positive behavioral ES in particular was responsible for the updating in that condition. This result is difficult to reconcile with dual-systems approaches positing differential sensitivity of implicit processing to ostensibly associative learning procedures (McConnell & Rydell, 2014; Rydell & McConnell, 2006; Strack & Deutsch, 2004), and demonstrates greater change than that observed in prior work on the updating of REP-based implicit evaluations through ES-based information (Gast & De Houwer, 2013; Hu et al., 2017b; Zanon et al., 2014). Instead, it is consistent with recent findings that diagnostic behavioral information can update implicit evaluations (Cone, Mann, & Ferguson, 2017), and extends that idea to suggest that positive behavior can update negative implicit evaluations more widely, even when the formation of those evaluations (A-REP) does not match the second learning experience (ES).

Experiment 2

The evidence for strong updating presented in Experiment 1 is consistent with previous work in which initial evaluations formed via evaluative statements were rapidly updated by new learning (ES; Cone & Ferguson, 2015; Mann, Cone, Heggeseth, & Ferguson, 2019; Mann & Ferguson, 2015, 2017; see Cone, Mann, & Ferguson, 2017). One chief finding from that body of work, however, is that positive ES that provide an explanation and reframing of earlier negative information were more effective than ES which presented unrelated positive information (Mann & Ferguson, 2015; Mann et al., 2019). Specifically, informing participants that a man who had previously been described breaking into the homes of his neighbors did so for the purpose of saving children from a fire produced a reversal in initial negative implicit evaluations, whereas

informing them that the man had previously saved a baby from getting hit by a train (which provides no explanation for breaking into the homes) did not.

Might it be the case that so too when initial evaluations form via A-REP, behavioral ES that provide an explanation of earlier learning will similarly be more effective in producing updating than behavioral ES that do not? Indeed, though the behavioral ES used in Experiment 1 did not explicitly instruct participants to draw a connection between that information and the A-REP experience in any way (and certainly did not provide an explanation or reinterpretation of the pairings), it did make use of an adapted version of the ES from Mann and Ferguson's (2015) work on reinterpretation: the target person ran into a home to save a baby from a fire. Although this statement was selected for its face-valid positivity and successful use in prior work on the updating of implicit evaluations, it is possible that its effectiveness in the present work was derived at least in part from spontaneous connections drawn by participants between this information and the screams; for example, perhaps the screams were reinterpreted as stemming from the fire, which reduced the A-REP effect by undercutting inferences that the target person was causing the screams. In the absence of such inferences, positive ES may be less effective.

In the present experiment, we tested this idea by comparing a positive behavior ES condition that explained the reason behind the earlier A-REP to a positive behavior condition that offered no such explanation. Importantly, beyond testing the effects of the distinction between information that is related vs. unrelated to initial REP, Experiment 2 also provides an opportunity to probe the robustness of the main finding from Experiment 1 to minor procedural variations.

Method

Participants. We aimed to recruit 500 participants from Prolific Academic (https://prolific.ac) in return for cash payment. Out of 538 participants who began the experiment, the following were excluded: 34 for not completing the experiment, 15 for responding faster than 300ms on more than 10% of trials of one of the two IATs, 6 for failing the manipulation check, 7 for reporting that the sound was off during the experiment, 3 for reporting that they could not hear the sounds, and 1 for whom the web server failed to record some data. Two of these participants were excluded for multiple reasons, yielding a final sample size of 474 for analysis (48% women, Age M = 33 years, SD = 12 years).

Materials and procedure. The procedure of Experiment 2 was identical to Experiment 1, with a few differences described here. As in Experiment 1, implicit and explicit evaluations of the target person were measured both before and after presentation of the ES. Participants were randomly assigned to receive one of two statements: A *positive-explanation* statement or a *positive-unrelated* statement. We opted not to include a control statement condition conveying neutral information about the target person in this experiment, because Experiment 1 had demonstrated that participants receiving neutral information did not display the same degree of updating as those receiving positive behavioral information, and the initial IAT continued to provide subject-level baselines for each participant.

In both the positive-explanation statement and positive-unrelated statement conditions, the image of the target person was displayed on-screen with an instruction to use the information contained in the statement to form an impression of him. Participants were instructed to "Please take some time to think about how (and if) this new information relates to what you have previously experienced with this person." The intention was that participants in the positive-explanation condition would appreciate that the new information undermined any negative

inferences drawn about the target person from the A-REP procedure, while participants in the positive-unrelated condition would be more likely to conclude that no clear relationship existed.

Positive-explanation condition. In the positive-explanation condition, a description of the target person's positive behavior was presented in a way that explicitly tied it to the previous A-REP experience. Specifically, participants read:

> "**The reason why** this person was previously paired with screams is that he is a social worker and counselor who has devoted his life and career to seeking out and helping victims of violence or other abuse. He has passionately and effectively pursued this work for many years, taking pay cuts and even putting himself in harm's way several times in order to help people in need."

By indicating that the target person was paired with the aversive stimulus during A-REP only because he was attempting to help, this condition bears similarities to the work of Moran and Bar-Anan (2013) on the effects of scream pairings on implicit evaluations (see also Moran & Bar-Anan, 2019; Moran et al. 2015, 2016). In that work, participants learned that some target stimuli were paired with an unpleasant scream because they stopped the scream. The findings indicated that the aversive pairings had an impact on implicit evaluations despite the "stop" relationship; however, unlike in the present study, that work left elaboration of the motive of the targets and full details of the manner in which they were related to the screams up to participant interpretation. By providing behavioral details to elaborate the nature of the connection between the target person and the screams, the present condition may be a stronger and more diagnostic version of the kind of relationship described by Moran and Bar-Anan (2013).

Positive-unrelated condition. In the positive-unrelated condition, we presented participants with positive behavior of the target person without explicitly linking it to the prior A-REP experience. Participants read:

> "This person is an animal welfare worker and veterinarian who has devoted his life and career to seeking out and helping abused or mistreated animals. He has passionately and effectively pursued this work for many years, taking pay cuts and even putting himself in harm's way several times in order to help animals in need."

The manipulation check in the final part of the experiment asked participants to select the ES that had been presented to them out of three options, including the two conditions described above along with one additional distractor.

Position of the sound belief question. In Experiment 1, the sound belief question was located just prior to the ES manipulation. Compared to placing this question at the end of the experiment, after the ES manipulation, this has the advantage of providing a more temporally proximal measure of participant beliefs during the A-REP experience and avoiding potential retroactive contamination by the ES manipulation. It is possible that this placement, however, enhances the enduring impact of sound-related beliefs even after participants learned the ES. In Experiment 2, we asked participants for their sound-related beliefs in the set of final questions at the end of the experiment, just prior to the manipulation check. (see Supplemental Material for analyses involving this measure.)

Results

IAT scores were assessed within a 2 (time: Time 1, Time 2) × 2 (statement: positive-unrelated, positive-explanation) mixed ANOVA, with time manipulated within subjects and statement manipulated between subjects. The results showed that initially negative implicit

evaluations were significantly updated by subsequent ES, regardless of whether the ES provided an explanation of the prior A-REP. Accordingly, we obtained a significant main effect of time, $F(1,472) = 156.02$, $p < .0001$, $\eta^2_G = .10$, but not of statement, $F(1,472) < 0.001$, $p = .98$, $\eta^2_G < .0001$, and no interaction between time and statement, $F(1,472) = 0.60$, $p = .44$, $\eta^2_G = .0004$, indicating that the effects of the two ES conditions were similar to each other (see Figure 2).

Given that ES condition did not produce a significant main effect or interaction, the analyses below focus exclusively on the effect of time (i.e., implicit evaluations before vs. after presentation of the ES). Specifically, a paired-samples t-test suggested that both types of ES resulted in effective updating of implicit evaluations over time, as implicit evaluations of the target person relative to the control person became more positive from Time 1 ($M = -0.23$, $SD = .45$) to Time 2 ($M = 0.05$, $SD = 0.39$), $t(473) = 12.51$, $p < .0001$, $d_z = 0.57$, $BF_{10} = 8.30 \times 10^{27}$. Replicating the initial negative formation effect of A-REP from Experiment 1, before the statement, IAT scores were significantly below zero overall, $t(473) = -10.98$, $p < .0001$, $d = 0.50$, $BF_{10} = 7.67 \times 10^{21}$. After the statement, on the other hand, IAT scores were no longer negative, replicating the overturning of initial negative implicit evaluations via positive behavioral ES in Experiment 1. As in Experiment 1, a t-test indicated that IAT scores had even become significantly positive on average, $t(473) = 2.90$, $p = .004$, $d = 0.13$, with the Bayes Factor suggesting that the evidence for this difference was moderate, $BF_{10} = 3.29$.

Discussion

Extending the main finding of Experiment 1, the current experiment found robust updating of initial negative implicit evaluations based initially in A-REP from subsequent behavioral ES. Where Experiment 1 found that a description of a highly heroic act of the target person was sufficient to overcome earlier scream pairings during A-REP, Experiment 2 showed

similar updating following two different descriptions of the altruistic career paths of the target person. Such diagnostic information was effective regardless of whether it provided clarity on the A-REP experience: A comparison of positive behavior conditions that did vs. did provide an explanation for the prior pairing of the target person with screams showed that this element was not impactful. Not only did participants update their implicit evaluations of the target person in the presence of an external indication that the new ES explained the prior aversive experience, but they updated to a similar extent in the absence of such an indication as well.[4]

This supports the idea that, unlike with updating of implicit evaluations based in ES (Mann & Ferguson, 2015, 2017), it may not be necessary to reframe earlier REP-based learning in order for implicit evaluations to become consistent with new, counter-attitudinal behavioral ES. In comparison to earlier work (Mann & Ferguson, 2015), the finding that non-explanatory positive information updated implicit evaluations raises the possibility that REP-based learning is in some sense more tenuous or of potentially lower weight than initial behavioral ES. This would be consistent with the general proposal that the relative weight of inferences drawn from learning experiences of any format – whether based in REP or ES – drives their contributions to implicit evaluation (Cone, Mann, & Ferguson, 2017; De Houwer, 2018).

One possibility, however, is that potential differences between explanatory and unrelated positive behavioral ES in this experiment were minimized by the lack of any direct statement within the unrelated information condition to inform participants that they should not draw any connection between the statement and the prior scream-based A-REP. Even though the unrelated information did not indicate that the information was connected to the A-REP, the instruction to think about whether it was related could have prompted participants to come up with a reason, and they may have done so very effectively. In addition to its main goal of testing the durability

of updated implicit evaluations, the next experiment will attempt to address this limitation by widening the gap between the explanation and unrelated ES conditions.

Experiment 3

Experiments 1 and 2 have provided clear evidence in favor of the idea that implicit evaluations initially formed via aversive repeated evaluative pairings can be successfully updated by providing participants with diagnostic behavioral information about the target. However, the question of whether such updating can endure over time is of particular importance in the present investigation, and is one that is rarely examined in the context of implicit evaluation acquisition and updating (for exceptions, see Devine, Forscher, Austin, & Cox, 2012; Forscher, Mitamura, Dix, Cox, & Devine, 2017; Kurdi & Banaji, 2019; Lai et al., 2016; Mann & Ferguson, 2015, 2017). When durability has been examined, it has sometimes been found that effects of ES after a delay of even minutes or days are reduced relative to their immediate impact (e.g., Lai et al., 2014, vs. Lai et al., 2016; Kurdi & Banaji, 2017, vs. Kurdi & Banaji, 2019).

It could be the case that a preferential sensitivity of implicit evaluations to aversive repeated evaluative pairings (A-REP; i.e. the screams), as predicted by dual-process accounts (McConnell et al., 2014; McConnell et al., 2008; Rydell & McConnell, 2006) would emerge most strongly after a delay. Compared to immediate measurement, delayed measurement may allow more time for the salience of the evaluative statements (ES) to wane, and for spontaneous recovery of the earlier conditioned response to reemerge if the memory traces formed by A-REP were not fully overridden by the ES (Bouton, 1994). One recent finding that could be seen as supporting this possibility is provided by Kurdi and Banaji (2019), who found that even after short temporal delays (over the course of a single IAT, and a 15-minute delay between learning and testing), the effects of REP persisted, while the effects of ES waned. As Kurdi and Banaji

(2019) noted, this finding is also consistent with previous work on more prolonged memory for directly experienced vs. indirectly learned events (e.g., Larsen & Plunkett, 1987; Toglia, Shlechter, & Chevalier, 1992), such that directly experienced learning may simply tend to be more enduringly impactful on memory, without a need to appeal to dual learning processes or systems. Either way, when directly experienced scream-based A-REP are followed by indirectly learned positive behavioral information about the target contained in statements, it may be the screams that have a more lasting impact. If A-REP were found to be simply more impactful than ES after a delay, then such a finding would represent an important limitation to claims of the apparent power of ES to meaningfully update A-REP based implicit evaluations.

On the other hand, it is possible that updated implicit evaluations following ES will endure, such that a delayed implicit measure will show similar evaluations to an immediate measure. Such a finding here would reinforce the conclusion that A-REP has no special influence on implicit evaluations, and that after clearly diagnostic information is presented about the target person, earlier impressions derived from scream-based A-REP are effectively set aside. This would support the proposal from the propositional perspective that inferences drawn from statement-based information can powerfully drive implicit cognition (De Houwer, 2014, 2018).

A final possibility is that only positive behavioral ES that fail to adequately explain and undermine the earlier A-REP will wane in effectiveness over time. If an explanatory evaluative statement succeeds in undermining the initial negativity implied by the scream pairings, its effects might better persist over a delay than statements that do not, because retrieval even of memory of the A-REP experience may be more likely to activate the positive information conveyed in the ES due to greater integration between the A-REP experience and the ES in the explanation case. As initial evidence to support this possibility, Mann and Ferguson (2015)

found that updated implicit evaluations from explanatory positive information – a man who broke into the homes of his neighbors had actually done so to rescue children from a fire – persisted over a 3-day period. Work that has presented various other forms of ES, on the other hand, has tended to find that effects can be short-lived (e.g., Kurdi & Banaji, 2019; Lai et al., 2016). Such work, however, differs in many ways from the current paradigm, and did not specifically compare temporal effects of positive behavioral ES that do vs. do not provide an explanation for earlier A-REP.

To test these competing possibilities, in this experiment, we directly examine the durability of updated implicit evaluations from behavioral ES following initial A-REP over the course of a 1-day delay, including both explanatory and unrelated positive behavioral ES conditions. Such a delay is far longer than the typical 30-minute extent of most laboratory experiments in social cognition, and is common in work on the alteration of consolidated memory representations (Elsey, Van Ast, & Kindt, 2018).

Method

Participants. We aimed to recruit 550 participants from Prolific Academic (https://prolific.ac) to participate in Session 1 of a two-part experiment.

Session 1. Out of 574 participants who began Session 1, the following were excluded: 31 for not completing all measures, 9 for responding faster than 300ms on more than 10% of trials of the IAT, 15 for failing the manipulation check, 10 for reporting that the sound was off during the experiment, 6 for reporting that they could not hear the sounds, and 2 for whom the web server failed to record IAT data. Five of these participants were excluded for multiple reasons, yielding a final sample size of 507 participants in Session 1 (Age M = 33.41 years, SD = 12.01 years; 52.29% men).

Session 2. Approximately 1 day after Session 1 was posted, participants who had completed Session 1 were invited on Prolific to complete Session 2 of the experiment, with the participation window left open for 24 hours. The retention rate was very high, with 87% of participants included in analyses of the first session completing Session 2 (N = 442). Of these, 1 participant responded faster than 300ms on more than 10% of trials on the Part 2 IAT, 7 participants did not accurately remember the final ES information presented to them during Session 1, and IAT data were not recorded for 1 participant. Twelve additional participants began but did not complete the second session. The exclusion of these participants left a final sample of 433 participants for analyses involving Session 2 (Age *M* = 33.83 years, *SD* = 12.07 years; 50.00% men). The average time between when these participants began the first and second sessions was 22.92 hours (*SD* = 3.47 hours, min = 20.30 hours, max = 45.27 hours).

Materials and procedure. The first session of the experiment (Session 1) paralleled Experiment 2, with a number of exceptions detailed below.

Initial learning phase (REP). To make it even clearer to participants that the faces presented during the evaluative conditioning task came from two (and only two) different individuals, we added the frontal images of both individuals to the instruction page of the A-REP task, and informed participants that they would see these two people during this part of the experiment. In addition, we narrowed the face set from the 4 white male images used in Experiments 1–2 to include only the two faces that showed A-REP effects of similar magnitude to one another at Time 1 in our previous experiments, to reduce extraneous variation attributable to the specific faces. For each participant, one image was randomly assigned to be the target person, and the other to be the control person.[5]

Evaluation measures. Given that Experiments 1–2 demonstrated that the A-REP task was successful in creating negative implicit evaluations of the target person relative to the control person on the IAT, the present experiment dropped the first implicit and explicit evaluation measures. Instead, it focused entirely on the pattern of evaluations *after* participants read the ES, both immediately and with a 1-day delay. With this change, the A-REP task was followed immediately by the ES manipulation.

In order to determine whether effects of time on the final (Session 2) IAT were attributable to the mere passing of time or to retesting, during Session 1 we manipulated whether participants completed the standard IAT (target vs. control faces) or a control IAT measuring implicit preference for blue shapes vs. green shapes (target stimuli included a square, circle, triangle, pentagon, and star-like shape of each color). This allowed us to examine whether practicing the expression of the evaluation of the target person immediately after learning the ES plays any role in enhancing or undermining its durability over the temporal delay between sessions. For a similar reason, we orthogonally manipulated whether participants completed the explicit evaluation measure during the first session as well.

Furthermore, given the relative nature of the IAT (i.e., it measures implicit evaluations of the target person relative to the control person), we added to the design of the present experiment an explicit scale measure of evaluations of the control person, so as to allow for tests of relative preference within each ES condition both implicitly and explicitly. The scale consisted of the same 6 items asked about the target person ($\alpha_{session1} = .95$, $\alpha_{session2} = .95$) and was only administered to participants who completed the corresponding scale for the target person. The explicit scale for the target person also continued to be reliable, for both Session 1 and Session 2

($\alpha_{session1}$ = .94, $\alpha_{session2}$ = .96). The presentation order of the two explicit scales (target and control) was counterbalanced between participants for each administration.

Evaluative statement conditions. This experiment retained the explanation and unrelated positive ES conditions used in Experiment 2, but added additional text to more strongly highlight the differences between them for participants and remove ambiguity. Most importantly, the explanation version now made explicit that the target person was attempting to help the people screaming, and the unrelated version made explicit that the information was unconnected to the screams. Both conditions omitted the instruction to deliberate on "how (and if) this new information relates to what you have previously experienced with this person", which could have induced unwanted confusion or inferences of explanatory connection in the "unrelated" positive ES condition. Instead, participants were informed that they should simply form an impression of the person from the information, and that he was one of the same people they saw during the sounds task.

Next, in the *explanation* condition, they read, "The reason why this person was previously paired with screams is that he was trying to **help** the people who were screaming," followed by the remainder of the text for the explanation condition from Experiment 2. In the *unrelated* condition, they instead read, "Though this has nothing to do with the screams you heard before, this person always tries to **help** animals in need," followed by the remainder of the text for the unrelated condition from Experiment 2.

Other questionnaire items. To the sound belief question from Experiments 1–2, we added a note that there was no right or wrong answer, to encourage participants to offer their "true" beliefs about the relationship between the target person and the sounds rather than take into account any perceived expectations of the researchers. In addition, to the question asking

participants to estimate the number of times they heard a scream paired with the target face, we added a parallel item regarding the control face (order randomized). Finally, the manipulation check was configured with 4 response options reflecting the wording of the two positive ES conditions used in the experiment, along with two positive ES distractors (see Open Materials).

Session 2. During the second session on the following day, participants were simply welcomed back to the experiment, with no reference made the scream-based A-REP or any of the tasks completed on the previous day. All participants were directed immediately to the IAT to test their implicit evaluations of the target face vs. the control face. The block order (target-good first vs. target-bad first) was manipulated orthogonally to the block order from Session 1. Next, participants completed the explicit evaluation scales for both faces, followed by the same questions about the number of times a scream was paired with each face, recall of what they believed the relationship was between the target face and screams during the A-REP task on the previous day, and a re-presentation of the same manipulation check. Following this, participants were debriefed. Any participants from Session 1 who had not returned to complete Session 2 by the time it concluded were sent the debriefing material in a message through the Prolific system.

Results

Implicit evaluations were assessed within a linear mixed-effects model that included fixed effects for statement condition (positive-unrelated vs. positive-explanation), time (Session 1 vs. Session 2), and their interaction, along with a random intercept for participants to account for participant-level dependencies. The data for Session 1 consisted of all included participants who were assigned to the faces IAT, and data for Session 2 consisted of all included participants who completed the second session.[6]

Results showed that the effects of both manipulations (related vs. unrelated statements) were similar at both time points (measured immediately and after a delay), in contrast to the significantly negative initial implicit evaluations that we observed in the previous two experiments immediately after the evaluative conditioning task. This was indicated by the lack of main effects and interactions: there was no main effect of time, $F(1, 356.24) = 0.022$, $p = .881$ (Satterthwaite degrees of freedom), or statement, $F(1, 460.32) = 2.52$, $p = .113$, and no interaction between time and statement, $F(1, 356.24) = 0.00$, $p = 1.00$ (Figure 3).[7]

Because effects of time, statement condition, and their interaction were not significant in the above analysis, we refit a simpler mixed model with only a fixed intercept and random intercept for participants to account for participant-level dependencies. This model indicated that after the ES were presented, implicit evaluations of the target person relative to the control person were significantly positive on average ($M = 0.08$, 95% CI: [0.049, 0.116]), despite the previous scream-based A-REP.

Discussion

Over the course of a 1-day delay, implicit evaluations of the target person remained remarkably consistent. Replicating the results of Experiment 2, participants showed positive implicit evaluations of the target person regardless of whether the information they learned in the ES provided an explanation of the earlier A-REP. Also regardless of ES condition, implicit evaluations toward the target remained positive one day after initial learning; in the explanation condition in particular, there was strong evidence that implicit evaluations were significantly greater than zero at both sessions, though the two ES conditions did not significantly differ.

This finding of the enduring effect of the ES on implicit evaluations joins a limited body of evidence on the durability of novel implicit evaluation formation and change (Kurdi & Banaji,

2019; Mann & Ferguson, 2015, 2017). The results clearly show that initial REP neither dominate implicit evaluations immediately, nor after a delay. In conjunction with earlier work on the durability of updated implicit evaluations over a span of days (Mann & Ferguson, 2015), this finding suggests that the strength and persistence of updated evaluations may hinge on the presentation of diagnostic behavioral details in the ES. By comparison, the decay in effects of ES found by Kurdi and Banaji (2019), for instance, may have been because the ES used in that work was an instruction about upcoming REP, rather than a description of individual behavior.

Experiment 4

Contrary to previous evidence and the predictions of dual-systems approaches to evaluation (Rydell & McConnell, 2006), Experiments 1–3 found evidence that evaluative statements (ES) could strongly update initial implicit evaluations produced via aversive repeated evaluative pairings (A-REP), and that such updated evaluations were durable over the span of at least 1 day. However, though the evidence so far supports that at least some positive ES outperformed neutral statements in producing updating, it remains unclear more generally what kinds of counter-attitudinal ES may have this effect. Though propositional theories make clear that propositional learning from ES can affect implicit cognition (De Houwer, 2014), no work has taken up the critical direction of comparing different types of ES within a common design. The goal of Experiment 4 was to take a first step toward understanding the conditions under which ES will vs. will not be effective, and to start isolating the features that determine its effectiveness. We hope that much future work will continue to map this theoretical landscape.

Was it necessary that the ES in Experiments 1–3 conveyed new counter-attitudinal behavioral information about the target person, or could any ES aimed at counteracting the initial negative evaluation of the target person be effective? Past research is decidedly mixed on the

topic, in no small part because variations in the ES presented to test updating have often been accompanied by variation in the procedures that produced the initial evaluations. This makes it difficult to know whether certain forms of ES may be more effective than others, or whether procedural details explain differences in effectiveness of updating.

For instance, past research has shown that instruction to suppose that previous information presented about the targets of initial learning were reversed did not lead to updating (Gregg et al., 2006). Unlike in the present work, however, in that experiment, initial evaluations were based on ES, and it is possible that participants would be impacted differently by a suppose-style condition after A-REP. In general, procedural differences may limit the ability to generalize a priori the effectiveness of different types of ES across paradigms. In our final experiment, then, we directly compared the impact of diagnostic behavioral ES (Cone, Mann, & Ferguson, 2017) with an abstract instruction to suppose that earlier learning had been reversed (Gregg et al., 2006) in a single paradigm in which initial implicit evaluations formed via A-REP.

Method

Participants. In accordance with our pre-registered plan, we planned to recruit participants in two stages. For reasons of cost, we planned first to recruit 450 participants from Prolific Academic (https://prolific.ac), and to test whether the Bayes Factor comparing IAT scores between the positive behavior and suppose statement conditions provided stronger than anecdotal evidence with regard to a difference ($BF_{10} > 3$) or lack of difference ($BF_{10} < 1/3$) between those critical two conditions. We determined that if the Bayes Factor were found to offer only anecdotal evidence ($1/3 < BF_{10} < 3$), an additional 300 participants would be recruited. Results after the first wave suggested strong evidence in favor of a condition difference, $BF_{10} = 24.32$, and so recruitment was terminated. Of 467 participants who began the experiment, the

following were excluded: 21 for not finishing all parts of the experiment, 11 for responding faster than 300ms on more than 10% of IAT trials, 36 for failing a manipulation check, 4 for reporting that the sound was off during the experiment, and 6 for reporting an inability to hear the sounds during A-REP. Eleven of these participants were excluded for multiple reasons, leaving a final sample of 400 participants (53% women; Age M = 33 years, SD = 11 years).

Materials and procedure. The procedure of this experiment was similar to the first session of the previous experiment, with a few exceptions. The largest difference was in the statement conditions. First, similar to Experiment 1, the present experiment included a neutral information condition. Because the two neutral statement conditions in Experiment 1 showed similar patterns of implicit evaluations, only the neutral-behavior condition was retained. Unlike in Experiments 1–2, which included the initial post-REP IAT, we chose to include a control statement condition here so as to provide evidence for successful negative formation from the A-REP task in a between-subjects manner. Second, in line with the theoretical focus of the present experiment on testing the effect of a counter-attitudinal statement that does not present diagnostic behavioral evidence about the target person, a new condition not included in Experiment 1 was added (*suppose* condition). In this statement condition, participants were shown the images of both individuals presented during the A-REP task, and asked to suppose that they had previously had a different experience with them. Specifically, they were instructed:

> "In particular, please suppose that the person who was previously <u>always</u> paired with screams had in fact <u>never</u> been paired with screams.
>
> In addition, please suppose that the person who was previously <u>never</u> paired with screams had in fact <u>always</u> been paired with screams."

The language in this condition was deliberately crafted to avoid directly pairing the target person with the concept of not being presented with screams (and the pairing of the control person with the concept of being paired with screams), relying on the reasoning of participants to properly parse the meaning of the instructions.

As in the other two statement conditions, participants were asked to use the new information to try to form an impression of the individuals. In addition, participants were instructed to "continue to suppose that this was the case while doing each of the remaining tasks, keeping the new ideas about these two people in mind until the very end of the experiment," using language adapted from Gregg et al. (2006).

Besides the ES conditions, the other differences between the present experiment and the main session of Experiment 3 were: 1) The manipulation check at the end of the experiment was updated to contain this new statement as a multiple-choice option, along with a tailored control; 2) the sound-belief question was placed just after the A-REP task, as in Experiment 1; and 3) target and control faces for each participant were randomly drawn from the full set of 4 face stimuli from Experiments 1–2.

Results

IAT scores were compared between the three ES conditions (neutral, positive behavior, and suppose) in a one-way between-subjects ANOVA analysis. Showing that implicit evaluations following the ES depended on the specific ES that had been presented, there was a significant effect of condition (Figure 4), $F(2,397) = 11.15$, $p < .0001$, $\eta^2_G = .05$. Providing evidence again for the success of the initial A-REP experience for instilling initial negative implicit evaluations of the target person, implicit evaluations were negative in the neutral statement condition ($M = -0.13$, $SD = 0.43$), $t(141) = -3.47$, $p < .001$, $d = 0.29$, $BF_{10} = 27.21$.

In line with the analysis presented above, post hoc pairwise comparisons between the ES conditions (Tukey adjusted) revealed that whereas the positive behavioral statement was effective in shifting implicit evaluations following the initial negative A-REP experience, the suppose condition produced no change. D-scores after the positive statement were significantly more positive than D-scores after the neutral statement, $t(397) = 4.52$, $p < .0001$, $d = 0.53$, $BF_{10} = 1.21 \times 10^3$, as well as after the suppose statement, $t(397) = 3.34$, $p = .003$, $d = 0.42$, $BF_{10} = 24.32$. Compared to neutral statement condition, however, the suppose condition did not produce more positive implicit evaluations, $t(397) = 0.93$, $p = .62$, $d = 0.12$, $BF_{01} = 4.73$.

Discussion

Experiment 4 directly compared the ability of two counter-attitudinal evaluative statement (ES) conditions to update implicit evaluations formed after aversive repeated evaluative pairings (A-REP), and found notable differences between them. Implicit evaluations were found to update only in the positive behavior condition, but not the suppose condition. The suppose condition was no more effective in updating implicit evaluations of the target person than the neutral information condition, with implicit evaluations in both cases remaining consistent overall with the A-REP experience. This finding replicates earlier work with a suppose condition that found no updating (Gregg et al., 2006). In addition, it is also consistent with findings that a similar type of ES condition – counter-conditioning instruction – is ineffective in eliminating or reducing implicit REP effects (Hu et al., 2017b; cf. Gast & De Houwer, 2013). Similar to the suppose condition in the present experiment, that work showed that informing participants that the novel stimulus will soon be paired with images opposite in valence to those used during REP was not effective in updating implicit evaluations.

By directly comparing an ES condition that presented instructions qualifying the relationship between the target person and aversive stimuli during REP (suppose) with a diagnostic behavioral ES about the target person, the present experiment provides particularly strong evidence that not all counter-attitudinal ES are effective in producing updating of A-REP based implicit evaluations. By comparing different kinds of ES within a single paradigm, this experiment is among the first to allow for stronger conclusions of the relative effectiveness of different ES by holding constant other elements of the design.

Strength of Updating Across Experiments 1–4

Experiments 1–4 offer strong support for the possibility that evaluative statements (ES) can update implicit evaluations that stemmed initially from aversive repeated evaluative pairings (A-REP), particularly if the ES suggest diagnostic positive inferences about the target person. Though the evidence for updating from initial negative evaluations after A-REP to positive evaluations after ES was strong, the evidence across experiments left it unclear as to whether implicit evaluations of the target person were fully *reversed* after a positive behavioral statement (i.e., had become significantly positive) vs. were only revised to *neutral* relative to the control face (i.e., were no different from zero).

Of course, the question of whether implicit evaluations were reversed (i.e., changed sign by crossing zero) from their levels immediately after REP to after ES is secondary to the evidence for substantial updating overall. Though the IAT has a leg up over many measures in psychology in having a meaningful zero due to being a difference score between controlled experimental conditions (Greenwald, Nosek, & Sriram, 2006), shifts in scores on the IAT may be meaningful regardless of whether they cross zero, which could depend on study-specific factors like the strength of initial formation or the length of delay between sessions. Still, because such

shifts provide evidence regarding with which source of information (REP or ES) implicit responses are currently aligned, and because there is broad interest in the qualitative sign on implicit measures following updating (e.g., Gawronski & Cesario, 2013; Hu et al., 2017b; Lai et al., 2014), it may be useful to test whether, at least under the conditions of the current work, initially negative A-REP based implicit evaluations can shift in sign following positive diagnostic behavioral ES.

To obtain a higher-powered analysis of the comparisons of implicit evaluations to zero, we conducted a combined analysis of data pooled across experiments. We fit a linear mixed model to the IAT scores of participants in all experiments within ES conditions that presented diagnostic positive behavior: the "fire rescue" positive behavior condition in Experiments 1 and 4, and the unrelated and explanation positive behavior conditions in Experiments 2 and 3 (setting aside the delayed IAT data in the case of the Experiment 3, for consistency). We included in the model a fixed factor for time (Time 1 after A-REP, and Time 2 immediately after ES) and a random intercept to capture variation in IAT scores attributable to participants.

The results were consistent with the finding across experiments that in the positive behavioral statement conditions, implicit evaluations updated substantially after presentation of the ES. This was revealed by a significant main effect of time, $F(1, 1108.9) = 234.87$, $p < .001$ (Satterthwaite), with a paired-samples effect size of $d_z = .50$, $BF_{10} = 1.67 \times 10^{40}$.[8] Analyses comparing mean IAT scores to zero both after A-REP and after ES showed that implicit evaluations were initially negative after A-REP, but had updated to positive after diagnostic behavioral ES. At Time 1, immediately after the A-REP task but prior to the positive behavioral ES, implicit evaluations of the target person were significantly below zero, $M = -.21$, $SE = 0.015$, 95% CI: [-0.236, -0.178], corroborated by a one-sample t-test, $t(852) = -13.20$, $p < .001$, $d = .45$,

$BF_{10} = 7.95 \times 10^{32}$. At Time 2, on the other hand, immediately after the positive behavioral ES, implicit evaluations of the target person had reversed to become significantly more positive than zero, $M = 0.04$, $SE = 0.011$, 95% CI: [0.019, 0.064], corroborated by a one-sample t-test, $t(1502) = 3.73$, $d = .10$, $p = .0002$, $BF_{10} = 29.06$.

The results were clear: Implicit evaluations were significantly negative after the initial A-REP, but had reversed to positive after presentation of subsequent positive diagnostic behavioral ES. The data across experiments, then, provide the first strong evidence that negative implicit evaluations based in A-REP can be fully *reversed* by ES (cf. Hu et al., 2017b; Gast & De Houwer, 2013; Zanon et al., 2014).

General Discussion

A prominent perspective on implicit evaluations is that they are "easier done than undone", in that they may readily arise from initial information, but be more resistant to updating from new learning than explicit evaluations (Gregg et al., 2006; see also Rydell & McConnell, 2006; Rydell et al., 2007; Wilson et al., 2000). From a dual-systems perspective, this may be particularly true for implicit evaluations that formed initially from repeated evaluative pairings (REP) followed by attempts to update them with information in evaluative statements (ES).

In contrast to this idea, the present project provides robust evidence, spanning four experiments, that negative implicit evaluations toward a novel target instilled via aversive repeated evaluative pairings (A-REP) can be updated by an ES conveying positive behavioral information about the target. Experiment 1 showed strong evidence for updating of REP-based implicit evaluations via positive behavioral ES in comparison to a neutral statement. Experiment 2 demonstrated that a positive behavioral ES did not need to provide an explanation of the target-sound pairings to be effective (cf. Mann & Ferguson, 2015), suggesting that a wide range of

positive behavioral ES might succeed in overturning REP-based implicit evaluations. Experiment 3 extended this finding even further, showing that updated implicit evaluations persisted over a 1-day delay regardless of whether the statement provided an explanation for the screams. Finally, Experiment 4 supported the specific importance of positive behavioral information by showing that an ES condition that presented counter-attitudinal relational information about the REP target pairings was ineffective in updating implicit evaluations.

Together, these results demonstrate that ES are capable of effectively overriding REP-based implicit evaluations (cf. Gast & De Houwer, 2013; Hu et al., 2017b; Zanon et al., 2014), both immediately and after a delay, and even using the IAT, which has shown more resistance to subsequent verbal information than other measures (Moran et al., 2017; Van Dessel, Ye, et al., 2019). This robust evidence for updating across experiments is inconsistent with the proposal that there is a special "fit" between implicit evaluations and learning procedures that rely on the repeated experience of non-verbal information like REP (cf. Rydell & McConnell, 2006), showing instead that implicit evaluations based in REP can be updated by diagnostic behavioral information contained in ES in a similar manner to implicit evaluations based in ES (Cone et al., 2017).

The similarity of the statements in the current work that effectively updated A-REP based implicit evaluations to those that successfully updated ES-based implicit evaluations (Cone et al., 2017) offers a possible explanation for why the present experiments showed more updating than has been observed in prior work on REP-based implicit evaluations, by suggesting that the relative informational value of the inferences drawn from new learning may generally predict updating regardless of the format in which initial evaluations were acquired. Cone et al. (2017) found that extreme behaviors that reflect clearly on the character of the target person (e.g. saving

children from a fire; mutilating animals) produce larger updating than less diagnostic interventions used in prior work, like a hypothetical supposition that two groups differed in character (Gregg et al., 2006) or less extreme behaviors (Rydell & McConnell, 2006).

Here too, the positive behavior ES – rescuing children from a fire, devoting one's life to helping people or animals – may provide more convincing evidence about the character of the targets than other ES conditions that showed weaker change, such as those that informed participants about changes in upcoming pairings (the "suppose" condition of the current work; Gast & De Houwer, 2013; Hu et al., 2017b; Zanon et al., 2014). As such, this work challenges the idea that implicit evaluations after REP are inherently less sensitive to subsequent ES due to a mismatch between statements and an associative-learning system (cf. McConnell & Rydell, 2014), and is consistent with the general ability for implicit evaluations to be updated.

A-REP: Using a Particularly Powerful Initial Learning Method

A strength of the current results is that the diagnostic behavioral ES were able to overcome initial implicit evaluations that were formed initially from a particularly powerful version of REP (A-REP), which made use of a human scream. Auditory stimuli have been found to be effective unconditioned stimuli (US) in conditioning research (Hofmann et al., 2010), and human screams in particular have been used for fear conditioning as a strong alternative to electric shocks (e.g., Glenn et al., 2011; Lau et al., 2008; Lau et al., 2011; Machlin, Miller, Snyder, McLaughlin, & Sheridan, 2019; Oyarzún et al., 2012; Schmitz et al., 2011). The visceral nature of this stimulus provides a particularly challenging test of the possibility that ES-based information could be successful—a challenge that positive behavioral statements met.

It is possible, of course, that the A-REP task employed here qualitatively differs from other forms of REP in ways that impact ES-based updating. Specifically, it may be that repeated

pairings of aversive auditory stimuli with a target stimulus, as in the A-REP task, produce particularly affective, high-arousal, and/or fear-based negative representations (Glenn et al., 2011; Lau et al., 2011). To the degree that the representations formed via different REP variants vary qualitatively, the procedures necessary for updating may vary as well. In fact, the growing body of research on memory reconsolidation indicates that representations arising via fear conditioning procedures may be especially amenable to updating under some conditions (Elsey, Van Ast, & Kindt, 2018; Lee, Nader, & Schiller, 2017; Schiller, Kanen, LeDoux, Monfils, & Phelps, 2013; Schiller et al., 2010). It is conceivable, then, that evaluations stemming from different forms of REP may be sensitive to different procedures of updating. Of course, in combination with other recent work on updating implicit evaluations that initially formed from statements (Cone et al., 2017; Ferguson et al., 2019), the present results imply that the amenability of initial implicit evaluations to updating may depend far more on the diagnostic value of the new information itself than on the specific method through which initial evaluations arose.

What Makes an Evaluative Statement (In)Effective?

In comparison to the success of the positive behavioral information statements in updating implicit evaluations across Experiments 1–4, the inability of the suppose condition to achieve similar updating (Experiment 4) raises a fundamental question: *why* are some ES effective, and some not? Theoretical discussion and evidence have often focused on *whether* ES can update implicit evaluations, rather than the types of statements that will vs. will not successfully do so (De Houwer, 2014). By comparing different kinds of ES within a common experimental paradigm, our work offers novel evidence on the differential impacts of different ES on implicit evaluation updating after REP. For what reason, though, were the positive

behavioral ES conditions successful, while the suppose condition was not? The positive behavior conditions and the suppose condition differed in a number of ways, so at present, we can only speculate on the reasons for the greater updating observed in the former than the latter.

One possibility is that the statement conditions that were effective—those presenting diagnostic behavioral information—prompted participants to form stronger inferences about the evaluative significance of the target person. This is consistent with the recent proposal that the strength of inferences drawn about the evaluative properties of targets from learning experiences (regardless of format) drives learning effects on implicit evaluation (Van Dessel, Hughes, et al., 2019). It is possible that in the positive behavioral statement conditions, participants may have thought that the information they learned conveyed more valuable information about the target compared to the earlier A-REP experience, because the evaluative significance of any REP procedure depends on additional premises about the meaning of co-occurrence and is open to interpretation (Hughes, Ye, & De Houwer, 2019; Hughes, Ye, Van Dessel, et al., 2019). In contrast, an ES instruction to suppose that A-REP contingencies had been reversed might not have lead participants to adjust their interpretation of the evaluative properties of the target, presumably due to the abstract and clearly hypothetical nature of the manipulation.

Alternatively, a co-occurrence of the target person with a direct activation of the new, intended evaluation – whether in the external statement or a mental simulation – may actually be critical in driving successful updating. In the "suppose" condition, participants learned the abstract information well enough for their impression of the target person to be impacted on an explicit measure (see Supplemental Material), but may not have stored the conclusion of this computation ("He is not bad" or "he is good") strongly enough to be activated on the implicit measure. This possibility is consistent with the idea that fast implicit responses may not be able

to reflect evaluative representations that require a large amount of online computation (Kurdi, Gershman, & Banaji, 2019). In the positive behavior ES conditions, the greater vividness of the new information may have helped participants to mentally simulate an image of the target person conducting specific positive behaviors and store a new impression of him in a direct fashion (see Lai et al., 2014; Wyer, 2010). Future studies could examine whether a delay, or other manipulations that allow for offline simulation, could allow for even indirect learning like in the "suppose" condition to have a greater impact on implicit evaluations by storing conclusions drawn from complex knowledge more directly (Gershman, Zhou, & Kommers, 2017). For example, if participants were prompted to actually simulate the co-occurrence of the control face with screams and the target face with silence, the suppose condition may have been more effective.

Consistency with Current Theories

Lively theoretical debates continue over the nature of REP-based learning (Bar-Anan & Balas, 2018; Corneille & Stahl, 2019) and the differentiation of implicit vs. explicit evaluation (Cone et al., 2017; De Houwer, 2014; De Houwer & Hughes, 2016; Gawronski & Bodenhausen, 2011; Hu et al., 2017a; 2017b). Because of the inherent flexibility of many theories and perspectives to accommodation of new evidence, the present findings do not and cannot provide conclusive support for any one position. In providing novel demonstrations of the ability for evaluative statements to update A-REP based implicit evaluations as well as direct comparisons of the effectiveness of different types of ES, they do, however, offer evidence that can further refine and constrain such theories. Regardless of theoretical perspective, the findings advance knowledge of the conditions under which implicit evaluations emerge and change.

The results seem most difficult to reconcile with strict dual-systems approaches to evaluation that posit that implicit evaluations will be preferentially sensitive to associative learning (McConnell & Rydell, 2014; Rydell & McConnell, 2006; Rydell et al., 2006; see also Smith & DeCoster, 2000; Strack & Deutsch, 2004). This perspective would predict that implicit evaluations would be selectively dependent on the initial A-REP tasks pairing the target person with aversive screams, and relatively insensitive to counter-attitudinal information provided in brief ES about the target (McConnell & Rydell, 2014; McConnell et al., 2008). Both immediately and even after a 1-day delay, however, updated implicit evaluations remained consistent with the ES presented after A-REP. Such findings would certainly require a much greater amount of flexibility and interaction between systems than has often been assumed.

The results seem much more consistent with propositional accounts of REP and implicit evaluation (De Houwer, 2009, 2014, 2018), and offer fruitful data to support the idea that the strength of inferences drawn from any learning experience may mediate its impact on implicit evaluations (Van Dessel, Hughes, et al., 2019). Specifically, positive behaviors seem to result in stronger inferences about the positivity of the target person than abstract supposition. The perspective that implicit evaluations and REP effects are mediated by propositional reasoning readily accommodates the finding that ES updated A-REP based implicit evaluations, and an additional finding that subjective beliefs about the sound-target pairings predicted implicit evaluations (see Supplemental Material). A critical idea in the field of impression formation is that the relative diagnosticity of information matters for impression change (e.g., Skowronski & Carlston, 1989), a principle that has yielded insights into both explicit and implicit evaluation updating (Cone, Mann, & Ferguson, 2017). This suggests that the difference between the positive behavior ES condition and the suppose condition (Experiment 4) might be attributable in

part to stronger inferences in the former than the latter, perhaps due to the hypothetical nature of the suppose instruction. From this point of view, the results suggest that even if implicit and explicit cognition both draw upon propositional knowledge, diagnostic behavioral information about a target may sometimes be prioritized in early evaluative processing to a greater degree than other evaluative statements like supposition or counter-conditioning instructions (Gregg et al., 2006; Hu et al., 2017b).

In a similar way, the current findings can advance memory-based models of evaluative learning (e.g., Stahl & Aust, 2018), which frame dissociations between implicit and explicit measures in terms of differential retrieval of common representations from episodic memory based on the different features of tasks. For instance, the speeded nature of the IAT may impair integration of retrieved information (e.g., relational qualifiers) prior to the response, while the longer duration of an explicit measure may allow for full integration. In both cases, a common representational store of episodic detail is drawn upon. Under this framework, the current results of counter-attitudinal behavioral information being particularly impactful on implicit measures after earlier A-REP might suggest that this format of information facilitates retrieval under the conditions that operate during implicit measurement.

Beyond the strict dual-systems approach of SEM (McConnell & Rydell, 2014; Rydell & McConnell, 2006), the present results can also be explained by dual-process approaches that posit greater interaction between associative and propositional processes in learning, such as the APE model (Gawronski & Bodenhausen, 2006, 2018) and Metacognitive Model (MCM; Petty et al., 2006, 2007). To the degree that propositional learning can influence the strength of associations (Gawronski & Bodenhausen, 2018) or the strength of negation associations (Petty et al., 2007), strong ES could overcome earlier A-REP based learning. The present results do,

however, strongly support the idea that propositional learning of relatively minimal ES can have large and durable effects, even in competition with earlier associative learning from A-REP. At the least, this qualifies dual-process theories that propose that ES may need to be elaborated or rehearsed in order to overcome earlier REP (Petty et al., 2006; see also Wyer, 2010, 2016).

Limitations and Future Directions

We hope that future studies will be able to address some issues left unexplored by the present work. For instance, participants in the 1-day delay study (Experiment 3) were aware that they were taking part in a two-session experiment, and it is possible that this knowledge artificially increased their tendency to retain memory of the ES (and/or its impact on their impressions). In the absence of a perceived continued relevance of the ES information presented in the first session, it is possible that updated implicit evaluations would decay. Research showing that active goals can constrain memory in a goal-consistent manner supports this possibility (e.g., Ferguson & Wojnowicz, 2011; Zeigarnik, 1927). Future work measuring retested implicit evaluations in the absence of such strong expectations could provide insight into the generalizability of the current findings of durability under a broader set of conditions. Relatedly, future work can examine whether a delay longer than a timescale of days—which may allow more time for explicit memory of the previous session to fade—diminishes the strength of the updated implicit evaluations. Though few studies have examined the durability of shifts in implicit evaluations over longer timescales, with delays of minutes to days being most frequent (e.g., Kurdi & Banaji, 2019; Lai et al. 2016; Mann & Ferguson, 2015; cf. Devine et al., 2012; Forscher et al., 2017), establishing the long-term perseverance of updated evaluations remains a critically important topic for research.

Another notable area for future research is the generalization of the current results to a broader set of implicit measures. We selected the IAT as the implicit measure in the present work both because of its prominent use in studies seen as supporting the dual-systems perspective (Rydell & McConnell, 2006) and due to evidence that updating may be particularly difficult with the IAT (Van Dessel, Ye, et al., 2019; see also Moran, Bar-Anan, & Nosek, 2017). However, generalization across diverse measures with distinct properties is, of course, desirable (De Houwer et al., 2009). At present, the current findings of substantial updating parallel robust updating in other paradigms using the Affect Misattribution Procedure (Cone & Ferguson, 2015; Mann, Cone, Heggeseth, & Ferguson, 2019; Mann & Ferguson, 2015, 2017) and Evaluative Priming Task (Van Dessel, Ye, et al., 2019). However, comparison to other measures invoking features of automaticity such as speeded self-report (e.g., Gawronski, Ye, Rydell, & De Houwer, 2014) and the use of formal process models (e.g., Conrey, Sherman, Gawronski, Hugenberg, & Groom, 2005) may help to provide further clarity on the conditions and processes of updating.

Examining a wider variety of ES conditions in future work promises to further increase our understanding of when and how implicit evaluations can be updated, whether based on REP or ES (see also Lai et al., 2014, 2016). Future work can continue to isolate the features of new ES, and other moderating conditions and mediating mechanisms, that make the updating of implicit evaluations as potent as possible across different learning tasks. After all, there seem to be plenty of cases in which seemingly compelling interventions fail to result in meaningful immediate or lasting change in impressions (see also Lai et al., 2016). To the extent that the present result can shed light on such cases, it may be that persistence of discredited impressions – and the "stickiness" of first impressions in general – has little to do with any inherent sensitivity of implicit evaluations for nonverbal learning such as REP or insensitivity to ES (see

also Cone, Mann, & Ferguson, 2017). Instead, such stickiness may hinge on other moderating conditions of memory and cognition that have yet to be identified. The difference reported here between the positive behavior and suppose conditions, and the potential reasons for this difference discussed in the last section, offer a roadmap for future work to bridge the gap between studies demonstrating effective change and those that do not.

Conclusion

Theories and evidence have sometimes suggested that after initial implicit evaluations arise, they are resistant to purely verbal information—especially to the extent that they are rooted in repeated evaluative pairings. The present results counteract this conclusion, providing initial evidence on the possibility of such updating and the types of evaluative statements that can effectively drive it. Extrapolating from these data to social learning in daily life, the current results support the optimistic view that learning about individuals and perhaps even social groups can be modified in the face of countervailing evidence. Social cognition is not set in stone; it is flexible in the face of new information. The surprising aspect of the present work is the degree to which this is true when relying on updating methods that capitalize on a distinctly human ability of language-based propositional thought, which we have shown can overturn existing evaluations even with minimal inputs, as long as they are diagnostic.

References

Ambady, N., Bernieri, F. J., & Richeson, J. A. (2000). Toward a histology of social behavior: Judgmental accuracy from thin slices of the behavioral stream. In M. P. Zanna (Ed.), *Advances in experimental social psychology, Vol. 32* (pp. 201-271). Elsevier.

Bar-Anan, Y., & Balas, R. (2018). Why does co-occurrence change evaluation? Introduction to a special issue on evaluative conditioning. *Social Psychological Bulletin*, *13*(3), 1–10. http://doi.org/10.5964/spb.v13i3.29154

Bates, D., Maechler, M., Bolker, B., & Walker, S. (2015). Fitting linear mixed-effects models using lme4. *Journal of Statistical Software*, *67*(1), 1-48. doi:10.18637/jss.v067.i01

Cacioppo, J. T., Gardner, W. L., & Berntson, G. G. (1997). Beyond bipolar conceptualizations and measures: The case of attitudes and evaluative space. *Personality and Social Psychology Review*, *1*(1), 3–25.

Cone, J., & Ferguson, M. J. (2015). He did what? The Role of Diagnosticity in Revising Implicit Evaluations. *Journal of Personality and Social Psychology*, *108*(1), 37–57. http://doi.org/10.1037/pspa0000014

Cone, J., Flaharty, K., & Ferguson, M. J. (2019). Believability of evidence matters for correcting social impressions. *Proceedings of the National Academy of Sciences*, *116* (20), 9802-9807. https://doi.org/10.1073/pnas.1903222116

Cone, J., Mann, T. C., & Ferguson, M. J. (2017). Changing our implicit minds: How, when, and why implicit evaluations can be rapidly revised. In J. Olson (Ed.), *Advances in Experimental Social Psychology, Vol. 56* (pp. 131-199). New York: Academic Press.

Corneille, O., & Stahl, C. (2019). Associative Attitude Learning: A Closer Look at Evidence and How It Relates to Attitude Models. *Personality and Social Psychology Review*, *38*,

108886831876326–29. http://doi.org/10.1177/1088868318763261

Darley, J. M., & Fazio, R. H. (1980). Expectancy confirmation processes arising in the social interaction sequence. *American Psychologist*, *35*, 867–881. http://dx.doi.org/ 10.1037/0003-066X.35.10.867.

De Houwer, J. (2006). Using the Implicit Association Test does not rule out an impact of conscious propositional knowledge on evaluative conditioning. *Learning and Motivation*, *37*(2), 176–187. http://doi.org/10.1016/j.lmot.2005.12.002

De Houwer, J. (2009). The propositional approach to associative learning as an alternative for association formation models. *Learning & Behavior*, *37*(1), 1–20. http://doi.org/10.3758/LB.37.1.1

De Houwer, J. (2014). A propositional model of implicit evaluation. *Social and Personality Psychology Compass*, *8*(7), 342–353. http://doi.org/10.1111/spc3.12111

De Houwer, J. (2018). Propositional models of evaluative conditioning. *Social Psychological Bulletin*, *13*(3), 49–21. http://doi.org/10.5964/spb.v13i3.28046

De Houwer, J., Gawronski, B., & Barnes-Holmes, D. (2013). A functional-cognitive framework for attitude research. *European Review of Social Psychology*, *24*(1), 252–287. http://doi.org/10.1080/10463283.2014.892320

De Houwer, J., & Hughes, S. (2016). Evaluative conditioning as a symbolic phenomenon: On the relation between evaluative conditioning, evaluative conditioning via instructions, and persuasion. *Social Cognition*, 1–26.

De Houwer, J., Teige-Mocigemba, S., Spruyt, A., & Moors, A. (2009). Implicit measures: A normative analysis and review. *Psychological Bulletin*, *135*(3), 347–368. http://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. *Psychological Bulletin*, *127*(6), 853–869.

Devine, P. G., Forscher, P. S., Austin, A. J., & Cox, W. T. L. (2012). Long-term reduction in implicit race bias: A prejudice habit-breaking intervention. *Journal of Experimental Social Psychology*, *48*(6), 1267–1278. http://doi.org/10.1016/j.jesp.2012.06.003

Elsey, J. W. B., Van Ast, V. A., & Kindt, M. (2018). Human memory reconsolidation: A guiding framework and critical review of the evidence. *Psychological Bulletin*, *144*(8), 797–848. http://doi.org/10.1037/bul0000152

Fazio, R. H., Eiser, J. R., & Shook, N. J. (2004). Attitude formation through exploration: Valence asymmetries. *Journal of Personality and Social Psychology*, *87*, 293–311.

Ferguson, M. J., & Wojnowicz, M. T. (2011). The when and how of evaluative readiness: A social cognitive neuroscience perspective. *Social and Personality Psychology Compass*, *5*(12), 1018–1038.

Forscher, P. S., Mitamura, C., Dix, E. L., Cox, W. T. L., & Devine, P. G. (2017). Breaking the prejudice habit: Mechanisms, timecourse, and longevity. *Journal of Experimental Social Psychology*, 72, 133-146. http://doi.org/10.1016/j.jesp.2017.04.009

Gast, A., & De Houwer, J. (2013). The influence of extinction and counterconditioning instructions on evaluative conditioning effects. *Learning and Motivation*, *44*(4), 312–325. http://doi.org/10.1016/j.lmot.2013.03.003

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. http://doi.org/10.1037/0033-2909.132.5.692

Gawronski, B., & Bodenhausen, G. V. (2011). The Associative–Propositional Evaluation Model: Theory, evidence, and open questions. *Advances in Experimental Social Psychology* (1st ed., Vol. 44, pp. 59–127). Elsevier Inc. http://doi.org/10.1016/B978-0-12-385522-0.00002-0

Gawronski, B., & Bodenhausen, G. V. (2018). Evaluative conditioning from the perspective of the associative-propositional evaluation model. *Social Psychological Bulletin*, *13*(3), 1–33. http://doi.org/10.5964/spb.v13i3.28024

Gawronski, B., & Cesario, J. (2013). Of mice and men: What animal research can tell us about context effects on automatic responses in humans. Personality and Social Psychology Review, 17(2), 187–215. http://doi.org/10.1177/1088868313480096

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. http://doi.org/10.1016/j.jesp.2014.05.010

Gershman, S. J., Zhou, J., & Kommers, C. (2017). Imaginative reinforcement learning: computational principles and neural mechanisms. *Journal of Cognitive Neuroscience*, *29*(12), 2103–2113. http://doi.org/10.1162/jocn_a_01170

Glenn, C. R., Klein, D. N., Lissek, S., Britton, J. C., Pine, D. S., & Hajcak, G. (2011). The development of fear learning and generalization in 8-13 year-olds. *Developmental Psychobiology*, *54*(7), 675–684. http://doi.org/10.1002/dev.20616

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

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. http://doi.org/10.1037/0022-3514.85.2.197

Greenwald, A. G., Nosek, B. A., & Sriram, N. (2006). Consequential validity of the Implicit Association Test: Comment on Blanton and Jaccard (2006). *American Psychologist*, *61*(1), 56–61. http://doi.org/10.1037/0003-066X.61.1.56

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

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.

Higgins, E. T. (1996). Knowledge activation: Accessibility, applicability, and salience. In E. T. Higgins & A. W. Kruglanski (Eds.), Social psychology: Handbook of basic principles (pp. 133–168). New York: The Guilford Press.

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. http://doi.org/10.1037/a0018916

Hu, X., Gawronski, B., & Balas, R. (2017a). 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*(1), 17–32. https://doi.org/10.1177/0146167216673351

Hu, X., Gawronski, B., & Balas, R. (2017b). Propositional versus dual-process accounts of evaluative conditioning: II. The effectiveness of counter-conditioning and counter-instructions in changing implicit and explicit evaluations. *Social Psychological and Personality Science*, *8*(8), 858–866. http://doi.org/10.1177/1948550617691094

Hughes, S., Ye, Y., & De Houwer, J. (2019). Evaluative conditioning effects are modulated by the nature of contextual pairings. *Cognition & Emotion*, *33*(5), 871–884. http://doi.org/10.1080/02699931.2018.1500882

Hughes, S., Ye, Y., Van Dessel, P., & De Houwer, J. (2019). When people co-occur with good or bad events: Graded effects of relational qualifiers on evaluative conditioning. *Personality and Social Psychology Bulletin*, *45*(2), 196–208. http://doi.org/10.1177/0146167218781340

Kunda, Z. (1990). The case for motivated reasoning. *Psychological Bulletin*, *108*(3), 480–498.

Kurdi, B., & Banaji, M. R. (2017). Repeated evaluative pairings and evaluative statements: How effectively do they shift implicit attitudes? *Journal of Experimental Psychology: General*, *146*(2), 194–213. http://doi.org/10.1037/xge0000239

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), 681-703. http://doi.org/10.1037/pspa0000151

Kurdi, B., Gershman, S. J., & Banaji, M. R. (2019). Model-free and model-based learning processes in the updating of explicit and implicit evaluations. *Proceedings of the National Academy of Sciences*, *116*(13), 6035–6044. http://doi.org/10.1073/pnas.1820238116

Kuznetsova A., Brockhoff, P. B., & Christensen, R. H. B. (2017). lmerTest package: Tests in linear mixed effects models. *Journal of Statistical Software*, *82*(13), 1–26. doi: 10.18637/jss.v082.i13

Lai, C. K., Marini, M., Lehr, S. A., Cerruti, C., Shin, J.-E. L., Joy-Gaba, J. A., et al. (2014). Reducing implicit racial preferences: I. A comparative investigation of 17 interventions. *Journal of Experimental Psychology: General*, *143*(4), 1765–1785. http://doi.org/10.1037/a0036260

Lai, C. K., Skinner, A. L., Cooley, E., Murrar, S., Brauer, M., Devos, T., et al. (2016). Reducing implicit racial preferences: II. Intervention effectiveness across time. *Journal of Experimental Psychology: General*, *145*(8), 1001–1016. http://doi.org/10.1037/xge0000179

Langner, O., Dotsch, R., Bijlstra, G., Wigboldus, D. H. J., Hawk, S. T., & van Knippenberg, A. (2010). Presentation and validation of the Radboud Faces Database. *Cognition & Emotion*, *24*(8), 1377–1388. http://doi.org/10.1080/02699930903485076

Larsen, S. F., & Plunkett, K. (1987). Remembering experienced and reported events. *Applied Cognitive Psychology, 1,* 15–26. http://dx.doi .org/10.1002/acp.2350010104

Lau, J. Y., Britton, J. C., Nelson, E. E., Angold, A., Ernst, M., Goldwin, M., . . . Pine, D. S. (2011). Distinct neural signatures of threat learning in adolescents and adults. *Proceedings of the National Academy of Sciences of the United States of America*, *108*, 4500–4505.

Lau, J. Y. F., Lissek, S., Nelson, E. E., Lee, Y., Roberson- Nay, R., Poeth, K., . . . Pine, D. S. (2008). Fear conditioning in adolescents with anxiety disorders: Results from a novel experimental paradigm. *Journal of the American Academy of Child and Adolescent Psychiatry*, *47*, 94–102.

Lee, J. L. C., Nader, K., & Schiller, D. (2017). An Update on Memory Reconsolidation Updating. *Trends in Cognitive Sciences*, *21*(7), 1–15. http://doi.org/10.1016/j.tics.2017.04.006

Levey, A. B., & Martin, I. (1975). Classical conditioning of human “evaluative” responses. *Behaviour Research and Therapy*, *13*(4), 221–226. http://doi.org/10.1016/0005-7967(75)90026-1

Machlin, L., Miller, A. B., Snyder, J., McLaughlin, K. A., & Sheridan, M. A. (2019). Differential associations of deprivation and threat with cognitive control and fear conditioning in early

childhood. *Frontiers in Behavioral Neuroscience*, 13. http://doi.org/10.3389/fnbeh.2019.00080

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*(6), 823–849. http://doi.org/10.1037/pspa0000021

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. http://doi.org/10.1016/j.jesp.2016.06.004

Mann, T. C., Kurdi, B., & Banaji, M. R. (2019, August 21). How effectively can implicit evaluations be updated? Using evaluative statements after aversive repeated evaluative pairings. Retrieved from https://osf.io/43an6

McConnell, A. R., & Rydell, R. J. (2014). The systems of evaluation model: A dual-systems approach to attitudes. In J. Sherman, B. Gawronski, & Y. Trope (Eds.), *Dual process theories of the social mind* (pp. 204–217). New York: Guilford.

McConnell, A. R., Rydell, R. J., Strain, L. M., & Mackie, D. M. (2008). Forming implicit and explicit attitudes toward individuals: Social group association cues. *Journal of Personality and Social Psychology*, *94*(5), 792–807. http://doi.org/10.1037/0022-3514.94.5.792

Mende-Siedlecki, P., & Todorov, A. (2016). Neural dissociations between meaningful and mere inconsistency in impression updating. *Social Cognitive and Affective Neuroscience*, 11(*9*), 1489-1500.

Mitchell, C. J., De Houwer, J., & Lovibond, P. F. (2009). The propositional nature of human associative learning. *Behavioral and Brain Sciences*, *32*(02), 183–64. http://doi.org/10.1017/S0140525X09000855

Moran, T., & Bar-Anan, Y. (2013). The effect of object–valence relations on automatic evaluation. *Cognition & Emotion*, *27*(4), 743–752. http://doi.org/10.1080/02699931.2012.732040

Moran, T., Bar-Anan, Y., & Nosek, B. A. (2015). Processing goals moderate the effect of co-occurrence on automatic evaluation. *Journal of Experimental Social Psychology*, 1–16.

Moran, T., Bar-Anan, Y., & Nosek, B. A. (2016). The assimilative effect of co-occurrence on evaluation above and beyond the effect of relational qualifiers. *Social Cognition*, *34*(5), 435–461.

Moran, T., Bar-Anan, Y., & Nosek, B. A. (2017). The effect of the validity of co-occurrence on automatic and deliberate evaluations. *European Journal of Social Psychology*, *46*(6), 1101–16. http://doi.org/10.1002/ejsp.2266

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

Oyarzún, J. P., Lopez-Barroso, D., Fuentemilla, L., Cucurell, D., Pedraza, C., Rodriguez-Fornells, A., & de Diego-Balaguer, R. (2012). Updating fearful memories with extinction training during reconsolidation: A human experiment using auditory aversive stimuli. *PloS One*, *7*(6), e38849.

Peters, K. R., & Gawronski, B. (2011). Are we puppets on a string? Comparing the impact of contingency and validity on implicit and explicit evaluations. *Personality and Social Psychology Bulletin*, *37*(4), 557–569. http://doi.org/10.1177/0146167211400423

Petty, R. E., Briñol, P., & DeMarree, K. G. (2007). The meta-cognitive model (MCM) of attitudes: Implications for attitude measurement, change, and strength. *Social Cognition*, *25*(5), 657–686.

Petty, R. E., Tormala, Z. L., Briñol, P., & Jarvis, W. B. G. (2006). Implicit ambivalence from attitude change: An exploration of the PAST model. *Journal of Personality and Social Psychology*, *90*(1), 21–41. http://doi.org/10.1037/0022-3514.90.1.21

Rydell, R. J., & McConnell, A. R. (2006). Understanding implicit and explicit attitude change: A systems of reasoning analysis. *Journal of Personality and Social Psychology*, *91*(6), 995–1008. http://doi.org/10.1037/0022-3514.91.6.995

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*(11), 954–958.

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. http://doi.org/10.1002/ejsp.393

Schiller, D., Kanen, J. W., LeDoux, J. E., Monfils, M. H., & Phelps, E. A. (2013). Extinction during reconsolidation of threat memory diminishes prefrontal cortex involvement. *Proceedings of the National Academy of Sciences*, *110*(50), 20040–20045. http://doi.org/10.1073/pnas.1320322110

Schiller, D., Monfils, M. H., Raio, C. M., Johnson, D. C., LeDoux, J. E., & Phelps, E. A. (2010). Preventing the return of fear in humans using reconsolidation update mechanisms. *Nature*, *463*(7277), 49–53.

Schmitz, A., Merikangas, K., Swendsen, H., Cui, L., Heaton, L., & Grillon, C. (2011). Measuring anxious responses to predictable and unpredictable threat in children and adolescents. *Journal of Experimental Child Psychology*, *110*, 159–170.

Skowronski, J. J., & Carlston, D. E. (1989). Negativity and extremity biases in impression formation: A review of explanations. *Psychological Bulletin*, *105*(1), 131-142.

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*(2), 108–131.

Stahl, C., & Aust, F. (2018). Evaluative conditioning as memory-based judgment. *Social Psychological Bulletin*, *13*(3), Article e28589. https://doi.org/10.5964/spb.v13i3.28589

Strack, F., & Deutsch, R. (2004). Reflective and impulsive determinants of social behavior. *Personality and Social Psychology Review*, *8*(3), 220–247.

Tamir, D. I., & Thornton, M. A. (2018). Modeling the predictive social mind. *Trends in Cognitive Sciences*, *22*(3), 201-212. https://doi.org/10.1016/j.tics.2017.12.005

Todorov, A., Olivola, C. Y., Dotsch, R., & Mende-Siedlecki, P. (2015). Social attributions from faces: Determinants, consequences, accuracy, and functional significance. *Annual Review of Psychology*, *66*(1), 519–545. http://doi.org/10.1146/annurev-psych-113011-143831

Todorov, A., & Uleman, J. S. (2002). Spontaneous trait inferences are bound to actors' faces: Evidence from a false recognition paradigm. *Journal of Personality and Social Psychology*, *83*(5), 1051–1065. http://doi.org/10.1037//0022-3514.83.5.1051

Toglia, M. P., Shlechter, T. M., & Chevalier, D. S. (1992). Memory for directly and indirectly experienced events. *Applied Cognitive Psychology, 6,* 293–306. http://dx.doi.org/10.1002/acp.2350060403

Van Dessel, P., De Houwer, J., Gast, A., & Smith, C. T. (2015). Instruction-based approach-avoidance effects. *Experimental Psychology*, *62*(3), 161–169. http://doi.org/10.1027/1618-3169/a000282

Van Dessel, P., Hughes, S., & De Houwer, J. (2019). How do actions influence attitudes? An inferential account of the impact of action performance on stimulus evaluation. *Personality and Social Psychology Review*, *23*(3), 267–284. http://doi.org/10.1177/1088868318795730

Van Dessel, P., Ye, Y., & De Houwer, J. (2019). Changing deep-rooted implicit evaluation in the blink of an eye. *Social Psychological and Personality Science*, 10(2), 266–273. http://doi.org/10.1177/1948550617752064

Whitfield, M., & Jordan, C. H. (2009). Mutual influence of implicit and explicit attitudes. *Journal of Experimental Social Psychology*, *45*(4), 748–759. http://doi.org/10.1016/j.jesp.2009.04.006

Willis, J., & Todorov, A. (2006). First impressions: making up your mind after a 100-ms exposure to a face. *Psychological Science*, *17*, 592–598.

Wilson, T. D., Lindsey, S., & Schooler, T. Y. (2000). A model of dual attitudes. *Psychological Review*, *107*(1), 101–126. http://doi.org/10.1037/0033-295X.107.1.101

Wojciszke, B. (2005). Morality and competence in person- and self-perception. *European Review of Social Psychology*, *16*(1), 155–188. http://doi.org/10.1080/10463280500229619

Wyer, N. A. (2010). You never get a second chance to make a first (implicit) impression: The role of elaboration in the formation and revision of implicit impressions. *Social Cognition*, *28*(1), 1–19.

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. http://doi.org/10.1016/j.jesp.2015.12.006

Zanon, R., De Houwer, J., Gast, A., & Smith, C. T. (2014). When does relational information influence evaluative conditioning? *The Quarterly Journal of Experimental Psychology*,

67(11), 2105–2122. http://doi.org/10.1080/17470218.2014.907324

Zeigarnik, B. (1927). Das Behalten erledigter und unerledigter Handlungen [The memory of completed and uncompleted actions]. *Psychologische Forschung, 9,* 1–85.

Footnotes

[1] We use the term *implicit evaluations* to refer to observable responses on indirect measures, to clearly avoid conflating responses with the kinds of mental representations posited to produce them, which the broader term *implicit attitudes* may sometimes include (see De Houwer, Gawronski, & Barnes-Holmes, 2013). However, usage of the term *implicit attitudes* to refer only to positive or negative responses would be equivalent to the sense in which we mean *implicit evaluations* here (e.g., Kurdi & Banaji, 2017). Relatedly, though implicit measures are broadly posited to measure responses under some conditions of automaticity (e.g., unintentional, fast; see De Houwer, Teige-Mocigemba, Spruyt, & Moors, 2009), due to the emphasis of the present work on identifying if and when responses on such measures can be updated rather than on characterizing the operating conditions of the measures themselves, we remain agnostic here on the automaticity feature(s) that may be most central to implicit measures.

[2] In using these atheoretical terms, we aimed once again to avoid conflating procedural features with mental constructs or processes posited to explain them (De Houwer et al., 2013), like *association* and *associative learning* vs. *proposition* and *propositional learning*. While remaining *a priori* agnostic on the processes and representations that underlie responses on implicit and explicit measures, we describe these ideas in the context of specific theories that use these distinctions to make predictions about the extent to which implicit evaluations will update under various conditions. The use of atheoretical terms also facilitates discussion of how the findings of the current work can be reconciled with diverse theories.

[3] Across Experiments 1, 2, and 4, there were consistent main effects of the specific identities of the target and control faces (out of the set of four) on implicit evaluations, but face factor never moderated any of the effects of time, statement condition, or the time*statement

interaction in any of the experiments. In addition, IAT block order assignment on the first and second IATs often produced main effects that also did not interact with the focal manipulations. Results controlling for face assignments and IAT block order produce patterns of results consistent with those reported in the main text; these analyses are available in the Supplemental Material.

[4] One limitation of Experiment 2 is that the positive-explanation and positive-unrelated conditions were not pretested on valence or extremity, such that one statement might generally convey more positive or more diagnostic information than the other. In our view, this concern is tempered by the lack of significant difference between the conditions. It is possible, however, that the positive-explanation information was less strong than the positive-unrelated condition, but was boosted in effectiveness by its explanatory value, resulting in the two conditions being equivalently effective in updating implicit evaluations. We thank a Reviewer for calling this to our attention.

[5] The A-REP formation effect at Time 1 was independently significant (i.e., below 0) in both Experiment 1 and 2 even when restricting the analysis to the subset of the participants for whom these two stimuli served as target and control faces (see Supplemental Material).

[6] Due to differences in sample across sessions (due to assignment to the control IAT in Session 1, attrition, and the other exclusion criteria), we also analyzed the effect of time using only those participants who completed the faces IAT during Session 1 and for whom data were included for both sessions. Results of this analysis were substantively very similar to the conclusions of the mixed model, and are presented in the Supplemental Material. Included therein is an analysis of the effects of completing the initial implicit and/or explicit measure during the first session on IAT scores in the second session, which produced only one marginally

significant effect, additional analyses of implicit evaluations within each session, and the correlations between implicit evaluations across sessions.

[7] These conclusions are equivalent to the results of chi-squared deviance tests derived from iteratively adding terms to a simple random intercept model in the following order: statement main effect, $\chi^2(1) = 2.71$, $p = .100$; time main effect, $\chi^2(1) = 0.02$, $p = .881$; interaction between statement and time, $\chi^2(1) < 0.001$, $p = .999$.

[8] Attempting to add an additional random intercept to this model to capture variation attributable to Experiment (with 4 levels) produced a singular model that did fit the data significantly better than the initial model, $\chi^2(1) < 0.001$, $p = 1$. A model that instead added a fixed main effect of Experiment did not fit the data significantly better than the initial model, $\chi^2(3) = 5.83$, $p = .12$, and led to the same conclusions. Furthermore, consistent with the similarity in the effects of positive behavioral statements across experiments, a model including a fixed main effect of evaluative statement (ES) condition did not fit the data significantly better than the initial model, $\chi^2(2) = 3.33$, $p = .189$, nor did a model including both fixed main effects and interactions of ES condition, $\chi^2(4) = 4.73$, $p = .316$.

Table 1

Evidence for Implicit Evaluation Updating, by Format of Initial Learning and New Learning

	New learning	
Initial Learning	Pairings (REP)	Statements (ES)
Pairings (REP)	Strong updating *Hu et al., 2017b* *Kurdi et al., 2019* *Rydell et al., 2006*	**Weak or no updating** *Gast & De Houwer, 2013* *Hu et al., 2017b* *Kurdi & Banaji, 2019* *McConnell & Rydell, 2014* *Petty et al., 2006* *Rydell et al., 2006* *Whitfield & Jordan, 2009* *Zanon et al., 2014*
Statements (ES)	Insufficient evidence	Strong updating *Cone & Ferguson, 2015* *Mann et al., 2019* *Mann & Ferguson, 2015* *Mann & Ferguson, 2017*

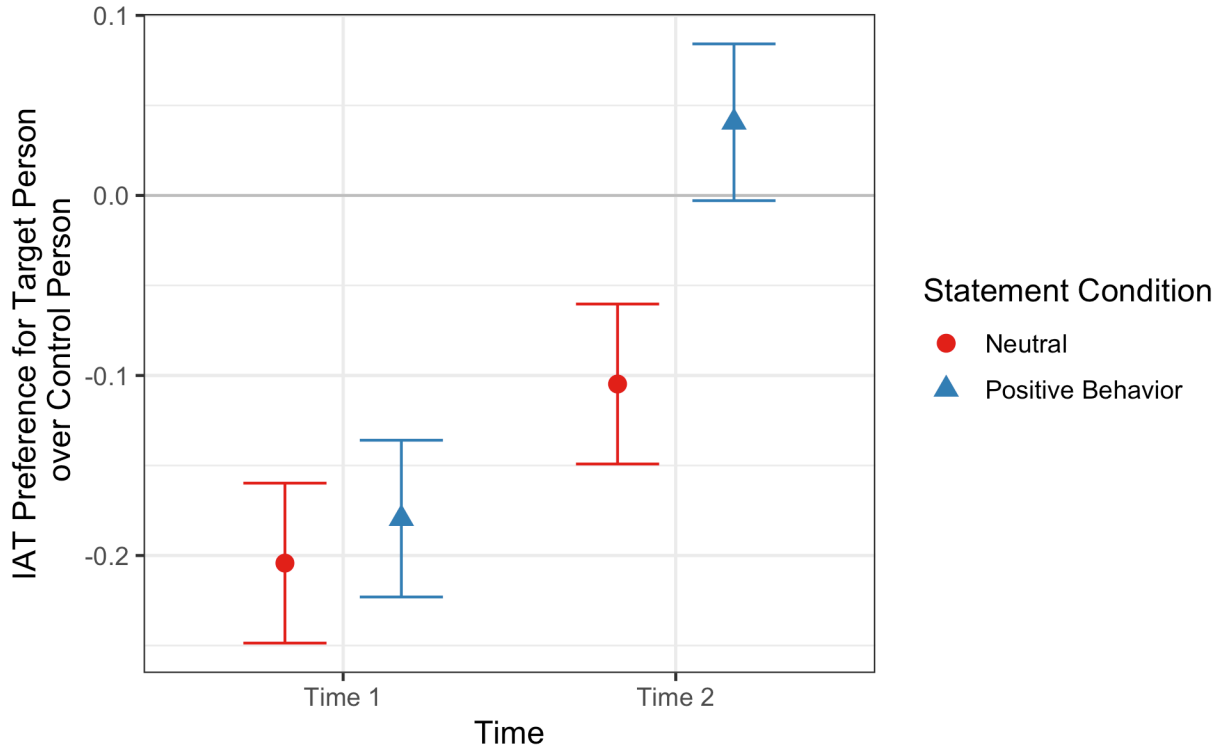

Figure 1. IAT D-score means by time and statement condition (Experiment 1). Higher values indicate greater relative evaluative preference for the target person over the control person, with zero indicating no relative preference. Error bars are 95% confidence intervals of the mean.

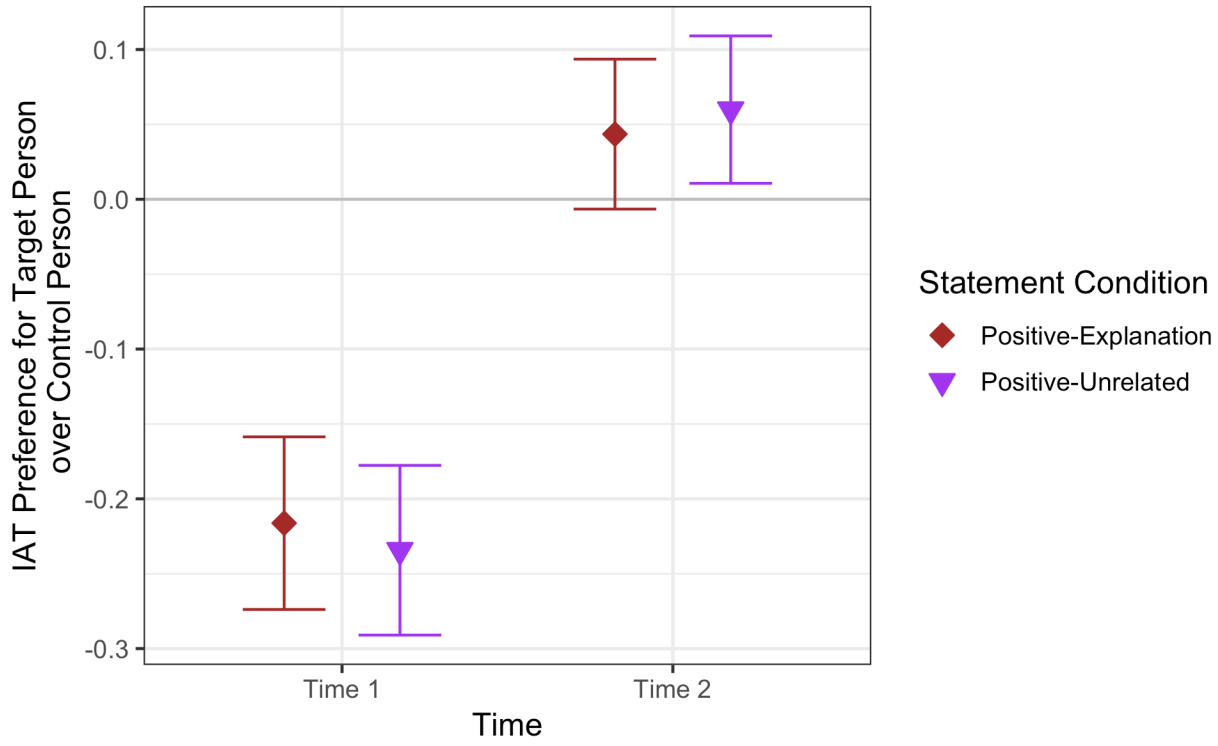

Figure 2. IAT D-score means, by time and statement condition (Experiment 2). Higher values indicate greater relative evaluative preference for the target person over the control person, with zero indicating no relative preference. Error bars are 95% confidence intervals of the mean.

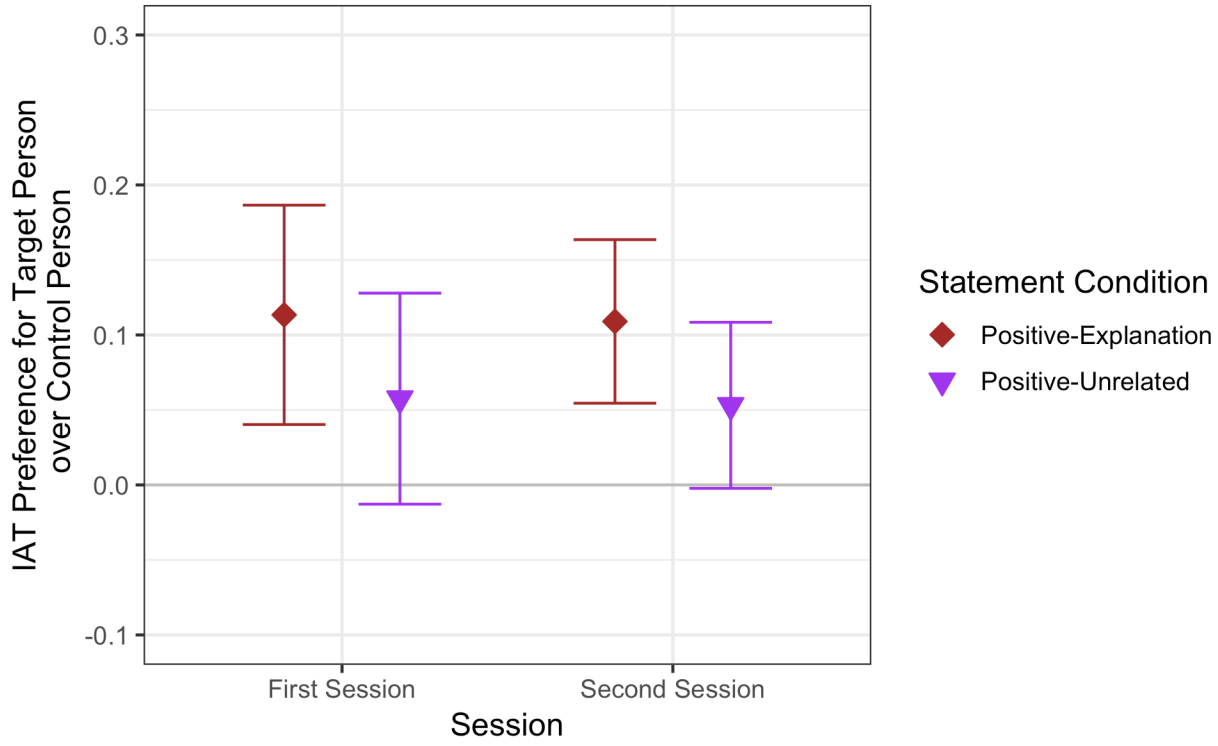

Figure 3. IAT D-score means, by session and statement condition (Experiment 3) estimated from the linear mixed effects model. Higher values indicate greater relative evaluative preference for the target person over the control person, with zero indicating no relative preference. Error bars are 95% confidence intervals of the mean.

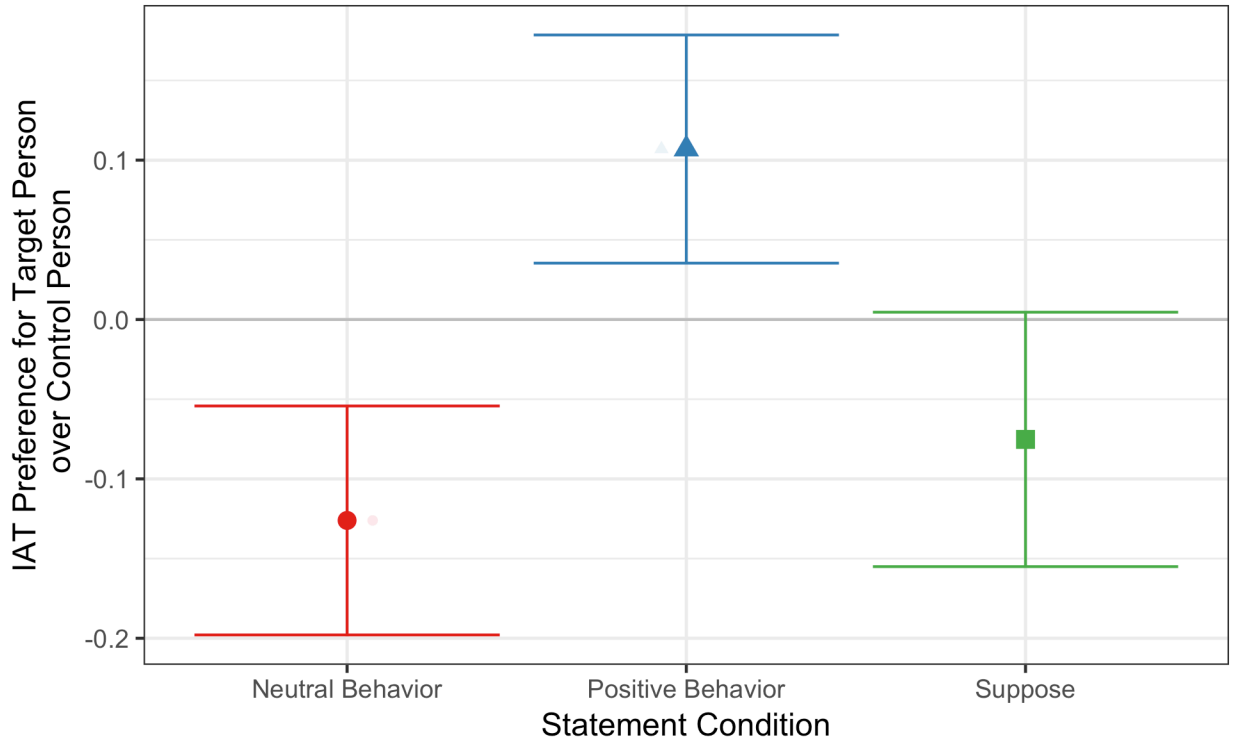

Figure 4. IAT D-score means by statement condition (Experiment 4). Higher values indicate greater relative evaluative preference for the target person over the control person, with zero indicating no relative preference. Error bars are 95% confidence intervals of the mean.