He Did What? The Role of Diagnosticity in Revising Implicit Evaluations

Jeremy Cone and Melissa J. Ferguson

Online First Publication, November 3, 2014. http://dx.doi.org/10.1037/pspa0000014

Journal of Personality and Social Psychology

© 2014 American Psychological Association
0022-3514/14/$12.00 http://dx.doi.org/10.1037/pspa0000014

He Did *What*? The Role of Diagnosticity in Revising Implicit Evaluations

Jeremy Cone
Yale University

Melissa J. Ferguson
Cornell University

Research suggests that implicit evaluations are relatively insensitive to single instances of new, countervailing information that contradicts prior learning. In 6 experiments, however, we identify the critical role of the perceived *diagnosticity* of that new information: Counterattitudinal information that is deemed highly *diagnostic* of the target's true nature leads to a complete reversal of the previous implicit evaluation. Experiments 1a and 1b establish this effect by showing that newly formed implicit evaluations are reversed minutes later with exposure to a single piece of highly diagnostic information. Experiment 2 demonstrates a valence asymmetry in participants' likelihood of exhibiting rapid reversals of newly formed positive versus negative implicit evaluations. Experiment 3 provides evidence that a target must be personally responsible for the counterattitudinal behavior and not merely incidentally associated with a negative act. Experiment 4 shows that participants exhibit revision only when they judge the target's counterattitudinal behavior as offensive and thus diagnostic of his character. Experiment 5 demonstrates the behavioral implications of newly revised implicit evaluations. These studies show that newly formed implicit evaluations can be completely overturned through deliberative considerations about a single piece of counterattitudinal information.

Keywords: implicit, attitude, evaluation, revision, diagnosticity

Throughout much of their over 30 years of marriage, Barbara Kuklinski would have described her life with her husband, Richard, in mostly idyllic terms. "We had what seemed to be the perfect life," she said. "We were the All-American family." It was thus quite a shock to Barbara and her children, on the day that Richard was arrested, when she abruptly discovered that he had been, over his 30-year tenure as a hit man for a number of mafia families in New Jersey, one of the most prolific contract killers in American history. The man whom she had known since she was 19 years old, whom she described as a romantic who would often leave flowers for her, and with whom she had raised three children was rather suddenly revealed to be a cold-blooded killer. He would later boast in interviews that he had been responsible for over 100 murders. Outside of what one detective described as a "normal, classic family existence," Kuklinski had earned himself the nickname *Iceman*—a reference to his technique of freezing his victims in an industrial-sized freezer to obscure the time of death. After his conviction, Kuklinski would spend the rest of his life in a maximum security prison, until his death in 2006 (Monet & Ginsberg, 2001).

When faced, like Barbara Kuklinski, with a revelation that is strikingly inconsistent with our well-learned and long-held beliefs about someone, to what extent can this single new piece of information change our evaluations of them? Contemporary research on evaluations suggests that minimal amounts of counterattitudinal information of this sort can quickly influence our evaluations in some ways but not others. Although *explicit,* self-reported evaluations can rapidly and readily incorporate revelations that overturn previous learning, some theory and evidence suggest that *implicit* (i.e., unintentional, spontaneous) evaluations are relatively less sensitive to such information (e.g., Gregg, Seibt, & Banaji, 2006; Rydell & McConnell, 2006; Rydell, McConnell, Mackie, & Strain, 2006; Rydell, McConnell, Strain, Claypool, & Hugenberg, 2007).

The claim that implicit evaluations are resistant to rapid revision is important because, in addition to suggesting that the processes underlying implicit versus explicit evaluations may differ (as we discuss in more detail below), it also has potential implications for our behavior. Indeed, implicit evaluations have been shown to uniquely predict our behavior in ways that explicit evaluations do not (Cameron, Brown-Iannuzzi, & Payne, 2012; Ferguson, 2007; Galdi, Arcuri, & Gawronski, 2008; Greenwald, Poehlman, Uhlmann, & Banaji, 2009; McNulty, Olson, Meltzer, & Shaffer, 2013; Perugini, Richetin, & Zogmaister, 2010; Towles-Schwen & Fazio, 2006; cf. Oswald, Mitchell, Blanton, Jaccard, & Tetlock, 2013). If implicit evaluations cannot be readily updated to accommodate a newly learned truth, they may ultimately guide us in ways that are inconsistent with our beliefs, resulting in potentially maladaptive behavioral outcomes.

In the present research, however, we show that implicit evaluations can, in fact, be highly responsive to single pieces of countervailing information, particularly when this information is extremely negative. We argue and provide evidence to suggest that this occurs because extreme negative information is judged as highly *diagnostic*—that is, especially revealing of a person's true

Jeremy Cone, Department of Psychology, Yale University; Melissa J. Ferguson, Department of Psychology, Cornell University.

This research was supported by National Institutes of Health Grant 1R21AG042662-01A1 awarded to Melissa J. Ferguson.

Correspondence concerning this article should be addressed to Jeremy Cone, 2 Hillhouse Avenue, Department of Psychology, Yale University, New Haven, CT 06511. E-mail: jeremy.cone@yale.edu

nature or character. Such information, we propose, can rapidly and fully reverse an otherwise well-learned and evaluatively uniform implicit evaluation toward an individual.

In what follows, we first review evidence that implicit evaluations are relatively insensitive to single instances of countervailing behavior and discuss how contemporary theories explain such insensitivity. Next, we consider the possibility that implicit evaluations may be more responsive to minimal amounts of countervailing information than current findings suggest, identifying the extremity and diagnosticity of new information as a factor that predicts strong revision.

Evidence of Insensitivity to Single Pieces of Countervailing Evidence

Several lines of work show that implicit evaluations become relatively insensitive to new information once established. First, some research suggests that implicit evaluations are relatively unresponsive to the negation of previously learned associations. In one study often cited as evidence of this assertion (see, e.g., Gawronski & Sritharan, 2010, p. 227), Gregg et al. (2006) had participants learn about two fictitious groups that strongly differed in their character—one group's actions were highly positive, while the other's were highly negative. Following this, participants' implicit evaluations of the groups were consistent with the induction. Next, Gregg and colleagues sought to undo this learning in various ways by telling participants reasons that the groups had changed character. These attempts ranged from assertions that the groups had changed to long and detailed narratives about how and why the groups had changed. Yet none of these strategies successfully altered implicit evaluations (despite the fact that they readily shifted explicit evaluations).

Second, implicit evaluations seem to respond slowly and gradually to new information, such that any single piece of information does not appear to exert a potent influence. For example, Rydell and colleagues (2007) gave participants extensive exposure to evaluatively consistent behavioral statements (100 trials) about a target individual named Bob (e.g., "Bob donates his time at the soup kitchen"). Afterward, participants were given varying amounts of counterattitudinal information (0, 20, 40, 60, 80, or 100 trials), followed by measures of their implicit and explicit evaluations of Bob. Whereas explicit evaluations were found to be highly responsive to the counterattitudinal information—exhibiting revision after just 20 trials—implicit evaluations were more resistant, responding incrementally and linearly to the *amount* of counterattitudinal information to which participants were exposed. It was not until participants had been exposed to 100 trials of counterconditioning that there was unequivocal evidence that implicit evaluations reflected the counterattitudinal information rather than the initial learning.

Some readers may be surprised by these results, given that implicit evaluations seem to shift rather easily in response to subtle contextual manipulations, such as being reminded of positive exemplars of an otherwise stigmatized group (e.g., Dasgupta & Greenwald, 2001; Ferguson & Bargh, 2004; Wittenbrink, Judd, & Park, 2001). However, there is an important distinction between effects of *contextualization* and changes of the sort we consider here. Contextualization reflects the activation of potentially well-established mental representations that were learned previously in different contexts. Because it is possible that these previously learned evaluations required a great deal of evaluative learning within a given context to develop, context dependence does not necessarily reveal anything about whether people can update their evaluation of a target based on learning new information (see also Fazio, 2007; Ferguson & Fukukura, 2012; Gregg et al., 2006, p. 16).

Why Might Implicit Evaluations Be Resistant to Minimal Counterevidence?

These lines of work indicate that implicit evaluations are less responsive to new information than explicit evaluations. This means that a revelation such as learning that one's husband is actually a contract killer for the mob could be expected to immediately alter one's self-reported evaluation but potentially have little effect on one's implicit evaluation. Why might this be? Whereas some theories have argued that such insensitivity is in line with assumptions about the operation of implicit and explicit evaluations, others argue that implicit evaluations may be more alterable than current evidence suggests.

Dual process models assert that such findings emerge because implicit and explicit evaluations are generated by two qualitatively distinct mental processes or systems (e.g., Bassili & Brown, 2005; Conrey & Smith, 2007; Gawronski & Bodenhausen, 2006, 2011; Rydell & McConnell, 2006; Rydell et al., 2006, 2007; Smith & DeCoster, 2000; Strack & Deutsch, 2004; Wilson, Lindsey, & Schooler, 2000). Implicit evaluations, on the one hand, are thought to emerge from associative mental processes that operate through the spreading activation of associations in memory. These mental associations typically develop and strengthen over time through repeated co-occurrence between an object and positively or negatively valenced experiences that occur in close spatial/temporal contiguity or that possess features that cue similarity between an object and other representations in memory. Importantly, associative processing is thought to operate independently of the perceived truth of the experiences one has with the object (Gawronski & Bodenhausen, 2006, 2011, 2014; Rydell & McConnell, 2006), meaning that one can hold an implicit evaluation completely at odds with one's explicit evaluation.

Explicit evaluations, on the other hand, are assumed to be governed by propositional processes that operate on the basis of rule-based logic, inference, and reasoning (Gawronski & Bodenhausen, 2006, 2011; Rydell & McConnell, 2006; Sloman, 1996; Strack & Deutsch, 2004). Propositional processes can operate abstractly and symbolically, giving rise to deliberative evaluations that may be either consistent or inconsistent with the spontaneous affective reactions that are the products of associative learning. In contrast to the associative mental processes underlying implicit evaluation, rule-based processes are capable of taking account of the validity of information and can thus negate past experiences if the consideration of current propositions warrants it. Thus, in contrast to implicit evaluation, explicit evaluations can more strongly and immediately reflect new information, even if this information is not the product of direct experience ("Did you hear what Jason did at the party Saturday?") and even if it is inconsistent with one's prior beliefs ("Oh, I had thought that Elaine was a nice person; I can't believe she did *that*!").

Some dual process models, relying on such assumptions, assert that implicit evaluations are insensitive to counterevidence once they are established (Rydell & McConnell, 2006; Rydell et al., 2006, 2007). For example, Rydell and McConnell's systems of evaluation model posits that implicit evaluations rely on a slow-learning associative system that is largely (but perhaps not completely; see McConnell & Rydell, 2014) incapable of rapidly incorporating new information about a target, particularly after a lot of prior information has already been learned.

Other models allow for greater interaction between the two processes. For example, according to Gawronski and Bodenhausen's (2006, 2011, 2014) associative–propositional evaluation (APE) model, the key distinction between implicit and explicit evaluation is whether implicit evaluations can incorporate validity information, particularly information that negates prior learning. Although new associations can be created that affirm the opposite of a previously held evaluation, once an association has been created, it cannot be negated by new information that challenges its validity (e.g., "That information was false"). Thus, counterevidence may be effective in changing implicit evaluations only when it affirms the opposite rather than negates prior learning (cf. Peters & Gawronski, 2011).

Empirical Evidence and Theory Consistent With Rapid Revision

Although there is evidence suggesting that implicit evaluations are relatively unresponsive to single pieces of countervailing information, some research suggests that information can be rapidly incorporated into implicit evaluations (see Mann, Cone, & Ferguson, in press). First, implicit evaluations can *form* toward novel targets extremely rapidly (Ashburn-Nardo, Voils, & Monteith, 2001; Castelli, Zogmaister, Smith, & Arcuri, 2004; De Houwer, 2006; Gregg et al., 2006; Otten & Wentura, 1999; Whitfield & Jordan, 2009)—sometimes in response to only a single piece of information about a target. For example, Gregg et al. (2006) found that when participants were first asked to merely suppose that novel, fictional groups possessed good or bad traits, they immediately exhibited implicit positivity toward the good versus bad group, suggesting that implicit evaluations were rather easy to create toward novel targets. Similarly, De Houwer (2006) found that when he informed participants that nonsense syllables would, later in the experiment, signal the immediate presentation of a positive or negative photo, they exhibited implicit evaluations in line with this information even before the pairings had occurred. If implicit evaluations can *form* so rapidly toward novel stimuli, as these studies demonstrate, it may also be that they can be *revised* rather quickly under certain circumstances.

Second, there is evidence that implicit evaluations can be at least somewhat modified by small amounts of behavioral information. For example, in one study, Whitfield and Jordan (2009, Study 3) had participants undergo an evaluative conditioning procedure and then exposed them to 13 behavioral statements that were in the opposing direction from the conditioning. They found that participants' relative implicit preference for one target over the other was responsive to these statements (though it was unclear from their analysis whether these 13 statements were enough to lead to full reversals of the previously held preferences). Third, recent evidence also suggests that implicit evaluations may be responsive to negations under some circumstances. When Peters and Gawronski (2011; see also Boucher & Rydell, 2012) exposed participants to information about a novel target (e.g., "Mike lent money to a friend in financial trouble") that was sometimes immediately revealed to be false, participants' implicit evaluations did, in fact, reflect the validity of the learned information—a finding that was inconsistent with Peters and Gawronski's expectations and with a number of current dual process models. Interestingly, however, when participants were notified of the truth or falsity of the behavioral information after a short delay, their implicit evaluations were less sensitive to this information (see also Foerde & Shohamy, 2011). Though it is unclear why implicit evaluations were responsive to the truth of the behavioral statements in one case but not the other, these results nonetheless raise the possibility that there are circumstances under which countervailing information can influence implicit evaluation formation and change.

Some contemporary models of implicit and explicit evaluation are consistent with this growing body of evidence demonstrating that implicit evaluations can sometimes exhibit responsiveness to minimal counterattitudinal information. For example, De Houwer and colleagues' single-process model rejects the existence of a (slow-learning) associative system altogether, positing instead that implicit and explicit evaluations both rely on the formation of propositions about a stimulus. On this view, once a proposition has been formed in memory, it can then be retrieved automatically (e.g., Bar-Anan, De Houwer, & Nosek, 2010; Hughes, Barnes-Holmes, & De Houwer, 2011; see also De Houwer, 2014), and because propositional processes are thought to operate rather quickly, implicit evaluations may thus also change rather quickly in response to new propositions.

As another example, the memory systems model (Amodio, 2014; Amodio & Ratner, 2011; see also Amodio & Devine, 2006) posits that, rather than implicit and explicit evaluations relying on a single process, implicit social cognition more generally is instead the product of multiple interacting systems that each have their own properties of acquisition and extinction. For instance, whereas (implicit) semantic learning of the sort that occurs when one learns to pair *doctor* with *nurse* may proceed slowly and extinguish only gradually, Pavlovian fear conditioning in animal models develops rapidly (e.g., Hermer-Vazquez et al., 2005; Yin & Knowlton, 2006), including during single trial episodes (e.g., see Cahill & McGaugh, 1990; Hilliard, Nguyen, & Domjan, 1997). The developmental and learning trajectories of implicit evaluations should be influenced by the type(s) of memory system(s) involved in the acquisition of new information (see also Ferguson, Mann, & Wojnowicz, 2014). If information is acquired through a (slower learning) semantic memory process, it may develop and change rather slowly; however, if some types of implicit learning proceed quite rapidly, they may be updated in response to minimal amounts of information.

There is also a burgeoning literature over the last decade suggesting that implicit evaluation measures such as the implicit association test (Greenwald, McGhee, & Schwartz, 1998; Greenwald et al., 2009) or affect misattribution procedure (AMP; Payne, Cheng, Govorun, & Steward, 2005) are unlikely to reflect just one type of process, whether associative or propositional. Sherman, Payne, and colleagues (e.g., Bishara & Payne, 2009; Conrey, Sherman, Gawronski, Hugenberg, & Groom, 2005; Sherman, 2006) have provided evidence that multiple processes underlie implicit evaluation measures—including processes that have been characterized as controlled in nature (Sherman, Klauer, & Allen, 2010). This work implies that implicit evaluations may not be under the sole purview of associative processes (see also Mitchell, De Houwer, & Lovibond, 2009), and it is up to researchers to identify and catalogue the features of the conditions under which implicit evaluations form and change in an attempt to better understand these processes. The present work is one attempt to this end.

The Role of Extremity and Diagnosticity

Our goal in the current studies was to identify one type of minimal, countervailing information that could lead to rapid revision of implicit evaluations—specifically, the extent to which new information is judged as especially diagnostic of the person's true nature. In line with prior research, in order to signal such diagnosticity, we presented behaviors that were *extreme* in valence (e.g., Fiske, 1980). We examine the effects of single instances of both extreme positive and extreme negative counterattitudinal verbal information. However, we predict and find evidence that extreme negative information is especially potent—a proposal in line with classic work on person perception suggesting that extreme negativity is particularly diagnostic of a person's disposition and likely future behavior (e.g., Knobe, 2003, 2006; Malle & Knobe, 1997; Mende-Siedlecki, Baron, & Todorov, 2013; Nadelhoffer, 2006; Reeder & Brewer, 1979; Reeder & Coovert, 1986; Reeder, Pryor, & Wojciszke, 1992; Trafimow & Schneider, 1994; Trafimow & Trafimow, 1999). In other words, we argue that extremely negative behavior is likely to overturn previously positive implicit evaluations because such behavior is often perceived as especially diagnostic of the person's true character.

Consider, for example, the revelation that someone has been convicted of a serious crime, not unlike the situation faced by Barbara Kuklinski. To the degree that one sees this revelation as revealing of the kind of person someone *truly* is, one may see a great deal of informational value in this single revelation, thus imbuing it with greater significance and ascribing it greater weight in one's evaluation of the target.

If the processes that govern implicit evaluation are entirely passive and responsive primarily to the amount of counterattitudinal information such that each piece of information is given approximately equal weighting, we should expect that a single piece of information—no matter how relevant—ought to influence a well-established implicit response only mildly, if at all. However, if one's assessment of the diagnosticity of each instance of behavior matters, then even a single piece of highly diagnostic information ought to rapidly undo one's prior implicit tendencies.

The Current Studies

Across six studies, we tested whether participants' implicit evaluations of a novel target were responsive to single, extreme, highly diagnostic verbal descriptions of behavior. To test this hypothesis, all of the experiments we report have a similar format. We first had participants undergo an extensive learning paradigm that consisted of 100 repeated co-occurrences of a target (referred to as Bob throughout the experiment) in conjunction with evaluatively consistent behavioral information—either positivity or negativity (e.g., Gawronski, Rydell, Vervliet, & De Houwer, 2010; Rydell & Gawronski, 2009; Rydell & McConnell, 2006; Rydell et al., 2006, 2007). We then exposed participants to their 101st instance of Bob's behavior; however, unlike the previous 100 behaviors, this one was preselected to be (a) inconsistent with the previous 100 instances of behavior and (b) particularly extreme in valence and thus especially high in perceived diagnosticity. We measured participants' evaluations toward the target before and after they were exposed to this single new piece of information, thus assessing whether their evaluations incorporated the new information. Importantly, throughout our experiments, we use novel objects and thus have control over participants' entire evaluative history with the object, reducing the likelihood that contextualization can explain the changes in task performance we observe across our studies (see Fazio, 2007; Ferguson & Fukukura, 2012; Gregg et al., 2006).

In Experiments 1a and 1b, we demonstrate that participants' implicit evaluations of Bob are sensitive to (two different) extreme negative behaviors. Experiment 2 extends our analysis by examining whether there is a valence asymmetry in participants' readiness to update their impressions in light of extremely positive versus extremely negative information. In Experiment 3, we show, however, that not all extremely negative behaviors are rapidly incorporated into participants' implicit responses. Rather, they must be clearly attributable to the target. To this end, we expose participants to two similarly extreme behaviors but implicate either the target or, alternatively, an individual who is merely incidentally associated with the target. We show revision in the former but not the latter case. In Experiment 4, we test our proposed mechanism of diagnosticity, providing evidence consistent with a mediational pathway in which those participants who judge an act to be extremely offensive are more inclined to view it as highly diagnostic of a person's true character and to then show the most evaluative updating. Finally, in Experiment 5, we test the behavioral implications of revised implicit evaluations.

Following Simmons, Nelson, and Simonsohn's (2012) recommendation, we report how we determined our sample size, all data exclusions (if any), all manipulations, and all measures in each of the six studies reported. With respect to sample size, in our first study we attempted to recruit approximately 30 subjects per cell. We then used the effect size in this experiment (1a) to determine that a sample size of 90 per cell would give us sufficient statistical power to detect a difference of equivalent size 90% of the time. We thus recruited 90 per cell for all remaining studies except for Experiment 4 (in which we test a mediational pathway and thus increased our intended sample size to the predetermined 240) and Experiment 5 (in which we test correlations with our implicit measures and thus increased our intended sample size to 300).

Experiment 1a: Who Is Bob, Really?

Method

Participants. Fifty-seven undergraduates participated in exchange for credit or $5. The responses from two participants were excluded from analyses due to a lack of variance (i.e., failing to follow instructions by pressing the same response key on all trials; see Payne et al., 2005, for similar exclusion criteria) on one or both implicit measures (described below). One participant pressed invalid response keys on an excessive number of trials (>10%) on both implicit measures and was thus also excluded from analyses. This left a final sample of 54 participants.

Procedure. Participants completed a learning paradigm in which they read a total of 100 behavioral statements about a target individual named Bob (e.g., "Bob gave a hitchhiker a ride to a shelter"), identified by a picture presented at the top of the screen on each trial (see Rydell & McConnell, 2006; Rydell et al., 2007).[1] Participants' task was to assess whether they thought each target behavior was characteristic or uncharacteristic of Bob by pressing the *c* key (characteristic) or *u* key (uncharacteristic). They received immediate feedback on their response, consisting of either the word *correct* printed in blue text or *incorrect* printed in red text, followed below it by a summary of the meaning of the feedback (e.g., "Giving a hitchhiker a ride to a shelter is characteristic of Bob"). To ensure that participants formed an initially positive and well-rehearsed evaluation toward Bob, all 50 positive behavioral statements were said to be characteristic of Bob, while all 50 negative statements were said to uncharacteristic of him. The order of the behavioral statements was randomly determined for each participant. These statements have been used previously in attitude formation research and are moderate in valence.

Assessing implicit and explicit evaluations. After the 100 learning trials, participants' implicit evaluations of Bob were assessed using an AMP (see Payne et al., 2005). Each of the 60 trials in the measure consisted of (a) a prime image (75 ms) of either Bob (30 trials) or one of six unfamiliar college-age White males (five trials each), (b) a Chinese pictograph (100 ms), and, (c) a backward mask (random black and white noise), which remained on the screen until participants provided a response. Participants' task was to indicate whether the Chinese pictograph was more or less pleasant than average. Following Payne et al.'s (2005) procedure, participants were warned that having just seen a positive or negative image immediately before the pictograph could sometimes bias their evaluation and that we were interested in their ability to ignore such influences and provide an objective assessment of each pictograph. Previous research (e.g., Payne et al., 2005) has shown that people misattribute their automatic evaluation of the prime to their feelings about the pictograph, thus providing a measure of people's spontaneous and unintentional evaluations of the primes. Although some recent evidence by Bar-Anan and Nosek (2012) raised the possibility that evaluations measured by the AMP may be the result of participants intentionally evaluating the primes (rather than unintentionally rating the Chinese pictographs), more recent work by Payne et al. (2013) has shown that implicit evaluations as measured by the AMP do seem to reflect unintentional responses to the prime stimuli and that participants' reported intentional evaluations of the primes appear to be post hoc confabulations.

Changing impressions of Bob. To assess the effects of extreme countervailing information on participants' implicit responses, participants were next told that they would learn a single new piece of information about Bob and that they should pay close attention because this new piece of information was more recent and could potentially violate the impression they had already formed. Participants in the *experimental* condition read the statement "Bob was recently convicted of molesting children." Participants in the *control* condition read the statement "Bob recently bought a soda." In both conditions, immediately after the statement, participants were reminded that this piece of information was characteristic of Bob and that they would be tested for their memory of it later in the experiment.

Participants then completed a Time 2 implicit evaluation measure, following the same protocol as the Time 1 measure except that a different set of 60 Chinese pictographs were presented during the AMP (to control for familiarity and mere exposure). Next, participants completed an explicit evaluation measure in which they judged Bob's likeability from 1 (*very unlikeable*) to 9 (*very likeable*) as well as how *bad–good*, *mean–pleasant*, *disagreeable–agreeable*, *uncaring–caring*, and *cruel–kind* they considered him to be, all on 9-point Likert-type scales (see Rydell et al., 2006, 2007). Because these items were highly intercorrelated ($\alpha = .912$), they were averaged into a single composite explicit evaluation measure.

Finally, participants completed a short demographics questionnaire and were debriefed, thanked, and dismissed.

Results

To create a measure of participants' implicit evaluations of Bob, we calculated, for each AMP, the proportion of times participants indicated that the target pictograph was more pleasant than average separately for Bob trials and neutral trials. We submitted these proportion-pleasant judgments to a 2 (time: 1 or 2) × 2 (condition: experimental or control) × 2 (prime: Bob or neutral) mixed-model analysis of variance (ANOVA) with time and prime as within-subjects factors and condition as a between-subjects factor. This analysis yielded the predicted three-way interaction, $F(1, 53) = 12.351$, $p < .01$, $\eta_p^2 = .189$ (see Figure 1).

Decomposing this interaction, in the experimental condition, there was the predicted significant interaction between time and prime, $F(1, 27) = 8.216$, $p < .01$, such that participants exhibited greater implicit positivity toward Bob ($M = .56$, $SD = .28$) than the unfamiliar faces ($M = .49$, $SD = .21$) at Time 1. Though this difference failed to reach significance in this experiment, $F(1, 27) = 1.327$, *ns*, in all subsequent experiments we find reliable evidence of formation at Time 1. At Time 2, as predicted, there was a complete reversal of the pattern at Time 1, such that participants exhibited significantly more implicit negativity toward Bob ($M = .39$, $SD = .22$) than toward the neutral targets ($M = .55$, $SD = .21$), $F(1, 27) = 8.384$, $p < .01$.

In the control condition, there was a predicted main effect of prime, $F(1, 26) = 14.267$, $p < .001$, such that participants exhibited greater implicit positivity toward Bob at both Time 1 (Bob: $M = .57$, $SD = .23$; Neutral: $M = .46$, $SD = .17$), $F(1, 26) = 4.590$, $p < .05$, and Time 2 (Bob: $M = .65$, $SD = .24$; Neutral: $M = .39$, $SD = .20$), $F(1, 26) = 14.786$, $p < .001$.[2]

[1] Following a similar protocol to Rydell and colleagues' (2006, 2007) procedure, one of six different pictures of White males (determined to be of similar physical attractiveness in a pretest) was used to represent Bob, randomly assigned for each participant. Whenever a particular target served as Bob, the other five targets served as neutral stimuli in the implicit measure. There were no effects of target picture in any of the studies except where noted.

[2] There was also an unpredicted interaction between time and prime, $F(1, 26) = 4.308$, $p < .05$, such that participants exhibited increased positivity toward Bob relative to unfamiliar faces at Time 2. (This unpredicted interaction did not emerge in any subsequent studies.)

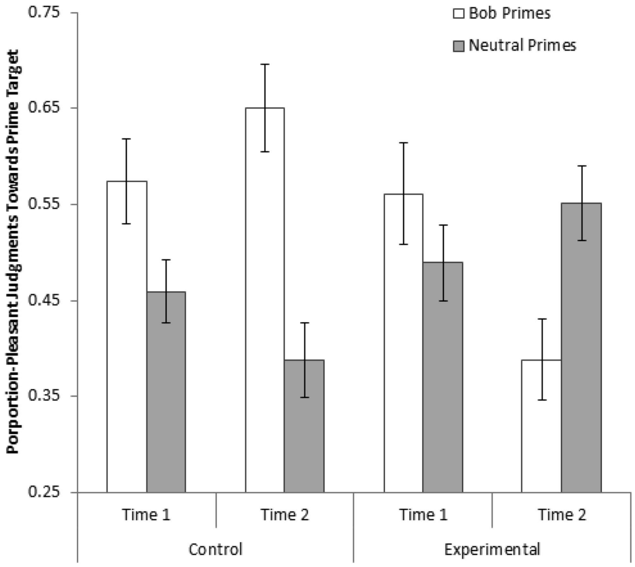

Figure 1. Participants' proportion-pleasant judgments toward Bob before (Time 1) and after (Time 2) learning either that he was convicted of molesting children (experimental) or that he bought a soda (control) in Experiment 1a. The error bars represent standard errors.

Explicit evaluations. As predicted, participants reported more favorable explicit evaluations of Bob after learning he had bought a soda ($M = 7.69$, $SD = 2.12$) versus had recently been convicted of molesting children ($M = 4.66$, $SD = 1.95$), $t(51) = 5.422$, $p < .001$. Thus, participants' explicit evaluations were highly sensitive to the new information learned at Time 2—a finding that is consistent with previous research (e.g., Gregg et al., 2006; Rydell et al., 2007).

Discussion

Following an extensive learning paradigm that has been shown in past research to result in relatively entrenched implicit evaluations (e.g., Rydell et al., 2007), participants exhibited implicit positivity toward a novel target. More importantly, whereas neutral information had little impact on participants' task performance on the Time 2 implicit measure, a single piece of diagnostic information that was inconsistent with prior learning fully reversed participants' Time 1 implicit responses. Whereas, at Time 1, individuals trended toward greater implicit positivity toward Bob than novel, neutral targets, they exhibited significantly *less* implicit positivity toward Bob than neutral targets at Time 2. This reversal resulted in a wide disparity between experimental and control participants' implicit evaluations toward Bob at Time 2, even though these implicit responses were essentially equivalent a minute earlier.

Though these results are consistent with the notion that implicit evaluations can be rapidly revised in light of just a single piece of countervailing information, there are, of course, many peculiarities of child molestation that may have made it a rather potent influence on task performance. To ensure that our findings were not unique to child molestation, in the next experiment we sought to replicate this pattern of results with a different, extremely negative behavior.

Experiment 1b: Who Is Bob? II

Method

Participants. One hundred and eighty Mechanical Turk workers (http://www.mturk.com) completed an online experiment on "learning about a new person." Five participants submitted a survey code but failed to complete all of the components of the study and were thus excluded from analyses. Responses from an additional 12 participants were excluded from analyses due to (a) a lack of variance in their responses on one or both of the implicit measures or (b) excessive pressing (>10% of trials) of incorrect response keys on the implicit measures (see Payne et al., 2005). This left a final sample of 163 participants.

Procedure. The procedure for Experiment 1b was identical to Experiment 1a with the following exceptions. First, rather than completing the experiment in a cubicle in the lab for academic credit or $5, workers on Amazon's Mechanical Turk service were directed to a survey link and completed the entire study from their browser in exchange for a small sum of money. This change in procedure also afforded us an opportunity to draw from a more diverse population in terms of ethnicity, age, and socioeconomic status than the undergraduate sample we drew from in Experiment 1a (e.g., Buhrmester, Kwang, & Gosling, 2011). Second, participants completed a Time 1 explicit evaluation measure immediately after completing the first AMP, following the same protocol as the measure in Experiment 1a (Time 1: $\alpha = .963$, Time 2: $\alpha = .986$). This allowed us to test implicit and explicit evaluation change simultaneously. Third, at Time 2, instead of learning that Bob had been convicted of molesting children, participants in the experimental condition were told that "Bob recently mutilated a small, defenseless animal." On the basis of a short pretest completed by 60 undergraduates that included a variety of different positive and negative behavioral statements, this behavior was determined to be moderately high in diagnosticity (on a scale from 1 = *not at all diagnostic* to 7 = *extremely diagnostic*; $M = 5.51$, $SD = 1.48$; for comparison, child molester item: $M = 6.63$, $SD = 1.03$) and of high negative valence (on a scale from −3 = *extremely negative* to 3 = *extremely positive*; $M = -2.42$, $SD = .72$; child molester: $M = -2.93$, $SD = .41$). Finally, we included an exploratory measure at the end of the experiment in which we assessed participants' explicit memory for behaviors they saw during the initial learning task by giving them a 10-item recognition test (e.g., "Did this behavior appear in the earlier task: 'Bob helped a foreign student find a place to live'?"). No effects emerged on this measure, and we thus do not discuss it any further.

After this exploratory measure, participants answered several demographics questions and were then given a code to enter on mturk to receive their payment.

Results

A 2 (time: 1 or 2) × 2 (condition: experimental or control) × 2 (prime: Bob or neutral) mixed ANOVA on participants' proportion-pleasant judgments again yielded the predicted three-way interaction, $F(1, 161) = 40.178$, $p < .001$, $\eta_p^2 = .200$ (see Figure 2).

Decomposing this interaction, in the experimental condition, there was a significant interaction between time and prime, F(1,

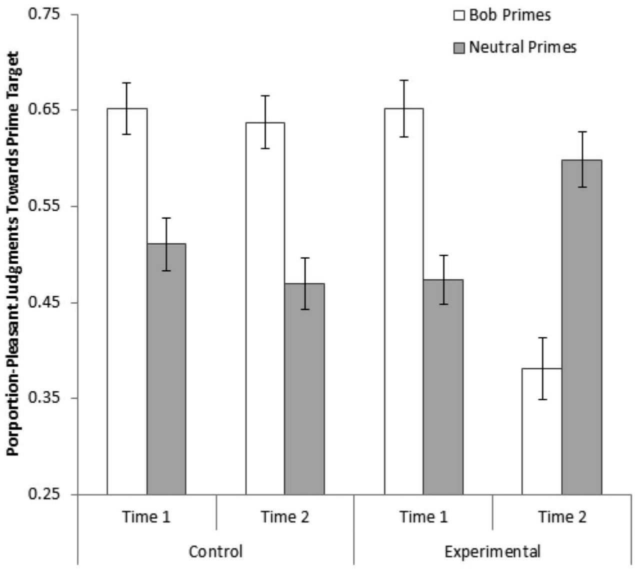

Figure 2. Participants' proportion-pleasant judgments toward Bob before (Time 1) and after (Time 2) learning either that he mutilated a small animal (experimental) or that he bought a soda (control) in Experiment 1b. The error bars represent standard errors.

80) = 41.156, $p < .001$, such that participants exhibited significantly greater implicit positivity toward Bob ($M = .65$, $SD = .27$) than the unfamiliar faces ($M = .47$, $SD = .23$) at Time 1, $F(1, 80) = 18.562$, $p < .001$. At Time 2, however, there was a full reversal of the pattern at Time 1, such that participants now exhibited significantly more negativity toward Bob ($M = .38$, $SD = .29$) than toward the neutral targets ($M = .60$, $SD = .26$), $F(1, 80) = 18.246$, $p < .001$. Interestingly, this effect was driven not just by (predicted) decreases in implicit positivity toward Bob from Time 1 ($M = .65$, $SD = .27$) to Time 2 ($M = .38$, $SD = .29$), $F(1, 80) = 40.937$, $p < .001$, but also partly by an unpredicted *increase* in implicit positivity toward *unfamiliar targets* from Time 1 ($M = .47$, $SD = .23$) to Time 2 ($M = .60$, $SD = .26$), $F(1, 80) = 16.621$, $p < .001$. This latter finding for unfamiliar targets, although unpredicted, emerged across the studies. It is intriguing evidence that in addition to implicit evaluations of Bob changing based on the newly learned behavior, participants also changed their implicit evaluations of the *control* faces based on a comparison with Bob. As Bob became much worse, these strangers become relatively more positive. However, this pattern might also have emerged simply as a response bias due to one of the two classes of prime stimuli in this task being rather highly affectively charged, thus serving as a kind of default that influenced responses to the other stimuli.

In the control condition, there was once again the predicted main effect of prime, $F(1, 81) = 16.120$, $p < .001$. Participants exhibited equivalent levels of implicit positivity toward Bob relative to neutral targets at both Time 1 (Bob: $M = .65$, $SD = .25$; Neutral: $M = .51$, $SD = .25$), $F(1, 81) = 13.421$, $p < .001$, and Time 2 (Bob: $M = .64$, $SD = .25$; Neutral: $M = .47$, $SD = .24$), $F(1, 81) = 15.592$, $p < .001$.

Explicit evaluations. Because Experiment 1b included a Time 1 explicit evaluation measure, we were able to test revision of explicit evaluations. As anticipated, an interaction between time and condition emerged, $F(1, 161) = 276.828$, $p < .001$, $\eta_p^2 = .635$ (see Figure 3), such that participants in the control condition exhibited equivalent positivity toward Bob at Time 1 ($M = 6.59$, $SD = .51$) and Time 2 ($M = 6.60$, $SD = .84$; $F < 1$, *ns*), whereas participants in the experimental condition expressed strong positivity toward Bob at Time 1 ($M = 6.72$, $SD = .58$) but negativity at Time 2 ($M = 3.13$, $SD = 1.82$), $F(1, 80) = 279.173$, $p < .001$.

Discussion

In Experiment 1b, we again found evidence that implicit evaluations were highly sensitive to a single piece of extremely negative information. Whereas participants' implicit responses in the control condition were largely unchanged from Time 1 to Time 2, those in the experimental condition exhibited a reversal, such that their positivity toward Bob at Time 1 was completely reversed at Time 2. Participants' explicit evaluations mirrored their implicit responses, exhibiting sensitivity to new, countervailing information. These results suggest that the pattern of rapid revision that we observed in Experiment1 was unique neither to the particular extreme behavior we chose nor to the undergraduate sample we utilized.

So far in our experiments, participants formed an evaluation on the basis of highly positive information and then learned a new piece of extremely negative information. We have not yet tested, however, whether a well-rehearsed negative evaluation would be sensitive to new pieces of highly positive information. Although it may be easy to undo one's good deeds with a single, particularly egregious act of wrongdoing, it may be that a consistent prickly and disagreeable disposition can be considerably less easy to shake with single, extreme acts of kindness.

There are strong theoretical reasons to suspect that a heavily entrenched negative evaluation should be relatively less sensitive to countervailing information than a similarly entrenched positive one. First, research shows that bad events generally exert a larger impact on individuals than equivalent positive events (e.g., Baumeister, Bratslavsky, Finkenauer, & Vohs, 2001; Rozin &

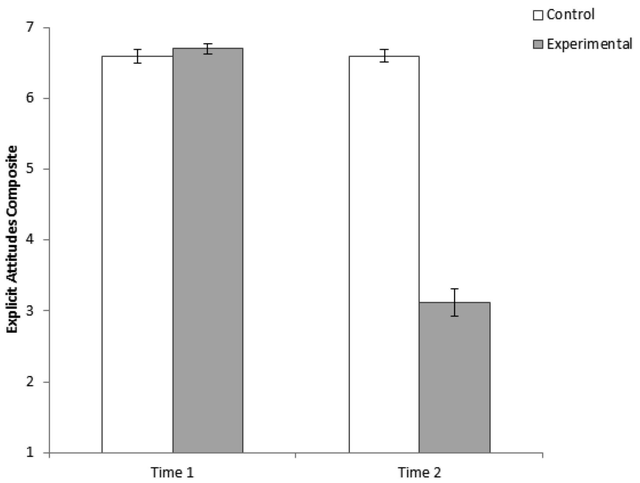

Figure 3. Participants' explicit evaluations of Bob before (Time 1) and after (Time 2) learning either that he mutilated a small animal (experimental) or that he bought a soda (control) in Experiment 1b. The error bars represent standard errors.

Royzman, 2001). As a result of this *negativity bias,* individuals exhibit stronger brain activation and behavioral reactions to negative events (Cacciopo, Gardner, & Berntson, 1997), have their attention drawn more quickly to negative versus positive words (Pratto & John, 1991), and learn more quickly from punishment versus reward (Vaish, Grossmann, & Woodward, 2008). In their classic article on the negativity bias, Rozin and Royzman (2001) argued that a beverage can be utterly tainted by brief contact with a single cockroach but that no amount of goodness can be added to a plate of cockroaches to make it even slightly more palatable. Perhaps so, too, a single instance of particularly egregious behavior can tarnish one's otherwise positive impression of Bob. However, many more good deeds may be necessary to change one's (implicit) mind about someone we have learned over time is distasteful.

Moreover, research on person perception has shown a valence asymmetry in response to learning about personality behavior. Whereas participants tend to view a single instance of morally bankrupt behavior as highly indicative of one's true character and to thus draw broad conclusions about one's likely future interactions with that individual, they are less likely to forgive past transgressions on the basis of a single instance of particularly altruistic behavior. This can occur, in part, because such behaviors are more likely to be discounted as deviations from an otherwise negative disposition, rather than as especially diagnostic of one's true nature (e.g., Malle & Knobe, 1997; Mende-Siedlecki, Baron, & Todorov, 2013; Reeder & Coovert, 1986; Reeder et al., 1992; Trafimow & Schneider, 1994; Trafimow & Trafimow, 1999). This work suggests that when an initially seemingly likeable Bob molests a child, he is not merely a good person who did a bad thing; rather, he is now (irreversibly) a bad person. However, when a consistently disagreeable and unsavory Bob decides to donate a kidney to a stranger, he does not simply become a good person; he is still ultimately a bad person who did a good thing. Such negative impressions may thus be especially difficult to shake precisely because good behaviors can be seen as less diagnostic of one's true nature.

To test for the possibility of this kind of valence asymmetry in implicit evaluative revision, in the next study, we included both positive and negative evaluation inductions at Time 1. At Time 2, we provided participants with either a neutral piece of information (i.e., "Bob bought a soda") or an evaluatively inconsistent piece of information—that is, a highly negative piece of information in one case or a highly positive piece of information in the other. The two primary goals of the study were to discern whether (a) an established negative implicit evaluation can be overturned by a single, extreme positive behavior and (b) the amount of change in the positive and negative directions reveals a valence asymmetry.

Experiment 2: The Nonequivalence of Molestation and Kidney Donation

Method

Participants. Three hundred and thirty Mechanical Turk workers completed the experiment online. Twenty-three participants submitted a survey code but failed to complete all of the components of the study and were thus excluded from analyses. The final sample was thus 307 participants.

Pretest. To select a behavior that was of the same extremity but opposite in valence from a child molestation conviction, we conducted a short pretest ($N = 30$) in which Mechanical Turk workers assessed the diagnosticity—on a scale from 1 (*not at all*) to 7 (*extremely*)—and valence—on a scale from -3 (*extremely negative*) to 3 (*extremely positive*)—of a collection of 10 different positive (e.g., "Bob helped a stranger on the road with a flat tire") and negative (e.g., "Bob lied to his boss about how much he contributed to a team project") behaviors.

On the basis of this pretest, we selected, for the positive counterinduction, the behavior "Bob recently donated one of his kidneys to a child in need he had never met before," which was equivalent but opposite in valence to child molestation ($|M_{\text{kidney}}| = 2.70$, $SD_{\text{kidney}} = 1.21$; $|M_{\text{molester}}| = 2.90$, $SD_{\text{molester}} = .548$; $t < 1$), as well as diagnosticity ($M_{\text{kidney}} = 6.47$, $SD_{\text{kidney}} = 1.25$; $M_{\text{molester}} = 6.53$, $SD_{\text{molester}} = 1.25$; $t < 1$).

Procedure. The study employed a 2 (time: 1 or 2) × 2 (Time 1 induction: positive or negative) × 2 (Time 2 information: control or counterattitudinal) × 2 (prime: Bob or neutral) mixed design. To manipulate the valence of participants' initial implicit evaluations toward Bob, we simply reversed the characteristic/uncharacteristic feedback for positive and negative induction participants—that is, all participants encountered the same 100 behaviors; however, participants in the negative induction received feedback that 100% of the positive statements were uncharacteristic of Bob and 100% of the negative statements were characteristic of him, while those in the positive induction received the opposite feedback (i.e., a learning induction procedure identical to Experiments 1a and 1b).

Following the induction, participants completed Time 1 implicit and explicit evaluation measures. At Time 2, participants either received counterattitudinal information (i.e., kidney donation following a negative induction or child molestation following a positive induction) or neutral information (i.e., "Bob bought a soda"). Across all conditions, this information was provided using the same protocol as the previous experiments. Next, participants completed the Time 2 implicit and explicit evaluation measures, followed by questions assessing participants' perceptions of the valence of the statement they read at Time 2 on a -3 (*extremely negative*) to 3 (*extremely positive*) scale and several demographics questions.

Results

We submitted participants' proportion-pleasant judgments to a 2 (Time: 1 or 2) × 2 (Time 1 induction: positive or negative) × 2 (Time 2 information: control or counterattitudinal) × 2 (prime: Bob or neutral) mixed ANOVA, which revealed the predicted four-way interaction, $F(1, 303) = 43.946$, $p < .001$, $\eta_p^2 = .127$ (see Figure 4).[3]

Breaking down this interaction, we explored the effects of time separately for the conditions in which neutral information was provided at Time 2 and the conditions in which counterattitudinal information was provided. In the control conditions consisting of neutral information at Time 2, there was a Time 1 Induction ×

[3] There was also, in this study, an unpredicted four-way interaction between time, prime type, reversal, and which image represented Bob, $F(5, 283) = 3.150$, $p < .01$.

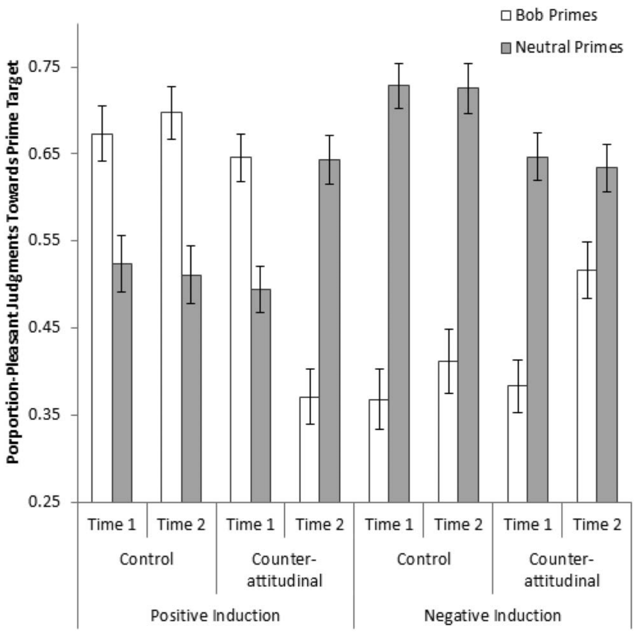

Figure 4. Participants' implicit evaluations of Bob before (Time 1) and after (Time 2) learning neutral (control) or counterattitudinal information in Experiment 2. The error bars represent standard errors.

Prime interaction, $F(1, 153) = 65.819$, $p < .001$, that was unqualified by time ($F < 1$)—that is, participants' implicit responses toward Bob were significantly affected by the type of initial learning they received, but there was no evidence of any changes over time after learning neutral information at Time 2. When participants underwent a positive induction, there was a main effect of prime, $F(1, 77) = 14.919$, $p < .001$, such that participants exhibited significantly more implicit positivity for Bob (Time 1: $M = .67$, $SD = .28$; Time 2: $M = .70$, $SD = .26$) than unfamiliar targets (Time 1: $M = .52$, $SD = .28$; Time 2: $M = .51$, $SD = .29$) at both time points. When they underwent a negative induction, there was also a main effect of prime, $F(1, 76) = 57.146$, $p < .001$. As expected, participants exhibited significantly *less* positivity for Bob (Time 1: $M = .37$, $SD = .30$; Time 2: $M = .41$, $SD = .32$) than unfamiliar targets (Time 1: $M = .73$, $SD = .23$; Time 2: $M = .73$, $SD = .26$) at both time points.

In contrast, in the experimental conditions, the Time 1 Induction × Prime interaction was significant, $F(1, 150) = 8.516$, $p < .01$, but was qualified by the predicted three-way interaction with time, $F(1, 150) = 54.276$, $p < .001$. When participants underwent a positive induction, there was a significant interaction between time and prime, $F(1, 75) = 38.397$, $p < .001$, such that participants exhibited significantly greater positivity toward Bob relative to unfamiliar targets at Time 1 ($M_{\text{Bob}} = .65$, $SD = .24$; $M_{\text{neutral}} = .49$, $SD = .23$), $F(1, 75) = 15.564$, $p < .001$, but exhibited the reverse pattern at Time 2 ($M_{\text{Bob}} = .37$, $SD = .28$; $M_{\text{neutral}} = .49$, $SD = .23$), $F(1, 75) = 35.079$, $p < .001$. Thus, participants in the positive induction fully reversed their implicit responses toward Bob after encountering diagnostic and extreme negative information about him at Time 2. This pattern replicates the results of the first two studies. Mirroring the pattern of results in Experiment 1b, there was also evidence that this reversal at Time 2 was driven both by the predicted decreases in implicit responses toward Bob, $F(1, 75) = 34.102$, $p < .001$, and by unpredicted *increases* in responses toward neutral targets, $F(1, 75) = 16.628$, $p < .001$.

When participants underwent a negative induction, there was also a significant interaction between time and prime, $F(1, 75) = 16.353$, $p < .001$. However, unlike the positive induction, there was no evidence of a full reversal—that is, at Time 1, participants exhibited significantly less positivity toward Bob ($M = .38$, $SD = .27$) than unfamiliar targets ($M = .65$, $SD = .24$), $F(1, 75) = 41.458$, $p < .001$, and although this difference was attenuated at Time 2, there was still evidence that participants evaluated Bob ($M = .52$, $SD = .28$) significantly less favorably than controls ($M = .63$, $SD = .24$), $F(1, 75) = 8.078$, $p < .01$. Thus, we did find evidence of moderate revision when the initial learning was negative, but no evidence of a reversal.

Explicit evaluations. Participants' responses were once again highly intercorrelated (Time 1: $\alpha = .996$, Time 2: $\alpha = .986$) and were thus combined into a single composite explicit evaluation measure at each time point. A 2 (time: 1 or 2) × 2 (Time 1 induction: positive or negative) × 2 (Time 2 information: control or counterattitudinal) mixed ANOVA on this measure revealed the predicted three-way interaction, $F(1, 303) = 503.559$, $p < .001$, $\eta_p^2 = .624$ (see Figure 5).

Decomposing this interaction, in the positive induction, participants who received counterattitudinal information at Time 2 initially exhibited high levels of positivity toward Bob at Time 1 ($M = 6.56$, $SD = .94$) but very strongly revised their initial impressions in light of the child molestation conviction ($M = 2.36$, $SD = 1.51$), $F(1, 75) = 461.627$, $p < .001$. When participants received neutral Time 2 information, they exhibited strong positivity toward Bob at both Time 1 ($M = 6.61$, $SD = .76$) and Time 2 ($M = 6.62$, $SD = .71$; $F < 1$, *ns*). Taken together, these results fully replicate the findings of the previous studies.

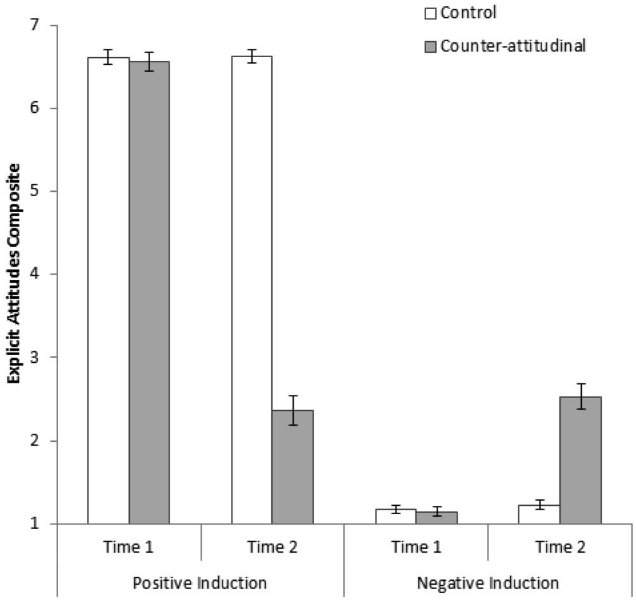

Figure 5. Participants' explicit evaluations toward Bob before (Time 1) and after (Time 2) learning neutral (control) or counterattitudinal information in Experiment 2. The error bars represent standard errors.

In the negative induction, participants who received counterattitudinal information at Time 2 initially exhibited high levels of negativity toward Bob ($M = 1.14$, $SD = .49$). At Time 2, they did indeed revise their initial impressions, developing increased positivity toward him in light of his altruistic behavior ($M = 2.36$, $SD = 1.51$), $F(1, 75) = 89.405$, $p < .05$, although not to the extent that they did in response to the counterattitudinal information in the positive induction—that is, just like their implicit evaluations, participants exhibited evidence of a valence asymmetry on their explicit evaluations. Participants who received neutral information at Time 2 remained unchanged in their negative opinions of Bob (Time 1: $M = 1.17$, $SD = .49$; Time 2: $M = 1.22$, $SD = .53$), $F(1, 76) = 1.646$, *ns*.

Perceptions of Time 2 behavior. Although we had chosen the counterattitudinal behaviors in this study on the basis of their equivalence in a pretest, we could examine whether participants saw the behaviors through a different lens given what they learned about Bob during the induction. Whereas the two behaviors were similar in absolute valence in the pretest, significant differences in absolute valence for the positive versus negative behavior did emerge during the full study such that the molester behavior was more extreme than the kidney donation behavior ($|M_{\text{kidney}}| = 2.63$, $SD_{\text{kidney}} = .73$; $|M_{\text{molester}}| = 2.93$, $SD_{\text{molester}} = .30$), $t(150) = 3.356$, $p < .001$.

There are two possibilities for these differences. The first is that there are significant differences in participants' perceptions of the two behaviors but that we lacked sufficient power to detect these differences in the pretest. The second, however, is that participants in the pretest and in the full study viewed these behaviors through different lenses. Indeed, our analysis posits that a particularly heinous or altruistic act is never evaluated in isolation but rather viewed in light of the previous impressions that one has already formed of the individual engaging in the behavior. Critically, we may be less inclined to see the donation of a kidney to a stranger, for instance, in highly positive terms if we already know of a great deal of Bob's previous transgressions. This would suggest that a prior, well-learned negative evaluation might cast a suspicious light on even highly positive behavior in a way that would not occur for a well-learned positive evaluation when one encounters a highly negative behavior (e.g., Reeder & Coovert, 1986; Trafimow & Trafimow, 1999). Indeed, this is the essence of the asymmetry between morally positive and negative reversals.

Discussion

The primary contributions of Experiment 2 are twofold. First, we replicated the finding that a single piece of extremely negative behavior can overturn previously positive implicit evaluations. Second, in line with research and theory on negativity dominance (Baumeister et al., 2001; Rozin & Royzman, 2001) and trait asymmetry (e.g., Reeder & Coovert, 1986; Reeder et al., 1992; Trafimow & Trafimow, 1999), we found evidence for a valence asymmetry in participants' receptiveness to new, countervailing information. Whereas a single egregious act gave rise to a full reversal of participants' positive implicit responses, a single particularly altruistic act was not enough to fully undo many instances of moderately negative behaviors.

It is important to note that participants *did* show evidence of being responsive to information that Bob had donated one of his kidneys to a stranger. That is, implicit responses toward Bob did *change*—a finding that is noteworthy given previous research suggesting that a single piece of information cannot easily change prior evaluations (e.g., Rydell et al., 2007). However, he was nonetheless still implicitly evaluated negatively relative to the control faces; rather than becoming *positive* in participants' eyes, he had instead merely become *less negative*.

One limitation of this study is that we established the equivalence of the extremity of kidney donation and child molestation using a self-report questionnaire. However, it is possible that people respond to these two behaviors differently at the implicit level. What we have shown here, then, is provisional evidence of a valence asymmetry (in line with other research in person perception). However, further research may reveal that the effects of positive versus negative behaviors on revision can sometimes be more symmetrical than what we observed here.

We now have evidence over multiple studies that learning about a single, counterattitudinal, extreme behavior results in significant implicit evaluative revision, particularly when this behavior is negative. We have argued that the reason for such sensitivity to a single piece of extreme information is because extreme behaviors of this sort have important implications for an individual's true nature or character—that is, they communicate important, diagnostic information about a target's likely future behavior, thus leading participants to imbue them with greater significance (e.g., Fiske, 1980). However, thus far, our evidence for diagnosticity has only been indirect. In the next two studies, we directly test our account. In Experiment 3, we test the hypothesis that the extremely negative behavior has to be attributed to the actor (Bob) rather than merely associated with him. In Experiment 4, we collect direct evidence that the perceived extreme negativity of the behavior informs the perceived diagnosticity of the behavior, which in turn predicts revision.

Experiment 3: The Company Bob Keeps

On the view of some associative accounts of implicit evaluation formation and change, implicit evaluation can often be susceptible to a kind of guilt by mere association (e.g., Walther, 2002). Because the associative processes that are thought (by some) to govern implicit evaluation formation and change operate primarily on the basis of similarity and temporal contiguity, situations in which either (a) a liked or disliked individual is paired in close temporal proximity with a target individual or (b) a liked or disliked individual shares superficial features of similarity with a target individual can potentially give rise to a *spreading attitude effect* in which one's evaluation toward one individual can be associatively transferred to the other (see also Skowronski, Carlston, Mae, & Crawford, 1998). This perspective would predict that an extreme negative behavior that is merely associated with Bob may be enough to lead to implicit evaluative revision.

However, if the perceived diagnosticity of the behavior is necessary for revision, then participants should show that they are able to discriminate between a very bad behavior that Bob performs versus a very bad behavior that someone superficially associated with Bob performs. We tested this question about guilt by mere association by adding a condition in which we exposed participants to the same negative behavior as in previous studies (i.e., child molesting) but implicated someone with whom Bob was merely incidentally associated—a coach at a high school in Bob's home town. If guilt by mere association can give rise to a spreading attitude effect, there ought to be strong revision in this condition, even though Bob was not responsible for the behavior. However, if participants consider the diagnosticity of the new information, they should be inclined to discount this piece of information in their implicit and explicit evaluations of him.

Method

Participants. Two-hundred and thirty mturk workers completed the experiment online. Fifteen workers submitted a survey code but failed to complete all components of the study and were thus excluded from analyses. The responses from an additional 14 participants were excluded from analyses due to either a lack of variance in their responses or excessive pressing of incorrect response keys on one or both of the implicit measures. This left a final sample of 201 participants.

Procedure. The procedure in Experiment 3 was essentially identical to the previous studies, but with the addition of a third condition in which the new piece of information provided at Time 2 implicated someone incidentally associated with Bob as a child molester rather than Bob himself. The Time 2 behavioral statement in this condition read, "The football coach at a high school in Bob's home town was recently convicted of child molestation." The statements in the Bob-molester condition and in the control condition were identical to Experiment 1a. The study was thus a 2 (time: 1 or 2) × 3 (Time 2 behavior: Bob molester, high school coach molester, or soda) × 2 (prime: Bob or neutral) mixed design.

Results

A 2 (Time: 1 or 2) × 3 (Time 2 information: Bob molester, coach molester, or soda) × 2 (prime: Bob or neutral) mixed ANOVA on participants' proportion-pleasant judgments once again yielded the predicted three-way interaction, $F(2, 198) = 13.998$, $p < .001$, $\eta_p^2 = .124$ (see Figure 6).[4]

Decomposing this interaction, in the Bob-molester condition, replicating the previous results, there was a significant interaction between time and prime, $F(1, 65) = 25.262$, $p < .001$, such that participants exhibited significantly greater implicit positivity toward Bob ($M = .60$, $SD = .26$) than the unfamiliar faces ($M = .50$, $SD = .22$) at Time 1, $F(1, 65) = 4.267$, $p < .05$. At Time 2, however, participants now exhibited significantly more negativity toward Bob ($M = .36$, $SD = .29$) than toward the neutral targets ($M = .62$, $SD = .25$), $F(1, 65) = 21.529$, $p < .001$. Like the previous studies, this interaction effect was driven, in part, by large decreases in implicit positivity toward Bob from Time 1 to Time 2, $F(1, 65) = 24.661$, $p < .001$, but also partly by an increase in implicit positivity toward *unfamiliar targets* from Time 1 to Time 2, $F(1, 65) = 11.493$, $p < .001$.

In the control condition, there was a main effect of prime, $F(1, 66) = 24.427$, $p < .001$, that was unqualified by an interaction with time ($F < 1$, *ns*). Participants exhibited greater implicit positivity toward Bob at both Time 1 (Bob: $M = .69$, $SD = .23$; Neutral: $M = .48$, $SD = .23$), $F(1, 66) = 24.315$, $p < .001$, and Time 2 (Bob: $M = .64$, $SD = .24$; Neutral: $M = .46$, $SD = .23$), $F(1, 66) = 15.870$, $p < .001$.

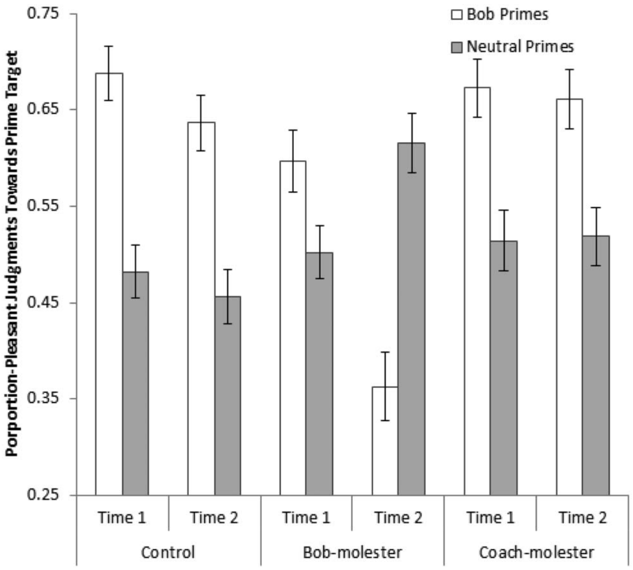

Figure 6. Participants' implicit evaluations of Bob before (Time 1) and after (Time 2) learning information about his behavior or the behavior of someone incidentally associated with him in Experiment 3. The error bars represent standard errors.

More important for the current study, the pattern of results in the coach-molester condition closely mimicked those in the control condition and not those in the Bob-molester condition. There was a predicted main effect of prime, $F(1, 67) = 13.451$, $p < .001$, that was unqualified by an interaction with time ($F < 1$, *ns*). Participants exhibited greater implicit positivity toward Bob at both Time 1 (Bob: $M = .67$, $SD = .24$; Neutral: $M = .51$, $SD = .26$), $F(1, 67) = 12.652$, $p < .001$, and Time 2 (Bob: $M = .66$, $SD = .26$; Neutral: $M = .52$, $SD = .25$), $F(1, 67) = 9.734$, $p < .01$.

Moreover, when testing the two-way simple effect between the coach-molester and control conditions, only a main effect of target emerged, $F(1, 133) = 36.741$, $p < .001$, that was unqualified by condition, $F(1, 133) = .036$, *ns*.

Explicit evaluations. On the composite measure of explicit evaluations (Time 1: $\alpha = .926$, Time 2: $\alpha = .983$), there was a predicted interaction between time and condition, $F(2, 198) = 168.737$, $p < .001$, $\eta_p^2 = .630$ (see Figure 7). Participants in the Bob-molester condition initially expressed strong positivity toward Bob at Time 1 ($M = 6.65$, $SD = .55$) but exhibited a reversal of their previous evaluations at Time 2 ($M = 2.65$, $SD = 1.67$), $F(1, 65) = 362.710$, $p < .001$. Participants in the control condition exhibited high levels of positivity toward Bob at both Time 1 ($M = 6.69$, $SD = .63$) and Time 2 ($M = 6.69$, $SD = .63$; $F < 1$, *ns*). Taken together, these results replicate the findings of the previous studies.

More interesting, however, were the effects of time on participants' explicit evaluations in the coach-molester condition. Here, participants were equivalently positive (relative to the other two conditions) toward Bob at Time 1 ($M = 6.60$, $SD = .85$) but

[4] There was also an unpredicted three-way interaction between time, prime type, and which image represented Bob, $F(5, 183) = 2.600$, $p < .05$.

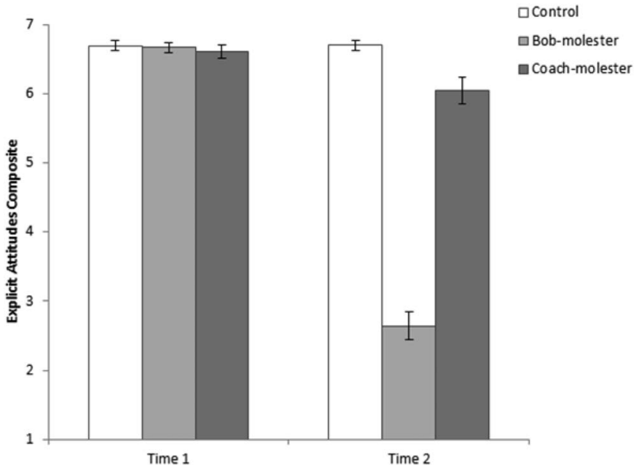

Figure 7. Participants' explicit evaluations of Bob before (Time 1) and after (Time 2) learning information about his behavior or the behavior of someone incidentally associated with him in Experiment 3. The error bars represent standard errors.

nonetheless exhibited a(n unpredicted) decrease in their positivity toward Bob at Time 2 ($M = 5.98$, $SD = 1.75$), $F(1, 67) = 10.093$, $p < .01$. There were also significant differences in participants' explicit evaluations at Time 2 among the three conditions, $F(2, 198) = 26.793$, $p < .001$. As predicted, post hoc comparisons on Time 2 explicit evaluations (Tukey's honestly significant difference) showed that the Bob-molester condition significantly differed from the coach-molester and control conditions but that the coach-molester and control conditions did not significantly differ from each other.

Discussion

As our account predicted, participants were mostly unmoved (both implicitly and explicitly) by the revelation that someone in Bob's hometown was convicted of molesting children, despite whatever incidental association existed between them and despite the fact that our explicit instructions presumably led participants to conclude that the information was important and relevant. Whereas those who learned that Bob had molested children exhibited similar updating of implicit and explicit evaluations as observed in our previous studies, participants in the coach-molester condition exhibited a pattern that was essentially equivalent to those who learned that Bob had bought a soda. Thus, Experiment 3 shows that diagnosticity can predict both large amounts of change on the basis of small amounts of information—as evidenced in the Bob-molester condition—as well as no change whatsoever despite encountering highly negative information—as evidenced by the coach-molester condition.

Participants' explicit evaluations became more negative from Time 1 to Time 2 after they had learned about the coach. Thus, although participants' implicit evaluations were largely unaffected by such uninformative information, their explicit evaluations were. It may be that as participants were self-reporting their evaluations, they considered the reasons why we were presenting them with this information (e.g., Gricean norms; see Grice, 1975), simply responding according to demand. That is, although the information itself was not informative about Bob, participants may have considered the presentational and demand pressures surrounding that information while reporting their evaluations, while such information did not affect their implicit evaluations. We note, however, that participants' explicit evaluations at Time 2 in the coach condition were not significantly different from those in the control condition and were also significantly more positive than those in the Bob-molester condition, as predicted.

The results from this study support the role of diagnosticity in implicit evaluative change. However, in the next study, we sought evidence that people think that extremely negative actions are important because they are diagnostic about the person's likely future actions.

Experiment 4: The Mediating Role of Diagnosticity

In this experiment, we tested whether extremity predicts revision via perceived diagnosticity. We contend that extreme negative behavior signals diagnosticity because extreme negative acts are relatively rare and are thus likely to be interpreted as being due to the person being dispositionally bad (e.g., see Mende-Siedlecki, Baron, & Todorov, 2013). Research on person perception, for example, shows that whereas people do not think a target's single honest act necessarily says anything about the target's stable disposition, a single dishonest act indicates that the target is fundamentally dishonest even if that person was otherwise well behaved (see Reeder & Coovert, 1986). We argue, then, that to the extent that people find animal mutilation as extremely offensive, they will see that act as saying something stable and true about the target. Such perceived diagnosticity should then lead them to weigh that important information over any number of earlier learned positive behaviors.

After the initial learning about Bob, we measured participants' beliefs about whether they considered animal mutilation offensive and worthy of condemnation or punishment. We then measured the extent to which they viewed the action as diagnostic of who Bob truly is and whether it would predict his future behavior.

Method

Participants. Two-hundred and forty Mechanical Turk workers completed an online experiment on "learning about a new person."

Procedure. The procedure was identical to Experiment 1b in which participants learned at Time 2 that Bob had recently mutilated a small, defenseless animal. Following the Time 2 implicit and explicit evaluation measures, participants completed a short questionnaire assessing (a) the offensiveness of the Time 2 behavior (four items; $\alpha = .879$; e.g., "To what extent do you feel it is morally wrong to harm animals?", "How offensive do you consider Bob's later behavior to be?"; on 1 = *not at all* to 7 = *extremely* Likert-type scales) and (b) the perceived diagnosticity of the Time 2 behavior for Bob's future actions (eight items; $\alpha = .765$; e.g., "To what extent did the later information change your overall impression of Bob?", "To what extent do you feel that the later information you learned about Bob is a reflection of his true nature or character?"; also on 1 = *not at all* to 7 = *extremely* Likert-type scales).

After this short questionnaire, participants provided demographic information and received a survey completion code used to receive payment.

Results

We first ensured that participants exhibited rapid implicit evaluative change in light of Time 2 information. To assess this, we conducted a 2 (time: 1 or 2) × 2 (prime: Bob or neutral) within-subjects ANOVA, which revealed the predicted significant Time × Prime interaction, $F(1, 239) = 53.848$, $p < .001$, $\eta_p^2 = .184$ (see Figure 8). At Time 1, participants initially exhibited greater positivity toward Bob ($M = .68$, $SD = .26$) than toward unfamiliar strangers ($M = .51$, $SD = .26$), $F(1, 239) = 60.944$, $p < .001$. At Time 2, participants exhibited a full reversal of their evaluations at Time 2, expressing lower levels of positivity toward Bob ($M = .49$, $SD = .31$) than neutral targets ($M = .59$, $SD = .28$), $F(1, 239) = 12.439$, $p < .001$. Moreover, a significant effect of time emerged for both Bob primes, $F(1, 239) = 62.747$, $p < .001$, and neutral primes, $F(1, 239) = 14.722$, $p < .001$, replicating the previous results.

On the composite measure of explicit evaluations (Time 1: $\alpha = .919$, Time 2: $\alpha = .964$), there was a large effect of time, $F(1, 239) = 659.074$, $p < .001$, such that participants expressed strong positivity toward Bob at Time 1 ($M = 6.63$, $SD = .74$) but strongly revised their impressions at Time 2 ($M = 3.57$, $SD = 1.78$).[5]

We anticipated that participants' perceptions of the offensiveness of animal mutilation would predict implicit evaluative change and that this effect would be mediated by participants' perceived diagnosticity of animal mutilation as an indicator of Bob's true character and, thus, his likely future behavior. To assess this, we conducted a bootstrapping analysis using the procedures outlined by Preacher and Hayes (2008, 1,000 resamples) for estimating direct and indirect effects of a potential mediator. First, we created a measure of participants' implicit preference for Bob at each time point by constructing a difference score between their proportion-pleasant judgments for Bob trials and proportion-pleasant judgments for neutral trials. Participants' Time 2 implicit Bob preference served as the dependent measure, with Time 1 implicit Bob preference serving as a covariate in the analysis. Offensiveness ($M = 5.69$, $SD = 1.35$) was entered as the predictor variable, and diagnosticity ($M = 4.50$, $SD = 1.57$) was entered as the proposed mediator.

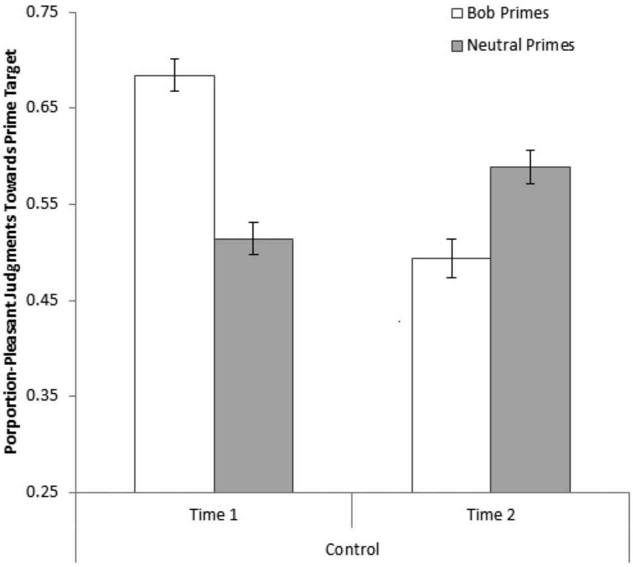

Figure 8. Participants' implicit evaluations of Bob before (Time 1) and after (Time 2) learning that he had mutilated a small animal in Experiment 4. The error bars represent standard errors.

This analysis indicated that the total effect of offensiveness on Time 2 implicit Bob preference (controlling for Time 1 Bob preference; total effect: −.07, $p < .001$; partial effect of Time 1 Bob preference on Time 2 Bob preference: −.10, *ns*) became nonsignificant when diagnosticity was included in the model (direct effect of offensiveness: −.03, *ns*). The total indirect effect of offensiveness on Time 2 implicit Bob preference was significant, with a point estimate of −.04 and a 95% confidence interval of −.07 and −.01. These findings are consistent with an account suggesting that the effect of offensiveness on Time 2 implicit Bob preference was meditated by perceived diagnosticity.

Correlations among measures. As predicted, offensiveness and diagnosticity were each strongly negatively correlated with Time 2 implicit and explicit evaluations—that is, the more participants saw Bob's Time 2 behavior as offensive and diagnostic of his character, the more negative their implicit and explicit evaluations at Time 2 (see Table 1 for correlations between the two composite measures of offensiveness and diagnosticity and the evaluation measures).

Discussion

Experiment 4 provides evidence that the offensiveness of an action is a key ingredient in rapid implicit evaluative change in response to single pieces of extreme information. Participants' perceptions of the offensiveness of Bob's Time 2 action were a strong predictor of their likelihood of exhibiting such rapid changes. More importantly, the findings were consistent with a meditational account for the role of perceived diagnosticity of that offensive behavior. The more participants saw Bob's animal mutilation as an especially immoral and offensive action, the more they saw that behavior as reflective of his true character and updated their implicit evaluations of him accordingly. Although, in this experiment, we did not manipulate diagnosticity and instead relied on correlational evidence, these findings, together with those from Experiment 3, strongly suggest that extreme negative behaviors lead to implicit revision through perceived diagnosticity.

Overall, the first five studies together provide strong and consistent evidence that a single, extreme behavior leads to rapid implicit evaluative change. One important outstanding question, however, is whether such newly revised implicit responses are predictive of participants' behavior. Rapid changes of the sort we

[5] To ensure that diagnosticity and offensiveness are indeed two different (though related) constructs, we conducted a factor analysis using oblimin rotation on the 12 individual items participants answered at the end of the experiment. This analysis produced a two-factor solution with eigenvalues of 6.524 and 2.017, with individual items loading onto these two factors as predicted. Thus, although the correlation between the offensiveness and diagnosticity composites was significant, $r(238) = .538$, $p < .001$, they were nonetheless empirically separable measures.

Table 1
Correlations Among Measures in Experiment 4

	Implicit Bob preference		Explicit composite	
Measure	Time 1	Time 2	Time 1	Time 2
Offensiveness	.00	−.22***	.13*	−.57***
Diagnosticity	.05	−.28***	−.04	−.57***

* $p < .05$. *** $p < .001$.

have observed in the preceding studies would be of little consequence if they did not ultimately guide our responses toward individuals. In our final experiment, we thus examined the consequences of single pieces of diagnostic, countervailing information with respect to participants' behavioral intentions. After undergoing a procedure similar to the previous studies, participants were asked about their behavioral intentions toward Bob. Critically, we expected that, if participants' evaluations had truly been updated, participants' Time 2 implicit evaluations of Bob (but not their Time 1 implicit evaluations) would predict their behavioral intentions.

Another important objective of this final experiment was to compare the predictive validity of revised implicit evaluations with that of revised explicit evaluations. Because both implicit and explicit evaluations have, across our studies, evidenced strong revision in response to a single, extreme behavior, one might wonder whether they are redundant with each other. In the next experiment, we examine this question by testing the predictive validity of participants' Time 1 and Time 2 implicit and explicit evaluations of Bob in subsequent intentions.

Experiment 5: The Effects of Revised Evaluations on Behavioral Intentions

Method

Participants. Three hundred Mechanical Turk workers completed an online experiment on "learning about a new person." Four participants submitted a survey code but failed to complete all of the components of the study and were thus excluded from analysis. Finally, in line with previous work (Payne et al., 2005), seven participants were excluded from analyses due to a lack of variance or one or both implicit measures. The final sample was 289 participants.

Procedure. The first part of the procedure in Experiment 5 was identical to the experimental condition in Experiment 1a. All participants underwent a positive induction and then completed Time 1 implicit and explicit evaluation measures. Next, all participants were told that Bob had recently been convicted of child molestation, and then they completed Time 2 implicit and explicit evaluation measures.

Following these tasks, participants answered a single-item measure of their behavioral intentions toward Bob, specifically: "How much would you try to organize your neighbors to try to prevent Bob from moving into your area?" on a Likert-type scale from 1 (*not at all*) to 9 (*a lot*).

Following this question, participants completed a short demographics questionnaire and were then given a survey completion code.

Results

First, to ensure that we had replicated our earlier findings, we conducted a 2 (time: 1 or 2) × 2 (prime: Bob or neutral) within-subjects ANOVA, which revealed the predicted significant Time × Prime interaction, $F(1, 289) = 114.707$, $p < .001$, $\eta_p^2 = .284$ (see Figure 9). At Time 1, participants initially exhibited greater positivity toward Bob ($M = .66$, $SD = .24$) than toward unfamiliar strangers ($M = .48$, $SD = .22$), $F(1, 289) = 85.938$, $p < .001$. However, at Time 2, participants exhibited a full reversal of their Time 1 preferences, now exhibiting less implicit positivity toward Bob ($M = .41$, $SD = .28$) than neutral targets ($M = .60$, $SD = .25$), $F(1, 289) = 56.850$, $p < .001$. As in the previous studies, there was evidence that this effect was driven both by shifts toward greater negativity toward Bob, $F(1, 289) = 122.388$, $p < .001$, as well as shifts toward greater positivity toward neutral targets, $F(1, 289) = 39.212$, $p < .001$.

On the composite measure of explicit evaluations (Time 1: $\alpha = .953$, Time 2: $\alpha = .933$), there was an effect of time, $F(1, 289) = 1390.608$, $p < .001$, such that participants expressed strong positivity toward Bob at Time 1 ($M = 6.67$, $SD = .73$) but strongly revised their impressions at Time 2 ($M = 3.35$, $SD = 1.36$).

To assess the effects of participants' newly revised implicit evaluations on their behavioral intentions, we created a difference score between participants' proportion-pleasant judgments toward Bob and toward neutral targets on the AMP at Time 1 and 2. This measure thus reflects participants' relative implicit preference at each time point. We then conducted a linear regression that included each of these relative implicit preference measures for Bob as well as participants' explicit evaluation composites at each time point as predictors of the extent to which they would try to organize their neighbors to prevent Bob from moving into their

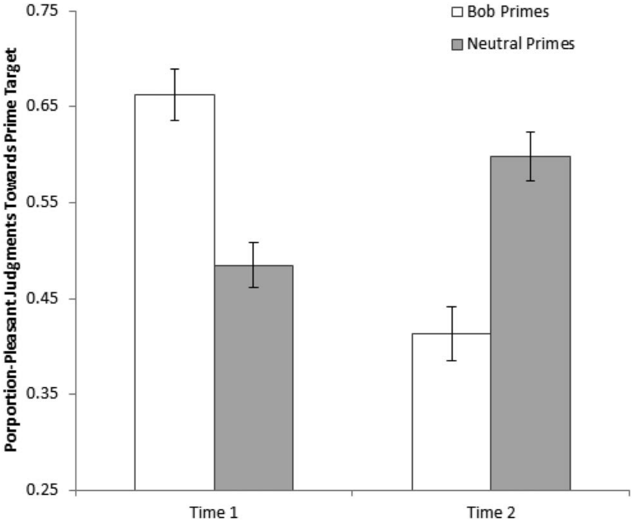

Figure 9. Participants' implicit evaluations of Bob before (Time 1) and after (Time 2) learning that he had mutilated a small animal in Experiment 5. The error bars represent standard errors.

area. In this analysis, participants' Time 2 relative implicit preference for Bob, $\beta = -.11$, $t(289) = -2.03$, $p < .05$, as well as participants' Time 2 explicit evaluations, $\beta = -.74$, $t(289) = -7.51$, $p < .001$, emerged as the only two significant predictors of participants' behavioral intentions, $\beta = -.16$, $t(100) = -2.040$, $p < .05$ (see Table 2).

Discussion

We found that participants' newly revised implicit responses predicted their behavioral intentions toward Bob such that the more they implicitly disliked Bob, the more they said they would try to organize their neighbors to prevent him from moving into their area. These assessments were uniquely predicted by their explicit and implicit Time 2 evaluations toward Bob.

We note, however, that the unique predictive validity of our implicit measure in predicting participants' behavioral intentions need not be a result of differences in the underlying *processes* that govern implicit and explicit evaluation. Payne, Burkley, and Stokes, (2008) argued that differences among implicit and explicit measures may be driven not by different processes but rather by differences in the *structure* of the measures used to assess these evaluations (e.g., kinds of stimuli, timing). It is noteworthy, however, that our measure of behavioral intentions shared considerably more structural properties with our explicit evaluation measure than with the implicit measure and yet implicit evaluations nonetheless emerged as a unique predictor.

General Discussion

The results of six experiments show that implicit impressions can be quickly reversed by a single piece of counterevidence (see also Mann et al., in press). After learning 100 unique positive pieces of behavioral evidence about a target, participants exhibited a positive implicit evaluation of that target. Yet, after learning just a single new piece of negative information (their 101st in the experiment), they immediately showed a complete reversal of their evaluations of the target.

We found evidence for this rapid revision in response to two extremely negative behaviors (child molestation: Experiments 1a, 2, and 3; animal mutilation: Experiments 1b, 4, and 5) and one extremely positive behavior (kidney donation to a stranger: Experiment 2). However, we also found support—as our analysis had predicted—for a valence asymmetry. Whereas participants readily incorporated negative diagnostic behaviors into their implicit evaluations of previously uniformly positive targets, they were less inclined to cast aside many initial instances of negative behavior in light of a single act of altruism (Experiment 2). Extreme negativity is especially effective as counterevidence because it is seen as especially diagnostic of a person's true character (e.g., someone who mutilates a small animal is not merely engaging in a bad behavior but rather is a bad *person*; Reeder & Coovert, 1986). In support of this, we showed in Study 4 that the more participants saw the negative act as offensive, the more they saw it as revealing something stable and predictive about the target, which in turn predicted the degree to which they updated their implicit impression of him.

Table 2
Regression Analysis in Experiment 5

Predictor	*B*	*SE B*	β
(Intercept)	(8.128***)	(1.191)	
Time 1 implicit Bob preference	.466	.384	.064
Time 2 implicit Bob preference	−.658*	.324	−.117*
Time 1 explicit evaluation	.128	.171	.039
Time 2 explicit evaluation	−.737***	.098	−.423***

Note. Implicit preference is a differences score between participants' proportion-pleasant judgments of Bob and their proportion-pleasant judgments of neutral targets.
* $p < .05$. *** $p < .001$.

We also identified situations in which extreme behaviors were *less* influential on revision. Though a strong associationistic view might suppose, for example, that implicit evaluations should promiscuously incorporate information about targets even when they are only incidentally associated with the behavior, we found no evidence, in the coach-molester condition in Experiment 3, to suggest that participants succumbed to mere guilt by association (cf. Walther, 2002).

Finally, we showed in Study 5 that revised implicit evaluations have unique predictive validity for behavioral intentions. Participants' revised implicit evaluations of Bob significantly predicted whether they would be inclined to mobilize with their neighbors to drive Bob out of the neighborhood should he decide to move in. Participants' implicit evaluation of Bob measured before the revision were not a predictor of this outcome, indicating that the revised implicit evaluation had effectively silenced the initial 100 behaviors participants had learned about Bob.

Reconciling These Findings With Current Theories

Although our results unequivocally show that implicit evaluations can be rapidly updated in light of minimal counterevidence, it is less clear what kind of underlying processing was influencing task performance on the implicit measures in our studies—that is, whether task performance was primarily influenced by associative processes, propositional processes, or some interaction between the two. As one of us has argued elsewhere (Ferguson et al., 2014; see also Moors, 2014), it is quite difficult, on the sole basis of behavioral evidence, to identify the representational format or algorithmic nature (Marr, 1982) of the underlying processes governing observed behaviors. Moreover, given past work by Sherman, Payne, and others (e.g., Bishara & Payne, 2009; Conrey et al., 2005; Payne, 2001; Sherman, 2006), task performance on the AMP (or any other implicit measure) is likely influenced by multiple processes—some of which may be propositional in nature—and the operating characteristics of such processes still remain unclear (e.g., Sherman, 2006).

Nevertheless, there is a precedent in the literature to classify the types of information we used in our learning paradigm—that is, verbal information requiring sentence comprehension—as propositional in nature (e.g., Gawronski & Bodenhausen, 2006, 2011, 2014; Rydell & McConnell, 2006; Rydell et al., 2006). If participants' processing of the behavioral statements did occur at least partially propositionally, then the rapid changes in task performance that we observed in our studies would appear to be inconsistent with those classes of dual process models that predict that task performance on implicit measures should be influenced only by (slow-learning) associative processes (Rydell & McConnell, 2006; Rydell et al., 2006, 2007; though see McConnell & Rydell, 2014, pp. 213–216).

Other dual process models, such as Gawronski and Bodenhausen's (2006, 2011) APE model, do allow for some kinds of interactions between associative and propositional processes in the formation and change of implicit responses. On this view, although it may be difficult to negate existing associations after they have been established, one might reflect new learning by creating new associations that affirm the opposite of a previously learned association (e.g., Gawronski, Deutsch, Mbirkou, Seibt, & Strack, 2008). If the counterevidence that we presented about Bob is an affirmation that he is a decidedly awful person—rather than a negation of the proposition that he is a relatively good person—then our findings may be consistent with such a model.

There are other theoretical approaches that allow for even stronger influences of propositional information on implicit evaluations, including those that do not assume that task performance on implicit measures like the AMP is solely under the purview of associative processes. For example, De Houwer's single-process model (De Houwer, 2014; Hughes et al., 2011) posits that both implicit and explicit responses are governed by a single propositional process. On this view, participants may have rapidly incorporated the new verbal information propositionally, thus allowing for rapid influences on their task performance on the implicit measure.

Of course, the information that we presented about Bob, though propositional in its format, was also highly affectively charged. It may be that learning that someone has recently been convicted of a serious crime has such strong affective connotations that it ultimately results in rapid changes in task performance through a relatively associative process. For example, Amodio and colleagues (Amodio, 2014; Amodio & Ratner, 2011; see also Amodio & Devine, 2006) argue that there are multiple memory systems involved in implicit processing that each possesses its own learning characteristics and trajectories. With this model in mind, perhaps participants formed their implicit responses toward Bob on the basis of a slower learning, implicit semantic system but then rapidly revised their implicit responses in light of new highly affective information about Bob on the basis of a faster learning, classical conditioning system.

Implicit Versus Explicit Evaluations

A number of current theories have suggested that the relative insensitivity of implicit evaluations to a new, single countervailing (and propositional) piece of evidence implies that implicit and explicit evaluations rely on different processes or systems (Gawronski & Bodenhausen, 2006, 2011, 2014; Rydell & McConnell, 2006; Rydell et al., 2006, 2007). After all, if implicit and explicit evaluations exhibit different learning trajectories, it could be due to different processes. The current results, then, might lead one to speculate that implicit and explicit evaluations share underlying processing more than current theories have assumed. For example, it could be that performance on both types of measures emerged entirely through associative processes, entirely through propositional processes, or through some interaction between associative and propositional processes.

However, just as it is difficult to identify (in algorithmic terms; Marr, 1982) the type of process(es) underlying implicit evaluations on the basis of patterns of change versus stability, it is similarly difficult to infer that a dissociation between implicit and explicit performance is due to different underlying processes or that equivalence in performance implies the same underlying processing. Without the use of corroborative computational modeling, which can explain behavioral evidence, it is hard to pinpoint the type of processing underlying task performance (see Ferguson et al., 2014). Moreover, there are numerous and important structural differences in implicit versus explicit evaluation measures (Payne et al., 2008), such as timing, type of response, type of stimuli, abstractness of stimuli, and so on, that could serve to inflate the dissociation between the two measures. Perhaps in some circumstances these structural differences are overwhelmed by some other variable that drives performance in similar ways across the measures. Research will continue to identify cases where implicit and explicit evaluations are similar versus dissimilar and inform an understanding of the processes at play for each type of measure.

Reconciling Our Findings With Previous Research

How do our findings speak to earlier work showing that implicit evaluations are, in the conception of Gregg and colleagues (2006), "easier done than undone"? One possibility is that the strategies that were used to induce revision in earlier work did not lead participants to view new information as highly diagnostic or, at least, to view it as no more diagnostic than the information they learned earlier in the experiment. For example, in Gregg and colleagues' Study 4, participants were provided with a narrative that described one fictitious group as decidedly negative and the other as highly positive. They were then given a narrative counterinduction in which the two groups were described as having switched characters over time. Gregg et al. found that this narrative counterinduction was not influential. From our perspective, one reason for this lack of revision may have been that the rationale provided for the role reversals in the counterinduction did not negate the groups' past behavior in participants' eyes; instead, they may have seen each group's earlier behaviors as better examples of the group's true natures than their later behaviors.

Similarly, in Rydell et al.'s (2007) research, in which participants' implicit evaluations changed only slowly in response to many instances of counterevidence, we would contend that no single behavior was sufficiently negative (or extreme) to trigger perceived diagnostic priority. When the initial behaviors and the new, countervailing behaviors are all moderate in their valence, there simply may not be a strong enough signal to trigger the kinds of diagnostic judgments that can alter people's implicit impressions as occurred in our studies

Were Participants' Time 1 Implicit Evaluations "Erased"?

An important consideration in the interpretation of our findings is whether the counterattitudinal information we provided at Time 2 erased (i.e., replaced) the information participants learned initially or merely created new associations that somehow overpowered initial learning. Research has shown that evaluatively inconsistent exposure to an object can often lead to contextualization—a process in which different aspects of a mental representation may be activated in response to different contextual cues in the environment (e.g., Gawronski et al., 2010; Rydell & Gawronski, 2009; see also Gawronski & Sritharan, 2010). If such a process were operating in our experiments, it would have important implications for the durability of the effects of the Time 2 information because it would suggest that the older implicit evaluations could perhaps resurface in certain contexts. Ultimately, we cannot distinguish, with our current data, between each of these two possibilities. However, there are a number of reasons why we suspect that contextualization is a less likely explanation for our findings. First, we made every effort to ensure that the circumstances under which participants learned Time 1 and Time 2 information were essentially identical and that no obvious contextual cues were available to allow participants to make sense of the evaluative inconsistencies to which we exposed them (cf. Gawronski et al., 2010; Rydell & Gawronski, 2009). Thus, even if contextualization processes could, in principle, have operated in the current studies, the necessary antecedents for such a mechanism may have been largely absent (but see AAB renewal effects in Gawronski et al., 2010).

Second, in contrast to the well-established and multifaceted objects that have been shown to give rise to contextual effects (in evaluation) in previous work, because participants were exposed to novel targets in our experiments, we know participants' entire evaluative history with Bob. What this evaluative history suggests is that even if Time 2 information gave rise to a contextualized, evaluatively inconsistent mental representation, the fact that a single new piece of propositional information was enough to rapidly give rise to this contextualization is a new finding that must be accommodated in this work.

Factors Affecting Assessments of Diagnosticity

Although we have documented one particular set of circumstances in which minimal counterevidence can result in implicit evaluative revision, there are likely to be a number of relevant moderators. First, it is unlikely that an assessment of the diagnosticity of a new piece of information occurs in complete isolation; a more relative assessment may occur in which individuals evaluate not just the diagnosticity of *current* information but also how later behavior compares to other behaviors that they have already learned (which may themselves be seen as diagnostic).

In practice, however, this relative judgment of diagnosticity may be less clear-cut than it was in our studies. Such assessments may require, for example, some explicit memory for particular instances of an individual's behavior that one can then compare and contrast with the current behavior—an assumption that may not necessarily be warranted (e.g., Dijksterhuis, 2004; Olson & Fazio, 2001, 2002; Rydell et al., 2006; see also Klein & Loftus, 1993; Tulving, 1993), particularly for well-established implicit responses that are the product of years of experience. Moreover, it is likely more often the case that participants come to the assessment that *both* earlier and later information is relevant to participants' understanding of a person's character, which could, in many cases, lead to feelings of implicit ambivalence (e.g., Petty, Tormala, Briñol, & Jarvis, 2006) rather than the full reversals we document in our studies.

Second, though we found that participants exhibited rapid reversals when they construed new information as highly diagnostic, we have not specified the conditions under which participants were likely to arrive at this assessment. We made the assumption, in line with previous work, that extreme, negative behaviors would be the most likely to trigger diagnostic assessments (Fiske, 1980; Knobe, 2003, 2006; Malle & Knobe, 1997; Mende-Siedlecki, Baron, & Todorov, 2013; Nadelhoffer, 2006; Reeder & Brewer, 1979; Reeder & Coovert, 1986; Reeder et al., 1992; Trafimow & Schneider, 1994; Trafimow & Trafimow, 1999). However, there are sure to be important factors that govern the extent to which participants see new behaviors as gleaning information about a person's true character or not.

Although we are suggesting that assessments of diagnosticity can influence the likelihood of implicit evaluation change, this is not to say that it must be an *unbiased* assessment. Consider, for example, a scenario in which you come home to discover your romantic partner in a compromising position with your best friend. Our analysis suggests that whether this single act induces change in your presumably heavily entrenched implicit (positive) evaluation of your partner will be driven by whether you deem this behavior as revealing of your partner's true nature. However, note that whether you ultimately come to this assessment or not will be influenced by various psychological mechanisms that serve to ameliorate the impact of negative events and threats to the self (e.g., Baumeister & Newman, 1994; Ditto & Lopez, 1992; Ditto, Scepansky, Munro, Apanovitch, & Lockhart, 1998; Dunning, 2012; Kunda, 1987, 1990; Lord, Ross, & Lepper, 1979)—mechanisms that may prevent you from seeing the situation objectively or clearly ("It was a one-time thing; it'll never happen again"; "I know that s/he actually really loves me") and thus alter your judgment of the diagnosticity of the behavior, with implications for the likelihood of rapid changes in implicit evaluations.

In short, one's own assessment of the diagnosticity of a behavior may be quite different from a more objective assessment provided by unbiased others, and there may be important barriers— barriers that were not present in the current studies—that prevent individuals from reasoning that a behavior is more important than past experience. What is clear from our results, however, is that *when* participants arrive at this assessment, they can readily update their implicit evaluations.

Limitations of the Current Work

We have presented evidence across multiple experiments that a single behavior can completely undo previously learned implicit evaluations. However, throughout these experiments, we measured implicit evaluations exclusively using the AMP (Payne et al., 2005). Although the most recent evidence suggests that the AMP measures unintentional evaluations of the prime stimuli (see Payne et al., 2013), it will be valuable to test in future work whether the same kind of rapid revision occurs in performance on other implicit measures, as we would expect. Additionally, we have used a specific learning format in these experiments where single behaviors are presented individually and sequentially. Whereas, in the current work, the relation among the behaviors—and especially the relationship between the old and new behavior—is never offered to participants, in future work, it will be critical to test whether different types of learning, perhaps via more ecological formats, would lead to a similar type of revision of implicit impressions.

Conclusion

According to some current evidence and theory, when we discover information that is highly inconsistent with our well-learned memories about someone, we face what would seem to be a rather troubling and highly maladaptive situation: Despite the implications that revelations of this sort may have for the validity of our past experiences, we may nonetheless harbor a lingering implicit evaluation of the individual that can continue to exert an influence on our behavior for some time.

Yet the experiments we report here suggest a more adaptive outcome: that we may be able—like the participants in these studies—to quickly incorporate minimal counterattitudinal information into our implicit evaluations. We may sometimes be able to cast earlier impressions aside and come to an arguably more adaptive and accurate implicit impression. We need only, on the basis of our analysis, see such revelations as especially diagnostic of who someone truly is. Far from leading us astray, implicit evaluations can uniquely shape and guide our behavioral choices in accord with what we learn to be true of the world.

References

Amodio, D. M. (2014). Dual experiences, multiple processes: Looking beyond dualities for mechanisms of the mind. In J. W. Sherman, B. Gawronski, & Y. Trope (Eds.), *Dual-process theories of the social mind* (pp. 560–577). New York, NY: Guilford Press.

Amodio, D. M., & Devine, P. G. (2006). Stereotyping and evaluation in implicit race bias: Evidence for independent constructs and unique effects on behavior. *Journal of Personality and Social Psychology, 91,* 652–661. http://dx.doi.org/10.1037/0022-3514.91.4.652

Amodio, D. M., & Ratner, K. G. (2011). A memory systems model of implicit social cognition. *Current Directions in Psychological Science, 20,* 143–148. http://dx.doi.org/10.1177/0963721411408562

Ashburn-Nardo, L., Voils, C. I., & Monteith, M. J. (2001). Implicit associations as the seeds of intergroup bias: How easily do they take root? *Journal of Personality and Social Psychology, 81,* 789–799.

Bar-Anan, Y., De Houwer, J., & Nosek, B. A. (2010). Evaluative conditioning and conscious knowledge of contingencies: A correlational investigation with large samples. *Quarterly Journal of Experimental Psychology, 63,* 2313–2335.

Bar-Anan, Y., & Nosek, B. A. (2012). Reporting intentional rating of the primes predicts priming effects in the affective misattribution procedure. *Personality and Social Psychology Bulletin, 38,* 1194–1208.

Bassili, J. N., & Brown, R. (2005). Implicit and explicit attitudes: Research, challenges and theory. In D. Albarracín, B. T. Johnson, & M. P. Zanna (Eds.), *Handbook of attitudes and attitude change* (pp. 543–574). Mahwah, NJ: Erlbaum.

Baumeister, R. F., Bratslavsky, E., Finkenauer, C., & Vohs, K. D. (2001). Bad is stronger than good. *Review of General Psychology, 5,* 323–370. http://dx.doi.org/10.1037/1089-2680.5.4.323

Baumeister, R. F., & Newman, L. S. (1994). Self-regulation of cognitive inference and decision processes. *Personality and Social Psychology Bulletin, 20,* 3–19. http://dx.doi.org/10.1177/0146167294201001

Bishara, A. J., & Payne, B. (2009). Multinomial process tree models of control and automaticity in weapon misidentification. *Journal of Experimental Social Psychology, 45,* 524–534. http://dx.doi.org/10.1016/j.jesp.2008.11.002

Boucher, K. L., & Rydell, R. J. (2012). Impact of negation salience and cognitive resources on negation during attitude formation. *Personality and Social Psychology Bulletin, 38,* 1329–1342. http://dx.doi.org/10.1177/0146167212450464

Buhrmester, M., Kwang, T., & Gosling, S. D. (2011). Amazon's Mechanical Turk: A new source of inexpensive, yet high-quality, data? *Perspectives on Psychological Science, 6,* 3–5. http://dx.doi.org/10.1177/1745691610393980

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,* 3–25. http://dx.doi.org/10.1207/s15327957pspr0101_2

Cahill, L., & McGaugh, J. L. (1990). Amygdaloid complex lesions differentially affect retention of tasks using appetitive and aversive reinforcement. *Behavioral Neuroscience, 104,* 532–543. http://dx.doi.org/10.1037/0735-7044.104.4.532

Cameron, C. D., Brown-Iannuzzi, J. L., & Payne, B. K. (2012). Sequential priming measures of implicit social cognition: A meta-analysis of associations with behavior and explicit attitudes. *Personality and Social Psychology Review, 16,* 330–350. http://dx.doi.org/10.1177/1088868312440047

Castelli, L., Zogmaister, C., Smith, E. R., & Arcuri, L. (2004). On the automatic evaluation of social exemplars. *Journal of Personality and Social Psychology, 86,* 373–387. http://dx.doi.org/10.1037/0022-3514.86.3.373

Conrey, F. R., Sherman, J. W., Gawronski, B., Hugenberg, K., & Groom, C. (2005). Separating multiple processes in implicit social cognition: The quad model of implicit task performance. *Journal of Personality and Social Psychology, 89,* 469–487. http://dx.doi.org/10.1037/0022-3514.89.4.469

Conrey, F. R., & Smith, E. R. (2007). Attitude representation: Attitudes as patterns in a distributed, connectionist representational system. *Social Cognition, 25,* 718–735. http://dx.doi.org/10.1521/soco.2007.25.5.718

Dasgupta, N., & Greenwald, A. G. (2001). On the malleability of automatic attitudes: Combating automatic prejudice with images of admired and disliked individuals. *Journal of Personality and Social Psychology, 81,* 800–814. http://dx.doi.org/10.1037/0022-3514.81.5.800

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,* 176–187. http://dx.doi.org/10.1016/j.lmot.2005.12.002

De Houwer, J. (2014). Why a propositional single-process model of associative learning deserves to be defended. In J. W. Sherman, B. Gawronski, & Y. Trope (Eds.), *Dual-process theories of the social mind* (pp. 542–559). New York, NY: Guilford Press.

Dijksterhuis, A. (2004). Think different: The merits of unconscious thought in preference development and decision making. *Journal of Personality and Social Psychology, 87,* 586–598. http://dx.doi.org/10.1037/0022-3514.87.5.586

Ditto, P. H., & Lopez, D. F. (1992). Motivated skepticism: Use of differential decision criteria for preferred and nonpreferred conclusions. *Journal of Personality and Social Psychology, 63,* 568–584. http://dx.doi.org/10.1037/0022-3514.63.4.568

Ditto, P. H., Munro, G. D., Apanovitch, A. M., Scepansky, J. A., & Lockhart, L. K. (2003). Spontaneous skepticism: The interplay of motivation and expectation in responses to favorable and unfavorable medical diagnoses. *Personality and Social Psychology Bulletin, 29,* 1120–1132. http://dx.doi.org/10.1177/0146167203254536

Ditto, P. H., Scepansky, J. A., Munro, G. D., Apanovitch, A. M., & Lockhart, L. K. (1998). Motivated sensitivity to preference-inconsistent information. *Journal of Personality and Social Psychology, 75,* 53–69.

Dovidio, J. F., Kawakami, K., & Gaertner, S. L. (2002). Implicit and explicit prejudice and interracial interaction. *Journal of Personality and Social Psychology, 82,* 62–68. http://dx.doi.org/10.1037/0022-3514.82.1.62

Dunning, D. (2012). *Self-insight: Roadblocks and detours on the path to knowing thyself.* New York, NY: Psychology Press.

Fazio, R. H. (2007). Attitudes as object-evaluation associations of varying strength. *Social Cognition, 25,* 603–637. http://dx.doi.org/10.1521/soco.2007.25.5.603

Ferguson, M. J. (2007). On the automatic evaluation of end-states. *Journal of Personality and Social Psychology, 92,* 596–611. http://dx.doi.org/10.1037/0022-3514.92.4.596

Ferguson, M. J., & Bargh, J. A. (2004). Liking is for doing: The effects of goal pursuit on automatic evaluation. *Journal of Personality and Social Psychology, 87,* 557–572.

Ferguson, M. J., & Fukukura, J. (2012). Likes and dislikes: A social cognitive perspective on attitudes. In S. Fiske & C. N. Macrae (Eds.), *The Sage handbook of social cognition* (pp. 165–190). Thousand Oaks, CA: Sage.

Ferguson, M. J., Mann, T. C., & Wojnowicz, M. T. (2014). Rethinking duality: Criticisms and ways forward. In J. W. Sherman, B. Gawronski, & Y. Trope (Eds.), *Dual-process theories of the social mind* (pp. 578–594). New York, NY: Guilford Press.

Fiske, S. T. (1980). Attention and weight in person perception: The impact of negative and extreme behavior. *Journal of Personality and Social Psychology 38,* 889–906. http://dx.doi.org/10.1037/0022-3514.38.6.889

Foerde, K., & Shohamy, D. (2011). Feedback timing modulates brain systems for learning in humans. *Journal of Neuroscience, 31,* 13157–13167. http://dx.doi.org/10.1523/JNEUROSCI.2701-11.2011

Galdi, S., Arcuri, L., & Gawronski, B. (2008, August 22). Automatic mental associations predict future choices of undecided decision-makers. *Science, 321,* 1100–1102. http://dx.doi.org/10.1126/science.1160769

Gawronski, B., & Bodenhausen, G. V. (2006). Associative and propositional processes in evaluation: An integrative review of implicit and explicit attitude change. *Psychological Bulletin, 132,* 692–731. http://dx.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, 44,* 59–127. http://dx.doi.org/10.1016/B978-0-12-385522-0.00002-0

Gawronski, B., & Bodenhausen, G. V. (2014). The associative–propositional evaluation model: Operating principles and operating conditions of evaluation. In J. W. Sherman, B. Gawronski, & Y. Trope (Eds.), *Dual-process theories of the social mind* (pp. 188–203). New York, NY: Guilford Press.

Gawronski, B., Deutsch, R., Mbirkou, S., Seibt, B., & Strack, F. (2008). When "just say no" is not enough: Affirmation versus negation training and the reduction of automatic stereotype activation. *Journal of Experimental Social Psychology, 44,* 370–377. http://dx.doi.org/10.1016/j.jesp.2006.12.004

Gawronski, B., Rydell, R. J., Vervliet, B., & De Houwer, J. (2010). Generalization versus contextualization in automatic evaluation. *Journal of Experimental Psychology: General, 139,* 683–701. http://dx.doi.org/10.1037/a0020315

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

Greenwald, A. G., McGhee, D. E., & Schwartz, J. L. (1998). Measuring individual differences in implicit cognition: The implicit association test. *Journal of Personality and Social Psychology, 74,* 1464–1480. http://dx.doi.org/10.1037/0022-3514.74.6.1464

Greenwald, A. G., Poehlman, T. A., Uhlmann, E. L., & Banaji, M. R. (2009). Understanding and using the implicit association test: III. Meta-analysis of predictive validity. *Journal of Personality and Social Psychology, 97,* 17–41. http://dx.doi.org/10.1037/a0015575

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

Grice, H. P. (1975). Logic and conversation. In P. Cole & J. L. Morgan (Eds.), *Syntax and semantics, 3: Speech acts* (pp. 41–58). New York, NY: Academic Press.

Hermer-Vazquez, L., Hermer-Vazquez, R., Rybinnik, I., Greebel, G., Keller, R., Xu, S., & Chapin, J. K. (2005). Rapid learning and flexible memory in "habit" tasks in rats trained with brain stimulation reward. *Physiology & Behavior, 84,* 753–759.

Hilliard, S., Nguyen, M., & Domjan, M. (1997). One-trial appetitive conditioning in the sexual behavior system. *Psychonomic Bulletin & Review, 4,* 237–241.

Hughes, S., Barnes-Holmes, D., & De Houwer, J. (2011). The dominance of associative theorising in implicit attitude research: Propositional and behavioral alternatives. *Psychological Record, 61,* 465–498.

Klein, S. B., & Loftus, J. (1993). The mental representation of trait and autobiographical knowledge about the self. In T. K. Srull & R. S. Wyer, Jr. (Eds.), *Advances in social cognition: Vol. 5. The mental representation of trait and autobiographical knowledge about the self* (pp. 1–49). Hillsdale, NJ: Erlbaum.

Knobe, J. (2003). Intentional action and side effects in ordinary language. *Analysis, 63,* 190–194. http://dx.doi.org/10.1093/analys/63.3.190

Knobe, J. (2006). The concept of intentional action: A case study in the uses of folk psychology. *Philosophical Studies, 130,* 203–231. http://dx.doi.org/10.1007/s11098-004-4510-0

Kunda, Z. (1987). Motivated inference: Self-serving generation and evaluation of causal theories. *Journal of Personality and Social Psychology, 53,* 636–647. http://dx.doi.org/10.1037/0022-3514.53.4.636

Kunda, Z. (1990). The case for motivated reasoning. *Psychological Bulletin, 108,* 480–498. http://dx.doi.org/10.1037/0033-2909.108.3.480

Lord, C. G., Ross, L., & Lepper, M. R. (1979). Biased assimilation and attitude polarization: The effects of prior theories on subsequently considered evidence. *Journal of Personality and Social Psychology, 37,* 2098–2109. http://dx.doi.org/10.1037/0022-3514.37.11.2098

Malle, B. F., & Knobe, J. (1997). The folk concept of intentionality. *Journal of Experimental Social Psychology, 33,* 101–121. http://dx.doi.org/10.1006/jesp.1996.1314

Mann, T., Cone, J., & Ferguson, M. J. (in press). Social-psychological evidence for the effective updating of implicit attitudes. *Behavioral and Brain Sciences*.

Marr, D. (1982). *Vision: A computational approach*. San Francisco, CA: Freeman.

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 Press.

McNulty, J. K., Olson, M. A., Meltzer, A. L., & Shaffer, M. J.(2013, November 29). Though they may be unaware, newlyweds implicitly know whether their marriage will be satisfying. *Science, 342,* 1119–1120. http://dx.doi.org/10.1126/science.1243140

Mende-Siedlecki, P., Baron, S. G., & Todorov, A. (2013). Diagnostic value underlies asymmetric updating of impressions in the morality and ability domains. *Journal of Neuroscience, 33,* 19406–19415.

Mitchell, C. J., De Houwer, J., & Lovibond, P. F. (2009). The propositional nature of human associative learning. *Behavioral and Brain Sciences, 32,* 183–198. http://dx.doi.org/10.1017/S0140525X09000855

Monet, G. (Producer), & Ginsberg, A. (Director). (2001). *The Iceman confesses: Secrets of a mafia hitman* [Motion picture]. United States: Home Box Office.

Moors, A. (2014). Examining the mapping problem in dual-process models. In J. W. Sherman, B. Gawronski, & Y. Trope (Eds.), *Dual-process theories of the social mind* (pp. 20–34). New York, NY: Guilford Press.

Nadelhoffer, T. (2006). Bad acts, blameworthy agents, and intentional actions: Some problems for jury impartiality. *Philosophical Explorations, 9,* 203–219. http://dx.doi.org/10.1080/13869790600641905

Olson, M. A., & Fazio, R. H. (2001). Implicit attitude formation through classical conditioning. *Psychological Science, 12,* 413–417. http://dx.doi.org/10.1111/1467-9280.00376

Olson, M. A., & Fazio, R. H. (2002). Implicit acquisition and manifestation of classically conditioned attitudes. *Social Cognition, 20,* 89–104. http://dx.doi.org/10.1521/soco.20.2.89.20992

Oswald, F. L., Mitchell, G., Blanton, H., Jaccard, J., & Tetlock, P. E. (2013). Predicting ethnic and racial discrimination: A meta-analysis of IAT criterion studies. *Journal of Personality and Social Psychology, 105,* 171–192. http://dx.doi.org/10.1037/a0032734

Otten, S., & Wentura, D. (1999). About the impact of automaticity in the minimal group paradigm: Evidence from affective priming tasks. *European Journal of Social Psychology, 29,* 1049–1071. http://dx.doi.org/10.1002/(SICI)1099-0992(199912)29:8<1049::AID-EJSP985>3.0.CO;2-Q

Payne, B. K. (2001). Prejudice and perception: The role of automatic and controlled processes in misperceiving a weapon. *Journal of Personality and Social Psychology, 81,* 181–192. http://dx.doi.org/10.1037/0022-3514.81.2.181

Payne, B. K., Brown-Iannuzzi, J., Burkley, M., Arbuckle, N. L., Cooley, E., Cameron, C. D., & Lundberg, K. B. (2013). Intention invention and the affect misattribution procedure: Reply to Bar-Anan and Nosek (2012). *Personality and Social Psychology Bulletin, 39,* 375–386. http://dx.doi.org/10.1177/0146167212475225

Payne, B. K., Burkley, M. A., & Stokes, M. B. (2008). Why do implicit and explicit attitude tests diverge? The role of structural fit. *Journal of Personality and Social Psychology, 94,* 16–31.

Payne, B. K., Cheng, C. M., Govorun, O., & Stewart, B. (2005). An inkblot for attitudes: Affect misattribution as implicit measurement. *Journal of Personality and Social Psychology, 89,* 277–293. http://dx.doi.org/10.1037/0022-3514.89.3.277

Perugini, M., Richetin, J., & Zogmaister, C. (2010). Prediction of behavior. In B. Gawronski & B. K. Payne (Eds.), *Handbook of implicit social cognition: Measurement, theory, and applications* (pp. 255–277). New York, NY: Guilford Press.

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,* 557–569. http://dx.doi.org/10.1177/0146167211400423

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,* 21–41. http://dx.doi.org/10.1037/0022-3514.90.1.21

Pratto, F., & John, O. P. (1991). Automatic vigilance: The attention-grabbing power of negative social information. *Journal of Personality and Social Psychology, 61,* 380–391. http://dx.doi.org/10.1037/0022-3514.61.3.380

Preacher, K. J., & Hayes, A. F. (2008). Asymptotic and resampling strategies for assessing and comparing indirect effects in multiple mediator models. *Behavior Research Methods, 40,* 879–891. http://dx.doi.org/10.3758/BRM.40.3.879

Reeder, G. D., & Brewer, M. B. (1979). A schematic model of dispositional attribution in interpersonal perception. *Psychological Review, 86,* 61–79. http://dx.doi.org/10.1037/0033-295X.86.1.61

Reeder, G. D., & Coovert, M. D. (1986). Revising an impression of morality. *Social Cognition, 4,* 1–17. http://dx.doi.org/10.1521/soco.1986.4.1.1

Reeder, G. D., Pryor, J. B., & Wojciszke, B. (1992). Trait–behavior relations in social information processing. In G. R. Semin & K. Fiedler (Eds.), *Language, interaction and social cognition* (pp. 37–57). Thousand Oaks, CA: Sage.

Rozin, P., & Royzman, E. B. (2001). Negativity bias, negativity dominance, and contagion. *Personality and Social Psychology Review, 5,* 296–320. http://dx.doi.org/10.1207/S15327957PSPR0504_2

Rydell, R. J., & Gawronski, B. (2009). I like you, I like you not: Understanding the formation of context-dependent automatic attitudes. *Cognition and Emotion, 23,* 1118–1152. http://dx.doi.org/10.1080/02699930802355255

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,* 995–1008. http://dx.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,* 954–958. http://dx.doi.org/10.1111/j.1467-9280.2006.01811.x

Rydell, R., 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,* 867–878. http://dx.doi.org/10.1002/ejsp.393

Sherman, J. W. (2006). Clearing up some misconceptions about the quad model. *Psychological Inquiry, 17,* 269–276. http://dx.doi.org/10.1207/s15327965pli1703_7

Sherman, J. W., Klauer, K. C., & Allen, T. J. (2010). Mathematical modeling of implicit social cognition: The machine in the ghost. In B. Gawronski & B. K. Payne (Eds.), *Handbook of implicit social cognition: Measurement, theory, and applications* (pp. 156–175). New York, NY: Guilford Press.

Simmons, J. P., Nelson, L. D., & Simonsohn, U. (2012). *A 21 word solution*. Retrieved from http://ssrn.com/abstract=2160588

Skowronski, J. J., Carlston, D. E., Mae, L., & Crawford, M. T. (1998). Spontaneous trait transference: Communicators take on the qualities they describe in others. *Journal of Personality and Social Psychology, 74,* 837–848. http://dx.doi.org/10.1037/0022-3514.74.4.837

Sloman, S. A. (1996). The empirical case for two systems of reasoning. *Psychological Bulletin, 119,* 3–22. http://dx.doi.org/10.1037/0033-2909.119.1.3

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

Strack, F., & Deutsch, R. (2004). Reflective and impulsive determinants of social behavior. *Personality and Social Psychology Review, 8,* 220–247. http://dx.doi.org/10.1207/s15327957pspr0803_1

Towles-Schwen, T., & Fazio, R. H. (2006). Automatically activated racial attitudes as predictors of the success of interracial roommate relationships. *Journal of Experimental Social Psychology, 42,* 698–705. http://dx.doi.org/10.1016/j.jesp.2005.11.003

Trafimow, D., & Schneider, D. J. (1994). The effects of behavioral, situational, and person information on different attribution judgments. *Journal of Experimental Social Psychology, 30,* 351–369. http://dx.doi.org/10.1006/jesp.1994.1017

Trafimow, D., & Trafimow, S. (1999). Mapping imperfect and perfect duties onto hierarchically and partially restrictive trait dimensions. *Personality and Social Psychology Bulletin, 25,* 687–697. http://dx.doi.org/10.1177/0146167299025006004

Tulving, E. (1993). What is episodic memory? *Current Directions in Psychological Science, 2,* 67–70. http://dx.doi.org/10.1111/1467-8721.ep10770899

Vaish, A., Grossmann, T., & Woodward, A. (2008). Not all emotions are created equal: The negativity bias in social-emotional development. *Psychological Bulletin, 134,* 383–403. http://dx.doi.org/10.1037/0033-2909.134.3.383

Walther, E. (2002). Guilty by mere association: Evaluative conditioning and the spreading attitude effect. *Journal of Personality and Social*

Psychology, 82, 919–934. http://dx.doi.org/10.1037/0022-3514.82.6.919

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

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

Wittenbrink, B., Judd, C. M., & Park, B. (2001). Spontaneous prejudice in context: Variability in automatically activated attitudes. *Journal of Personality and Social Psychology, 81,* 815–827. http://dx.doi.org/10.1037/0022-3514.81.5.815

Yin, H. H., & Knowlton, B. J. (2006). The role of the basal ganglia in habit formation. *Nature Reviews Neuroscience, 7,* 464–476.

Received March 21, 2014
Revision received August 26, 2014
Accepted August 29, 2014 ■