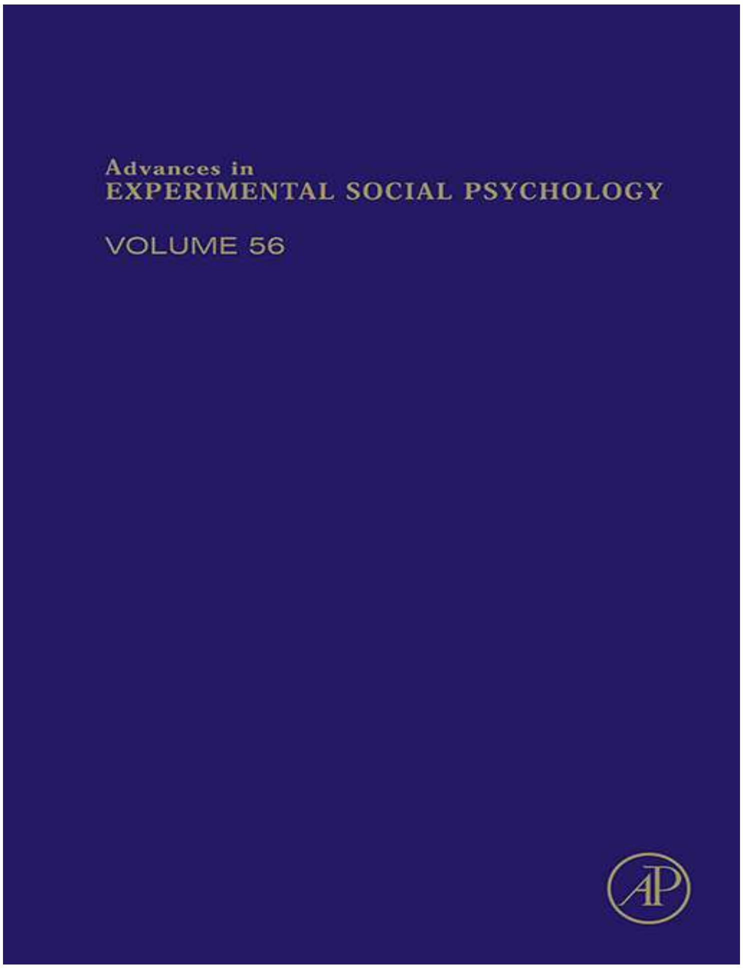

From Jeremy Cone, Thomas C. Mann and Melissa J. Ferguson, Changing Our Implicit Minds: How, When, and Why Implicit Evaluations Can Be Rapidly Revised. In: James M. Olson, editor, *Advances in Experimental Social Psychology, Vol. 56*, Burlington: Academic Press, 2017, pp. 131-199.
ISBN: 978-0-12-812120-7
© Copyright 2017 Elsevier Inc.
Academic Press

CHAPTER THREE

# Changing Our Implicit Minds: How, When, and Why Implicit Evaluations Can Be Rapidly Revised

**Jeremy Cone**[*,1], **Thomas C. Mann**[†], **Melissa J. Ferguson**[†]
[*]Williams College, Williamstown, MA, United States
[†]Cornell University, Ithaca, NY, United States
[1]Corresponding author: e-mail address: jdc2@williams.edu

## Contents

*Advances in Experimental Social Psychology*, Volume 56
ISSN 0065-2601
http://dx.doi.org/10.1016/bs.aesp.2017.03.001

## Abstract

How easily can our first impressions of others be updated when we learn new, contradictory evidence? We review recent work in the social cognition literature on the ways in which implicit evaluations can be updated in a durable and robust manner. These findings show that implicit evaluations of novel individuals can be revised when the new information is believable and diagnostic, and if it reinterprets the evaluative meaning of the original information. We discuss implications of this evidence for the degree to which evaluative memories can be updated, as well as new directions for theories of human evaluation and implicit cognition.

Here are the facts of the case: On July 27, 1996, three pipe bombs stuffed inside a backpack were detonated in Centennial Olympic Park in Atlanta, Georgia, killing one bystander and injuring over 100 others who had gathered to attend a concert as part of celebrations for the 1996 summer Olympic Games. A security guard at the park, Richard Jewell, purportedly discovered the bomb just minutes before it detonated, successfully alerted authorities, and cleared the area of many innocent people who otherwise would have been in the blast radius, thus saving their lives.

But just as quickly as he had been branded a hero by news anchors and media personalities across the country, the FBI investigation of the incident suddenly took a dark turn. What if Jewell had actually planted the bomb and had been attempting to save others at the last minute as an act of subterfuge? Perhaps the media's rush to characterize him as a hero had blinded them to his sinister, ulterior motives. Within days of the bombing, he was reported by anonymous sources to be a prime suspect. The media that had lauded Jewell's altruism just days earlier now began vilifying him, aggressively attacking his character, and digging into various details of his past that may have portended a descent into radicalization and terrorism. His home was searched by authorities. His friends and acquaintances were extensively questioned by FBI agents. He was under constant surveillance by law enforcement.

By the time the true culprit of the attack—Eric Rudolph, who was hoping to force officials to cancel the summer games for what he saw as an unacceptable tolerance of abortion in the United States—was apprehended 6 years later (after having successfully carried out three additional bombings), the damage to Jewell's reputation had been done.

As the twists and turns of Jewell's harrowing story exemplify, the task we face when developing impressions of others is fraught with error, uncertainty, and complexity. Not only must we come to an assessment about what someone is like, but also we must revise and update those impressions as new information comes to light and as new discoveries or interpretations about past insights are made. Although humans and other animals are thought to be cognitively geared to attempt to predict the future (e.g., Bar, 2011; Edelman, 2008; Gregory, 1980; Helmholtz, 1860), Jewell's case reminds us that we are imperfect at this task and are sometimes—even startlingly—surprised by the world around us. How often and to what degree are our expectations violated? How many of our moment-to-moment interactions in daily life contain revelations that we absorb as corrections to our prior beliefs? How many of us have experienced even life-altering revelations that categorically change our understanding of others? Though most of us can probably identify many such moments in our own lives, their true frequency remains unknown.

To be sure, we are learning and updating every second of our lives. Every time we walk into our offices and see the same furniture, the same view through the window, we are strengthening—and therefore updating—these memories (Hebb, 1949). Indeed, such updating is such an important and regular feature of our lives that it occurs even unconsciously, including during sleep (e.g., Hu et al., 2015). In this chapter, however, we are interested not in this kind of ubiquitous updating, but instead in situations in which we face some appreciable degree of inconsistency between current and past memory of other people. How effectively can we update our thoughts, feelings, judgments, and behaviors when we have experiences that are unambiguously inconsistent with our prior learning and expectations about others? How—and how easily—does the mind incorporate new social information from our environments?

We first discuss what we mean by "first impressions," followed by a brief review of findings in social psychology demonstrating both the importance of them, and the processes that can cause them to be rather difficult to change. Next, we consider the role of evaluative impressions, in particular, and review evidence pointing to dissociations in the learning characteristics of implicit and explicit evaluations. We then outline theoretical frameworks that attempt to explain these differential learning patterns, before turning to a discussion of our own approach and a variety of lines of evidence from our labs that suggest that implicit evaluations may be more capable of rapid revision than current theories posit.

# 1. WHAT COUNTS AS A FIRST IMPRESSION?

Some have defined first impressions as the "initial perception and formation of thoughts about another" (Rule & Ambady, 2008b, p. 35). Although it might seem easy to determine exactly what ought to count as an "initial perception," the concept of first impressions has been operationalized broadly in the person perception literature, conceived of as everything from "thin slices" of exposure—including even just a 100-ms exposure to a static image of a face (e.g., Rule & Ambady, 2008a; Tabak & Zayas, 2012; Willis & Todorov, 2006)—to reading about a single behavior while encountering an image of someone's face (Olivola & Todorov, 2010; Todorov & Uleman, 2002, 2004), to reading about a person's 100 different behaviors (e.g., Cone & Ferguson, 2015) or a person's 26 related behaviors (Mann & Ferguson, 2015). Thus, before we turn our attention to evaluative change specifically, it is worth taking a brief pause to consider both what is meant by "first" and what is meant by "impression."

## 1.1 What Is "First"?

It might seem as though the mind creates a series of discrete impressions that can easily be distinguished from one another. However, contemporary cognitive and social science suggests that the mind accrues information in an incremental, continuous, and dynamic fashion (e.g., Balcetis & Lassiter, 2010; Ehret, Monroe, & Read, 2015; Freeman & Ambady, 2011; Freeman & Johnson, 2016; Kunda & Thagard, 1996; Read & Miller, 1998; Schröder & Thagard, 2013; Schwarz, 2007; Song & Nakayama, 2009; Spivey, 2008; Wojnowicz, Ferguson, Dale, & Spivey, 2009). Thus, any impressions we develop of others are heavily influenced by both the time and context of the measurement, including, for example, by the orientation of someone's face (Todorov & Porter, 2014). Even in the absence of any new information about a target, a perceiver's impressions may fluctuate due to a variety of extraneous factors, such as a perceiver's mood (e.g., Forgas & Bower, 1987) or incidental exposure to other information in the environment (e.g., Anderson, 1971; Hamilton & Zanna, 1974; see also Schwarz & Sudman, 1992).

The continuous nature of person perception thus represents a conceptual challenge for researchers studying first impressions. Our own view is that initial perceptions dynamically emerge after some amount of relatively *minimal* and *uniform* evidence. Consider, for example, a woman who has developed a positive view of her new coworker, Jeff, over the past several weeks. When she learns that Jeff has been arrested for tax fraud, it is perhaps conceptually useful to consider her initial (uniformly positive) impression of Jeff as her "first impression" and to consider how this new information has modified that initial perception. Now consider instead a woman who learns that her husband of 30 years secretly has another family in an adjacent state. Would we consider the presumably positive impression developed over 30 years her "first" impression? Perhaps not, although this could be, in part, because we do not imagine these 30 years to be *actually* uniformly positive, but rather possibly more nuanced, multifaceted, or ambivalent (see Zayas & Shoda, 2015). Thus, we consider first impressions as those that are evaluatively uniform and that are formed toward a relatively novel target about whom we have only minimal information (though "minimal," of course, is a subjective term influenced by the goals of the researcher, which could encompass everything from very rapid exposures to more repeated exposure to many evaluatively consistent behaviors).

## 1.2 What Is an "Impression"?

Our impressions of others are not merely a disjointed collection of memories of instances of learning about them. Rather, we integrate these discrete behaviors and pieces of information into broader impressions that allow us to more effectively navigate our interactions with a target. Much of the current literature has focused on *attitudes* or *evaluations*, in particular—that is, evaluative summaries of the positivity or negativity associated with someone or something (for review of evaluations, see Albarracin, Johnson, & Zanna, 2005; Ferguson & Wojnowicz, 2011; Ferguson & Zayas, 2009; Maio & Haddock, 2015). Evaluations are a general currency by which we can successfully generate predictions about what others are likely to do, whether good or bad, and are thus broadly applicable across a wide variety of situations (though, of course, with such generality comes ample opportunities for errors, see Banaji & Greenwald, 2013).

A variety of different kinds of cues can serve as the basis for an evaluation, including faces (Alaei & Rule, 2016; Bonnefon, Hopfensitz, & De Neys, 2015; Gunaydın, Selcuk, & Zayas, 2017; Olivola, Funk, & Todorov, 2014; Rule & Alaei, 2016; Rule, Ambady, Adams, & Macrae, 2008; Todorov, Funk, & Olivola, 2015; Todorov, Olivola, Dotsch, & Mende-Siedlecki, 2015), visually apparent group membership (age, race, gender, etc.), nonverbal bodily behavior (Aviezer, Trope, & Todorov, 2012;

Tickle-Degnen & Rosenthal, 1990), verbal behavior (Ireland et al., 2011), speech cues (Berry, 1992; Schroeder & Epley, 2015, 2016), height (Harrison & Saeed, 1977; Lerner & Moore, 1974), weight (McConnell, Rydell, Strain, & Mackie, 2008), attractiveness (Lorenzo, Biesanz, & Human, 2010), clothing (Mills & Aronson, 1965), and, importantly, behavioral information (Fiske, 1980; Schneid, Carlston, & Skowronski, 2015; Skowronski & Carlston, 1989; Winter & Uleman, 1984), among many others. All of these sources of information come together to form what we might refer to as "impressions" and, in particular, evaluative impressions that can serve as guides of our behaviors toward others.

## 1.3 How Easily Can We Update First Impressions?

Conventional wisdom suggests that you "never get a second chance to make a first impression," and there is no shortage of self-help books or professional tips on how to make a good first impression (e.g., Boothman, 2008) or to change a bad one (e.g., Grant, 2015). Indeed, at least four lines of empirical evidence lend credence to the idea that first impressions can exert a disproportionate influence on judgments and behaviors toward someone.

First, research suggests that first impressions, especially negative ones, can impact subsequent behavior in ways that prevent the opportunity to update or overturn them. Fazio and Eiser and colleagues have found, for example, that in attitude formation tasks that involve learning about the evaluative connotations of beans that vary in terms of their shape and other aspects of their appearance, participants who have a false expectation that certain types of beans will result in bad outcomes are less likely to approach them and thus discover that their expectation was wrong (Eiser, Fazio, Stafford, & Prescott, 2003; Eiser, Stafford, & Fazio, 2008; Fazio, Eiser, & Shook, 2004; Fazio, Pietri, Rocklage, & Shook, 2015). Thus, if an initial impression is sufficiently negative, it may lead to avoidance behavior that halts the acquisition of any further information about someone.

Second, even if we interact with someone or something long enough to encounter inconsistent information, research suggests that our consideration of new information will be seen through the lens of our initial impressions and thus color our perceptions. Because social information is often ambiguous (see Higgins, 1996), first impressions act as a filter through which we interpret and understand subsequent behavior (see also Bruner, 1957), which can lead people to cast inconsistent information aside or discount it in ways that preserve an initial impression.

Third, first impressions can create expectations that then influence our subsequent information gathering about someone or something in ways that confirm our initial hypotheses or beliefs—that is, they can serve as the basis for powerful confirmation biases (e.g., Darley & Fazio, 1980; Snyder & Stukas, 1999; Snyder, Tanke, & Berscheid, 1977). By learning that one's coworker seems to be extraverted, for example, it establishes an expectation that can lead us to both notice and actively solicit additional examples of extraversion, but fail to notice or solicit ways in which the person might be introverted.

Finally, our impressions of, and expectations about, someone can lead us to behave toward him or her in ways that unwittingly elicit the very behavior we expect—that is, they can result in self-fulfilling prophecies (Snyder et al., 1977). Thus, if we expect someone to be attractive and likeable, we may behave more warmly and friendly toward him or her (Snyder et al., 1977), whereas if we expect someone to be hostile or aggressive, we might ourselves respond more aggressively or antagonistically (Chen & Bargh, 1997)—the very behaviors that may elicit attractive or aggressive behaviors, respectively, from our interaction partners, thus confirming and more deeply entrenching our initial impressions of them.

## 1.4 How Easily Can We Update Explicit vs Implicit Evaluations?

Thus, several lines of research suggest that first impressions can sometimes linger even in spite of evidence that would appear to be inconsistent with them. However, of course, there *are* occasions when we encounter new, conflicting information about someone that is unambiguous and difficult to discount or discredit. We might discover that a seemingly likable acquaintance engaged in behaviors that we find morally repugnant, or we might learn that a seemingly prickly and disagreeable coworker donates half of her salary to charitable organizations every month. In short, we are sometimes surprised by others in ways that can override the otherwise powerful processes that tend to confirm our first impressions. Can we successfully update them in these kinds of cases?

The answer to this question depends, in part, on the kind of evaluative impression we are considering. Research has pointed to an important distinction between explicit evaluations, on the one hand—that is, those that are measured directly by asking how someone feels about someone or something deliberately and intentionally—and implicit evaluations, on the other hand—that is, those that are measured indirectly, meaning without the person's intention or endorsement of its measurement, and are thus spontaneous and relatively less intentional (Ferguson & Fukukura, 2012). A number of measures have been developed to assess implicit evaluations behaviorally, including the Implicit Association Test (IAT; e.g., Greenwald, McGhee, & Schwartz, 1998), the sequential evaluation priming paradigm (e.g., Fazio, Sanbonmatsu, Powell, & Kardes, 1986), and the Affect Misattribution Procedure (AMP; e.g., Payne, Cheng, Govorun, & Stewart, 2005), as well as measures that assess brain activation in response to exposure to attitude objects that can assess unintentional responding to affectively charged stimuli (e.g., Cunningham, Raye, & Johnson, 2004).

A number of lines of evidence suggest that whereas explicit evaluations can be very easily and rapidly updated in light of new information, implicit evaluations are relatively harder to modify once established, even for novel attitude objects for which we have relatively little information (e.g., Gregg, Seibt, & Banaji, 2006; Rydell & McConnell, 2006; Rydell, McConnell, Mackie, & Strain, 2006; Rydell, McConnell, Strain, Claypool, & Hugenberg, 2007; see also Wilson, Lindsey, & Schooler, 2000). For example, Gregg et al. (2006) introduced participants to two novel groups of people, referred to as the *Niffites* and *Luupites*. One group was described as brutal, violent, and mean, whereas the other group was described as gentle, peaceful, and loving. Immediately afterward, participants showed evidence of having successfully formed both implicit and explicit evaluative impressions in line with this information. Next, the researchers made a variety of attempts to overturn participants' initial impressions by providing information that was evaluatively inconsistent with their earlier-formed attitudes. In one attempt, participants were provided with a detailed narrative that indicated that the moral character of the two groups had changed over time, such that the violent group eventually saw the errors of their ways, seeking redemption by becoming peaceful and benevolent. At the same time, the other group became angry and bitter as a result of the oppression they experienced at the hands of the other group and became violent and terrible as a consequence. Although explicit evaluations successfully incorporated this reversal of the groups' characters, implicit evaluations showed very little evidence of revision, suggesting perhaps that they can be formed rather easily, but, once established, they become heavily entrenched and largely unmovable.

In another influential demonstration of the relative lack of malleability of implicit evaluations, Rydell et al. (2007) had participants read 100 positive behavioral statements about a novel target named Bob (e.g., "Bob donates his time at a soup kitchen"). After this initial impression had been formed, they then provided participants with varying amounts of counterattitudinal evidence about Bob—0, 20, 40, 60, 80, or 100 instances of negative behavioral statements. Whereas participants' explicit evaluations were immediately responsive to counter-evidence and exhibited significant revision even after just 20 statements, implicit evaluations responded much more slowly, changing in a linear manner as a function of the amount of counter-evidence offered—and it was not until fully 100 instances of negative behaviors were provided that participants exhibited a significant reversal of their initial impressions (see also McConnell & Rydell, 2014). This idea of slow change can also be seen in the dominant approach to changing implicit evaluations, via evaluative conditioning paradigms in which participants are exposed to a large number of repeated pairings between attitude objects and positivity or negativity (e.g., Karpinski & Hilton, 2001; Kawakami, Dovidio, Moll, Hermsen, & Russin, 2000; Kawakami, Dovidio, & van Kamp, 2005; Kawakami, Phills, Steele, & Dovidio, 2007; Olson & Fazio, 2006). Overall, then, many investigations have suggested that implicit evaluations are relatively less capable of rapid revision than explicit evaluations, except in response to extensive counter-conditioning (and even in such cases, the changes observed may be relatively ephemeral, at least for well-established attitude objects; see Lai et al., 2016).

These dissociations in the learning characteristics of implicit and explicit evaluative impressions are important because there is growing evidence that implicit evaluations uniquely predict a number of downstream judgments and behaviors (for review, see Cameron, Brown-Iannuzzi, & Payne, 2012; Friese, Hofmann, & Schmitt, 2008; Greenwald, Banaji, & Nosek, 2015; Greenwald et al., 2015; Greenwald, Poehlman, Uhlmann, & Banaji, 2009; cf. Oswald, Mitchell, Blanton, Jaccard, & Tetlock, 2015), including the domains of political behavior (Arcuri, Castelli, Galdi, Zogmaister, & Amadori, 2008; Friese, Smith, Plischke, Bluemke, & Nosek, 2012; Galdi, Arcuri, & Gawronski, 2008; Payne & Lundberg, 2014), self-control and regulation (Critcher & Ferguson, 2016; Ferguson, 2007, 2008; Friese, Hofmann, & Wänke, 2008b; Hofmann & Friese, 2008; Nederkoorn, Houben, Hofmann, Roefs, & Jansen, 2010; van Deursen et al., 2015), relationships (McNulty, Baker, & Olson, 2014; McNulty, Olson, Meltzer, & Shaffer, 2013), and intergroup behavior (e.g., Banaji & Greenwald, 2013; Towles-Schwen & Fazio, 2006). Thus, if implicit evaluations are incapable of successfully incorporating new information once formed, then they may be unreliable or maladaptive guides of our thoughts, feelings, or behaviors toward someone or something.

Some readers may be puzzled by the assertion that implicit evaluations are relatively difficult to change. After all, it would seem that there are a dizzying number of demonstrations in the literature of implicit evaluations appearing to be highly malleable in response to features of the current context or one's current goals (for review, see Blair, 2002; Ferguson & Fukukura, 2012; Gawronski & Sritharan, 2010). For example, implicit evaluations of Black people differ depending on whether they are presented in front of a threatening or nonthreatening background image (Maddux, Barden, Brewer, & Petty, 2005), whether participants first imagine counter-stereotypical exemplars (Blair, Ma, & Lenton, 2001), or whether the experimenter displays signs of having egalitarian ideals (Sinclair, Lowery, Hardin, & Colangelo, 2005). However, these are not demonstrations of *change* in the way we mean them here; instead, they may reflect transient activation patterns of particular aspects of a multifaceted mental representation, rather than any evidence of learning per se (i.e., changing the memories associated with the attitude object). All of the many possible evaluative responses to an image of Black people may simply reflect many different prior learning episodes, each of which may have taken a long time to develop. We return to this idea and discuss it in much more detail in Section 5.

This distinction—between change in the evaluative representation of a person in memory on the one hand, and transient reactivation of previously learned subsets of a multifaceted person impression on the other—provides a useful lens for viewing the results of a recent metaanalysis that found little evidence that shifts in implicit bias meaningfully impact behavior (Forscher et al., 2016). Although many studies have found evidence that implicit impressions predict behavior across a variety of domains (with some examples briefly described above), Forscher and colleagues find little evidence to suggest that revision of implicit impressions has corresponding substantive impacts on behavior. If revisions to implicit impressions do not correspond to behavioral change, it undermines the utility of studying such revision and raises concerns about the validity of implicit impressions as a theoretical construct. Their findings are also consistent with recent work suggesting that when shifts in implicit bias do occur, they may not be durable over days or even hours (Lai et al., 2016).

In our view, one explanation for the lack of both the durability and predictive validity of shifts in implicit impressions found in these investigations concerns an important distinction between the *context dependence* of impressions vs the more durable *revision* of impressions. All of the studies included in both Forscher and colleagues' (2016) metaanalysis and Lai et al. (2016) investigation consisted of known targets, about whom participants likely possessed an immense amount of previously acquired knowledge and evaluative content. The attempts to "change" impressions of these groups could easily have consisted merely of reactivating previously learned, context-dependent content, and not resulted in any real learning. If so, if these attempts at revision were simply temporarily highlighting past learning, without adding any new content, it does not seem surprising that such revision attempts did not map onto behavior and were not that durable.

Imagine, for example, that you possess 10 memories of interactions with elderly people that are, on the whole, quite negative—say, 8 of them are negative valenced and 2 of them are positively valenced. In most contexts, we should expect that your evaluative reaction to exemplars from this group will be negative. However, this is not to say that it will always be negative. For example, if you were told something decidedly positive about the group immediately before we elicited your implicit response, this may temporarily make the two positive memories relatively more accessible than the otherwise larger number of negative memories. Yet, even if you were to exhibit a positive evaluation in this temporary context, nothing has necessarily *changed* about your mental representation of elderly people. As a result, later that day, if you were to encounter an elderly person, your evaluation may quickly revert back to being negative. This is quite different from what occurs during a more durable learning experience in which you acquire new information about the group (i.e., new, positive memories about them) or in which the earlier negative memories are altered to become more positive. When we provide new information about well-established groups or otherwise attempt to change evaluations of these targets, then, because we are uncertain of the content of people's mental representations prior to the experiment, it is ambiguous as to which of these processes—contextualization or more durable change—has occurred. Of course, our interpretation of the recent work with established targets by Forscher and Lai and others remains untested, and the number of studies demonstrating durability and behavioral consequences of change with novel targets remains small, so much more work clearly remains to be done assessing the extent to which implicit updating is both durable and predictive of behavior.

# 2. THEORETICAL PERSPECTIVES ON IMPLICIT IMPRESSION CHANGE

A variety of theories have been developed to motivate the experiments discussed earlier and to account for the collective body of evidence on the updating of implicit and explicit evaluations. Many of these theories have explained dissociations in the rate and conditions of change in implicit and explicit impressions by proposing dual-mode models, which hold that the two forms of evaluation are implemented through distinct processes, systems, and/or representations (Gawronski & Bodenhausen, 2006, 2011; Petty, Briñol, & DeMarree, 2007; Petty, Tormala, Briñol, & Jarvis, 2006; Rydell & McConnell, 2006; cf. De Houwer, 2014). We briefly summarize some of the central claims of these perspectives, with particular attention to how they account for the empirical evidence discussed in the previous section, before turning to a discussion of our alternative approach to conceptualizing implicit evaluation updating.

## 2.1 Associative and Propositional Processes

Though the details of specific theories differ in important ways, a commonality among many dual-mode theories of implicit and explicit impressions is the proposal that whereas implicit impressions are the products of associative processing, explicit impressions are instead (or additionally) the products of propositional processing. The systems of evaluation (SEM) model, for example, draws on Sloman's (1996) model of systems of reasoning to explain implicit and explicit evaluations as the product of two independently operating mental systems, with the implicit system implementing associative processing and the explicit system implementing rule-based processing (McConnell & Rydell, 2014; Rydell & McConnell, 2006). According to SEM, implicit evaluations are generated through the activation of associative representations according to principles such as similarity and temporospatial contiguity, whereas explicit evaluations are produced through the rule-based activation of propositional beliefs.

Other theories view the distinction not in terms of dissociable mental systems and types of representations, but interacting processes. The Associative-Propositional Evaluation Model (APE; Gawronski & Bodenhausen, 2006, 2011) and Meta-Cognitive Model (MCM; Petty et al., 2007, 2006) each view both implicit and explicit evaluations as emerging from a common associative representational store, and propose that associative processing can drive both forms of evaluation. When a stimulus is perceived, a process of pattern completion in the associative network (constrained by stimulus features, context, and prior activity) drives an activation pattern that can result in an implicit or explicit evaluation. Explicit evaluation, however, can be additionally impacted by propositional processing, including the affirmation or negation (rejection) of currently active associations.

Finally, De Houwer and others have proposed that implicit and explicit evaluations can both be produced through the activation of propositional representations, rather than associations (De Houwer, 2014; Hughes, Barnes-Holmes, & De Houwer, 2011), with the central distinction between implicit and explicit evaluations resting with the level of automaticity with which they are expressed (De Houwer, Teige-Mocigemba, Spruyt, & Moors, 2009). This perspective emphasizes that most concept-level knowledge is relational, such that simple associations have limited explanatory power; even the products of "associative learning" tasks can be understood as directional representations—such as that a CS signals an impending US, but not vice versa—reflecting more meaning than simple associations would encode (Mitchell, De Houwer, & Lovibond, 2009).

## 2.2 Sources of Dissociations and Routes for Change Under Current Models

The theories discussed earlier offer various perspectives on why dissociations between implicit and explicit impressions might occur, with corresponding implications for how such dissociations can be overcome—i.e., how implicit impressions can be brought in line with explicit beliefs. In SEM, representations in the implicit system form and change through associative learning, responsive only to the repeated pairing (i.e., coactivation) of evaluative information with the target. The associative system is considered relatively insulated from effects of propositional reasoning, such as judgments about the validity or invalidity of information; because of this, implicit evaluations should not change in response to a decision to disbelieve or reject a prior impression. Explicit evaluations, on the other hand, are generated through the activation of representations within a rule-based propositional system. This allows explicit evaluations to be sensitive to judgments of validity, allowing for the rapid updating of explicit beliefs. Because the two systems operate independently (though some interaction may be possible; see

McConnell & Rydell, 2014), dissociations can be readily explained: Explicit evaluations may change quickly through propositional rejection of the prior impressions, but because the associative system is not impacted by this rejection, implicit evaluations change only slowly, after repeated exposure to many counterattitudinal pieces of information (Rydell & McConnell, 2006; Rydell et al., 2006, 2007).

Under the dual-process approaches of APE and MCM, implicit and explicit evaluations both draw upon the activation of associations in memory, but dissociations can emerge when explicit evaluations draw additionally upon propositional processes. For example, processing an elderly face may result in the activation of a learned negative association (based on prejudice) that drives a negative response on an implicit measure. However, the person might reject this active negative association to an elderly face as inconsistent with her values. In that case, she would show a dissociation such that an implicit measure reflects the "elderly bad" association, because it is driven directly by associative processing regardless of the negation, whereas an explicit measure reflects the propositional rejection of the association.

What does the greater interaction between associative and propositional processing in these models (compared with SEM) imply about how implicit impressions can be revised? Though they both propose routes through which propositional processing can shift implicit impressions, the APE model and MCM differ in their details. Under APE, the propositional validity of information (whether it is considered true or false) is not reflected in associative processing. However, propositional processing can shift the currently active pattern of associations, such as when rejection of the proposition "Elderly is bad" leads the perceiver to remember positive elderly exemplars and endorse an "Elderly is good" proposition, thereby shifting the pattern of associative activity from "Elderly bad" to "Elderly good." This newly active "Elderly good" association may then drive implicit responses, and such activation can strengthen that associative connection, resulting in durable change (Gawronski & Bodenhausen, 2006, 2011).

One important prediction of the APE model is that although propositional processes can impact implicit impressions, propositional *affirmation* will generally be much more effective in doing so than propositional *negation*. APE specifies that negating (rejecting) the validity of a proposition generated by associative patterns (e.g., rejecting an "Elderly is bad" prejudiced belief that arises from an active "Elderly bad" association) should not shift implicit evaluations, because a mere negation has no effect on the pattern of currently active associations. That is, a propositional rejection of "Elderly is bad" simply maintains the activation of an "Elderly bad" association. Only when propositional processing shifts mental content by affirming a different proposition—such as affirming that "Elderly is good," which will activate an "Elderly good" association—will implicit evaluations shift as well. The APE model thus accounts for some findings that suggest that negations alone are ineffective in driving shifts in implicit evaluations (Gawronski, Deutsch, Mbirkou, Seibt, & Strack, 2008), at least if the negation does not immediately follow learning (Peters & Gawronski, 2011). These considerations open the door for propositional assessments of information, such as perceived diagnosticity and believability, to play a role in implicit updating to the degree that they promote the activation and strengthening of different associative connections.

The MCM is similar to APE in many respects, but unlike APE, it posits that validity (assessments of information as true or false) can come to be directly reflected in the associations that drive implicit evaluations (Petty et al., 2007, 2006). When a prior impression is judged to be false, under MCM, this creates a false "tag" associated with that rejected content. The strength of this association is generally thought to be weak initially relative to the past impression itself, becoming stronger over time through cognitive elaboration and repeated practice. MCM posits that implicit measures are more likely to draw upon stronger associations, such that a new (and relatively weakly encoded) validity tag may not initially be able to prevent the activation and expression of a rejected impression on an implicit measure. On the other hand, the slower and more effortful retrieval afforded by explicit measures make activation of the validity tag more likely. Nonetheless, because the MCM allows for the possibility that the strength of encoding of a validity tag influences its likelihood of activation and expression, it is possible under this model that highly compelling new information could lead to rapid strengthening of such validity associations, though the exact conditions under which this will occur remain little defined (though see later discussion of Wyer, 2010, 2016). Indeed, evidence suggests that negations might be more effective in shifting implicit responses if made more salient (Boucher & Rydell, 2012) or if made more meaningful (Johnson, Kopp, & Petty, in press).

Finally, out of all of the theories discussed, the propositional view (De Houwer, 2014) most easily accommodates the possibility of rapid revision in implicit impressions through propositional reasoning, because of the direct correspondence between the propositional form of that processing and the resulting propositional representational structure. In other words, if one rejects the validity of "Elderly people are bad" (a validity-based, and thus propositional process), then this rejection can be immediately reflected in memory in a way that can affect implicit (or explicit) measures, because, under this theory, implicit measures are driven by propositional beliefs. Under this view, dissociations between implicit and explicit measures are attributable to different propositions becoming active under different automaticity conditions, rather than the activation of representations with different formats or through different processes (e.g., propositional vs associative). This perspective on dissociations (that they need not reflect fundamental differences in the nature of underlying representations of processes) is shared by dynamical approaches to evaluation, which describe how dissociations can emerge from continuous evaluative processes under different inputs, contexts, and processing conditions like speed (Cunningham, Zelazo, Packer, & Van Bavel, 2007; Ferguson, Mann, & Wojnowicz, 2014; Ferguson & Wojnowicz, 2011; Van Bavel, Jenny Xiao, & Cunningham, 2012; Wojnowicz et al., 2009). The propositional approach to implicit cognition has already proven useful in highlighting the importance of capturing the particular relations encoded within implicit beliefs (e.g., Remue, De Houwer, Barnes-Holmes, Vanderhasselt, & De Raedt, 2013; Tibboel, De Houwer, Dirix, & Spruyt, 2017) and readily accommodates the strong effects of instructed knowledge on implicit evaluations (e.g., Kurdi & Banaji, 2017; Van Dessel, De Houwer, Gast, & Smith, 2015; Van Dessel, De Houwer, Gast, Smith, & De Schryver, 2016; Van Dessel, Gawronski, Smith, & De Houwer, 2017).

## 3. OUR APPROACH

Although the evidence and theoretical perspectives just summarized might appear to suggest that implicit impressions cannot be easily updated or revised, there are at least two reasons why we predict that implicit evaluations may be capable of rapid revision. First, evidence for dissociations of the sort summarized earlier does not necessarily imply that implicit and explicit attitudes are governed by multiple processes or systems. Second, even if implicit evaluations are solely under the purview of implicit processes, recent work suggests that implicit memory is not a single, monolithic entity. Rather, it is composed of multiple memory systems that each exhibit their own unique learning characteristics and developmental trajectories. We discuss these two reasons in more detail below.

## 3.1 Reason #1: Dissociation Does Not Imply Dual Processes

Although the studies described in the introduction—in which implicit and explicit evaluations appear to be differentially sensitive to different kinds of information—might seem to imply that implicit and explicit evaluations rely on different processes, which then result in marked differences in their learning characteristics and developmental trajectories, such an inference is not necessarily warranted. This is because, although these kinds of dissociations are consistent with the notion of separate processes or systems, they are also consistent with many other explanations, including single process models (Bedford, 2003; Chater, 2003; Dunn & Kirsner, 2003; Ferguson et al., 2014; Keren & Schul, 2009; Kruglanski, Erb, Pierro, Manetti, & Chun, 2006; Plaut, 1995). Indeed, no measure is "process pure"; instead, measures are likely to involve several different processes that each contribute to the ultimate response that participants produce.

Moreover, there are many differences that exist between implicit and explicit evaluations in both their conceptualization and measurement that may potentially explain empirical dissociations. For example, Payne, Burkley, and Stokes (2008) have argued that the dissociations observed between implicit and explicit evaluations may be partly explained by differences in the structure of the measures used to assess each type of evaluation. Indeed, the measures differ on a number of important dimensions, including the type of response that is required of participants, the timing of the measure, the type of stimuli that are used, and many others. In a provocative and influential set of studies, Payne et al. (2008) demonstrated that when these differences in structure were reduced between the two measures (i.e., by mimicking the properties of an implicit measure in explicit evaluations and vice versa), dissociations were also reduced. Thus, the gulf between implicit and explicit evaluations may be more apparent than real.

Our own view is that it is impossible to determine the underlying processes involved in performance on a particular task or measure without the use of computational modeling. Because computational models that specify underlying processes can then be tested for their fit with resulting behavioral data, they can better elucidate the underlying mechanisms that drive behavioral results than evidence for dissociations based on behavioral data alone (see Ferguson et al., 2014). Thus, although evidence of disparities in performance on measures of implicit and explicit evaluations is intriguing, they do not necessarily have implications for their respective learning characteristics.

## 3.2 Reason #2: Implicit: One Process or Many?

One consequence of a dual process or dual systems view of implicit and explicit evaluation is that potential complexity in the processes involved *within* each process, especially on the implicit side, may be deemphasized. In particular, a great deal of evidence suggests that implicit responses may exhibit a variety of different learning characteristics under different conditions, suggesting that implicit cognition may draw upon not one but rather a host of different implicit processes or systems (Amodio & Ratner, 2011; Conrey, Sherman, Gawronski, Hugenberg, & Groom, 2005; Sherman et al., 2008). For example, some types of implicit learning such as fear conditioning can occur exceedingly rapidly (Hermer-Vazquez et al., 2005; Yin & Knowlton, 2006)—very often within a single trial (e.g., see Cahill & McGaugh, 1990; Hilliard, Nguyen, & Domjan, 1997)—and, at the same time, may extinguish much more slowly than other kinds of implicit associative learning such as instrumental learning (Yin & Knowlton, 2006) or semantic learning (McClelland & Rumelhart, 1985).

Thus, from a memory systems perspective, implicit evaluative impressions may reflect the interaction of several implicit processes rather than merely a single, slow associative learning system. Consistent with this idea, Amodio and Devine (2006) found that implicit evaluations and implicit stereotypes of the same targets were not significantly correlated, and predicted different outcomes, with implicit stereotypes predicting stereotype-related performance expectations and implicit evaluations predicting approach/avoidance behavior. Amodio and Ratner (2011) argued that these findings were consistent with the notion that implicit evaluations draw differentially upon the affective memory system, whereas measures of implicit stereotypes draw differentially on the semantic memory system. Thus, to the degree that implicit impressions have substantial evaluative components that tap into a fast-learning affective system—a system recent work suggests can accommodate rapid updating (Agren et al., 2012; Schiller, Kanen, LeDoux, Monfils, & Phelps, 2013)—or a fast-learning and fast-updating instrumental system, rapid revision may be more likely or possible than a perspective informed exclusively by semantic memory research would suggest.

## 4. THREE ROUTES TO RAPID REVISION OF IMPLICIT EVALUATIONS

Based on the reasoning described earlier, we have uncovered at least three distinct mechanisms that can lead to rapid revision of implicit responses: when information is highly *diagnostic*; when information causes a *reinterpretation* of prior learning; and when information is particularly *believable*, it can rapidly overturn prior learning. Though we have pursued each of these mechanisms in separate lines of work, which we discuss in turn below, we view them as strongly interrelated.

### 4.1 He Did *What*? Diagnostic Revelations Result in Rapid Evaluative Change

Paul De Man was a well-respected literary critic who held coveted positions at Johns Hopkins, Cornell, and Yale. At the time of his death in 1983, he was one of the most prominent in his field, well-respected and admired as a scholar of French and comparative literature. Then, in 1987, not long after his death—in a stunning revelation that shook literary criticism to its core—it was discovered that De Man had been living a double life: he had written for two Nazi-controlled newspapers in Belgium in the 1940s and had appeared, after additional digging through Belgian archives, to be an unflinching and publicly proud fascist in his former life (Menand, 2014).

How might this new revelation influence mental representations of De Man? Should we expect that explicit evaluations will rapidly incorporate this knowledge, while implicit evaluations will be largely uninfluenced? One reason why we might expect this kind of revelation to resonate at the implicit level is that such information is decidedly *extreme* and, thus, more likely to be imbued with significance and weighted more heavily in overall impressions and evaluations. The discovery that De Man was a Nazi sympathizer is not merely another piece of information that could potentially color one's impressions, to be placed alongside everything else one knows about him. Rather, it reveals something much more important about his true nature—something about who he *really* is—and should thus carry much more weight in one's overall evaluation of him. When impression-inconsistent information has this property of revealing important aspects of a person's enduring traits, dispositions, or character—a feature we call *diagnosticity*—we expect it should be more rapidly and successfully incorporated into people's overall implicit impressions, and can thus rapidly overturn many prior instances of learning about someone.

From this perspective, one important reason why previous work has appeared to suggest that implicit evaluations cannot be easily "undone" by new information is that prior tests of this hypothesis may not have used sufficiently diagnostic revelations. The empirical strategies used by Gregg et al. (2006) to assess the malleability of already-formed implicit evaluations, for example, did not necessarily suggest anything especially *diagnostic* of either the *Niffites*' or the *Luupites*' true nature or fundamental essence, nor did they necessarily challenge any inferences about each group's character that had already been learned during their initial impression formation earlier in the experiment. Even in the case in which participants were given an extensive counter-narrative to attempt to reverse their implicit impressions, the information that participants learned—that the formerly morally corrupt aggressors had recognized the error of their ways and had sought to make amends for past wrongs, whereas the formerly morally virtuous, oppressed people had developed an animosity toward the aggressors and had sought vengeance—was, at most, equally diagnostic relative to the information they had already learned.

Similarly, in implicit evaluation change research employing the Bob paradigm (e.g., Gawronski, Rydell, Vervliet, & De Houwer, 2010; Rydell & Gawronski, 2009; Rydell et al., 2006, 2007), the information that has generally been used to attempt to reverse prior implicit learning has exclusively been of the same diagnosticity as earlier information learned during initial impression formation. Under these circumstances, previous theorizing that suggests that an equivalent *amount* of impression-inconsistent information may be necessary to undo prior learning (Rydell et al., 2006, 2007) could be valid. Yet, these findings do not necessarily speak to whether it is possible for implicit evaluations to change more quickly when exposed to more extreme diagnostic countervailing information.

#### 4.1.1 Empirical Evidence for the Role of Diagnosticity

Several lines of work now support the contention that rapid revision of implicit evaluations can occur when new impression-inconsistent information is sufficiently diagnostic—even with just a single, highly diagnostic revelation. In one line of work (Cone & Ferguson, 2015), participants completed a variant of the Bob paradigm described earlier (Boucher & Rydell, 2012; Gawronski et al., 2010; Gawronski, Ye, Rydell, & De Houwer, 2014; Rydell & Gawronski, 2009; Rydell & McConnell, 2006;

Rydell et al., 2006, 2007) in which they acquired many instances of evaluatively consistent information about a white college-aged male target. Next, participants were provided with just one additional piece of information about him that was inconsistent with the impression they had formed earlier. However, this new piece of information was specifically chosen to be especially extreme and highly diagnostic of Bob's character (e.g., "Bob was recently convicted of molesting children"). To assess the extent to which this information could successfully reverse participants' initial impressions, we measured implicit attitudes using an AMP (as well as explicit evaluations) using self-report measures immediately before and immediately after they encountered it. Across six studies, we consistently found that highly diagnostic revelations of this sort could indeed successfully revise implicit evaluations. At the same time, however, we also found evidence for a valence asymmetry (Cone & Ferguson, 2015, Study 2), such that highly negative diagnostic revelations (e.g., "Bob mutilated a small, defenseless animal") were more impactful on implicit positive evaluations than highly positive diagnostic revelations (e.g., "Bob donated his kidney to a stranger in need") were on implicit negative ones (see Fig. 1). Whereas diagnostic negative information consistently led to full reversals of implicit evaluations—causing significantly positive implicit impressions to become significantly negative—diagnostic positive information merely attenuated implicit negativity rather than fully reversing it. Although we might be willing to decide that someone that seemed positive is, in fact, negative, we appear to be much more reticent to decide that someone who seems negative is, in fact, positive. This asymmetry fits with a large body of work on the dominance of negative information across domains, including in person impressions (e.g., Baumeister, Bratslavsky, Finkenauer, & Vohs, 2001; Roese & Olson, 2007; Rozin & Royzman, 2001; Skowronski & Carlston, 1989). A chief reason for this asymmetry is likely that negative (particularly immoral) behaviors are seen as especially revealing about a person's dispositional character (e.g., Reeder & Coovert, 1986; Skowronski & Carlston, 1989).

A number of additional lines of evidence further support the contention that these results are the product of the relative diagnosticity of the new information, rather than some other incidental feature of the information that participants learned. First, when the very same revelation was provided to participants but was described as an action performed not by Bob but rather by someone incidentally associated with Bob (i.e., the high school football coach in Bob's hometown), no changes were observed in either implicit or explicit evaluations (Cone & Ferguson, 2015, Study 3). Thus,

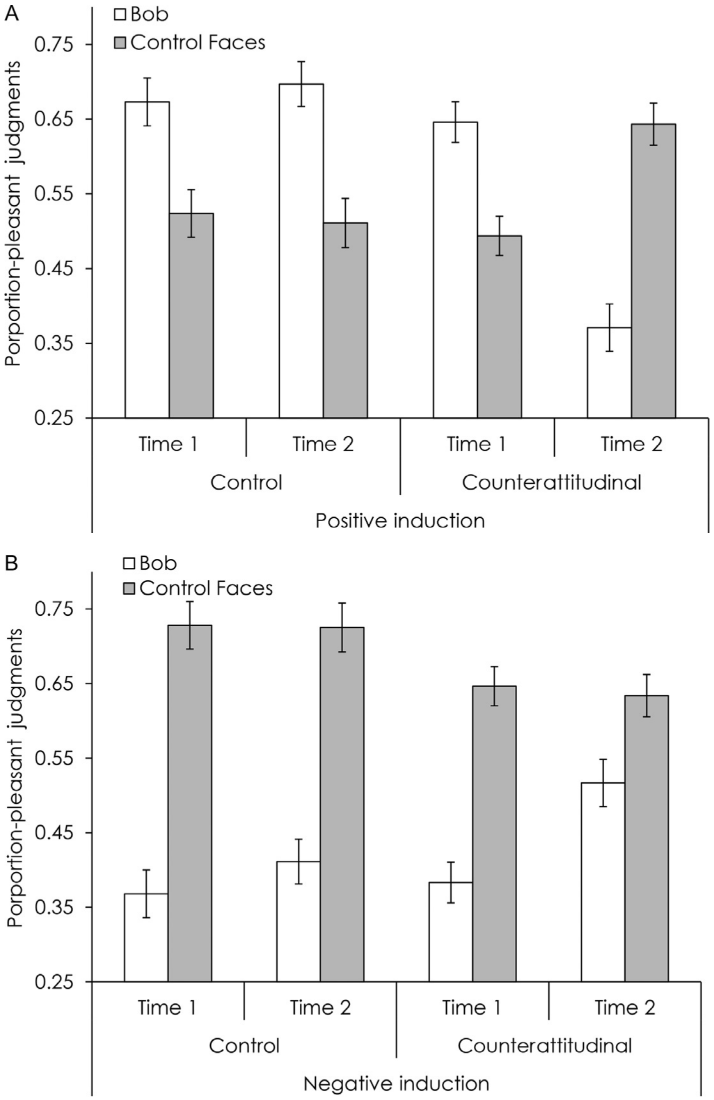

**Fig. 1** Implicit evaluations as a function of positive induction (A) or negative induction (B) in a learning paradigm, followed by either control information or counterattitudinal diagnostic information. In the positive induction, this information was that "Bob was recently convicted of child molestation," whereas in the negative induction, this information was that "Bob recently donated one of his kidneys to a child in need he had never met before." *Adapted from Cone, J., & Ferguson, M. J. (2015). He did what? The role of diagnosticity in revising implicit evaluations.* Journal of Personality and Social Psychology, 108*, 37–57, Study 2.*

extreme, highly diagnostic information that occurs in close temporal proximity to exposure to Bob was not enough to induce changes in implicit responses; instead, the actions only resonated at the implicit level when they were attributed to him. Second, a direct self-report measure of people's beliefs about the diagnosticity of new revelations (e.g., "To what extent do you feel that the later information you learned about Bob is a reflection of his true nature or character?") mediated the relation between extremity and implicit evaluation change: as assessments of new information became more negative, people's beliefs about the diagnosticity of the new information simultaneously increased, and these beliefs then ultimately predicted how much change was observed in people's implicit responses (Cone & Ferguson, 2015, Study 4).

### 4.1.2 Causal Attribution Processes as a Route to Rapid Revision

Though, as our earlier work would suggest, it is often the case that especially positive or negative behaviors can be revealing of a person's character, not all extreme behaviors are necessarily seen as diagnostic in this way. The public declarations of a prisoner of war made under duress or in a life-threatening situation tell us very little about the kind of *person* that he or she is, and even if his or her actions might, on the surface, be considered rather extreme or morally questionable, they would not necessarily be the sorts of revelations that ought to lead to rapid changes in our impressions of him or her. An interesting implication of the fact that information that is highly diagnostic can be one route to rapid revision of implicit responses is that when participants encounter a behavior that is inconsistent with their first impressions, it ultimately matters *why* they think the person engaged in the behavior—in other words, causal attribution processes may affect not only our explicit impressions of others (Jones & Harris, 1967; Kelley, 1967), but also implicit evaluative change. The very same behavior may or may not lead to rapid reversals of implicit responses depending on a more sophisticated understanding of the underlying causes of the behavior rather than the mere association of someone with extreme positivity or negativity.

In a conceptual replication of a classic demonstration of these kinds of causal attribution and person perception processes (Jones & Harris, 1967), Fourakis and Cone (in preparation, Study 1) told participants that a novel stranger had been given an opportunity to write an essay either supporting or condemning animal mutilation. In a 2 (Situational Constraint: Freely Chosen, No Choice) × 2 (Time: 1, 2) × 2 (Target: Bob, Control Faces) design, all participants learned that the target had ultimately written an essay in *support* of animal mutilation. However, half the participants were told that the target had no choice in writing the essay—that he or she was required to do so as part of an experiment—whereas others were told that the target could have written about either topic and freely chose to write his supporting argument. When the target freely chose to write the essay, participants naturally saw the behavior as highly negative and diagnostic of his or her character, and, just as our earlier work would suggest, their implicit evaluations (as measured using an AMP) immediately and significantly became more negative (as did their explicit attitudes). However, when the essay was a requirement of the experiment and thus not necessarily revealing of the person's true feelings about animal mutilation, implicit evaluations were essentially unchanged in response to the revelation, even though participants nonetheless saw the action itself as highly negative. Moreover, the effects of the situational constraint information on implicit evaluative change were mediated by participants' beliefs about the diagnosticity of the information they considered; when the essay-writing behavior was situationally compelled, participants imbued it with less overall diagnosticity, which then predicted less overall implicit evaluative change as a consequence.

We also found evidence to suggest that people can successfully modify their impressions of others even when people learn about one of their behaviors and only *later* discover the circumstances that elicited it (see also Peters & Gawronski, 2011; cf. Gawronski & Bodenhausen, 2006, 2011). In one study (Fourakis & Cone, in preparation, Study 2), participants were first told that a novel stranger had written an essay supporting neo-Nazism. Afterward, participants' implicit and explicit evaluations were assessed. As predicted, all participants developed highly negative implicit and explicit impressions of the target. Afterward, they were reminded about the essay-writing behavior and were only *then* told that it had been either situationally compelled or freely chosen. Despite the fact that they had already developed a negative implicit impression, participants nonetheless updated their assessment in light of the modification of their understanding of the causes of the earlier behavior, with participants exhibiting significant positive changes in their implicit evaluations when they learned the behavior was compelled by the situation, but maintaining highly negative implicit impressions when they discovered it had been freely chosen. There was also evidence to suggest that the extent to which participants reinterpreted the earlier information (see below for a more detailed discussion) was a significant mediator of the extent to which they revised their implicit attitudes in light of the situational constraint information.

Perhaps even more interestingly, when situational constraint information was provided before the essay as part of the impression formation task, implicit evaluations were affected *less* than explicit evaluations. Although the highly negative behavior still influenced explicit evaluations even when the person had no choice but to engage in the behavior (as would be anticipated by Jones & Harris's and others' classic work; see also, Gilbert, 1998; Ross, 1977), there was no statistically significant change in implicit evaluations, suggesting that implicit impressions actually reflected more appropriate use of situational constraint information than explicit impressions. This occurred despite the fact that there was marked evidence of implicit and explicit evaluative change when the essay-writing topic was freely chosen, thus indicating that participants were highly responsive to the information when it revealed something important about the person's character.

It is worth mentioning, of course, that participants may view highly negative behaviors as morally troublesome even when they are situationally constrained. Even if an essay-writing topic is technically a requirement of an experiment, some participants may nonetheless feel that writing an essay in support of animal mutilation is unacceptable and may still reveal important facets of a person's character (e.g., the target should have refused to write the essay). Thus, the extent to which evaluations *should* change, in a prescriptive sense, in response to situationally compelled behaviors is open to debate. Nonetheless, the fact that implicit evaluations appear to exhibit less responsiveness than explicit evaluations under these circumstance is a testament to the power of diagnosticity as a determinant of implicit (non)change.

What this suggests, then, is that the extent to which a particular behavior is ultimately seen as diagnostic is not just a feature of the action itself, but also an inference of what compelled it. Diagnosticity is thus not only a product of the behavior itself, but also the context and the circumstances surrounding its occurrence.

### 4.1.3 The Role of Visual Cues

Examining implicit evaluative change through the lens of diagnosticity can also help to unravel how and why implicit evaluations appear to be more sensitive to certain kinds of information relative to others. Consider, for example, the story of Jeremy Meeks, who quickly became an internet sensation very shortly after his decidedly photogenic mugshot was posted to the Stockton County Police Department's Facebook page in 2014. Now known to the world as the "Hot Felon," Meeks's mother successfully raised several thousand dollars for his legal defense, donated by dozens of random strangers around the world who saw his mugshot on a GoFundMe page. He was also offered a modeling contract by White Cross Management—an opportunity of which he was sadly unable to take immediate advantage because he was still in prison at the time of the offer—and all of this despite the fact that he had engaged in sufficiently illegal and morally questionable behavior to land him in jail—facts that ought to have colored perceptions of his trustworthiness and inferences about the quality of his character.

Under what circumstances can initial impressions override salient visual information (such as physical attractiveness, being overweight, or having facial scars or other physical deformities) and successfully incorporate relevant behavioral information? The case of Meeks—not to mention several lines of prior research (Blair, Chapleau, & Judd, 2005; McConnell et al., 2008; Rule, Tskhay, Freeman, & Ambady, 2014; Sritharan, Heilpern, Wilbur, & Gawronski, 2010)—might suggest that implicit evaluations are often unduly influenced by salient associative visual cues, perhaps even to the exclusion of other kinds of relevant information. In one of the first investigations of the interaction between visual cues and behavioral information on impressions, McConnell et al. (2008) consistently found that, although explicit evaluations were quite responsive to behavioral information about a person and largely unresponsive to visual cues such as whether a social target was physically attractive or overweight, implicit evaluations exhibited precisely the opposite pattern. They proposed that these findings were consistent with the SEM Model (McConnell & Rydell, 2014; Rydell & McConnell, 2006) in which implicit evaluations are largely under the purview of associative processes and less responsive to rule-based or propositional information derived from behavioral statements (see our earlier discussion).

However, a reconsideration of these findings through the lens of diagnosticity suggests that this may not always be the case and there may be certain kinds of behavioral information that can indeed override even very salient visual information in people's implicit evaluations. Just as individuals evaluate behavioral information in terms of its diagnosticity, so too does visual information carry with it an associated measure of its importance for understanding what a person is like. Physical attractiveness, for example, leads to other positive trait ascriptions, such as inferences of increased kindness, trustworthiness, and social competence (e.g., Dion, Berscheid, & Walster, 1972; Eagly, Ashmore, Makhijani, & Longo, 1991). Thus, visual cues may independently lead to diagnostic inferences about a person's character that may or may not be at odds with diagnosticity assessments derived from other sources of information.

How do these two kinds of diagnosticity assessments interact to influence the extent to which implicit evaluations respond to one or the other kind of information? To investigate this question, we had participants complete a variant of McConnell et al. (2008) procedure in which they acquired behavioral information about a woman named "Bobbie," whose photo was randomly assigned to be either attractive or unattractive (Cone, Mann, Meagher, & Ferguson, in preparation), in a 2 (Bobbie: Attractive, Unattractive) × 2 (Time 2 Information: Neutral, Diagnostic) ×2 (Time: 1, 2) ×2 (Target: Bobbie, Julia) design. To assess baseline implicit evaluations of Bobbie (as well as a neutral control, Julia, who was included in all conditions), we had participants familiarize themselves with each of their images by having them complete a learning paradigm in which they received 10 innocuous pieces of information about each of the two targets. These behavioral statements did not convey any meaningful information about them (e.g., "Bobbie buys groceries at a nearby store"). Immediately after this task, participants exhibited significantly greater implicit positivity on an AMP toward the attractive version of Bobbie than they did to the unattractive version (Cone et al., in preparation, Studies 1 and 2).

At Time 2, we exposed participants to additional behavioral information, but, unlike McConnell and colleagues' work, this information was highly diagnostic: "Bobbie was convicted of drowning her children in the bathtub." Also unlike McConnell and colleagues' work, we found that this single behavioral statement was effective in changing implicit evaluations, and it was *equally* effective (that is, led to equivalent amounts of evaluative change) irrespective of whether Bobbie was attractive or not. In other words, highly diagnostic behavioral information was able to successfully override the marked differences in evaluations derived from the visual cues at baseline (see Fig. 2).

These findings suggest that the undue influence of visual information on implicit evaluations observed by McConnell and colleagues is not inevitable. Rather, behavioral information can sometimes carry more weight than visual information when the details are so extreme that participants become unwilling (or unable) to ignore or discount them.

Of course, it is unlikely to be the case that either visual information or behavioral information will completely dominate one's implicit impressions of someone, and, in another study (Cone et al., in preparation, Study 3), we found evidence to suggest that these sources of information sometimes

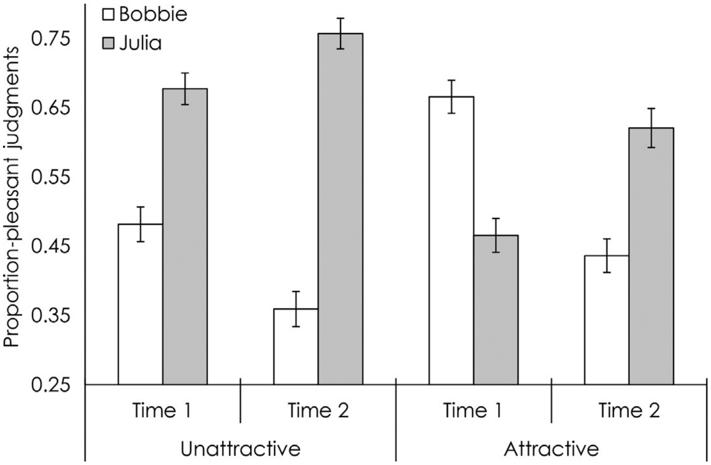

**Fig. 2** Implicit evaluations of Bobbie and Julia (control) as a function of whether Bobbie was attractive or unattractive and whether evaluations were obtained before (Time 1) or after (Time 2) learning that she drowned her children in a bathtub.

interact to determine which dominates an implicit impression. To assess individual differences in participants' beliefs about the diagnosticity of a visual cue (in this case, overweight), we contacted participants 1 week prior to the full study and had them complete a short measure assessing their beliefs about whether being overweight communicates information about a person's traits, personality, or disposition (e.g., "Some people are overweight because they have no willpower"). One week later, like the studies manipulating physical attractiveness described earlier, participants developed initial impressions of two individuals based on photos of their faces—one manipulated using computer software to be extremely overweight and the other selected to be normal weight—as well as exposure to 10 innocuous behavioral statements about each. Next, participants learned that the overweight individual had engaged in highly altruistic behavior: "Bob recently donated bone marrow to his friend with cancer."

Our main question of interest was how the behavioral statement would impact implicit evaluations as a function of participants' beliefs about the diagnosticity of the visual cue. If participants see being overweight as a moral failing or character flaw, they may imbue it with greater meaning, thus causing the highly altruistic behavioral information to carry less weight in their overall evaluation. However, if participants see being overweight as being caused more by situational or environmental factors (e.g., living in an area without easy access to healthful foods such as fruits, vegetables, or whole grains) than by features of a person's personality or disposition, then they may become more sensitive to behavioral information and thus exhibit greater implicit change. As Fig. 3 shows, this is precisely what we found. Whereas those whose overweight diagnosticity beliefs were low exhibited a reversal of their implicit responses in light of the behavioral information, those whose diagnosticity beliefs were higher exhibited much less change in response to the altruistic behavior. Thus, the contributions of salient visual cues and behavioral information to evaluative updating may be influenced by assessments of their *relative* diagnosticity and the extent to which each kind of information leads to inferences—whether positive or negative—about a person's underlying disposition, not necessarily by the *format* or *type* of information (i.e., visual or behavioral).

In another recent line of work, we have tested whether new diagnostic information can overturn negative implicit impressions that stem from facial features suggesting an undesirable character trait—specifically, untrustworthiness (Shen, Mann, & Ferguson, in preparation). Perceivers readily infer a variety of impressions about the character of an individual from facial structure (e.g., Eagly et al., 1991; Rule, Ambady, & Adams, 2009; Todorov, Olivola, et al., 2015), including levels of trustworthiness or untrustworthiness (Rule, Krendl, Ivcevic, & Ambady, 2013). These

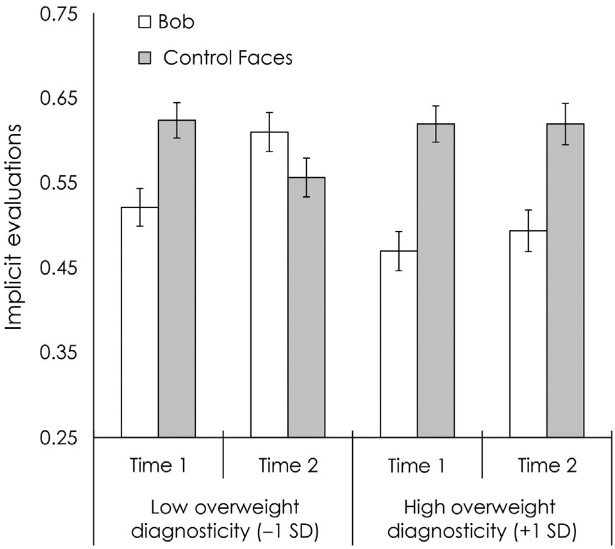

**Fig. 3** Implicit evaluations of Bob (overweight) and Joe (normal weight) as a function of participants' beliefs about the extent to which being overweight is a cue to a person's disposition or character, obtained before (Time 1) or after (Time 2) they learned that Bob had engaged in a highly altruistic behavior. Values are based on a regression equation, depicted at ±1 SD from the mean on overweight diagnosticity.

face-based inferences are predictive of behavior toward targets across different domains (e.g., Rezlescu, Duchaine, Olivola, & Chater, 2012; Tingley, 2014; van't Wout & Sanfey, 2008; Wilson & Rule, 2015). Given the behavioral consequences and ubiquity of such inferences, it is important to know whether negative implicit evaluations stemming from untrustworthiness cues can be overcome by highly diagnostic information that the person is, in fact, trustworthy.

Shen et al. (in preparation) found in a series of experiments that such face-based impressions can indeed be shifted by relevant propositional information. For example, in one study, participants learned a variety of neutral behaviors performed by a target individual named Joe who had a computer-generated face designed in FaceGen 3.1 to contain strong cues to untrustworthiness or to appear neither trustworthy nor untrustworthy, as well as a neutral control face named Scott who was visually neither trustworthy nor untrustworthy named Scott (Todorov, Dotsch, Porter, Oosterhof, & Falvello, 2013; Todorov & Oosterhoof, 2011). After this familiarization procedure, we assessed participants' implicit impressions of the two targets using an AMP. Participants then learned behavioral information about the individuals that was either neutral or indicated trustworthiness before assessing their implicit impressions a second time. (Neutral information was also provided about the control individual.) The study thus had a 2 (Time: 1, 2) × 2 (Target: Joe, Scott) × 2 (Joe Facial Cues: Untrustworthy, Neutral) × 2 (Additional Information: Trustworthy, Neutral) design. At Time 1, when participants' evaluations were based exclusively on each target's facial appearance, they exhibited responses that successfully incorporated the trustworthiness information: those for whom Joe appeared to be visually neutral did not exhibit any implicit preferences between Joe and Scott, whereas those for whom Joe appeared visually untrustworthy exhibited significantly negative implicit evaluations of him.

At Time 2, when participants had learned neutral behavioral information about Joe, they continued to exhibit the same pattern of implicit responses as they had at Time 1. However, when they learned that Joe had gone out of his way to secure a neighbor's home against an impending hurricane (i.e., behavior that strongly implied trustworthiness), they showed a significant positive shift in their implicit evaluations of him relative to Scott. Notably, the extent to which participants exhibited responsiveness to the behavioral information indicating trustworthiness was *not* influenced by the extent to which Joe visually appeared untrustworthy (see Fig. 4); as a result, neutral

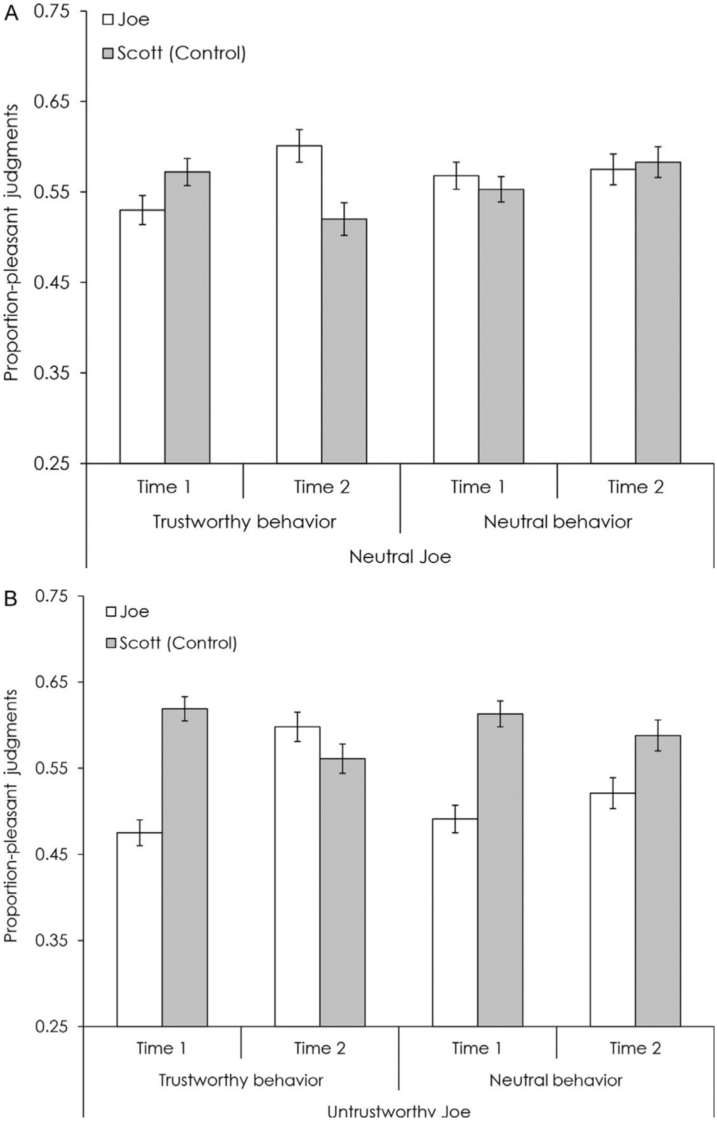

**Fig. 4** Implicit evaluations of Joe (target face) and Scott (control face) after learning neutral information about both (Time 1) and after learning additional information about Joe that was either trustworthy or neutral (Time 2). (A) Depicts the results for those participants in the condition in which Joe had a neutral appearance. (B) Depicts the results for participants in the condition in which Joe had an untrustworthy appearance.

Joe became significantly more positive than Scott after the additional trustworthy information, while untrustworthy Joe became directionally more positive than Scott. In fact, final implicit impressions did not significantly differ between untrustworthy and neutral-looking Joe, even though implicit impressions of the two versions of Joe did differ *prior* to the additional information.

## 4.2 Rewriting the Past: Revising Implicit Impressions Through Reinterpretation

Encountering new, important information about other people often leads us not only to incorporate that new knowledge into our impressions of them, but also to shift our understanding of what we have previously learned about them. Some revelations about a person—such as that he or she had an ulterior motive for past behavior, that earlier information was incomplete, or that our inferences were erroneous—demand such revision to past beliefs. We might find it to be an insufficient adjustment, for example, to continue to credit a high school coach for his work on the field after learning that he regularly abused children and used his coaching position to find victims. Likewise, it might seem inadequate to continue to begrudge a colleague for neglecting to contribute to a work fundraiser after learning that she quietly donates the majority of her salary to charity each year, leaving too little leftover for other donations. These shifts in the lens through which past events are understood not only add critical new information to our knowledge of a person, but also expunge earlier conclusions. A rational treatment of the information seems to compel a shift in understanding the past. This intuition suggests that *reinterpretation* may be particularly effective in updating initial implicit impressions—even (or especially) if those impressions are negative. As described earlier, Cone and Ferguson (2015) found evidence for a valence asymmetry such that positive implicit impressions were fully undone by diagnostic positive information but negative implicit impressions were merely attenuated by positive diagnostic information.

If first impressions arising from negative behavioral information are difficult to overturn with new, unrelated behavioral information due to the continued dominance of those earlier, diagnostic details, then might negation of the earlier details themselves prove more effective? Various theories and evidence suggest that merely negating (i.e., rejecting) information is not generally effective in changing implicit responses (Gawronski & Bodenhausen, 2006, 2011; Gawronski et al., 2008; Rydell & McConnell, 2006; see also Deutsch, Gawronski, & Strack, 2006; cf. De Houwer, 2014), at least when the negations come with anything more than a momentary delay after the initial information (Peters & Gawronski, 2011; cf. Boucher & Rydell, 2012)—though they may sometimes be effective if they are strongly elaborated (Petty et al., 2007), or when they are made more meaningful and important (Johnson et al., in press). Moreover, a strategy of attempting to merely "undo" prior learning may backfire due to ironic processing (Wegner, 1994); in other words, efforts to put the prior impression out of mind may counterproductively maintain it.

It may be the case, however, that a *combination* of negating an earlier impression and replacing it with a new one will be far more effective. Information that reinterprets the very basis of the initial impression may be particularly well suited for causing implicit revision, then, by undermining the initial conclusions (thus "rediagnosing" their meaning for understanding the target) and providing an updated impression, thereby allowing one to avoid dwelling on the negated earlier understanding of the events. This combination of "subtracting" an initial impression and "adding" a replacement simultaneously through reinterpretation is consistent with work on correcting effects of misinformation in memory that emphasizes the importance of countering misinformation and filling the "gap" that it leaves behind by providing an alternative (Lewandowsky, Ecker, Seifert, Schwarz, & Cook, 2012). It is also consistent with work exploring the reappraisal of negative emotions, which does not attempt to negate or counter-argue the emotion but instead attempts to replace it with one that is physiologically plausible (e.g., matches it in level of arousal; e.g., Brooks, 2014; see Gross, 2015 for review). Thus, a reinterpretation route to revising negative implicit impressions would suggest that when new information compels a reinterpretation of earlier behaviors, revision in the negative-to-positive direction may be robust. The important triggering condition for this mechanism, then, is to provide information which is not merely highly diagnostic, but also *strongly relates back* to earlier information that one had used in forming the initial impression.

Some evidence in support of this possibility comes from a study reported by Wyer (2010). Participants were presented with a story about a person whose actions seemed to conform to the stereotype of skinheads (e.g., tiredness ostensibly from partying, irritability), leading participants to form a negative implicit impression of the individual. Participants who later learned that this individual was in fact a cancer patient, and that the prior actions could be attributed to this fact, showed attenuation of their implicit evaluations, but only if they first had a chance to review the earlier story.

However, the results were limited in that revision occurred only after participants were given the opportunity to review the early details of the story, suggesting that change through reinterpretation may require elaboration (Petty et al., 2007, 2006). This leaves it an open question as to whether implicit revision could occur merely after learning the revelation (see also Wyer, 2016). Furthermore, because of the lack of comparison conditions or measures of reinterpretation, it remains unclear the extent to which participants in this study showed change due to reinterpretation in particular.

Several lines of our recent work have systematically examined the possibility that reinterpretation can drive fast revision in initially negative implicit evaluations, even without the opportunity to revisit the earlier learning (Kimball, Mann, & Ferguson, in preparation; Mann & Ferguson, 2015, 2017; Mann, Ferguson, & Axt, in preparation). In a set of seven experiments, Mann and Ferguson (2015) provided an initial demonstration that when new information prompts a reinterpretation of the events upon which a negative first impression of a person was based, implicit evaluations of that person can immediately reverse. Participants in these studies read a narrative about a man named Francis West who was described as breaking into and causing extensive damage to the homes of two of his neighbors when he noticed that they were not home, including breaking doors and windows, treading mud across the rug, and removing "precious things" from the bedrooms. This information produced correspondingly negative implicit impressions of Francis West relative to a set of control faces about whom participants had not learned anything. Next, in Studies 1a (using an AMP) and 1b (using an IAT), they either learned (in the control condition) one more action consistent with their earlier impression (he chucked rocks at the houses) or learned (in the critical condition) that the houses had in fact been on fire, and Francis broke in to find and rescue the children who he knew to be trapped inside—the "precious things" mentioned in the body of the story. This heroic detail not only provided information about Francis of a valence opposite to what they read before, but also fundamentally reframed those earlier details, casting seemingly negative behaviors in a positive light. Results showed that after new information, negative implicit impressions persisted in the control condition, but reversed in the reinterpretation condition—such that Francis West became significantly more implicitly positive than control faces.

Like the results reported by Wyer (2010), these results alone do not cleanly pinpoint the role of reinterpretation. To more finely isolate reinterpretation as a central element driving revision, in Study 2, Mann and Ferguson (2015) manipulated whether Francis was described performing heroics that reinterpreted the events in the earlier story (the fire scenario from Experiments 1a and 1b), or heroics that were unrelated (i.e., that he had jumped down onto subway tracks to rescue a fallen baby a few moments before the train would have struck them both). This resulted in a 2 (Time: 1, 2) ×2 (Target: Francis West, Control Faces) × 3 (Story Condition: Control, Unrelated Positive, Reinterpretation) design, and responses on an AMP served as the dependent measure. Crucially, though both heroic actions had been pretested as equally positive, only the heroic actions that provided a reinterpretation of the earlier details resulted in a reversal of implicit evaluations (see Fig. 5). This suggested that participants were processing the relation between the new information and the old, rather than just incorporating new, positive details into their prior impression of Francis. In particular, they were drawing inferences about the (in)validity of their prior assessments of the original information in light of the new evidence.

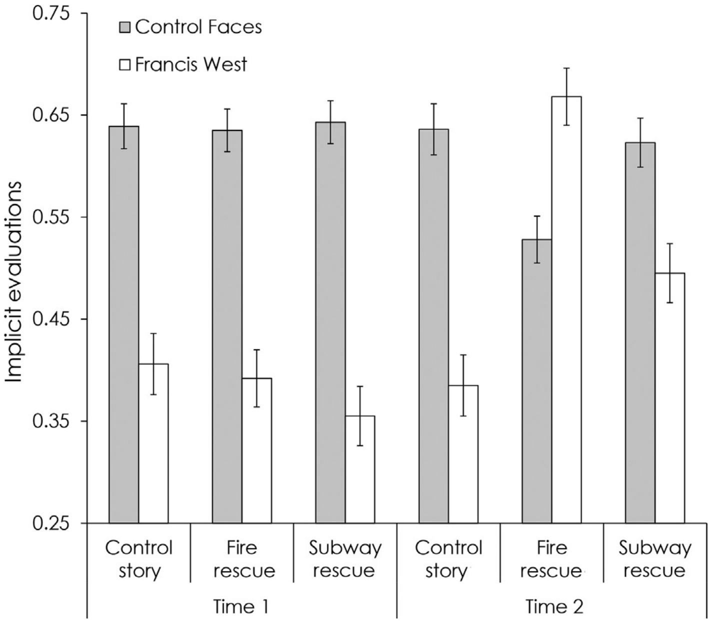

**Fig. 5** Implicit evaluations (AMP) by time, story condition, and target. *Adapted from Mann, T. C., & Ferguson, M. J. (2015). Can we undo our first impressions? The role of reinterpretation in reversing implicit evaluations.* Journal of Personality and Social Psychology, 108*(6), 823–849. http://doi.org/10.1037/pspa0000021. Experiment 2.*

In a further experiment (Study 3), we corroborated our view that these relational and validity assessments result from an active process that requires integrating new information into prior knowledge, finding that when participants learned the fire rescue information while rehearsing an eight-digit number (high load; Gilbert & Osborne, 1989), they revised their implicit impressions (on an AMP) far less than those under minimal load (two-digit number) or no load (no memorization).

Two other studies (Studies 4 and 5) took a different approach to examining whether reinterpretation best characterizes the nature of the processing that occurs in the fire rescue condition, relying on subjective introspection. To measure reinterpretation, participants responded to the question, "When you got the new information about Francis West a moment ago, how much did this new information *change the meaning* of Francis West's earlier actions?" Among participants receiving the "fire rescue" information, there was a strong correlation between subjective reinterpretation and revised, positive implicit impressions of Francis West on an AMP, even when controlling for the extent to which participants elaborated on the information more generally (Study 4). Furthermore, this self-reported reinterpretation mediated the difference in effectiveness of the subway rescue (unrelated positive) and fire rescue (reinterpretation) conditions (Study 5).

If reinterpretation is effective because it not only provides new countervailing information but also simultaneously undermines the negative interpretation of the earlier events, other interventions that both negate *and* replace earlier learning might be capable of producing similar revision. In a recent set of studies, Mann et al. (in preparation) tested this possibility by adding a new condition to the Francis West paradigm described earlier. In a new "negation + replacement" condition, participants were presented with the same subway rescue scenario from the study discussed earlier, describing heroic actions performed by Francis that were unrelated to his prior negative behaviors. Immediately above the text describing those heroic actions, however, was a new note informing participants that all of the information they had previously read about Francis was entirely false; Francis did not perform any of the actions (breaking-and-entering) they had read about before. This new condition, combining negation of earlier events with a counterattitudinal replacement, was compared to four others: a condition presenting one action consistent with the earlier negative impression (control), the fire rescue condition (reinterpretation), the subway rescue condition (replacement alone), and a condition containing only the statement that the earlier actions were in fact false (negation alone). Consistent with other work, the negation alone and replacement alone conditions did not lead to a reversal of implicit impressions (as measured with an AMP) of Francis (e.g., Gawronski et al., 2008; Peters & Gawronski, 2011). The combined negation-plus-replacement condition, however, produced a reversal of implicit evaluations, and did so just as strongly as the reinterpretation condition. A follow-up study using multinomial modeling replicated the equivalent levels of implicit revision in these two conditions and also found that change in both conditions produced similar shifts in the likelihood of activating a positive (vs negative) response to the face of Francis West when primed during the implicit measure.

Reinterpretation likely differs in many ways from the piecemeal negation and replacement examined by Mann et al. (in preparation), not the least of which is that the former seems to involve changing the meaning of earlier events, whereas the latter involves rejecting that those events ever took place. Nonetheless, the similarity of the results of these two approaches in reversing initially negative implicit evaluations of Francis West suggests that they may be members of a broader family of effective strategies for impression revision, ones that engage in some way with the meaning, interpretation, believability, or diagnosticity of earlier information *and* concurrently provide new information to fill any "gaps" that are left (or opened) by doing so (consistent with broader perspectives on debunking misinformation; see Lewandowsky et al., 2012).

If reinterpretation is to be understood as a change in the meaning of earlier information, another critical issue is the question of how much of the initial information actually needs to be retrieved, and in what detail. When encountering new details that reframe earlier learning about a target, a person might be able to recall nothing at all, the gist of earlier details in the absence of specifics, or all of the minute elements of those prior behaviors. If the ability for reinterpretation to reverse implicit evaluations hinges on reframing the low-level details of what was previously learned, this would require the ability to retrieve those details in the first place—something which may become particularly difficult as more time passes since the initial learning. Even with a loss of explicit memory for impressions or their sources, work shows that impressions can continue to exert influence on responses (Castelli, Zogmaister, Smith, & Arcuri, 2004; McCarthy & Skowronski, 2011; Todorov & Uleman, 2002). Little work has examined the potential for new implicit impressions to be undone after delays of any length, but Peters and Gawronski (2011) showed that even mere minutes made a difference in whether negations of earlier details led to reversals or smaller shifts in implicit evaluations. Longer delays may make retrieval of finer details even more difficult, as memory becomes more schematic and abstracted (e.g., Bartlett, 1932; Klein, Loftus, Trafton, & Fuhrman, 1992; Sherman & Klein, 1994). By providing time for memory stabilization to occur through consolidation (Dudai, 2004; McGaugh, 2000), longer delays may pose a further challenge for revision even if successful retrieval were to occur. On the other hand, recent work on memory reconsolidation has found that successful retrieval can allow even consolidated memories to be updated with new information (Hardt, Einarsson, & Nader, 2010; Lane, Ryan, & Nadel, 2015; Schiller et al., 2010, 2013; see also Mann, Cone & Ferguson, 2015).

Because engaging with earlier information seems to be a critical part of reinterpretation, we tested whether reinterpretation with the fire rescue information would still reverse implicit evaluations in the Francis West paradigm after a delay of 2 days (Mann & Ferguson, 2017). Additionally, we examined whether levels of recall for the details of the prior story on a 10-item quiz would be related to the size of revision, to test the possibility that revision would be the strongest among those with strong memory for the earlier details (allowing them to fully reframe those previously negative events to support a new, positive interpretation), and the alternative possibility that revision would be the strongest among those with weak memory for the earlier details (perhaps signifying a weaker negative impression in the first place). Finally, we chose to manipulate between subjects whether participants would even be presented with the recall quiz, to test whether merely reexposing participants to cues about the prior story would be enough to enhance the effectiveness of the reinterpretation-related new information.

Importantly, participants in the fire rescue condition showed significant reversal of their initially negative implicit evaluations (on an AMP) after encountering the new details 2 days later. The size of this revision was in the same range as that observed in our earlier Francis West studies (Mann & Ferguson, 2015) and was not moderated by whether the participants took the recall quiz, or the level of recall they evidenced on the quiz if they did take it (see Fig. 6). These results support the conclusion that implicit first impressions can still be reversed quickly by a small amount of new information days after initial learning, at least if the new information provides a reinterpretation of the basis of that impression. Interestingly, the extent of revision after reinterpretation was unrelated to recall. One reason why this may have occurred is that the reinterpreting information itself contained

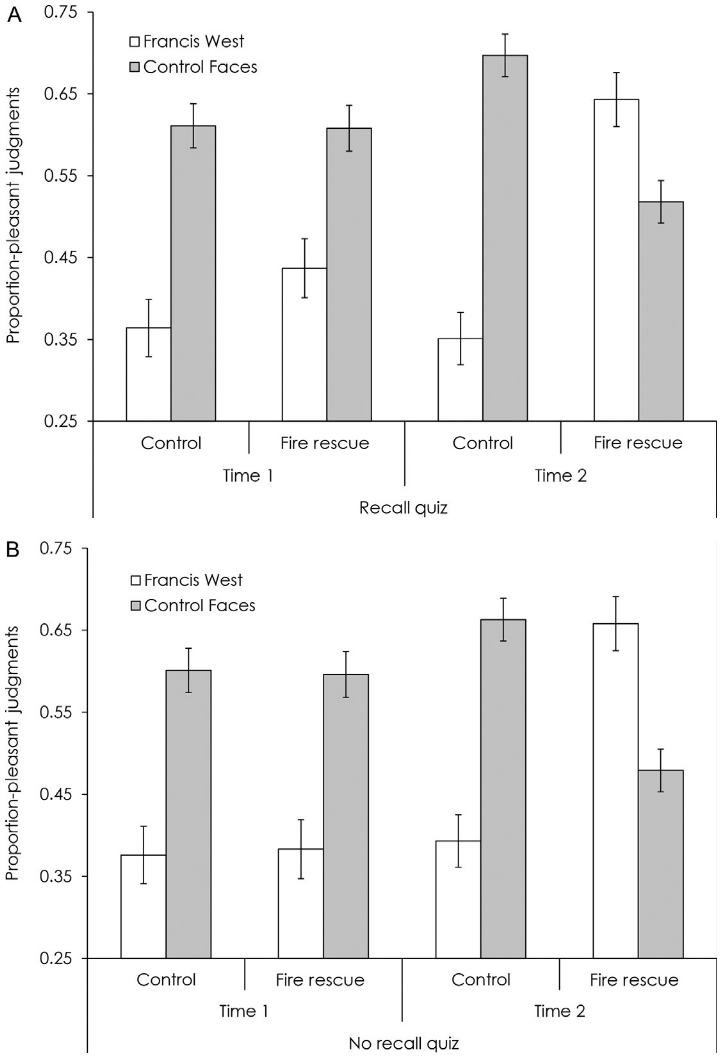

**Fig. 6** Implicit evaluations of Francis West and Control Faces immediately after learning the initial information about Francis West (Time 1) and 2 days later, immediately after learning the neutral or reinterpreting information during the second session (Time 2). (A) Depicts the results for those that completed a recall quiz. (B) Depicts the results for those that did not complete the recall quiz. *Adapted from Mann, T. C., & Ferguson, M. J. (2017). Reversing implicit first impressions through reinterpretation after a two-day delay.* Journal of Experimental Social Psychology, 68*, 122–127.* http://doi.org/10.1016/j.jesp.2016.06.004.

enough of a recall cue to remind the majority of participants of the gist of their earlier learning. In other words, the reinterpretation information described why Francis "broke into the adjoining … homes" and clarified the nature of the "precious things" that he removed, which may have itself reminded participants about details of the earlier story that they had previously failed to recall. If so, this suggests that the retrieval brought about by the new information itself may be sufficient for undoing first impressions through reinterpretation. Such built-in reminders may generally occur when reinterpretation-type information is encountered even after longer delays, but for now this hypothesis remains untested. It may also be the case that no recall occurs (or is needed) when the reinterpretation-related information is encountered; instead, the framing of the new information itself—in terms suggesting that it should take precedence over earlier impressions—may lead participants to replace their earlier impressions even in the absence of any memory for the sources of those impressions.

In our earlier section on diagnosticity, we described evidence that diagnostic behavioral information can overturn implicit first impressions that formed initially from visual cues in the face (e.g., obesity, attractiveness). Similarly, we have found in recent work that reinterpretation can also overturn first impressions formed on visual cues (Kimball et al., in preparation; Mann, Cone & Ferguson, in preparation). Evidence previously reviewed (Cone et al., in preparation) showed that if new behavioral information is seen as highly diagnostic, even negative implicit impressions formed from visual cues (e.g., obesity) can be overturned. Applying these ideas to reinterpretation, we have found that an individual with a prominent facial scar is initially evaluated implicitly as negative, but that evaluations flip to positive when it is revealed that the scar was acquired in the course of a heroic act (Mann et al., in preparation). However, similar to negative impressions formed from an individual's behaviors (Cone & Ferguson, 2015, Study 2; Mann & Ferguson, 2015), an *un*related heroic act (that the man saved children from getting hit by a train, unrelated to his facial scars) significantly shifted, but did not succeed in reversing, implicit impressions of the individual with the facial scar. However, evidence in this line of work is still accumulating, such that the magnitude of implicit updating in these two heroic behavior conditions has sometimes been equivalent. Further research will reveal whether reinterpretation—vs just extreme diagnosticity—is necessary for reversing implicit impressions of visual cues such as a scarred face. This provides further evidence of the idea that contrary to proposals of the special intransigence of impressions based on visual cues (McConnell et al., 2008), implicit first impressions based on faces may be malleable. Though speculative, this may be due to the low assessed diagnosticity of visual information. It is when negative first impressions are formed from (more diagnostic) behaviors that the need to directly engage with the earlier information (i.e., reframe or otherwise negate it) may be most pressing, because participants may be more hesitant to set aside immoral behaviors compared to negative visual cues (or positive behaviors; Cone & Ferguson, 2015; Skowronski & Carlston, 1989).

In another line of studies, we have examined whether reinterpretation can be effective at reversing implicit impressions formed from a *combination* of behavioral and visual information. Using the Francis West paradigm, we found that a reversal of initial negative implicit evaluations still occurred even when the image of the character clearly identified him as a black male, a group typically stereotyped as hostile and criminal, consistent with the typical interpretation of the character's initial behaviors (Kimball et al., in preparation). Across two studies, the race of the target (white or black) did not affect the size of revision, and, in both cases, final implicit evaluations of the character reversed to become significantly positive relative to control faces on an AMP. This was true even when the story appealed directly to black stereotypes of criminality, by ending with the revelation that the police arrested the character for an outstanding warrant upon arriving at the scene of the fire, despite his heroic actions. When, in a third study, the motives of the hero were called into question by revealing that he was later found at a pawn shop selling an expensive piece of jewelry removed from the house, revision was curtailed, but for both white and black versions of the character. These results suggest that when new information prompts a reinterpretation of prior behavior, revision of implicit impressions of the target can occur regardless of whether the initial behaviors were consistent with prior stereotypes. Importantly, this is the case even when the initial impression is supported by both the initial behaviors *and* stereotypes based on group memberships that are apparent from visual cues present in the face.

To sum up, across these multiple lines of work, reinterpretation seems to be an effective way to reverse implicit evaluations. It seems to be a particularly useful strategy when a prior impression was formed through the negative *behaviors* of the target, which may continue to be seen as diagnostic even in light of unrelated positive actions, unless directly addressed by the new learning.

## 4.3 Change We *Must* Believe In

A final characteristic of information that plays an important role in rapid revision of implicit impressions is the extent to which information is seen as believable. Information that we acquire about others carries with it a perception of its certainty or veracity. Although we are confident in some of our inferences about others (e.g., you discover incontrovertible evidence that your romantic partner is having an extramarital affair), other kinds of revelations lead us to harbor lingering doubts (e.g., your romantic partner appears to be lying about why he has arrived home late several nights this past week). Extensive work in the persuasion literature indicates that beliefs about the credibility and validity of persuasive messages is a key determinant of (explicit) attitude change (Petty & Cacioppo, 1986). In this line of work (Cone & Ferguson, in preparation), we assess the extent to which such validity assessments similarly play a role in implicit change (Briñol, Petty, & McCaslin, 2009).

To the extent that such inferences do indeed play a role in implicit impression updating, another potential reason why previous work has failed to uncover evidence for rapid revision is that the information carried a degree of uncertainty that prevented it from fully resonating with participants. In Gregg et al.'s (2006) work that had participants form impressions of the *Niffites* and *Luupites*, for example, the strategies that were used to attempt to overturn prior learning also communicated important information about the believability of the story. In the "mistake" paradigm, for example, the participants learned that they were randomly assigned to the incorrect condition: whereas they learned that the *Niffites* were good and the *Luupites* were bad, they were meant to learn that the *Niffites* were bad and the *Luupites* were good. Yet, learning this switch in group assignment not only communicates that the groups have switched roles, but also that the assignment to condition is *arbitrary*, determined entirely randomly by the experimenters rather than because of anything meaningful about the characters of the groups. Moreover, this also reveals that the groups—or at least the names assigned to them—are fictional and thus less meaningful than participants may have suspected.

Similarly, in another version of their experiment, Gregg and colleagues provided participants with a counter-narrative that suggested reasons why the groups may have switched characters. Although this counterattitudinal information failed to change implicit evaluations, it is not clear if this failure occurred because such information *cannot* cause rapid changes, or if the participants were reticent to accept that the switches in character described in the narrative had actually occurred—whether it was *plausible* that the *Niffites* had originally been morally good but, through a series of events that shaped their character as a group, later became morally corrupt. Thus, it is not clear whether participants actually believed the revisions they were asked to learn, and perhaps this is an important prerequisite for changes to resonate more thoroughly in implicit evaluations.

One interesting implication of this line of reasoning is that reversals of implicit evaluations are more likely for attitude objects that we believe can rapidly change character or evaluative connotations. Video game characters, for example, are sometimes programmed to exhibit changes in their roles depending on the state of the game. Consider Pac-Man, in which, in the regular mode of the game, ghosts are adversarial and must be avoided by the player's character, whereas after the player's character consumes a power pellet, the ghosts instantly become an opportunity to earn extra points by chasing and eating them. Notice that, contrary to these changes being random or hypothetical, the types of changes in role in this game are quite meaningful and influence how participants interact with the attitude objects. If people readily accept that these kinds of changes in role can occur, can their implicit evaluations of these attitude objects rapidly shift in line with these changes?

To test this possibility, we (Cone & Ferguson, in preparation, Study 1) described a version of a video game to participants in which they were told that they would encounter novel objects called *wugs*, which looked like a triangle with eyes in which one half was white and the other half was blue (see Fig. 7).

In a 2 (Round Order: Approach First, Avoid First) × 2 (Round: Approach, Avoid) × 2 (Target: Wugs, Neutral) preregistered design, the wugs were first described as either adversarial or helpful. We then assessed

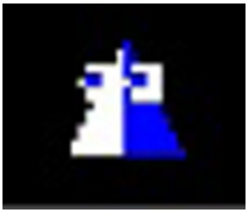

**Fig. 7** A *wug*.

participants' implicit evaluations of the *wug* objects using an AMP. Next, participants were told that the second round of the game would have them interact again with wugs, but that the meaning of the wugs would change: if they initially learned that the wugs were positive, they were told that in the next round they would instead be negative and vice versa. Whereas earlier studies making use of novel attitude objects have generally found that implicit evaluations are more easily done than undone, we found instead that participants' evaluations of the wugs were both easily done *and* undone: participants exhibited very rapid shifts in their implicit evaluations in line with the information that they learned about the wug objects at each time point.

Another important implication of this line of reasoning is that there may be individual differences in the extent to which information resonates at the implicit level as a function of the extent to which people see the information as believable or not. Those that readily accept new information about someone as plausible or true ought to exhibit very rapid changes in their implicit responses, whereas those who have reservations about the veracity of something they have learned ought to exhibit relatively less implicit change. When participants learned the story of the *Niffites* and *Luupites* used in Gregg et al. (2006) research, we (Cone & Ferguson, in preparation, Study 2) found that the extent to which new information suggesting the groups had changed character was rapidly incorporated into their implicit responses was predicted by a measure of their beliefs about whether the events described in the narrative could have happened in real-life, seemed plausible, and made sense (see Fig. 8). Interestingly, however, participants' more general beliefs about whether it is possible, in principle, for groups to change character—rather than the extent to which they believed that this had happened with the Niffites and Luupites, specifically—did *not* predict the extent to which they readily formed and changed their implicit responses in light of the new information. Thus, it is important not only to believe in the capability of groups to exhibit changes in their character, in the abstract, but also to believe that *these particular groups* have, in fact, changed—though, of course, the former belief may be a prerequisite for the latter.

A final important implication is that manipulations of the believability of the very same information ought to impact the extent to which rapid implicit revision occurs. In particular, we have tested how the *way* that we acquire information about someone can influence the extent to which that information causes rapid revisions in implicit responses. In one study

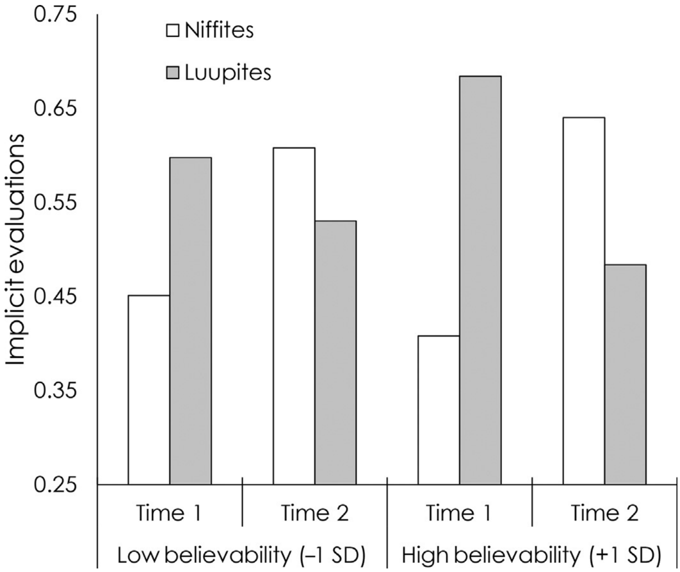

**Fig. 8** Implicit evaluations (AMP) based on a regression equation prediction in which participants' assessments of the believability of the narrative were entered as a predictor of implicit evaluations as a function of time and target. (Niffites, in this study, were always initially described as negative.) Evaluations are shown for believability values corresponding to −1 SD (4.03) and +1 SD (6.26). Higher scores indicate a more positive implicit impression of the target.

(Cone & Ferguson, in preparation, Study 3), just as in our earlier work, participants completed a paradigm in which they learned 100 positive pieces of information about a novel stranger, and were then asked to consider one additional piece of information about him: he had been arrested for domestic abuse several years ago. Some participants were asked to imagine that this information came from a reliable source and there had been independent verification of the arrest (i.e., a police report). Other participants were asked to imagine that one of their coworkers had told them this information and that she had an ulterior motive: her friend had been dating Kevin and the relationship had ended in a messy break-up, so she may have aimed to spread a false rumor about him. Otherwise, the details of the information they learned were identical. Thus, the study was a 2 (Time: 1, 2) × 2 (Target: Kevin, Control Faces) × 2 (Believability: Reliable Source, Rumor) design.

Because word-of-mouth communications are seen as less reliable (DiFonzo & Bordia, 2007; Schachter & Burdick, 1955), information that is explicitly described as a rumor becomes less believable (Kamins, 1997). If the believability of information plays a role in implicit attitude change, then when the information is described as gossip, it should have less impact on implicit evaluations than when it is described as originating from a reliable, verifiable source. And, indeed, this is precisely what we found. Whereas those who thought Kevin's arrest for domestic abuse came from a reliable source exhibited implicitly neutral evaluations of him (again, on an AMP), those who learned it was a rumor continued to exhibit significantly positive implicit evaluations of him, despite learning something that they would have viewed as extremely negative if true.

As further support that these effects were caused by perceptions of the believability of the information, using a multiple mediator model (Hayes, 2013), we found that measures of believability (e.g., "How likely do you think it is that Kevin actually engaged in this behavior?") and diagnosticity each independently predicted the extent of evaluative change observed in the two conditions. We also conceptually replicated this finding in a preregistered design using a more extreme Time 2 behavior in which participants learned that, instead of being arrested for domestic abuse, Kevin had been arrested and convicted of child molestation (Cone & Ferguson, in preparation, Study 4). Like the previous study, we found that a self-report measure of believability successfully explained the relation between the condition manipulation and the extent of evaluative change at Time 2. However, contrary to the previous study, diagnosticity did not also independently mediate these effects, suggesting that believability has an especially important role to play for very extreme impression-inconsistent revelations about someone (see Fig. 9).

It is worth mentioning that even when information was described to our participants as unreliable, it still had a significant impact on participants' implicit evaluations. Thus, open questions remain about the extent to which the truth value of information matters for implicit revision. Nonetheless, these findings present a challenge to attitude theories that propose that implicit evaluations are largely incapable of successfully responding to the truth value of the knowledge we acquire about others (Gawronski & Bodenhausen, 2006, 2011; McConnell & Rydell, 2014; cf. Peters & Gawronski, 2011). Whereas some work has suggested that implicit evaluations can be impacted by information irrespective of its truth and falsity, our findings suggest that believability may play a key role in whether and when implicit revision occurs.

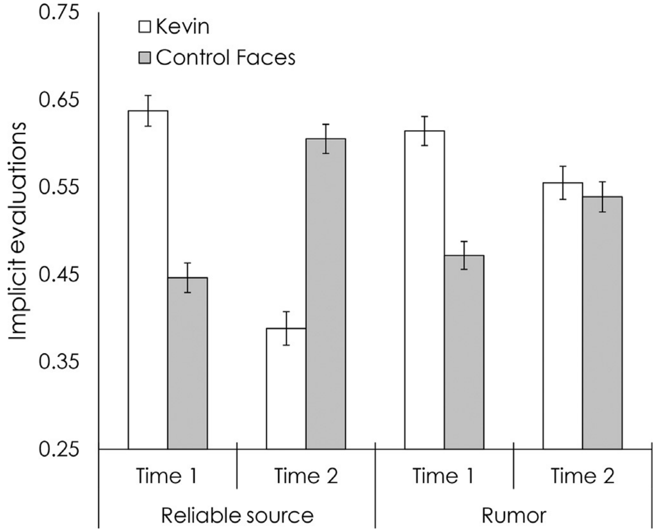

**Fig. 9** Implicit evaluations (AMP) of Kevin and Control Faces as a function of time and whether the Time 2 behavior (conviction of child molestation) was described as originating from a reliable source or from word-of-mouth communication.

# 5. COMMON QUESTIONS, MISPERCEPTIONS, AND THEORETICAL ISSUES

All of these lines of work converge on the notion that first impressions may be more capable of rapid revision at the implicit level than previous theorizing has posited. In this section, we discuss some common questions about our approach and some key theoretical issues that our work highlights.

## 5.1 Three Routes to Change or One Route Under Different Guises?

Though we have pursued each of these three different mechanisms of rapid change independently and have provided evidence for each of their unique contributions to the extent to which minimal counter-evidence can rapidly change implicit evaluations, they are clearly not entirely distinct processes. For example, although we found that highly diagnostic revelations led to rapid reversals of implicit evaluations, this is likely predicated on the fact that participants did not question the veracity of the Time 2 information; we took pains to communicate to participants that the information might be inconsistent with their previously developed impressions of the target, but that it was nonetheless characteristic of him. Indeed, these mechanisms are likely correlated with one another. After all, as information becomes more implausible, it will also be seen as less revealing of a person's true nature or character. Similarly, one reason why reinterpretation had such a powerful impact on implicit evaluative change is that the original (highly diagnostic) information about the target was successfully supplanted by a new (highly diagnostic, but in the opposing direction) understanding of the meaning of that information.

As alluded to in the previous section, in some of our recent work (e.g., Cone & Ferguson, in preparation), we have measured more than one of these three mechanisms simultaneously and have found them to be empirically separable, thus suggesting that they are indeed distinct processes. In our work focusing on the role of believability of information, for example, we measured not just participants' assessments of whether the information was true, but also simultaneously their beliefs about the diagnosticity of the revelations. Not surprisingly, believability and diagnosticity were positively correlated: as information became less plausible to participants, it also simultaneously was seen as less reflective of the target's character. Yet, there was still evidence of independent contributions of each process to implicit updating.

Similarly, in the line of work exploring the role of causal attribution processes and situational constraint information in implicit evaluative change described earlier, we simultaneously measured participants' beliefs about the diagnosticity of the essay-writing behavior and their reports of the extent to which learning that the essay-writing behavior had been situationally constrained or not led them to reinterpret the meaning of the earlier information. In this line of work, interestingly, diagnosticity beliefs and assessments of the extent to which they reinterpreted the target's behavior were not significantly correlated, suggesting perhaps that they made independent contributions to people's beliefs about the new information, at least in the context of this paradigm. Ultimately, more research will be necessary to disentangle the interrelations among these different processes and their relative contributions to implicit attitude change.

## 5.2 Are Attitudes Really *Changing*?

As we mentioned earlier, a well-established idea in the literature is that implicit evaluations are heavily context dependent (for review, see Blair, 2002; Gawronski & Sritharan, 2010), exhibiting momentary shifts as a function of the particular facets of a mental representation of an attitude object that are activated at the time of elicitation. For example, because "liking is for doing" (Ferguson & Bargh, 2004; see also Moors & De Houwer, 2001), implicit evaluations are strongly influenced by one's current needs and goals: smokers implicitly evaluate cigarettes more positively immediately before smoking relative to immediately after, when their nicotine needs have been fulfilled (Sherman, Presson, Chassin, Rose, & Koch, 2003); drinkers implicitly evaluate thirst-quenching stimuli such as bottles of water more positively when they are thirsty than when quenched (Ferguson & Bargh, 2004); and eaters implicitly evaluate food stimuli more positively when hungry than when sated (Seibt, Häfner, & Deutsch, 2007). Similarly, implicit evaluations have also been shown to be highly sensitive to recently primed thoughts, such as calling to mind concepts related to egalitarianism vs ingroup loyalty (Zogmaister, Arcuri, & Castelli, 2008).

A fair question, then, concerns the extent to which our findings are merely additional empirical demonstrations of this contextualization process. Could we have already predicted that implicit evaluations would rapidly shift in line with new evaluatively inconsistent revelations on the basis of this prior work? This is a particularly important consideration for our theorizing because it has strong implications for the extent to which implicit evaluations have indeed *changed*—that is, whether the implicit evaluations formed on the basis of earlier information have been "erased" or could instead resurface again in a future context that is similar to the one in which initial learning took place.

Consider how the process of contextualization unfolds as we learn additional information about someone. An important assumption that is made in theories of contextualization is that when new impression-inconsistent information is encountered, a first recourse is to seek out salient context cues that can help to explain the discrepancy (Gawronski et al., 2010). If Jeff is consistently prickly and disagreeable at the office, you may be surprised to discover that he is consistently a charming conversational partner at Happy Hour. To help explain this expectancy violation, we might pay closer attention to the context in which each behavior occurs, identifying the essential differences in the two contexts that may explain the different behaviors, and conclude that "Work Jeff" is negative and undesirable, whereas "Social Jeff" is positive and desirable. Gawronski et al. (2010; see also Gawronski & Cesario, 2013; Gawronski et al., 2014; Rydell & Gawronski, 2009) have demonstrated that this kind of process can lead to multifaceted, contextualized mental representations in which implicit evaluations elicited in the work context will be, on the whole, negative, whereas implicit evaluations elicited in the pub context will be, on the whole, positive.

This line of reasoning has important implications for our assessments of the speed with which revision can occur. It is still unclear how much prior learning is necessary for a multifaceted, contextualized mental representation of the sort studied by Gawronski and colleagues to emerge. If it is only through extensive exposure to Jeff in *both* work and social contexts that we come to understand when we are likely to have positive or negative interactions with him, then we would not necessarily say that implicit evaluations have developed or changed *rapidly* in the way that we mean here, nor would we say that if we elicited evaluations of Jeff immediately before and after a change in context—say, at 4:58 and 5:02 pm—that the implicit evaluations elicited in these contexts are the products of rapid learning or that the mental representation of Jeff after 5 pm is now categorically different than what it was prior to 5 pm. In these kinds of situations, it would *appear* that evaluations of Jeff are changing rapidly, but such a conclusion would be unwarranted, because such seemingly rapid shifts would have been possible only due to extensive prior learning—learning that may indeed operate solely on the basis of slow-learning associative mechanisms, as many researchers have posited (e.g., Rydell & McConnell, 2006; see Sloman, 1996; Strack & Deutsch, 2004).

Thus, to more adequately test whether rapid change is possible, it is crucial to assess the extent to which there may be unseen (slow-learning) mechanisms that could have guided people's implicit responses toward the targets in our studies. An especially important aspect of our methodology that, in our view, helps to shed light on this question is that all of our paradigms make use of novel attitude objects. Because we carefully control exposure to the targets in our studies, we have a greater understanding of individuals' entire evaluative history with the attitude objects, meaning that when participants learn something decidedly negative about him or her, we can be relatively certain that this piece of information is the *only* negative thing that they have encountered about him or her. Thus, if participants immediately express implicit negativity, we can be more certain that it is the product of the recently encountered information and, thus, more likely to reflect rapid learning processes. This methodological feature of our work helps to reduce the likelihood that our results are unwittingly the product of unseen prior slow-learning mechanisms (for similar argument, see Fazio, 2007; Fazio & Olson, 2003; Gregg et al., 2006).

Intuitively, it would also be especially surprising if the processes that operate in contextualization effects of the sort studied in prior work also applied to the kinds of revelations used in our paradigms. In a sense, almost by definition, what it means for a revelation to be diagnostic is that we do *not* go hunting for explanations to explain the inconsistencies in our evaluative exposure to someone. Rather, the earlier information immediately becomes irrelevant to our implicit impressions and is thus heavily discounted. If we find Jeff to be very likeable when we have encountered him at parties but later discover that he regularly abuses his children, we would not search for context cues that explain when Jeff's behavior has positive vs negative evaluative connotations. He is a bad *person*, not someone who contains inconsistencies that demand explanation. Similarly, information that causes prior learning to be reinterpreted is not evaluated in isolation, nor does it lead to expectancy violation; rather, the reinterpreting information itself provides a plausible explanation for why the inconsistency arose in the first place: the new information casts the earlier information in a different light and leads to the realization that it was incorrect or based on faulty inferences.

Recent work by Brannon and Gawronski (in press) has directly tested the role of contextualization processes in two of our rapid revision paradigms. These researchers conducted direct replications of both Cone and Ferguson (2015) and Mann and Ferguson (2015), in which they independently manipulated the background color during the initial learning task, during the presentation of subsequent diagnostic or reinterpreting information, and during measurement. If the changes induced in our paradigms are the result of contextualization processes, then, replicating prior work, the implicit evaluation that is elicited at the end of the experiment should be influenced by the background color during initial learning and during measurement. However, Brannon and Gawronski (in press) found no evidence of any effects of context on any of the changes induced across the two experiments, indicating that the rapid revisions observed were context independent.

Thus, all of the evidence to date suggests that contextualization is not a key driver of the changes observed in our paradigms, with important implications for inferences about the speed with which implicit evaluations can be updated and the robustness of the changes induced.

## 5.3 Are the Changes Induced in Our Paradigms Durable?

If the changes that are induced by the information provided in our paradigms are not caused by context effects, then we should also expect that the information we provide in our experiments will result in more sweeping, longer-term changes in people's implicit responses toward someone—changes that are context independent. Empirical investigations of the durability of changes to implicit responses bear this out. Mann and Ferguson (2015, Experiment 6) directly examined the effects of reinterpretation on implicit updating. Participants read through the story of Francis West, forming a negative implicit impression of him from the collection of seemingly negative behaviors, as measured on a subsequent AMP. They then received either the fire rescue reframing (reinterpretation) or control information, and had their implicit impressions measured again, showing the same reversal of implicit impressions in the reinterpretation condition as participants did in the earlier experiments.

Unlike the previous experiments, however, participants in this study were recontacted approximately 3 days after the conclusion of their previous session and invited to take part in a brief follow-up for bonus compensation. If they agreed, they were reminded that they completed a study 3 days earlier about a man named Francis West, which they would be asked to answer some questions about in a few minutes. After this brief reminder, they were then immediately prompted to complete the AMP one final time, measuring implicit evaluations of Francis West and Control Faces. The results revealed that in both the control and reinterpretation information conditions, implicit evaluations had persisted across the 3-day span: In the control condition, Francis West remained implicitly negative relative to control faces, and in the reinterpretation condition, Francis West remained implicitly positive relative to control faces.

This initial evidence of durability over at least a 3-day period suggests, in our view, the durability of implicit revision. Of course, future work will be required to systematically examine the varied conditions and contexts under which a revised implicit impression will (and will not) impact judgments and behavior, and any potential upper limits on the temporal durability of these novel implicit impressions (over even longer spans of weeks, months, or even years).

## 5.4 Practical Applications

The findings that we have outlined indicate that implicit evaluations may be more easily revised and updated than previously assumed—indeed, sometimes exhibiting full reversals immediately after considering just a single piece of countervailing information—and thus join other recent work that suggests that implicit and explicit evaluations may not inevitably exhibit distinct learning characteristics (see Ferguson et al., 2014; Ferguson & Wojnowicz, 2011; Wojnowicz et al., 2009).

However, despite these important theoretical advances, there are two unique features of our work that might seem to prevent their generalization to the kinds of evaluative impressions that dominate everyday life: our reliance on novel targets, and our use of extreme behavioral information. We consider each in turn.

### *5.4.1 Novel Attitude Objects*

Our research to date has made exclusive use of novel attitude objects with which participants have had no prior exposure. This was a theoretically necessary feature of our experimental designs in order to ensure that the changes we observed at different time points could not be attributed to elements of prior learning. However, an important practical question concerns the extent to which the inferences we have drawn on the basis of our paradigms using novel targets are generalizable to more familiar people or groups.

Of course, claims about the inescapability of first impressions—even those based on little exposure or information—are ubiquitous in both popular culture and in recent empirical research (e.g., Gunaydın et al., 2017). Thus, our contention is that our findings do indeed overturn some important notions about the nature of first impressions and the extent to which they exert an undue influence on our interactions with others; perhaps all of our efforts to carefully control the first impressions others have of us are less important than colloquialisms, self-help books, and other aspects of popular culture might suggest.

More broadly, we are cautiously optimistic that our work can be successfully generalized to well-known targets. Although it may be maladaptive for single pieces of information to completely overturn implicit evaluative impressions toward those for whom we have very large amounts of prior experience, our theorizing nonetheless suggests that information that possesses the kinds of properties we have outlined earlier—diagnosticity, reinterpretability, and believability—ought to be more potent than information that fails to possess these properties. For example, finding out that a highly familiar, disliked, obnoxious, and aggressive coworker was ruthlessly bullied as a child and now donates his time to bullied youth may be enough to cause a reinterpretation of a wide array of past behaviors, thus resulting in durable updating—a possibility that we are exploring in ongoing work (Mann & Ferguson, in progress). Whereas we might expect first impressions toward novel attitude objects to be overturned with just a single instance of behavior that captures these properties, it could be that more well-established attitude objects require a *series* of pieces of these kinds of information to result in significant change. If well-established impressions are largely evaluatively mixed (even for highly positive targets in our lives such as our romantic partners; see Zayas & Shoda, 2015), it could be that a powerful series of evaluatively consistent pieces of information can be rather consequential.

To be sure, one important distinction between novel and well-established attitude objects is that encountering examples of behaviors that we deem to be both counterattitudinal and highly diagnostic may be decidedly unlikely to occur. This is, in part, because we may *want* to maintain particular attitudes toward those we (think we) know well, and, as we have argued elsewhere (Cone & Ferguson, 2015), elements of motivated cognition (e.g., see Sharot & Garrett, 2016) may serve to prevent new information from being imbued with a sufficient amount of diagnosticity to result in rapid implicit revision (“Well, I know she *means* well”).

### 5.4.2 Extremity of Information

Another aspect of our work that is perhaps unlike the typical kinds of information we encounter in our everyday lives is that most of the counterattitudinal behaviors we use in our work are decidedly extreme and perhaps less like the kinds of information that we typically encounter about others. Still, we *do* experience situations in which we encounter surprising or unexpected revelations about someone—say, politicians who we later learn have been framed, or convicts who are later exonerated with incontrovertible DNA evidence—and our work would suggest that these kinds of situations can exert a powerful and durable impact on people’s implicit responses. Still, an open question concerns the extent to which extremity is a necessary condition of implicit evaluative change. Although our work has been focused on achieving rapid *reversals* of implicit responses—that is, changes in the sign or valence of the implicit response—there is no reason to think that we should not expect smaller-scale but nonetheless significant changes in response to less extreme information. Indeed, changes in implicit impressions in response to somewhat less extreme kinds of information may be a more frequent occurrence in daily life, relative to more dramatic (but rarer) reversals.

### 5.5 Do Newly Revised Implicit Evaluations Successfully Predict Behavior?

Even if implicit evaluations are capable of rapid, durable revision, as observed on indirect measures of attitudes, they would be of little use to us if they did not also have implications for our subsequent behavior. As we indicated earlier, our view is that recent reviews of the effects of implicit revision on behavior have some important limitations (Forscher et al., 2016; Lai et al., 2016). We think it is more informative to test whether actual learning leads to behavioral change. Our initial work on this question has suggested that recently revised implicit evaluations have implications for our behavioral intentions toward social targets. Cone and Ferguson (2015, Study 5) had participants engage in an initial learning task followed by the diagnostic revelation that Bob had been recently convicted of child molestation. After this discovery, participants were asked to imagine that Bob was considering moving into their neighborhood and were asked to voice their behavioral intentions to organize their neighbors to prevent it from happening. When entered simultaneously into a regression, both participants' implicit and explicit evaluations of Bob were significant (unique) predictors of their answer to this question. Notably, however, their initial (positive) implicit and explicit impressions of him were *not* significant predictors, suggesting that participants' updating of their impressions also successfully led to updates in their intended behavior toward Bob.

These initial findings thus suggest that impression-inconsistent revelations about someone can be rapidly incorporated into implicit (and explicit) evaluations of that person and then have immediate implications for one's behavioral intentions toward him or her. However, a more systematic investigation of when revised implicit impressions do, and do not, lead to changes in behavior is one of the most pressing next steps in this line of work.

## 6. SUMMARY AND CONCLUSIONS

For most of the last 2 decades, social cognition research has suggested a striking divergence in the malleability of our first impressions of individuals and groups. Many findings have indicated that whereas our *explicit* evaluations are immediately sensitive to, and reflective of, our current, most up-to-date beliefs, our *implicit* evaluations are stubbornly entrenched in prior learning, resistant to new facts and evidence. This widespread claim about the malleability of evaluations—one of the most central constructs in the field of social psychology—has had several implications. First, it is aligned with the dominant view of evaluations and attitudes as emerging from two fundamentally different kinds of processes or systems. Automatic, implicit, or "System I" processes are assumed to be resistant to change, whereas deliberate, explicit, or "System II" processes are assumed to be immediately alterable and reflective of recent learning. From this perspective, empirical evidence of the "stickiness" of implicit evaluations has appeared to provide confirmation for dualist approaches to evaluation and cognition more generally.

A second major implication has been pessimism about being able to change prejudice and bias against outgroups (see Bargh, 1999). Considerable evidence suggests that we commonly implicitly evaluate outgroup members in a negative manner. We can be strongly implicitly biased against other people's group identities, roles, and characteristics even as we consciously and vigorously embrace and strive for egalitarianism. Findings showing that implicit evaluations cannot be (easily) changed (e.g., see Lai et al., 2016) seem to foster a curious divide between what we earnestly and effortfully want and what our more immediate impulses reveal. This divide has been characterized as one of the most pressing, difficult problems to solve in contemporary society (Banaji & Greenwald, 2013).

A third implication of the claim that implicit evaluations are resistant to change is a challenge to functional views of human memory. How could it be that we would have impressions that are so resistant to relevant new facts? How could it be functional to be unable to easily update these impressions in the face of strong and clear counter-evidence, especially assuming that implicit evaluations are partly generated in the evolutionarily older parts of the brain (and thus have been fine-tuned over millennia and likely conserved across species)? This poses a dilemma in accounting for the utility of such a seemingly error-prone and inflexible manner of responding rapidly to the world.

The work we have reviewed here—much of it published or conducted over the last several years—suggests a strikingly different picture of how our evaluative impressions of others can be updated. Although this work is still in its infancy, we have summarized evidence showing that there are circumstances in which implicit evaluations can be instantly, durably, and robustly updated—even reversed in valence—given that the new information about the target is believable and diagnostic, and if it undoes the original evaluative meaning of the evidence. This has important implications for the three issues we described earlier. First, collectively, these findings challenge strict dualist approaches that assume certain learning characteristics of implicit vs explicit evaluations. To be sure, our findings could be explained by dualist approaches. But they might also be explained with single-process models that do not invoke any assumptions about differences in representational format, process, or structure between the two types of evaluations (e.g., Ferguson et al., 2014; Wojnowicz et al., 2009). We view this as an exciting opportunity to further refine theoretical explanations for human evaluative processes. The most pressing topic for us in this regard is to explain why explicit and implicit evaluations sometimes diverge so markedly. If they do not necessarily emerge from different processes or systems, why are they sometimes so strongly dissociated?

With regard to the second issue, there is still much work to be done in examining how implicit bias toward known groups can be reversed. These empirical demonstrations have exclusively made use of novel targets, rather than with well-established groups, which are already associated with rich and extensive arrays of memories. There has been good reason for this focus given the priority to show actual new learning rather than the mere reactivation of previously learned information. However, we acknowledge that the potential for implicit updating may be limited to very specific conditions with novel targets. And recent findings suggest the difficulty in changing implicit evaluations toward known groups (e.g., Lai et al., 2016; see also Cao & Banaji, 2016). Yet, we find reason to be hopeful. If there are newly discovered ways of overturning implicit evaluations toward novel targets, this means that there is no inherent process-based limitation in updating implicit evaluations per se. It could be that applying the lessons from this recent work—on the importance of believability, diagnosticity, and reinterpretation—will open new doors to the prospect of changing implicit biases toward known groups as well.

Finally, although this work is still in its early stages, it raises the promise of a more functional memory and evaluative system than has sometimes been assumed based on the last 2 decades of research. Rather than leading us astray, and guiding us to act on outdated and rejected beliefs, implicit evaluations might be more fine-tuned to the *meaningfulness* of evidence around us than we previously realized. Far from being chronically untethered from our conscious thoughts and beliefs, implicit evaluations may at times be remarkably well calibrated with careful, deliberate, and precise estimates of what we believe to be true of the world.

## REFERENCES

Agren, T., Engman, J., Frick, A., Bjorkstrand, J., Larsson, E. M., Furmark, T., et al. (2012). Disruption of reconsolidation erases a fear memory trace in the human amygdala. *Science*, *337*(6101), 1550–1552. http://doi.org/10.1126/science.1223006.

Alaei, R., & Rule, N. O. (2016). Accuracy of perceiving social attributes. In J. A. Hall, M. Schmid Mast, & T. V. West (Eds.), *The social psychology of perceiving others accurately* (pp. 125–142). Cambridge, United Kingdom: Cambridge University Press. http://dx.doi.org/10.1017/CBO9781316181959.006.

Albarracin, D., Johnson, B. T., & Zanna, M. P. (2005). *The handbook of attitudes*. Mahwah, NJ: Erlbaum Publishers.

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.

Amodio, D. M., & Ratner, K. G. (2011). A memory systems model of implicit social cognition. *Current Directions in Psychological Science*, *20*(3), 143–148.

Anderson, N. H. (1971). Integration theory and attitude change. *Psychological Review*, *78*(3), 171–206.

Arcuri, L., Castelli, L., Galdi, S., Zogmaister, C., & Amadori, A. (2008). Predicting the vote: Implicit attitudes as predictors of the future behavior of decided and undecided voters. *Political Psychology*, *29*, 369–387. http://dx.doi.org/10.1111/j.1467-9221.2008.00635.x.

Aviezer, H., Trope, Y., & Todorov, A. (2012). Body cues, not facial expressions, discriminate between intense positive and negative emotions. *Science*, *338*, 1225–1229. http://dx.doi.org/10.1126/science.1224313.

Balcetis, E., & Lassiter, G. D. (Eds.), (2010). *The social psychology of visual perception*. New York, NY: Psychology Press.

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

Bar, M. (2011). The proactive brain. In M. Bar (Ed.), *Predictions in the brain: Using our past to generate a future* (pp. 13–26). New York, NY: Oxford University Press.

Bargh, J. A. (1999). The cognitive monster: The case against the controllability of automatic stereotype effects. In S. Chaiken & Y. Trope (Eds.), *Dual-process theories in social psychology* (pp. 361–382). New York, NY: Guilford Press.

Bartlett, F. C. (1932). *Remembering*. London: Cambridge University Press.

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.

Bedford, F. L. (2003). More on the not-the-liver fallacy: Medical, neuropsychological, and perceptual dissociations. *Cortex*, *39*(1), 170–173. http://doi.org/10.1016/S0010-9452(08)70094-7.

Berry, D. S. (1992). Vocal types and stereotypes: Joint effects of vocal attractiveness and vocal maturity on person perception. *Journal of Nonverbal Behavior*, *16*, 41–54. http://dx.doi.org/10.1007/bf00986878.

Blair, I. V. (2002). The malleability of automatic stereotypes and prejudice. *Personality and Social Psychology Review*, *6*(3), 242–261. http://doi.org/10.1207/S15327957PSPR0603_8.

Blair, I. V., Chapleau, K. M., & Judd, C. M. (2005). The use of Afrocentric features as cues for judgment in the presence of diagnostic information. *European Journal of Social Psychology*, *35*(1), 59–68. http://doi.org/10.1002/ejsp.232.

Blair, I. V., Ma, J. E., & Lenton, A. P. (2001). Imagining stereotypes away: The moderation of implicit stereotypes through mental imagery. *Journal of Personality and Social Psychology*, *81*(5), 828–841.

Bonnefon, J.-F., Hopfensitz, A., & De Neys, W. (2015). Face-ism and kernels of truth in facial inferences. *Trends in Cognitive Sciences*, *19*(8), 421–422. http://doi.org/10.1016/j.tics.2015.05.002.

Boothman, N. (2008). *How to make people like you in 90 seconds or less*. New York, NY: Workman.

Boucher, K. L., & Rydell, R. J. (2012). Impact of negation salience and cognitive resources on negation during evaluation formation. *Personality and Social Psychology Bulletin*, *38*, 1329–1342.

Brannon S.M. and Gawronski B., A second chance for first impressions?: Exploring the context-(in)dependent updating of implicit evaluations, *Social Psychological and Personality Science*, in press. http://doi.org/10.1177/1948550616673875.

Briñol, P., Petty, R. E., & McCaslin, M. J. (2009). Changing attitudes on implicit versus explicit measures: What is the difference? In R. E. Petty, R. H. Fazio, & P. Briñol (Eds.), *Attitudes: Insights from the new implicit measures* (pp. 285–326). New York, NY: Psychology Press.

Brooks, A. W. (2014). Get excited: Reappraising pre-performance anxiety as excitement. *Journal of Experimental Psychology. General*, *143*(3), 1144–1158. http://doi.org/10.1037/a0035325.

Bruner, J. S. (1957). On perceptual readiness. *Psychological Review*, *64*, 123–152.

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 evaluations. *Personality and Social Psychology Review*, *4*, 330–350.

Cao, J., & Banaji, M. R. (2016). The base rate principle and the fairness principle in social judgment. *Proceedings of the National Academy of Sciences*, *113*(27), 7475–7480. http://doi.org/10.1073/pnas.1524268113.

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.

Chater, N. (2003). How much can we learn from double dissociations. *Cortex*, *39*, 167–169.

Chen, M., & Bargh, J. A. (1997). Nonconscious behavioral confirmation processes: The self-fulfilling consequences of automatic stereotype activation. *Journal of Experimental Social Psychology*, *33*, 541–560.

Cone, J., & Ferguson, M. J. (2015). He did what? The role of diagnosticity in revising implicit evaluations. *Journal of Personality and Social Psychology*, *108*, 37–57.

Cone, J., & Ferguson, M. J. (in preparation). *Change we must believe in: The role of believability in rapid revision of implicit evaluations*. Manuscript in preparation.

Cone, J., Mann, T. C., Meagher, B., & Ferguson, M. J. (in preparation). *Changing our implicit mind: Revision of implicit first impressions based on visual cues*. Manuscript in preparation.

Conrey, F. R., Sherman, J. W., Gawronski, B., Hugenberg, K., & Groom, C. J. (2005). Separating multiple processes in implicit social cognition: The quad model of implicit task performance. *Journal of Personality and Social Psychology*, *89*(4), 469–487.

Critcher, C. R., & Ferguson, M. J. (2016). "Whether I like it or not, it's important": Implicit importance of means predicts self-regulatory persistence and success. *Journal of Personality and Social Psychology*, *110*(6), 818–839.

Cunningham, W., Raye, C., & Johnson, M. (2004). Implicit and explicit evaluation: fMRI correlates of valence, emotional intensity, and control in the processing of attitudes. *Journal of Cognitive Neuroscience*, *16*(10), 1717–1729.

Cunningham, W. A., Zelazo, P. D., Packer, D. J., & Van Bavel, J. J. (2007). The iterative reprocessing model: A multi-level framework for attitudes and evaluation. *Social Cognition*, *25*, 736–760.

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

De Houwer, J. D. (2014). A propositional model of implicit evaluation. *Social and Personality Psychology Compass*, *8*(7), 342–353.

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

Deutsch, R., Gawronski, B., & Strack, F. (2006). At the boundaries of automaticity: Negation as reflective operation. *Journal of Personality and Social Psychology*, *91*(3), 385–405. http://doi.org/10.1037/0022-3514.91.3.385.

DiFonzo, N., & Bordia, P. (2007). *Rumor psychology: Social and organizational approaches*. Washington, DC: American Psychological Association.

Dion, K., Berscheid, E., & Walster, E. (1972). What is beautiful is good. *Journal of Personality and Social Psychology*, *24*, 207–213.

Dudai, Y. (2004). The neurobiology of consolidations, or, how stable is the engram? *Annual Review of Psychology*, *55*(1), 51–86. http://doi.org/10.1146/annurev.psych.55.090902.142050.

Dunn, J. C., & Kirsner, K. (2003). What can we infer from double dissociations. *Cortex*, *39*, 1–7.

Eagly, A. H., Ashmore, R. D., Makhijani, M. G., & Longo, L. C. (1991). What is beautiful is good, but…: A meta-analytic review of research on the physical attractiveness stereotype. *Psychological Bulletin*, *110*(1), 109–128.

Edelman, S. (2008). *Computing the mind: How the mind really works*. Oxford: Oxford University Press.

Ehret, P. J., Monroe, B. M., & Read, S. J. (2015). Modeling the dynamics of evaluation: A multilevel neural network implementation of the iterative reprocessing model. *Personality and Social Psychology Review*, *19*(2), 148–176.

Eiser, J. R., Fazio, R. H., Stafford, T., & Prescott, T. J. (2003). Connectionist simulation of attitude learning: Asymmetries in the acquisition of positive and negative evaluations. *Personality and Social Psychology Bulletin*, *29*, 1221–1235.

Eiser, J. R., Stafford, T., & Fazio, R. H. (2008). Expectancy-confirmation in attitude learning: A connectionist account. *European Journal of Social Psychology*, *38*, 1023–1032.

Fazio, R. H. (2007). Evaluations as object-evaluation associations of varying strength. *Social Cognition*, *25*, 603–637.

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

Fazio, R. H., & Olson, M. A. (2003). Implicit measures in social cognition research: Their meaning and use. *Annual Review of Psychology*, *54*(1), 297–327. http://doi.org/10.1146/annurev.psych.54.101601.145225.

Fazio, R. H., Pietri, E. S., Rocklage, M. D., & Shook, N. J. (2015). Positive versus negative valence: Asymmetries in attitude formation and generalization as fundamental individual differences. In J. M. Olson & M. P. Zanna (Eds.), *Advances in experimental social psychology: Vol. 51* (pp. 97–146). London, UK: Elsevier.

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

Ferguson, M. J. (2007). On the automatic evaluation of end-states. *Journal of Personality and Social Psychology*, *92*(4), 596–611. http://doi.org/10.1037/0022-3514.92.4.596.

Ferguson, M. J. (2008). On becoming ready to pursue a goal you don't know you have: Effects of nonconscious goals on evaluative readiness. *Journal of Personality and Social Psychology*, *95*(6), 1268–1294. http://doi.org/10.1037/a0013263.

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*(5), 557–572. http://doi.org/10.1037/0022-3514.87.5.557.

Ferguson, M. J., & Fukukura, J. (2012). Likes and dislikes: A social cognitive perspective. In S. Fiske & C. N. Macrae (Eds.), *SAGE handbook of social cognition* (pp. 165–189). Los Angeles: SAGE.

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

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

Ferguson, M. J., & Zayas, V. (2009). Automatic evaluation. *Current Directions in Psychological Science*, *18*(6), 362–366.

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.

Forgas, J. P., & Bower, G. H. (1987). Mood effects on person-perception judgments. *Journal of Personality and Social Psychology*, *53*(1), 53–60.

Forscher, P. S., Lai, C. K., Axt, J. R., Ebersole, C. R., Herman, M., Devine, P. G., et al. (2016). *A meta-analysis of change in implicit bias*. Unpublished manuscript. Retrieved from. https://osf.io/dv8tu.

Fourakis, E., & Cone, J. (in preparation). *Explaining why: Causal attribution processes affect implicit impression formation and updating*. Manuscript in preparation.

Freeman, J. B., & Ambady, N. (2011). A dynamic interactive theory of person construal. *Psychological Review*, *118*, 247–279.

Freeman, J. B., & Johnson, K. L. (2016). More than meets the eye: Split-second social perception. *Trends in Cognitive Sciences*, *20*(5), 362–374. http://doi.org/10.1016/j.tics.2016.03.003.

Friese, M., Hofmann, W., & Schmitt, M. (2008a). When and why do implicit measures predict behaviour? Empirical evidence for the moderating role of opportunity, motivation, and process reliance. *European Review of Social Psychology*, *19*, 285–338. http://dx.doi.org/10.1080/10463280802556958.

Friese, M., Hofmann, W., & Wänke, M. (2008b). When impulses take over: Moderated predictive validity of explicit and implicit attitude measures in predicting food choice and consumption behaviour. *British Journal of Social Psychology*, *47*, 397–419. http://dx.doi.org/10.1348/014466607X241540.

Friese, M., Smith, C. T., Plischke, T., Bluemke, M., & Nosek, B. A. (2012). Do implicit attitudes predict actual voting behavior particularly for undecided voters? *PLoS One*, 7(8), e44130. http://dx.doi.org/10.1371/journal.pone.0044130.

Galdi, S., Arcuri, L., & Gawronski, B. (2008). Automatic mental associations predict future choices of undecided decision-makers. *Science*, *321*(5892), 1100–1102. http://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 evaluation change. *Psychological Bulletin*, *132*, 692–731.

Gawronski, B., & Bodenhausen, G. V. (2011). The associative-propositional evaluation model: Theory, evidence, and open questions. In M. P. Zanna (Ed.), *Advances in experimental social psychology: Vol. 44* (pp. 59–127). New York: Academic Press.

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

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.

Gawronski, B., Rydell, R. J., Vervliet, B., & De Houwer, J. (2010). Generalization versus contextualization in automatic evaluation. *Journal of Experimental Psychology: General*, *139*(4), 683.

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.

Gawronski, B., Ye, Y., Rydell, R. J., & De Houwer, J. (2014). Formation, representation, and activation of contextualized attitudes. *Journal of Experimental Social Psychology*, *54*, 188–203.

Gilbert, D. T. (1998). Speeding with Ned: A personal view of the correspondence bias. In J. M. Darley & J. Cooper (Eds.), *Attribution and social interaction: The legacy of Edward E. Jones* (pp. 5–66). Washington, DC: American Psychological Association. xxxviii, 550 pp. http://dx.doi.org/10.1037/10286-001.

Gilbert, D. T., & Osborne, R. E. (1989). Thinking backward: Some curable and incurable consequences of cognitive busyness. *Journal of Personality and Social Psychology*, *57*, 940–949. http://dx.doi.org/10.1037/0022-3514.57.6.940.

Grant, H. (2015). A second chance to make the right impression. *Harvard Business Review*, *93*(1/2), 108–111. https://hbr.org/2015/01/a-second-chance-to-make-the-right-impression.

Greenwald, A. G., Banaji, M. R., & Nosek, B. A. (2015). Statistically small effects of the Implicit Association Test can have societally large effects. *Journal of Personality and Social Psychology*, *108*(4), 553–561. http://doi.org/10.1037/pspa0000016.

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

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*(1), 17–41.

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

Gregory, R. L. (1980). Perceptions as hypotheses. *Philosophical Transactions of the Royal Society, B: Biological Sciences*, *290*, 181–197.

Gross, J. J. (2015). Emotion regulation: Current status and future prospects. *Psychological Inquiry*, *26*(1), 1–26. http://doi.org/10.1080/1047840X.2014.940781.

Gunaydın, G., Selcuk, E., & Zayas, V. (2017). Impressions based on a portrait predict, one-month later, impressions following a live interaction. *Social Psychological and Personality Science*, *8*, 36–44.

Hamilton, D. L., & Zanna, M. P. (1974). Context effects in impression formation: Changes in connotative meaning. *Journal of Personality and Social Psychology*, *29*, 649–654. http://dx.doi.org/10.1037/h0036633.

Hardt, O., Einarsson, E. Ö., & Nader, K. (2010). A bridge over troubled water: Reconsolidation as a link between cognitive and neuroscientific memory research traditions. *Annual Review of Psychology*, *61*(1), 141–167. http://doi.org/10.1146/annurev.psych.093008.100455.

Harrison, A. A., & Saeed, L. (1977). Let's make a deal: An analysis of revelations and stipulations in lonely hearts advertisements. *Journal of Personality and Social Psychology*, *35*, 257–264.

Hayes, A. F. (2013). *Introduction to mediation, moderation, and conditional process analysis*. New York: Guilford Press.

Hebb, D. O. (1949). *The organization of behavior: A neuropsychological theory*. New York: Wiley.

Helmholtz, H. V. (1860). *Treatise on physiological optics*. New York: Dover.

Hermer-Vazquez, L., Hermer-Vazquez, R., Rybinnik, I., Greebel, G., Keller, R., Xu, S., et al. (2005). Rapid learning and flexible memory in "habit" tasks in rats trained with brain stimulation reward. *Physiology & Behavior*, *84*(5), 753–759.

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

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

Hofmann, W., & Friese, M. (2008). Impulses got the better of me: Alcohol moderates the influence of implicit attitudes toward food cues on eating behavior. *Journal of Abnormal Psychology*, *117*, 420–427. http://dx.doi.org/10.1037/0021-843X.117.2.420.

Hu, X., Antony, J. W., Creery, J. D., Vargas, I. M., Bodenhausen, G. V., & Paller, K. A. (2015). Unlearning implicit social biases during sleep. *Science*, *348*(6238), 1013–1015.

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.

Ireland, M. E., Slatcher, R. B., Eastwick, P. W., Scissors, L. E., Finkel, E. J., & Pennebaker, J. W. (2011). Language style matching predicts relationship initiation and stability. *Psychological Science*, *22*, 39–44. http://dx.doi.org/10.1177/0956797610392928.

Johnson I.R., Kopp B.M. and Petty R.E., Just say no! (and mean it): Meaningful negation as a tool to modify automatic racial attitudes, *Group Processes & Intergroup Relations*, in press. http://doi.org/10.1177/1368430216647189.

Jones, E. E., & Harris, V. A. (1967). The attribution of attitudes. *Journal of Experimental Social Psychology*, *3*(1), 1–24.

Kamins, M. A. (1997). Consumer responses to rumors: Good news, bad news. *Journal of Consumer Psychology*, *6*(2), 165–187.

Karpinski, A., & Hilton, J. L. (2001). Attitudes and the Implicit Association Test. *Journal of Personality and Social Psychology*, *81*(5), 774–788.

Kawakami, K., Dovidio, J. F., Moll, J., Hermsen, S., & Russin, A. (2000). Just say no (to stereotyping): Effects of training in the negation of stereotypic associations on stereotype activation. *Journal of Personality and Social Psychology*, *78*(5), 871–888.

Kawakami, K., Dovidio, J. F., & van Kamp, S. (2005). Kicking the habit: Effects of nonstereotypic association training and correction processes on hiring decisions. *Journal of Experimental Social Psychology*, *41*(1), 68–75.

Kawakami, K., Phills, C. E., Steele, J. R., & Dovidio, J. F. (2007). (Close) distance makes the heart grow fonder: Improving implicit racial attitudes and interracial interactions through approach behaviors. *Journal of Personality and Social Psychology*, *92*(6), 957–971. http://doi.org/10.1037/0022-3514.92.6.957.

Kelley, H. H. (1967). Attribution theory in social psychology. In D. Levine (Ed.), *Nebraska symposium on motivation: Vol. 15* (pp. 192–238). Lincoln: University of Nebraska Press.

Keren, G., & Schul, Y. (2009). Two is not always better than one: A critical evaluation of two- system theories. *Perspectives on Psychological Science*, *4*, 533–550.

Kimball, D., Mann, T. C., & Ferguson, M. J. (in preparation). *Race, reason, and revision: Does target race influence the updating of implicit impressions?* Manuscript in preparation.

Klein, S. B., Loftus, J., Trafton, J. G., & Fuhrman, R. W. (1992). Use of exemplars and abstractions in trait judgments: A model of trait knowledge about the self and others. *Journal of Personality and Social Psychology*, *63*(5), 739.

Kruglanski, A. W., Erb, H. P., Pierro, A., Manetti, L., & Chun, W. Y. (2006). On parametric continuities in the world of binary either ors. *Psychological Inquiry*, *17*, 153–165.

Kunda, Z., & Thagard, P. (1996). Forming impressions from stereotypes, traits, and behaviors: A parallel-constraint-satisfaction theory. *Psychological Review*, *103*, 284–308.

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

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

Lane, R. D., Ryan, L., & Nadel, L. (2015). Memory reconsolidation, emotional arousal and the process of change in psychotherapy: New insights from brain science. *Behavioral and Brain Sciences*, *38*, 1–64.

Lerner, R. M., & Moore, T. (1974). Sex and status effects on perception of physical attractiveness. *Psychological Reports*, *34*, 1047–1050.

Lewandowsky, S., Ecker, U. K. H., Seifert, C. M., Schwarz, N., & Cook, J. (2012). Misinformation and its correction: Continued influence and successful debiasing. *Psychological Science in the Public Interest*, *13*(3), 106–131. http://doi.org/10.1177/1529100612451018.

Lorenzo, G. L., Biesanz, J. C., & Human, L. J. (2010). What is beautiful is good and more accurately understood: Physical attractiveness and accuracy in first impressions of personality. *Psychological Science*, *21*, 1777–1782. http://dx.doi.org/10.1177/0956797610388048.

Maddux, W. W., Barden, J., Brewer, M. B., & Petty, R. E. (2005). Saying no to negativity: The effects of context and motivation to control prejudice on automatic evaluative responses. *Journal of Experimental Social Psychology*, *41*(1), 19–35. http://doi.org/10.1016/j.jesp.2004.05.002.

Maio, G. R., & Haddock, G. (2015). *The psychology of attitudes and attitude change*. Thousand Oaks, CA: Sage.

Mann, T. C., Cone, J., & Ferguson, M. J. (2015). Social-psychological evidence for the effective updating of implicit attitudes. *Behavioral and Brain Sciences*, *38*. e15.

Mann, T. C., & Ferguson, M. J. (2015). Can we undo our first impressions? The role of reinterpretation in reversing implicit evaluations. *Journal of Personality and Social Psychology*, *108*(6), 823–849. http://doi.org/10.1037/pspa0000021.

Mann, T. C., & Ferguson, M. J. (in progress). *Reinterpretation in the real world: The effects of arguments on implicit evaluations of known targets*. Work in progress.

Mann, T. C., & Ferguson, M. J. (2017). Reversing implicit first impressions through reinterpretation after a two-day delay. *Journal of Experimental Social Psychology*, *68*, 122–127. http://doi.org/10.1016/j.jesp.2016.06.004.

Mann, T. C., Ferguson, M. J., & Axt, J. R. (in preparation). *How do we reverse our first impressions?: Mechanisms of reinterpretation-driven changes in implicit impressions*. Manuscript in preparation.

McCarthy, R. J., & Skowronski, J. J. (2011). What will Phil do next? Spontaneously inferred traits influence predictions of behavior. *Journal of Experimental Social Psychology*, *47*(2), 321–332. http://doi.org/10.1016/j.jesp.2010.10.015.

McClelland, J. L., & Rumelhart, D. E. (1985). Distributed memory and the representation of general and specific information. *Journal of Experimental Psychology: General*, *114*, 159–188.

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

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

McGaugh, J. L. (2000). Memory—A century of consolidation. *Science*, *287*(5451), 248–251. http://doi.org/10.1126/science.287.5451.248.

McNulty, J. K., Baker, L. R., & Olson, M. A. (2014). Implicit self-evaluations predict changes in implicit partner evaluations. *Psychological Science*, *25*(8), 1649–1657.

McNulty, J. K., Olson, M. A., Meltzer, A. L., & Shaffer, M. J. (2013). Though they may be unaware, newlyweds implicitly know whether their marriage will be satisfying. *Science*, *342*, 1119–1120.

Menand, L. (2014). *The De Man Case: Does a critic's past explain his criticism?* (Vol. 90 (5)). New York: The New Yorker. Retrieved from http://www.newyorker.com/magazine/2014/03/24/the-de-man-case.

Mills, J., & Aronson, E. (1965). Opinion change as a function of the communicator's attractiveness and desire to influence. *Journal of Personality and Social Psychology*, *1*, 173–177. http://dx.doi.org/10.1037/h0021646.

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

Moors, A., & De Houwer, J. (2001). Automatic appraisal of motivational valence: Motivational affective priming and Simon effects. *Cognition & Emotion*, *15*(6), 749–766. http://doi.org/10.1080/02699930143000293.

Nederkoorn, C., Houben, K., Hofmann, W., Roefs, A., & Jansen, A. (2010). Control yourself or just eat what you like? Weight gain over a year is predicted by an interactive effect of response inhibition and implicit preference for snack foods. *Health Psychology*, *29*, 389–393.

Olivola, C. Y., Funk, F., & Todorov, A. (2014). Social attributions from faces bias human choices. *Trends in Cognitive Sciences*, *18*(11), 566–570. http://doi.org/10.1016/j.tics.2014.09.007.

Olivola, C. Y., & Todorov, A. (2010). Fooled by first impressions? Re-examining the diagnostic value of appearance-based inferences. *Journal of Experimental Social Psychology*, *46*, 315–324.

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

Oswald, F. L., Mitchell, G., Blanton, H., Jaccard, J., & Tetlock, P. E. (2015). Using the IAT to predict ethnic and racial discrimination: Small effect sizes of unknown societal significance. *Journal of Personality and Social Psychology*, *108*(4), 562–571. http://doi.org/10.1037/pspa0000023.

Payne, B. K., Burkley, M., & Stokes, M. B. (2008). Why do implicit and explicit evaluation 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 evaluations: Affect misattribution as implicit measurement. *Journal of Personality and Social Psychology*, *89*, 277–293.

Payne, K., & Lundberg, K. (2014). The affect misattribution procedure: Ten years of evidence on reliability, validity, and mechanisms. *Social and Personality Psychology Compass*, *8*(12), 672–686. http://doi.org/10.1111/spc3.12148.

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.

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

Petty, R. E., & Cacioppo, J. T. (1986). The elaboration likelihood model of persuasion. In *Communication and persuasion* (pp. 1–24). New York: Springer.

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

Plaut, D. C. (1995). Double dissociation without modularity: Evidence from connectionist neuropsychology. *Journal of Clinical and Experimental Neuropsychology*, *17*, 291–321.

Read, S. J., & Miller, L. C. (Eds.), (1998). *Connectionist models of social reasoning and social behavior*. Mahwah, NJ: Erlbaum.

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.

Remue, J., De Houwer, J., Barnes-Holmes, D., Vanderhasselt, M.-A., & De Raedt, R. (2013). Self-esteem revisited: Performance on the implicit relational assessment procedure as a measure of self- versus ideal self-related cognitions in dysphoria. *Cognition and Emotion*, *27*(8), 1441–1449. http://doi.org/10.1080/02699931.2013.786681.

Rezlescu, C., Duchaine, B., Olivola, C. Y., & Chater, N. (2012). Unfakeable facial configurations affect strategic choices in trust games with or without information about past behavior. *PLoS One*, 7(3), e34293–e34296. http://doi.org/10.1371/journal.pone.0034293.

Roese, N. J., & Olson, J. M. (2007). Better, stronger, faster: Self-serving judgment, affect regulation, and the optimal vigilance hypothesis. *Perspectives on Psychological Science*, *2*, 124–141.

Ross, L. (1977). The intuitive psychologist and his shortcomings: Distortions in the attribution process. In L. Berkowitz (Ed.), *Advances in experimental social psychology: Vol. 10* (pp. 173–220). New York: Academic Press.

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.

Rule, N. O., & Alaei, R. (2016). Gaydar: The perception of sexual orientation from subtle cues. *Current Directions in Psychological Science*, *25*, 444–448. http://dx.doi.org/10.1177/0963721416664403.

Rule, N. O., & Ambady, N. (2008a). Brief exposures: Male sexual orientation is accurately perceived at 50-ms. *Journal of Experimental Social Psychology*, *44*, 1100–1105. http://dx.doi.org/10.1016/j.jesp.2007.12.001.

Rule, N. O., & Ambady, N. (2008b). First impressions: Peeking at the neural underpinnings. In N. Ambady & J. Skowronski (Eds.), *First impressions* (pp. 35–56). New York, NY: Guilford.

Rule, N. O., Ambady, N., & Adams, R. B., Jr. (2009). Personality in perspective: Judgmental consistency across orientations of the face. *Perception*, *38*, 1688–1699. http://dx.doi.org/10.1068/p6384.

Rule, N. O., Ambady, N., Adams, R. B., Jr., & Macrae, C. N. (2008). Accuracy and awareness in the perception and categorization of male sexual orientation. *Journal of Personality and Social Psychology*, *95*, 1019–1028. http://dx.doi.org/10.1037/a0013194.

Rule, N. O., Krendl, A. C., Ivcevic, Z., & Ambady, N. (2013). Accuracy and consensus in judgments of trustworthiness from faces: Behavioral and neural correlates. *Journal of Personality and Social Psychology*, *104*(3), 409–426. http://doi.org/10.1037/a0031050.

Rule, N. O., Tskhay, K. O., Freeman, J. B., & Ambady, N. (2014). On the interactive influence of facial appearance and explicit knowledge in social categorization. *European Journal of Social Psychology*, *44*(6), 529–535. http://doi.org/10.1002/ejsp.2043.

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

Rydell, R. J., & McConnell, A. R. (2006). Understanding implicit and explicit evaluation change: A systems of reasoning analysis. *Journal of Personality and Social Psychology*, *91*(6), 995–1008.

Rydell, R. J., McConnell, A. R., Mackie, D. M., & Strain, L. M. (2006). Of two minds: Forming and changing valence-inconsistent implicit and explicit evaluations. *Psychological Science*, *17*(11), 954–958.

Rydell, R. J., McConnell, A. R., Strain, L. M., Claypool, H. M., & Hugenberg, K. (2007). Implicit and explicit evaluations respond differently to increasing amounts of counterattitudinal information. *European Journal of Social Psychology*, *37*(5), 867–878.

Schachter, S., & Burdick, H. (1955). A field experiment on rumor transmission. *Journal of Abnormal and Social Psychology*, *50*, 363–371.

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

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

Schneid, E. D., Carlston, D. E., & Skowronski, J. J. (2015). Spontaneous evaluative inferences and their relationship to spontaneous trait inferences. *Journal of Personality and Social Psychology*, *108*(5), 681–696. http://doi.org/10.1037/a0039118.

Schroeder, J., & Epley, N. (2015). The sound of intellect: Speech reveals a thoughtful mind, increasing a job candidate's appeal. *Psychological Science*, *26*, 877–891.

Schroeder, J., & Epley, N. (2016). Mistaking minds and machines: How speech affects dehumanization and anthropomorphism. *Journal of Experimental Psychology General*, *145*, 1427–1437.

Schröder, T., & Thagard, P. (2013). The affective meanings of automatic social behaviors: Three mechanisms that explain priming. *Psychological Review*, *120*(1), 255–280. http://doi.org/10.1037/a0030972.

Schwarz, N. (2007). Attitude construction: Evaluation in context. *Social Cognition*, *25*(5), 638–656.

Schwarz, N., & Sudman, S. (Eds.), (1992). *Context effects in social and psychological research*. New York: Springer Verlag.

Seibt, B., Häfner, M., & Deutsch, R. (2007). Prepared to eat: How immediate affective and motivational responses to food cues are influenced by food deprivation. *European Journal of Social Psychology*, *37*, 359–379.

Sharot, T., & Garrett, N. (2016). Forming beliefs: Why valence matters. *Trends in Cognitive Sciences*, *20*(1), 25–33.

Shen, X., Mann, T. C., & Ferguson, M. J. (in preparation). *Can we get past an untrustworthy face? Evidence of implicit updating about untrustworthy targets*. Manuscript in preparation.

Sherman, J. W., Gawronski, B., Gonsalkorale, K., Hugenberg, K., Allen, T. J., & Groom, C. J. (2008). The self-regulation of automatic associations and behavioral impulses. *Psychological Review*, *115*(2), 314–335. http://doi.org/10.1037/0033-295X.115.2.314.

Sherman, J. W., & Klein, S. B. (1994). Development and representation of personality impressions. *Journal of Personality and Social Psychology*, *67*(6), 972–983.

Sherman, S. J., Presson, C. C., Chassin, L., Rose, J. S., & Koch, K. (2003). Implicit and explicit attitudes toward cigarette smoking: The effects of context and motivation. *Journal of Social and Clinical Psychology*, *22*, 13–39.

Sinclair, S., Lowery, B. S., Hardin, C. D., & Colangelo, A. (2005). Social tuning of automatic racial attitudes: The role of affiliative motivation. *Journal of Personality and Social Psychology*, *89*(4), 583.

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

Sloman, S. A. (1996). The empirical case for two systems of reasoning. *Psychological Bulletin*, *119*(1), 3–22.

Snyder, M., & Stukas, A. A. (1999). Interpersonal processes: The interplay of cognitive, motivational, and behavioral activities in social interaction. *Annual Review of Psychology*, *50*, 273–303. http://dx.doi.org/10.1146/annurev.psych.50.1.273.

Snyder, M., Tanke, E. D., & Berscheid, E. (1977). Social perception and interpersonal behavior: On the self-fulfilling nature of social stereotypes. *Journal of Personality and Social Psychology*, *35*, 656–666. http://dx.doi.org/10.1037/0022-3514.35.9.656.

Song, J. H., & Nakayama, K. (2009). Hidden cognitive states revealed in choice reaching tasks. *Trends in Cognitive Sciences*, *13*, 360–366.

Spivey, M. J. (2008). *The continuity of mind*. New York: Oxford University Press.

Sritharan, R., Heilpern, K., Wilbur, C. J., & Gawronski, B. (2010). I think I like you: Spontaneous and deliberate evaluations of potential romantic partners in an online dating context. *European Journal of Social Psychology*, *40*, 1062–1077.

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

Tabak, J. A., & Zayas, V. (2012). The roles of featural and configural face processing in snap judgments of sexual orientation. *PLoS One*, 7(5), 1–7, e3667. http://doi.org/10.1371/journal.pone.0036671.

Tibboel, H., De Houwer, J., Dirix, N., & Spruyt, A. (2017). Beyond associations: Do implicit beliefs play a role in smoking addiction? *Journal of Psychopharmacology*, *31*, 43–53. http://doi.org/10.1177/0269881116665327.

Tickle-Degnen, L., & Rosenthal, R. (1990). The nature of rapport and its nonverbal correlates. *Psychological Inquiry*, *4*, 285–293. http://dx.doi.org/10.1207/s15327965pli0104_1.

Tingley, D. (2014). Face off: Facial features and strategic choice. *Political Psychology*, *35*(1), 344–368.

Todorov, A., Dotsch, R., Porter, J., Oosterhof, N., & Falvello, V. (2013). Validation of data-driven computational models of social perception of faces. *Emotion*, *13*, 724–738.

Todorov, A., Funk, F., & Olivola, C. Y. (2015). Response to Bonnefon et al.: Limited 'kernels of truth' in facial inferences. *Trends in Cognitive Sciences*, *19*(8), 422–423.

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

Todorov, A., & Oosterhoof, N. N. (2011). Modeling social perception of faces. *IEEE Signal Processing Magazine*, *28*, 117–122.

Todorov, A., & Porter, J. (2014). Misleading first impressions: Different for different facial images of the same person. *Psychological Science*, *25*, 1404–1417.

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

Todorov, A., & Uleman, J. S. (2004). The person reference process in spontaneous trait inferences. *Journal of Personality and Social Psychology*, *87*(4), 482–493. http://doi.org/10.1037/0022-3514.87.4.482.

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.

Van Bavel, J. J., Jenny Xiao, Y., & Cunningham, W. A. (2012). Evaluation is a dynamic process: Moving beyond dual system models. *Social and Personality Psychology Compass*, *6*(6), 438–454. http://doi.org/10.1111/j.1751-9004.2012.00438.x.

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

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

Van Dessel, P., Gawronski, B., Smith, C. T., & De Houwer, J. (2017). Mechanisms underlying approach-avoidance instruction effects on implicit evaluation: Results of a preregistered adversarial collaboration. *Journal of Experimental Social Psychology*, *69*, 23–32. http://doi.org/10.1016/j.jesp.2016.10.004.

van Deursen, D. S., Salemink, E., Boendermaker, W. J., Pronk, T., Hofmann, W., & Wiers, R. W. (2015). Executive functions and motivation as moderators of the relationship between automatic associations and alcohol use in problem drinkers seeking online help. *Alcoholism, Clinical and Experimental Research*, *39*(9), 1788–1796. http://doi.org/10.1111/acer.12822.

van't Wout, M., & Sanfey, A. G. (2008). Friend or foe: The effect of implicit trustworthiness judgments in social decision-making. *Cognition*, *108*, 796–803. http://dx.doi.org/10.1016/j.cognition.2008.07.002.

Wegner, D. M. (1994). Ironic processes of mental control. *Psychological Review*, *101*(1), 34–52.

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

Wilson, T. D., Lindsey, S., & Schooler, T. Y. (2000). A model of dual evaluations. *Psychological Review*, *107*(1), 101–126.

Wilson, J. P., & Rule, N. O. (2015). Facial trustworthiness predicts extreme criminal-sentencing outcomes. *Psychological Science*, *26*(8), 1325–1331. http://doi.org/10.1177/0956797615590992.

Winter, L., & Uleman, J. S. (1984). When are social judgments made? Evidence for the spontaneousness of trait inferences. *Journal of Personality and Social Psychology*, *47*(2), 237.

Wojnowicz, M. T., Ferguson, M. J., Dale, R., & Spivey, M. J. (2009). The self-organization of explicit attitudes. *Psychological Science*, *20*(11), 1428–1435. http://doi.org/10.1111/j.1467-9280.2009.02448.x.

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

Wyer, N. A. (2016). Easier done than undone… by some of the people, some of the time: The role of elaboration in explicit and implicit group preferences. *Journal of Experimental Social Psychology*, *63*, 77–85. http://doi.org/10.1016/j.jesp.2015.12.006.

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

Zayas, V., & Shoda, Y. (2015). Love you? Hate you? Maybe it's both: Evidence that significant others trigger bivalent-priming. *Social Psychological and Personality Science*, *6*(1), 56–64. http://doi.org/10.1177/1948550614541297.

Zogmaister, C., Arcuri, L., & Castelli, L. (2008). The impact of loyalty and equality on implicit ingroup favoritism. *Group Processes & Intergroup Relations*, *11*(4), 493–512. http://doi.org/10.1177/1368430208095402.